

Data Analytics-backed Vehicular Crowd-sensing for GPS-less Tracking in Public Transportation

by

Cem Kaptan

Thesis Supervisors: Dr. Burak Kantarci and Dr. Azzedine Boukerche

A thesis
presented to the University of Ottawa
in fulfillment of the
thesis requirement for the degree of
Master of Computer Science

School of Electrical Engineering and Computer Science
Faculty of Engineering
University of Ottawa

© Cem Kaptan, Ottawa, Ontario, Canada, 2018

I hereby declare that I am the sole author of this thesis. This is a true copy of the thesis, including any required final revisions, as accepted by my examiners.

I understand that my thesis may be made electronically available to the public.

Abstract

The widespread availability of sensors, improved computing, and storage capabilities, and ubiquity of networking services have led to the transformation of the conventional transportation services. Achieving the smart transportation goal can be either via dedicated or non-dedicated methods. The former denotes utilization of sensors that are explicitly deployed and configured for pre-defined sensing tasks whereas the latter exploits opportunistic and participatory sensing paradigms. With that motivation in mind, the contributions of this thesis are three-fold: (i) Sensor emulation, (ii) Crowd-sensing based GPS-less tracking, and (iii) Reliable data acquisition in crowd-sensed GPS-less tracking. Emulating non-dedicated sensors in a simulation environment enables us to perform large-scale crowd-sensing tasks. We introduce a variety of vehicular crowd-sensing-based frameworks to track public transportation vehicles that move over static routes in a smart city setting without *GPS*-enabled devices because of the major downsides of *GPS* (e.g. high energy consumption, inaccurate localization in certain environments such as indoor, and privacy violation due to direct location sharing). To this end, we propose a novel framework, in our initial approach, to emulate the functionality of a sensor by using multiple available soft sensors and machine intelligence algorithms. As a case study, the localization of city buses in a smart city setting is investigated by using the accelerometer and microphones of the passengers and supervised machine intelligence running in the cloud. In this application, the *GPS* functionality is emulated by using these two soft sensors. We evaluate our proposed scheme through simulations and show that the proposed framework can operate with more than 90% accuracy in estimating the location of public buses while preserving the actual location privacy of the smartphone users. This approach results in smartphone battery energy savings of 38–46% (as compared to *GPS*-based approaches) due to the elimination of the power-hungry *GPS* devices. Additional sensor recruitment schemes with various sensor combinations of accelerometer and microphone are developed in an extension work to examine both localization accuracy and energy consumption performance. Furthermore, in a subsequent work, crowd-sensed data undergoes an unsupervised machine learning module that estimates the location of the vehicle. We evaluate our proposed scheme through simulations and show that the proposed framework can operate with 95% accuracy in estimating the location of public vehicles in the best case and with 80% accuracy in the worst case. Since data trustworthiness is vital when data is crowd-solicited, assessment and quantification of the trustworthiness of participating sensors play a key role in the accuracy of the acquired information. To this end, we propose a reliability-aware participant recruitment scheme that assesses the trustworthiness of individual participants. Our simulation results lead to approximately 96% localization accuracy under the reliability-aware recruitment scheme compared to 93% localization accuracy the reliability-unaware participant recruitment scheme. Moreover, we introduce two trustworthiness-aware recruitment schemes in another work: Reliability-driven naive recruitment (*RDNR*) and Reliability-driven exclusive recruitment (*RDER*). We evaluate the performance of *RDNR*, *RDER*, and non-restrictive recruitment (reliability-unaware). Through simulations, we show that over 85% and 98% accuracy can be achieved in the worst and best cases, respectively, while consuming less energy than *GPS*-based localization approaches.

Acknowledgements

I would first like to express my sincere gratitude to my supervisors Dr. Burak Kantarci and Dr. Azzedine Boukerche for their continuous support throughout my Master study. They helped me to discover the research field that suits my future ambitions. Their motivations, guidance, and always-aiming-for-the-best-possible attitude pushed me to publish a plethora of works that contribute to this thesis. I could not have imaged having better supervisors and mentors for my Master study.

I remember the first e-mail I wrote to Dr. Burak Kantarci regarding a research position in his team and how prompt I received an affirmative response back. I have started working with him more than year ago and half a year ago with Dr. Azzedine Boukerche. Their deep technical knowledge in my field of research always shed a light upon my path to see clearly my path and destination. Thanks to them, this fruitful collaboration will be persistent in my Ph.D. study.

I also would like to thank those who made my researches possible, especially NSERC CREATE project TRANSIT and also my father and Dr. Burak Kantarci who spared hours in their busy schedule to help me collect the data that I needed for my experiments.

Last but not the least, I must express my very profound gratitude to my parents for believing in me that I can accomplish anything and for supporting me in every step of my life. I am so grateful to have such parents who provide me with their unfailing support and continuous encouragement throughout my years of study and through the process of researching and writing this thesis. This accomplishment would not have been possible without them. Thank you.

Table of Contents

List of Tables	viii
List of Figures	ix
1 Introduction	1
1.1 Motivation	1
1.2 Contributions	3
1.3 Outline	5
2 Literature Study	6
2.1 Enabling Communication Infrastructures and Technologies for Smart Transportation	7
2.1.1 Wireless Communication Technologies	7
2.1.2 Vehicular Ad Hoc Networks	10
2.2 Smart Transportation Applications and Services	12
2.2.1 Traffic Management	12
2.2.2 Public Transportation Management	13
2.2.3 Emergency Management	14
2.2.4 Information Management	14
2.2.5 Service Provisioning Methodologies in Smart Transportation	15
2.3 Vehicular Crowd-Sensing for Smart Transportation Applications and Services	16
2.3.1 Traffic Management	16
2.3.2 Public Transportation Management	18
2.3.3 Emergency Management	19
2.3.4 Information Management	20

2.4	Data Analytics for Smart Transportation Application and Services	21
2.4.1	Machine and Deep Learning	21
2.4.2	Signal Processing	34
3	Public Transport Tracking via Vehicular Crowd-Sensing: Proposed Supervised Methodology	40
3.1	Introduction	40
3.1.1	Problem Description	40
3.1.2	Proposed Solution	40
3.2	Proposed System Overview	41
3.3	Proposed Methodologies	41
3.3.1	Data Acquisition	41
3.3.2	Data Aggregation	44
3.3.3	Probabilistic Prediction	44
3.3.4	Analytics	46
3.4	Performance Evaluation	46
3.4.1	Data Acquisition and Simulation Settings	46
3.4.2	Numerical Results	50
3.5	Concluding Remarks	55
4	Public Transport Tracking via Vehicular Crowd-Sensing: Proposed Unsupervised Methodology	56
4.1	Introduction	56
4.1.1	Problem Description	56
4.1.2	Proposed Solution	57
4.2	Proposed System Overview	57
4.3	Proposed Methodologies	58
4.3.1	Data Acquisition	58
4.3.2	Data Aggregation	60
4.3.3	Data Analytics	62
4.4	Performance Evaluation	65
4.4.1	Data Acquisition and Simulation Settings	65
4.4.2	Numerical Results	67
4.5	Concluding Remarks	72

5	Ensuring Trustworthiness of Crowd-Sensed Vehicle Tracking Service	74
5.1	Introduction	74
5.1.1	Problem Description	74
5.1.2	Proposed Solution	74
5.2	Proposed System Overview	75
5.3	Proposed Methodologies	77
5.3.1	Non-Restricted Recruitment	77
5.3.2	Reliability-driven Naive Recruitment	77
5.3.3	Reliability-driven Exclusive Recruitment	80
5.4	Performance Evaluation	81
5.4.1	Simulation Settings	81
5.4.2	Numerical Results	82
5.5	Concluding Remarks	87
6	Conclusions, Open Issues, and Future Directions	89
	References	92

List of Tables

2.1	Overview of smart transportation services and referenced applications. Services are s1: traffic management, s2: public transportation management, s3: emergency management, and s4:information management.	39
3.1	Notations for GPS-less localization architecture based on supervised machine learning. The variables and constants are grouped in three categories: The first category (top section of the table) denotes the objects, i.e., sensors, sensory detection zones and coordinates, second category includes the variables regarding aggregated sensed values, the third category is related to the variables regarding communications, and the fourth category (bottom section of the table) is related to participant recruitment scenarios.	43
3.2	General settings	47
3.3	Accuracy settings	48
4.1	The constants that are used in formulas such as number of vehicles and clusters, sensor array, and so on.	59
4.2	The parameters that are used to formulate the $k - means$ algorithm	60
4.3	The parameters that are associated with the communication channel	60
4.4	The descriptions for participant recruitment scenarios	60
4.5	Global simulation settings	66
4.6	Simulation settings for feasibility analysis	67
4.7	Simulation settings for $\eta - rec.$ and $\alpha - rec.$	67
5.1	List of constants that are used in the formulas and throughout the Chapter 5	77
5.2	The parameters that are used to formulate the user recruitment scenarios . .	78
5.3	List of auxiliary functions that are used in the equations	78
5.4	Description for the participant activity profiles	78
5.5	Simulation settings	82

List of Figures

3.1	GPS-less localization architecture based on supervised machine learning. . .	42
3.2	A generic flowchart for <i>GPS</i> -less vehicle localization.	42
3.3	The concept of sliding window of stationary regions.	45
3.4	Accuracy performance under three varying simulation settings (a) The effect of window size on accuracy (b) The effect of route length on accuracy (c) The effect of Gaussian noise on accuracy.	50
3.5	Energy performance per vehicle under different scenarios and recruitment schemes (a) Best case energy consumption comparison of $\eta - GPS$, $\eta - nGPS$, $\alpha - GPS$ and $\alpha - nGPS$ (b) Worst case energy consumption comparison of $\eta - GPS$, $\eta - nGPS$, $\alpha - GPS$ and $\alpha - nGPS$ (c) Best case of $\eta - nGPS$ energy vs accuracy vs activity (d) Worst case of $\eta - nGPS$ energy vs accuracy vs activity.	52
3.6	Energy performance per vehicle with four different sensor array pool schemes under best and worst simulation settings respectively (a) Best case energy consumptions of identical sensor array pool (i.e., combo), non-identical sensor array pool (i.e., combo) - 1, 2, and 3 (b) Worst case energy consumptions of identical sensor array pool, non-identical sensor array pool - 1, 2, and 3. . . .	53
3.7	Energy (per vehicle) and accuracy performance with identical sensor array pool scheme under best and worst simulation settings (a) Best case energy consumption vs accuracy vs activity (b) Worst case energy consumption vs accuracy vs activity.	53
3.8	Energy (per vehicle) and accuracy performance with three different non-identical sensor array pool (i.e., combo) schemes (non-identical sensor array pool - 1, 2, and 3) under best and worst simulation settings. Best case energy consumption vs accuracy vs activity with (a) non-identical sensor array pool - 1 (c) non-identical sensor array pool - 2 (e) non-identical sensor array pool - 3. Worst case energy consumption vs accuracy vs activity with (b) non-identical sensor array pool - 1 (d) non-identical sensor array pool - 2 (f) non-identical sensor array pool - 3.	54

4.1	GPS-less localization architecture based on unsupervised machine learning. .	58
4.2	Application of the crowd-sensing concept to a public transportation network containing m vehicles each of which moves on a static route of n stops / stations.	62
4.3	Accuracy performance under three varying simulation settings (a) The effect of cluster size on accuracy (b) The effect of route length on accuracy (c) The effect of sensing noise (n_a) on accuracy.	69
4.4	The effect of nearly even participant recruitment on localization accuracy under different sensor usage distribution scenarios and cluster sizes: (a) scenario-1 ($\eta - rec.1$) (b) scenario-2 ($\eta - rec.2$) (c) scenario-3 ($\eta - rec.3$).	70
4.5	The effect of uneven participant recruitment on localization accuracy under different sensor usage distribution scenarios and cluster sizes: (a) scenario-1 ($\alpha - rec.1$) (b) scenario-2 ($\alpha - rec.2$) (c) scenario-3 ($\alpha - rec.3$).	71
5.1	Trustworthiness-driven <i>GPS</i> -free localization architecture.	76
5.2	A generic flowchart for <i>RDNR</i> and <i>RDER</i> . <i>RDNR</i> and <i>RDER</i> have their own reliability assessment methods explain in detail in Section 5.3.2 and 5.3.3 respectively.	76
5.3	Accuracy performance of reliability-aware and reliability-unaware sensor recruitment policies under two varying parameters: (a) The effect of route length on accuracy (b) The effect of noise factor, $q_{\theta}^{u_p}$, on accuracy.	83
5.4	The effect of $l=10$ on accuracy with increasing number of participants. . . .	84
5.5	The effect of sensing noise ($q_s^{u_v}$) on accuracy (a) $q_s^{u_v} = 1.0$ (b) $q_s^{u_v} = 2.0$ (c) $q_s^{u_v} = 3.0$	85
5.6	Comparison of recruitment probabilities of <i>RDNR</i> and <i>RDER</i> participant recruitment scenarios.	86
5.7	Energy performance per vehicle under <i>RDER</i> participant recruitment scenario with different participant profiles (a) α -participants profile (b) η -participants profile.	87

Nomenclature

ITS: Intelligent Transportation System is application for enhancing the safety, mobility, and efficiency of ground transportation systems via sensing, analysis, control, and communication technologies.

WLAN: Wireless Local Area Network is a computer network to connect two or more devices via wireless communication within a restricted area such as home.

WPAN: Wireless Personal Area Network is a network to connect the devices centered around a person via wireless communication.

WWAN: Wireless Wide Area Network is a network to link the separate areas of coverage or cells wirelessly in order to service a large geographical region.

HetNet: Heterogeneous Network is a network that uses various communication technologies to provide service.

VSN: Vehicular Sensor Network encompasses vehicles that is capable of offering sensing services.

VANET: Vehicular Ad Hoc Network consists of vehicles communicating among themselves.

DSRC: Dedicated Short Range Communication is a two-way short- to- medium-range wireless communication methodology.

RSU: Road Side Units are the stations situated near the roads in order to carry information between vehicles to cloud center, and also store, process information.

BS: Base Station is a stationary point of communication.

V2X: Vehicle-to-X communication enables vehicular communication with anything such as an infrastructure, and vehicle.

IoV: Internet of Vehicle integrates three networks: an inter-vehicle network, an intra-vehicle network, and vehicular mobile Internet to provide large-scale distributed wireless communication for vehicles.

MCS: Mobile Crowd-sensing is a technique to acquire sensor data from a large group of individuals through the sensors on their mobile devices.

VCS: Vehicular Crowd-sensing is a technique to acquire sensor data from the sensors of individuals' mobile devices located in a large group of ground vehicles.

Non-dedicated Sensors: are the sensors owned and maintained by the participants of a smart city application and used in multiple smart city applications on demand.

CrowdSemSim: is a discrete-event simulator developed to be used in the context of crowd-sensing.

SDZ: Sensory Detection Zone is the probability of a vehicle to be present in a stationary region.

SVM: Support Vector Machine is supervised machine learning algorithm to classify the given dataset.

k-means: is an unsupervised machine learning algorithm to cluster the given dataset.

Window Size: is a set of landmarks on a bus route where *SVM* analyzes acquired sensor data.

Cluster Size: is a set of labels that *k-means* uses to group the acquired sensor data.

AWGN: Additive White Gaussian Noise is basic noise model based on normal distribution to simulate the effect of many random processes occurring in nature.

Identical Sensor Array Pool: is combination of sensors equally distributed in a vehicle.

Non-identical Sensor Array Pool: is combination of sensors unequally distributed in a vehicle.

Active Participant: is an individual who uses his/her smart device for both data acquisition and entertainment.

Inactive Participant: is an individual who uses his/her smart smart device for only data acquisition.

NRR: Non-restricted Recruitment accepts sensor readings of a participant without applying any trustworthiness assessment.

RDNR: Reliability-driven Naive Recruitment accepts sensor readings of a participant based on the participant's trust score and a random value generated for the participant.

RDER: Reliability-driven Exclusive Recruitment accepts sensor readings of a participant based on participant's trust score and average trust score of the vehicle where participant is located.

Publications of the candidate during MCS studies

Publications that are the direct outcomes of the thesis:

- C. Kaptan, B. Kantarci, T. Soyata, and A. Boukerche, “Emulating smart city sensors using soft sensing and machine intelligence: a case study in public transportation”, **accepted** in IEEE International Conference on Communications (ICC) 2018,
- C. Kaptan, B. Kantarci, and A. Boukerche, “Reliability-driven crowd-sensed localization of public transportation vehicles”, **accepted** in IEEE Global Communication Conference (GLOBECOM) 2018,
- C. Kaptan, B. Kantarci, and A. Boukerche, “Vehicular crowd-sensing for GPS-less tracking of public transportation vehicles”, **revision requested** by IEEE Transactions on Vehicular Technology,
- C. Kaptan, B. Kantarci, and A. Boukerche, “On the dependability of vehicular crowd-sensing data”, **to be submitted to** IEEE Transactions on Vehicular Technology 2018,
- C. Kaptan, B. Kantarci, A. Boukerche, and T. Soyata, “GPS-less location verification framework for public transportation vehicles against vehicle hijacking”, **to be submitted** to IEEE Transactions on Intelligent Transportation Systems 2018.

Publications that are co-authored and related to the thesis:

- H. Habibzadeh, T. Soyata, B. Kantarci, A. Bourkerche, and C. Kaptan “Sensing, communication, and security planes: a new challenge for a smart city system design”, **minor revision requested** by Elsevier Computer Networks.
- H. Habibzadeh, C. Kaptan, T. Soyata, B. Kantarci, and A. Bourkerche, “Smart city system design: a comprehensive study of the application and data planes”, **under review** in ACM Computing Surveys,

Chapter 1

Introduction

1.1 Motivation

Vehicular sensing, evolving communication technologies, and cloud-assisted computing pave the way for the development of smart city applications. Smart transportation is among various applications under the smart city concept. Smart transportation services include autonomous vehicles, as well as applications involving vehicular sensing and vehicular crowd-sensing. Vehicular sensing is to use vehicles on the road for collecting, processing, and sharing sensory data associated with location (e.g., road condition, traffic flow) [86]. Mobility of vehicles and their capability to acquire multimodal data make vehicular sensing attractive for smart transportation applications [130]. While vehicular sensing is capable of providing useful insights [87] to a broad spectrum of smart city applications for various purposes such as saving substantial amount of travel time through traffic congestion alert systems [167] or autonomous car parking systems [41] to facilitate the hassle of parking in big cities, it also offers cost-effective solutions for monitoring physical phenomena [80, 157].

Forming Vehicular Sensor Networks (VSNs) in smart transportation applications to share sensor data in urban environments is an emerging concept. Data from variety of sources in VSNs is sensed and transmitted among the vehicles and other entities such as Road Side Units (RSUs) and Base Stations (BSs). VSN takes benefit of the intelligent wireless information infrastructure and resource pooling to facilitate smart sensing, transmission and processing of sensory data [159]. While VSNs enable ubiquitous collection of sensory data with smart-devices and/or vehicles, the coverage in VSNs is affected by vehicle mobility. Hence, as a time-variant factor, *opportunistic coverage* must be taken into account in VSNs [179].

Moreover, huge volume of data exchange at significantly low latency and at higher reliability with the involvement of 5G network architectures paves the way for wide adoption of the Internet of Vehicles (IoV) concept [37]. Cooperative transmission with the help of cooperating small cell Base Stations makes it possible to improve the transmission capacity as well as reliability in 5G networks [52]. In the IoV concept, cooperation among the vehicles

is of paramount importance. With the advent of 5G mobile cellular networks, scalability and latency challenges experienced by IEEE 802.11p and Long-Term Evolution (LTE)-enabled vehicular networks will be addressed, particularly for real time multimedia sharing [44]. Evolving 5G communication models have enabled offloading and opportunistic spectrum sharing for vehicular nodes to provide real time infotainment services more efficiently [65].

The advent of mobile sensing and communication leads to effective and affordable solutions to large-scale urban sensing problems. Sensing in such way is often called by other researches as opportunistic sensing [32, 84], participatory sensing [84], mobile crowd-sensing [51, 125, 128], or people/human-centric sensing [138, 32, 84]. Smart-devices, such as smartphones and tablets, with variety of built-in environment and motion sensors (e.g., accelerometer, GPS, and microphone) are strong candidates of non-dedicated sensors in the context of smart city applications under VSNs. Furthermore, smart-devices offer multiple wireless and cellular interfaces, which makes them suitable for participatory vehicular sensing [31, 106, 103] and/or vehicular crowd-sensing [156]. A typical mobile crowd-sensing architecture consists of a sensing application running on the participants' smart-devices, a back-end server for aggregation and analytics of participants' sensed data [33, 124]. Indeed, the crowd-sensed data may also be pre-processed and analyzed at the device level, as well. Furthermore, crowd-sensing campaigns may require explicit involvement of participants such as uploading multimedia images/videos or providing direct feedback to real time queries [67].

Vehicular sensing and VSNs along with the integration of sensors and embedded computing devices have paved the way for the development of vehicular crowd-sensing services [51]. Crowd and vehicular sensing services supported by communication technologies allow individuals collaboratively gather/share data and also obtain information to measure and map phenomena of common interest [112]. Individuals can utilize mobile devices for their needs and let mobile crowd-sensing services develop into a wide range of sensing standards. For instance, the recent version of an iPhone (either 8 or X) is equipped with large amount of sensors (e.g., camera, microphone, accelerometer, GPS and so on) to collect and share sensory data with interested parties [169]. Furthermore, information is obtained from raw data through a plethora of analytics techniques as crowd-sensed data undergoes machine or deep learning processes.

Vehicles that can access vast amount of sensors (either embedded in the vehicle or carried by the individuals) enable the development of various services such as *GPS*-less vehicle localization. In public transportation, crowd-sensing can be exploited when estimating the approximate locations of vehicles. Indeed, a vehicle can be located by the Global Navigation Satellite System (*GNSS*) technology (e.g., *GPS* (the most commonly used)) attached to the vehicle. However, performance of GPS-based localization systems may be poor in the case of insufficiently weak satellite signals in the vicinity of the vehicle. Under such circumstances, GPS devices may perform poorly and consume more energy than they used to consume, which will eventually drain the battery of the device in which it is integrated.

Moreover, exact location of the device that hosts the GPS sensors that receives and transmits the satellite signals are likely to be disclosed, which raises privacy concerns. Location

approximation through *GPS*-less methods could solve the issues observed with the *GPS*-based localization. On the other hand, solutions that provide crowd-sensing-based *GPS*-less localizations may suffer from low accuracy due to distributed and heterogeneous nature of crowd-sensing system, and low dependability of the data during the analysis stage. Unless there exists a mechanism that detects and eliminates such kind of data, performance degradation in terms of location accuracy will be inevitable and observed. Trustworthiness concerns may generally raise in a crowd-sensing-based application due to either or combination of the following reasons: (i) malfunctioning sensors, (ii) malicious/untrustworthy participants.

In this thesis, we propose a series of frameworks that enables tracking a public transport vehicle without the need for a *GPS* device. The elimination of *GPS*-enabled devices will further help us overcome the issues (described above) introduced by *GPS*. We particularly study our proposed approach and analyze its effectiveness on public transportation vehicles as our simulation environment, *CrowdSenSim* [49], allows seamless integration of crowd-sensing technique on a public transportation network which we have designed in a way that vehicles only moves on unidirectional routes (meaning while a vehicle is moving on its route, no other vehicle is moving on the opposite direction and passing by that vehicle). The proposed architecture incorporates crowd-sensing approach with non-dedicated sensors (specifically accelerometer and microphone data) to acquire sensor readings from selected participants (a.k.a. participants) located in the public transport vehicles. To analyze the crowd-sensed multi-dimensional non-dedicated sensor data for localizing a vehicle in the vehicular network, we have adopted both supervised and unsupervised machine intelligence methodologies and compared their performance in terms of localization accuracy. Moreover, we have also addressed accuracy degradation problem introduced by the crowd-sensed data due to various reasons explained above. To this end, we have developed different reliability-aware participant recruitment techniques and compare the performance of each technique with reliability-unaware technique to see their impact on localization accuracy. The following Section 1.2 lists contributions of the thesis in detail.

1.2 Contributions

We have provided major issues that *GPS* might possibly introduce into a system, and recent enhancement in the communication technologies, large-scale sensing made possible by crowd-sensing methodology in the previous Section 1.1. Moreover, we have introduced the objective of this thesis at the very last paragraph, which is briefly to provide an efficient framework to approximate the location of a public transportation vehicle without using *GPS* data.

Our contributions to the similar studies in the literature could be summarized as follows:

- The first and most important contribution of this thesis is the development of various *GPS*-less public transport vehicle localization frameworks. Our proposed frameworks adopt crowd-sensing method to acquire non-dedicated sensor data for tracking the

location of a public transport vehicle. Non-dedicated sensor readings are collected from participants on-board in vehicles. Furthermore, acquired data is analyzed with machine intelligence methodologies to finalize the process of approximating a vehicle's location.

- Non-dedicated sensors: (i) accelerometer, and (ii) microphone are selected to acquire data from the participants. When selecting aforementioned sensors, we have considered their ease of availability on smart devices currently in the market. Upon analyzing the acquired data transmitted from participants in vehicles, we were able to locate the whereabouts of vehicles.
- *GPS* shares direct location of individuals with the platform it is integrated into. Therefore, elimination of *GPS*-enabled devices also promotes privacy protection of the participants, who are willingly contributing their sensible information with the platform. In the proposed architecture, we provide the privacy protection by acquiring data from non-dedicated sensors that do not contain any information violating the individual privacy.
- Elimination of *GPS* does not only promote individual privacy, but also provides less energy consumption, which becomes very crucial if we consider *GPS*-enabled devices run on batteries. We have conducted extensive experiments on energy consumption of our proposed non-dedicated sensors. Several sensor recruitment techniques (encompasses various combinations of proposed non-dedicated sensors) have been tested and results prove us that all sensor recruitment techniques consume less energy when compared with the energy consumption of *GPS* performing the same task for the same duration.
- Acquired crowd-sensed, non-dedicated sensor readings might contain untrustworthy data, which causes degradation in location accuracy. To ensure the reliability of participants, we have introduced a reliability-aware participant recruitment scheme in [75]. Moreover, we have extended this framework and introduced an additional reliability-aware participants recruitment scheme in [76]. We have compared the results of reliability-aware user recruitment schemes with reliability-unaware user recruitment scheme to determine the performance improvement in terms of location accuracy.
- A simulation environment is set-up to conduct the experiment on large-scale scenario, where multiple public transport vehicles moving on multiple routes. Non-dedicated sensor readings are collected in real-time from one vehicle with multiple devices and they are used to emulate the sensor behaviors in the simulation environment. It is also worth noting that the communication flow in our simulation environment is unidirectional between the participants in public transport vehicles to data center.

1.3 Outline

Organization of this thesis is as follows. In Chapter 2, we survey enabling communication infrastructure and technologies for smart transportation in Section 2.1 (more specifically wireless communication technologies in Sub-Section 2.1.1, and Vehicular Ad Hoc Networks in Sub-Section 2.1.2, we introduce smart transportation services in Section 2.2 (traffic management service in Sub-Section 2.2.1, public transportation management in Sub-Section 2.2.2, emergency management in Sub-Section 2.2.3, and information management in Sub-Section 2.2.4) and service provisioning methodologies (Sub-Section 2.2.5), we list some vehicular crowd-sensing applications for smart transportation services in Section 2.3, and finally we categorized smart transportation applications under the specified services with regard to machine learning, deep learning, and signal processing methodologies. In the subsequent three chapters (Chapter 3-5), we introduce our proposed methodologies to overcome aforementioned issues in Section 1.1.

In Chapter 3, we propose a public transport localization via vehicular crowd-sensing with a supervised machine learning method: Section 3.1 defines the problem and propose a solution, Section 3.2 introduces the high level architecture of the proposed supervised public transport localization approach, Section 3.3 describes the methodologies adopted in detail, Section 3.4 presents the experiments results, and Section 3.5 concludes the chapter with final remarks.

In Chapter 4, we design a public transport tracking scheme via vehicular crowd-sensing with an unsupervised machine learning approach: Section 4.1 introduces the problem and proposes a solution, Section 4.2 describes the proposed unsupervised tracking architecture, Section 4.3 explains the techniques involved to build the proposed system in detail, Section 4.4 presents the results of performance analysis, and Section 4.5 wraps up the chapter with final remarks.

In Chapter 5, we propose a reliability assessment mechanisms to the existing unsupervised architecture to ensure trustworthiness of crowd-sensed vehicle tracking service: Section 5.1 introduces the problem and proposes a solution, Section 5.2 illustrates and defines the high level architecture, Section 5.3 describes methodologies to ensure trustworthiness in detail, Section 5.4 presents the results of experiments, and Section 5.5 concludes the chapter with final evaluations.

Finally in Chapter 6, we wrap up the thesis with our final comments, discuss about the issues that still needs to be addressed, and give future directions.

Chapter 2

Literature Study

Developing automated vehicle functions and connectivity between the vehicles and the surrounding environment took attention starting from 2000s [72] and today there exist adequate amount of technologies emerged from these efforts. *IoT*-enabled Intelligent Transportation Systems (*ITS*) are some of the products of these endeavors. The main building blocks of *IoT* encompasses cloud-based host center, gateways, routers, access points, and switches. These devices facilitate and provide system interoperability with the methodologies such as the vehicle-to-x (*V2X*), which comprises of vehicle-to-vehicle (*V2V*) and vehicle-to-infrastructure (*V2I*) methods. People, machines, and devices can also be connected to the network and be a part the network. They are equipped with the tools and technologies enabling them to exchange data without any human or computer interaction. Data transaction between objects in the network in an *IoT* system is enabled by utilizing several communication technologies.

Mobility in smart transportation solutions is the key factor when determining the suitable communication technologies that can be leveraged in the context of *ITS*. The wireless networks for *ITS* empower fast communications between vehicles and the roadside infrastructure and buildings. Therefore, they are more convenient to integrate in smart transportation solutions. On the other hand, the coverage range of the wireless devices may induce network availability problems. However there exist various solutions to overcome these type of problems such as installing multiple roadside nodes to let the information pass from one node to another and be delivered to its destination [10].

Communication among nodes *IoT*-based projects including *ITS* is crucial. Without transferring data from one end to the other, it is impossible to provide interoperability among the devices, applications, and services, which leads an overall non-functional system. In the following sub-sections, we will investigate the enabling communication infrastructures and technologies behind *ITS*, smart transportation applications and services, and finally vehicular crowd-sensing and data analytics for smart transportation applications and services.

2.1 Enabling Communication Infrastructures and Technologies for Smart Transportation

Communication is an integral component of smart transportation applications to provide their services as designed. It is also a fundamental criterion when assuring the *QoS*, especially for mission critical applications such as safe and security applications. In this section, we will explain the following communication methodologies: Wireless, Vehicular Ad Hoc Network, and Dedicated Short Range Communication. Furthermore, we will list various communication standards/protocols under each methodology.

2.1.1 Wireless Communication Technologies

Communication technologies that are being used in smart cities can be categorized into three general categories with regard to their coverage area: **(i)** Personal Area Network (PAN), **(ii)** Local Area Network (LAN), and **(iii)** Wide Area Network (WAN). *PAN* is a computer network that has the minimum coverage capacity among the others in the list. It is used for data transmission and communication among devices such as phones, tablets, computers, and gateways or for connecting to higher level network such as WAN via a gateway in its coverage area. *(LAN)* establishes connection among devices within a limited area such as residence, school, and office, etc. *Ethernet* and *WiFi* are two well-known technologies of this network architecture. *WAN* is a private communication network interconnecting multiple LANs scattered across cities, states, or countries. A router or other multifunction device is used to connect a LAN to a WAN. The most well-known example of *WAN* is the *Internet*. Moreover, Virtual Private Network (*VPN*) technology is adopted to provide a secure and protected communication.

If the data transmission or communication is conveyed wirelessly, those technologies are considered as Wireless Personal Area Network (*WPAN*), Wireless Local Area Network (*WLAN*), and Wireless Wide Area Network (*WWAN*) respectively. Therefore, communication technologies that can be adopted in smart transportation application and service can be categorized into four categories: **(i)** Wireless Personal Area Network (*WPAN*), **(ii)** Wireless Local Area Network (*WLAN*), **(iii)** Wireless Metropolitan Area Network (*WMAN*), and *(iv)* Cellular Network.

WPAN: *WPAN* is based on IEEE 802.15 and is used for interconnecting the individuals workplace devices. Bluetooth, Zigbee and Z-Wave are some examples of *WPAN*-based technology. Bluetooth (802.15.1), and Zigbee (802.15.4) are considered as low-range communication technologies and more suitable for personal-area-network-based applications. On the other hand, Z-Wave is a low-power RF communications technology designed for home automation [170].

WLAN: *WLAN* enables mobile users to connect *LAN* through a wireless medium. *IEEE 802.11* is set of standards outlining the technologies for *WLAN*s. **WiFi** is a *WLAN*

technology offering wireless connection to devices equipped with *WiFi* adapters in hotspot locations (where *WiFi* access is available). Data transmission frequency is at the level of 2.4 GHz or 5 GHz. Data transmission frequency level and maximum data transmission rate per second (i.e., bandwidth) determine the type of *802.11* standards:

- **802.11a:** transmits the data at the frequency level of 5 GHz. Maximum data rate is 54 megabits of data per second. Higher frequency compare to other standards under the same family decreases the range of service area and increases the possibility of having data transmission difficulties due to walls and obstacles present in the environment.
- **802.11b:** transmits the data at the frequency level of 2.4 GHz. Maximum data rate is 11 megabits of data per second. Low frequency increases the range, provides a robustness against the obstacles around, and decreases the production cost.
- **802.11g:** transmits the data at the frequency level of 2.4 GHz with a maximum data rate of 54 megabits of data per second. Good service range along with a better maximum speed and obstacle penetration makes *802.11g* favorable alternative to *802.11a* and *802.11b*. *802.11g* is backwards compatible with *802.11b*, which means that it is designed to work with *802.11b* devices.
- **802.11n:** transmits the data at the frequency level of 5 GHz with a maximum data rate of 300 megabits of data per second. It provides the fastest maximum data transmission rate, the best signal rate, and more resistant to signal interference among the others under the family of *802.11*. It is backwards compatible with *802.11 b/g* devices.
- **802.11ac:** leverages dual band wireless technology, which enables simultaneous connections on both the 2.4 GHz and 5 GHz *WiFi* bands. It can provide data transmission rate up to 1300 megabits per second on the 5 GHz band and 450 megabits per second on 2.4 GHz band. It is backwards compatible with *802.11 b/g/n* devices.

Although the *WiFi* protocol stack (*802.11a/b/g/n/ac*) is not architecture to use in the vehicular mobility applications and services, numerous modifications and enhancement to the *IEEE 802.11* standard enables *WiFi* to be adopted in the vehicular mobility context such as rapid re-authentication is enabled with *IEEE 802.11r* [107], improved security is provided with *IEEE 802.11i* [40], protected frame management is provided with *IEEE 802.11w*, and Quality of Service (*QoS*) is enhanced with *IEEE 802.11e*.

WWAN: *WWAN* is a wide area network in which separate areas of coverage are connected wirelessly to provide service to a large geographic area. Its services are provided by cellular service providers to the devices that are equipped of *WWAN* cards to communicate with the other devices. 3rd Generation Partnership Project (3GPP) and WirelessMAN (i.e., *IEEE 802.16*) are two big families specifying the standards for *WWAN*.

- **3GPP:** is a collaborative project involving set of telecommunication companies whose goal is to develop specifications for third generation (3G) mobile systems accepted

across the globe. The scope of *3GPP* is later broadened to encapsulate the development and maintenance of *GSM* and related 2G and 2.5G standards including GPRS and EDGE, (Universal Mobile Telecommunications System (UMTS)), which is based on *GSM*, and related 3G standards including HSPA, *LTE* and related 4G standards including *LTE Advanced* and *LTE Advanced Pro*, and *Next Generation* and related 5G standards.

- **WirelessMAN:** is an array of *IEEE 802.11* wireless broadband standards. *802.11* has developed with the aim of offering a high-speed, high-capacity, low-cost, scalable solution to wireless network needs. *Worldwide Interoperability for Microwave Access (WiMAX)* is commercialized name of this set of standards. *WiMAX* provides some great benefits to satisfy the mobile user needs. However its drawbacks can also be observed when it is being used by mobile users such as range limitation (closer to a (WiMAX) source, better the data transmission rate) and performance issues when many users connected to the same *WiMAX* station.

Heterogeneous Network: Heterogeneous Networks (*HetNets*) are composed of multiple types of access nodes in a wireless network. To provide wireless coverage in various environments (e.g., indoor locations such as homes, offices and malls; underground locations such as subways and tunnels; and outdoor locations such as roads, and city centers.), a *WAN* can take advantage of macrocells, picocells, and/or femtocells. From an architectural point of view, *HetNets* are built on top of the functionalities offered by conventional macro radio access network (*RAN*), *RAN* transport capability, small cells, and *WiFi*.

Vehicular networks, such as *VANETs*, are powered by a range of wireless communication methodologies. However, high velocity, smaller coverage range, and mobility patterns along with explosion in mobile traffic cause major challenges for *VANETs* when offering a seamless handover, resource allocation, and moving the service from the previous *RSU* to the new *RSU*. Hence, it is crucial to establish a proactive handover and resource allocation models for *VANET* systems to provide a stable framework for cooperative applications using *HetNets* [54].

5G: The recent enhancements on the *5G* network are anticipated to provide support for a large amount of data traffic and a massive number of wireless connections, and to attain better efficiency in terms of cost and energy while offering a better *QoS* with regard to communication delay, reliability, and security. To this end, employing new developments both in deployment of super-dense and heterogeneous cells, and massive antenna arrays (massive multiple-input, multiple-output (*MIMO*) technologies) and in usage of higher frequencies – especially millimetre-wave (*mmWave*) – are required to achieve in the *5G* wireless networks. In [25], authors explain the potential advantageous and issues of the *5G* wireless *HetNet* integrating massive *MIMO* and *mmWave* technologies. On the other hand, network operators have to cope with many serious challenges in terms of operation, management, cost-effective small cell deployment, and inter-cell interface mitigation in order to provide a large-scale *HetNets* (e.g., a massive *VANET*) functioning steadily, and robustly. The authors in [176] propose a cloud-based platform that solves the aforementioned challenges. The

proposed platform aims to simplify the deployment, operation, and management, and ease round-the-clock optimization of the network in order to mitigate the development of 5G.

2.1.2 Vehicular Ad Hoc Networks

Vehicular Ad Hoc Network (*VANET*) creates a mobile network that uses moving vehicles and roadside infrastructures (i.e., Roadside Units – *RSUs*) as a node. This mobile network is capable of covering a wide area. Driver’s safety and mobility can be enhanced by direct or indirect communication between two vehicles or between vehicles and *RSUs*. Other applications that are leveraging the communication techniques used in *VANET* could be cooperative driver assistance, decentralized probe vehicles, and user and information communications [98]. *VANET*s consist of two communication methods that are based on wireless local area network (*WLAN*) technologies: **(i)** Vehicle-to-vehicle (*V2V*), and **(ii)** vehicle-to-infrastructure (*V2I*) [66]. In general, they are under the category of vehicle to everything (*V2X*) infrastructure [164]. This communication flow enables information sharing among vehicles and/or *RSUs*, which makes the vehicles more knowledgeable about the surrounding and nearby vehicles, as well as potential risks. Therefore, this can particularly improve accident avoidance, congestion awareness, and efficient operation of vehicles in the network of transportation.

The vehicles in the *VANET* (e.g., cars, trucks, and buses) with the help of *V2V* infrastructure are able to communicate with each other and exchange data (e.g., direction, speed, acceleration, vehicle size, etc.) in order to forecast and notify the safety threats and potential accidents such as collisions [64]. The vehicles may launch series of protective actions in accordance with the predictions well ahead of human response in some extreme cases for the purpose of avoiding (at the best case scenario) or minimizing (at the worst case scenario) the risk of injury and property loss. To provide those benefits, each vehicle should be equipped with transceiver to become a part of the *VANET*. Vehicles equipped with transceivers and sensors can collect information from multiple sensors, transmit and receive sensory data, and response to the actions (as described above).

The vehicles and *RSUs* (e.g., traffic signals, road signs, road lights, etc.) with the *V2I* infrastructure are capable of exchanging information. The *RSUs* as infrastructure elements can be used to assist the data transmission among vehicles, which helps vehicles to be aware of their driving environment and extends the communication range as well. Moreover, information collections along the roadway (such as warning of curve speed, signal violation, and the acknowledgement of road signs) can be facilitated for users with the help of *V2I* infrastructure. There are various physical components (e.g., routers, gateways, etc.) and communication technologies (e.g., WiFi, cellular networks, Bluetooth, etc.) behind *VANET* network. In the next sub-subsections, we will cover the common communication technologies available today.

Dedicated Short Range Communication

Dedicated Short Range Communication (*DSRC*) is used to make two-way wireless connections between vehicles and *RSUs*. It is operating in the 5.8 GHz in the U.S. and 5.9 GHz in Japan and Europe. There exists three ways of data exchange among *DSRC* devices: (i) on-board units (*OBUs*) that are situated inside the vehicles, (ii) *RSUs* that are situated on the side of the road, or (iii) hand-held devices that are carried by the pedestrians. The spectrum of *DSRC* is particularly convenient for the communication that takes place in the *V2X* network as it is capable of supporting very low-latency, secure data transmissions, fast network acquisition, and rapid/frequent handovers which are part of vehicular environments; while maintaining its robustness under adverse weather conditions and being tolerant of multi-path data transmission. The listed advantages make *DSRC* an eligible and favorable communication method for many applications in smart transportation systems. Those applications include but not limited to vehicle safety, traffic management, and infotainment [68, 165, 116, 60, 17].

Communications in *V2X* networks are based on of the two main technologies: *DSRC* and cellular networks. However, a variety of expected *V2X* applications for large number of vehicles should be able to support both technologies in order to offer a better network infrastructure. For this matter, inter-working between *DSRC* and cellular network technologies for efficient *V2X* communications has been proposed in the survey paper [4]. Authors have been studied potential *DSRC* and cellular inter-working solutions to discover effective *V2X* communication techniques.

LTE-based Communication

V2V communication is one of the main components within the *LTE*-based *V2X* communications framework. *LTE*-based *V2V* communications can be achieved by employing a cellular interface between the network and the user equipment (*UE* which is often referred as up-link and down-link or *Uu* interface) or direct radio interface between the *UEs* (which is often referred as *LTE* device-to-device, side-link, or PC5 interface), or by employing both of these methodologies. Cloud and *IP*-based *V2X* services can further be enabled with the *eNB* being connected to the evolve packet core (*EPC*). This powerful and rich capability of *LTE* network provides support for a wide range of *V2X* use cases with various distinctive requirements on data rates, latency, reliability, coverage, and so on.

The access layers of *DSRC* are established on the *IEEE 802.11* standards [58] which consists of an amendment addressing *V2X* communications. This series of improvements on *IEEE 802.11* for *V2X* communications are referred as *IEEE 802.11p*. In [24], authors present a comparison between *LTE* side-link-based and *DSRC*-based *V2V* where authors explained that the former outperforms the latter with regard to packet reception ratio versus communication range in typical scenarios. In addition, authors in the article [160] have studied the performance of different options using *LTE Uu* interface and analyze the results by comparing them to *LTE* side-link-based and *DSRC*-based *V2V* communications.

2.2 Smart Transportation Applications and Services

The US Department of Transportation (DoT) [2] described the key services that the smart transportation system should consider in order to offer an efficient and safe smart transportation system: (i) travel and traffic management, (ii) public transportation management, (iii) emergency management, and (iv) information management. Providing these services can improve the quality of transportation in terms of various ways such as travel safety, reduction of travel time and gas use, etc. In the following sub-sections, we will examine case studies (i.e., applications) conducted under specific service categories.

2.2.1 Traffic Management

A privacy-preserving route reporting scheme for traffic management using *VANETs* has been proposed by the authors in the paper [126]. Two different approaches have been provided to meet the cases where the *VANET* formation and management with *RSUs* and where no *VANET* infrastructure is provided and the *VANET* has to self-organize. Vehicles' routes are aggregated because the traffic management systems essentially need to know the total number of vehicles in each road segment and do not necessarily need to know the vehicles' individual routes. Their first design of the proposed approach is called Privacy Preserving Route Reporting for Infrastructure-based *VANETs* (*PRIB*). According to this design, each vehicle uses an encryption technique called homomorphic encryption to encode its route. Then, vehicles transmit their ciphertext to an *RSU* in their range and the *RSU* aggregates the route of the vehicles and conveys the aggregated ciphertext to the traffic management center (*TMC*). The *TMC* is responsible for decoding the ciphertext and compiling traffic guidance reports generated with the information of the total number of vehicles in each segment. Then, it transmits the generated reports to the vehicles. Their second approach encompasses two different designs related with the self-organization of *VANETs*: (i) elliptic curve cryptography (*ECC*) point-addition-based route sharing scheme for self-organizing *VANETs* (*EPSO*) and (ii) homomorphic encryption-based route sharing scheme for self-organizing vehicles (*HESO*). The former scheme forms a ring with a group of vehicles in which each vehicle shares keys (which are used to obfuscate routes) with neighboring vehicles. The latter scheme takes advantage of homomorphic encryption without the need of sharing keys with neighboring vehicles. Creation of a platoon of vehicles is done upon the request of vehicles chosen as a leader.

On the other hand, authors have proposed a traffic management system by employing a method that leverages the usage of Unmanned Aerial Vehicles (*UAVs*) [100]. *ITS UAVs* can offer an efficient mechanism by enforcing the traffic rules and supporting the traffic police on the ground, and as well as contributing road users' information related with traffic such as intelligent traffic management. *DSRC* interface can be used to enable the operations of *ITS UAVs*. *DSRC* interface provides information exchange via *V2V* and *V2X* communications. *ITS UAVs* can communicate through a direct wireless link with the vehicles in the range with

the enablement of *DSRC* technology, which in turn better enforces road safety and provides traffic efficiency. Applications that can be made possible by *ITS UAVs* include, but are not limited to, the following: flying accident report agents, flying *RSUs*, flying speed cameras, flying police eyes, and flying dynamic traffic signals. These examples incorporate more than one *UAV* to fly jointly, cooperate, and coordinate to execute a specific mission. When operating a task collectively, they consume energy all together as well. However, *UAVs* could effectively use their energies by perfectly sharing the missions among all *UAVs* and hence overcome the issues related with the battery limitation with optimal coordination algorithms. For example, in the case of a traffic accident, one of the *UAVs* could fly to the accident location and prepare a report/alarm (e.g., video) and send the created report/alarm to the other *UAVs* via device-to-device (D2D) multihop communications. Meanwhile, another close by *UAV* or *RSU* that has network access via cellular wireless networks (e.g., 4G, LTE) can transmit the report to the relevant entity. One of the challenges that needs to be accomplished when integrating *UAVs* into *ITS* is to design an optimal coordination algorithms.

2.2.2 Public Transportation Management

Management of public transportation systems can be facilitated via surveillance systems. Another application of *UAV* in the area of *ITS* for the purpose of understanding the current state of a transportation system has been studied in [136]. An urban surveillance system with monitoring drones has been proposed by the authors. Support for urban computing architecture and on-demand execution of the distribution of data to the intended end-users can be realized with the coalescence of *ITS* with drones. However, energy needs for the operations of drones and their communications with the sensors at the ground may rise a problem and the designed *ITS* system with drones may result in obstructed transmissions. All the resources of the system may drain in an expedited manner in such a situation and the system may fail eventually. Energy overhead is resolved by integrating *LoRaWAN* protocol and using drones as on-demand *LoRaWAN* gateways to fortify energy efficient surveillance in *ITS*. Moreover, the lifetime of the network with a fail-safe and resilient strategy enabling the usage of drones in an energy-deficit urban environment is improved with the proposed approach.

Another service that should be offered under public transportation management is the payment management through mobile payment methods or smart card fare collection. In [89], authors propose a mobile payment protocol which considers limited connectivity for *VANET* applications. The proposed payment protocol aims to hold following four requirements: (i) adaptation to the real communication scenario, (ii) computation- and communication-wise performance efficient, (iii) fair exchange rates and anonymity of users, and (iv) secure payment in terms of confidentiality, non-repudiation of origin, robustness to traditional security attacks, and preventing over and double spending. A self-certified key agreement for producing the symmetric encryption keys is adopted in this work in order to incorporate the advantages of both public-key cryptography and symmetric-key cryptography. Moreover,

a new session key generation is triggered after every payment transaction to provide more robust security protection.

2.2.3 Emergency Management

Emergency management is an essential service that needs to be addressed in the context of smart transportation. With the recent advent on the vehicular communication, and the integration of machine intelligence methodologies into vehicular environment, applications such as traffic accident warning and prediction provide secure and safe driving experience on the road. Emergency management is an essential service that needs to be addressed in the context of smart transportation. With the recent advent on the vehicular communication, and the integration of machine intelligence methodologies into vehicular environment, applications such as traffic accident warning and prediction provide secure and safe driving experience on the road. The study [1] proposes a data collection, and analytics application that is capable of predicting the behavior of traffic users. Moreover, the proposed application is empowered with traffic evacuation and sending emergency vehicles (e.g., ambulances, fire trucks, etc.) to help people that are in need such assistance. Importance of such an application becomes more crucial considering the fact that accidents may happen at any time at any spot in the traffic due to various reasons such as fires, terrorist attacks, and earthquakes. These reasons not only lead to traffic accidents, but also cause panic behavior and unexpected traffic congestion. The architecture of proposed system encompasses three layers: (i) Sensor Processing (*SP*) layer, where all possible real-time data feed into the transport system, (ii) Intelligent Data Processing (*IDP*) layer, where real-time data collected from a diameter of approximately 1 km is initially processed (*IDPs* are equipped with small but computationally powerful units), and (iii) Cloud Processing (*CP*) layer, where the services in different cloud servers are provided. Furthermore, single point of failure risk is minimized by stationing servers at different locations in the city. *IEEE 802.11p WAVE* is used as a communication standard. While *RSUs* acquires vehicular data, *IDPs* acquires both vehicular data and data collected from on-road sensors, public transport vehicles, and local authorities such as traffic centers, transport coordination units of cities. Planting *IDPs* not just on roads but also at public transport stations enables the collection of these broad-spectrum of data.

2.2.4 Information Management

Vast amount of information acquired from vehicles flows from one end to another in smart transportation systems. Vehicles share sensitive information with the applications that offer variety of services. Therefore, one of the fundamental tasks of an information management service is to protect the data carried with any vehicular communication means and stored in the data centers. The paper [88] introduces a novel privacy-preserving incentive announcement network established on Blockchain – a decentralized storage method – called CreditCoin. The proposed network encompasses two parts, (i) announcement protocol, and

(ii) incentive mechanism. The proposed network is a pioneer privacy-preserving Blockchain-based incentive network in the context of *VANETs*, which is capable of building trust in communications of smart vehicles. While efficient and privacy-preserved announcement forwarding is attainable with the Echo-Announcement protocol, users control their reputation points with earning and spending coins as incentives with the Blockchain incentive mechanism designed. CreditCoin is capable of proving the aforementioned services while preserving privacy, achieving anonymity, and blocking security attacks. The proposed methodology is developed in a network simulation environment – Network Simulator 2 – that runs on Java Runtime Environment 1.8. Efficiency and practicality of CreditCoin in smart transportation and smart vehicles applications is proven with the simulations results.

2.2.5 Service Provisioning Methodologies in Smart Transportation

The fundamental services aforementioned in Section 2.2 act as a guideline when transforming a city’s conventional transportation network into a smart transportation network. The functionality of each service varies from one another depending on the way it is provided such as content sharing between two cooperating vehicles, autonomous driving of a vehicle that is fully aware about its environment thanks to vehicular communication between other vehicles and roadside infrastructures in the perimeter. Moreover, vehicles equipped with the wireless communication capabilities (e.g., WiFi, cellular networks) are increasing in number as a result of rapid development of vehicular cyber-physical systems (*VCPS*s), which leads such vehicles to access to the numerous mobile services or applications including safety and multimedia content sharing applications.

Vehicles that are part of a smart transportation network are also equipped with mobile storage and/or processing unit, which makes them capable of offering services to the other applications. However, involvement of many challenges – introduced by the mobile nature of the vehicular networks and lack of control over the data stored and processed in the cloud environment – requires extra attention to the privacy, service cost and latency aspect of the service offered by vehicles in a smart transportation network [133, 180].

In order to address the privacy, service cost, and latency issues in the vehicular network, an auction-driven, multi-objective provisioning framework with the existing trusted third party approach and the Quality-of-Experience (*QoE*) model has been proposed in [8]. However, auction model proposed in this work is disadvantageous in terms of delay and action complexity due to its $n \times n \times n$ modelling structure. Delay caused by the action model increases the services charges of a driver as how much a driver will pay is determined with elapsed time that the driver used the specific service. The study [9] introduces game theory concepts with a *QoE*-awareness system model to overcome the issue of latency, cost, and privacy. Proposed Multiagent/Multiobjective Interaction Game System (*MMIGS*) framework offers on-demand provisioning services in a vehicular environment at low latency, minimum cost, and minimal driver information disclosed. In each game, collected requirements and

preferences by the *QoE*-awareness system creates a *QoE* value for every service provider, trusted third party and driver.

Miscellaneous services provisioned by heterogeneous vehicular networks are targeted in [181]. A generic *HetVNET* framework based on service and user cases analysis is offered after the evaluation of numerous possible networking techniques. Number of typical application scenarios are examined with the proposed framework, which leads to the discovery of key solutions and challenges of the Medium Access Control (*MAC*) and networking layers along with several open research issues for future investigations in *HetVNET*s.

Due to the swift growth in mobile data traffic, an extreme drop in the quality of service provision is observed [161]. To overcome this problem and maintain an effective quality of service (*QoS*), cellular network capacity was increased by adopting variety of techniques such as developing base stations with a smaller cell size (e.g., picocell and femtocell) enabling us to install more base stations, and deploying 4G or LTE networks [127], which provide better *QoS*. However, extensive usage of advanced vehicular application such as video sharing, cloud applications, and big data services when combined with vehicular capabilities will boost mobile traffic data even more. Therefore, the need for offloading solutions is growing with the boost in mobile data traffic, which pushes operators to provide various offloading schemes in *VCPS*s.

Authors of the paper, [161], proposes a different approach to provide mobile data traffic offloading in *VCPS*s. What makes their approach different from the previous studies is the removal of the assumption of offloading capacity homogeneity and the global *QoS* guarantee of *QoS*-aware service provision. Initially, two famous offloading models, (i) Wi-Fi, and (ii) VANET, are formed and the offloading capacity of each *Wi-Fi AP* or vehicle is quantified. A formulation for multi-objective combinatorial optimization problem – which intends to simultaneously optimize possibly conflicting objectives including efficient use of multidimensional resources, achieving satisfactory global *QoS*, and lowering mobile data traffic – is, then, derived from the mobile data traffic offloading. In order to obtain global optimal solutions within a reasonable runtime, a mixed-integer programming search method is used.

2.3 Vehicular Crowd-Sensing for Smart Transportation Applications and Services

2.3.1 Traffic Management

Acceleration in the development of plethora of technologies and the design of intelligent system promote cleaner, safer, and more efficient transportation. The realization of Internet of Vehicles (*IoV*) is crucial for clean traffic environment in the future. Even trustworthy forecast of future traffic condition could potentially save travel time, decrease pollution, and enhance traffic efficiency. Traffic congestion can be mitigated with the power of Mobile Crowd-Sensing (*MCS*) methodology, which leverages the potential of mobile devices (e.g.,

smartphones and/or sensor-equipped vehicles) in the context of vehicular environment. Traffic monitoring systems enabled with such sensing paradigm acquire the traffic data from the mobile devices of participants (e.g., drivers) and analyze the traffic data to notify drivers or relevant traffic authorities about the upcoming traffic situation. Furthermore, with the involvement of vehicular communication, real-time road status and vehicle status can be obtained from the roadside units and this information can help estimating the average speed of vehicles and other miscellaneous useful information.

The study in [155] introduces a technique that uses data acquired from connected vehicles in order to enhance transportation efficiency. New algorithms are applied to estimate and forecast the travel times conveniently. In addition, forecasting the accidents enables dynamic route choices for drivers to prevent traffic congestion. Furthermore, proposed architecture can be expanded to operate on a larger scale with the power of smartphones and/or sensor-equipped vehicles (that provide *MCS*). To give an example to solidify the benefits of *MCS*, traffic data can be collected for detecting the congestion level by utilizing the smartphones carried by drivers instead of planting cameras and loop detectors on road-sides. This way, building specialized sensing infrastructure will be alleviated and the cost of such infrastructure will be decreased. The research in [156] also leverages the power of vehicular crowd-sensing in the context of evaluation of traffic condition. Estimating the average travel speed for road sections is advantageous in assessing the management of transportation, and so on. Real-time locations and velocities of vehicles can help estimate the travel speed of a vehicle and acquisition of such data is possible through vehicular crowd-sensing. Since vehicular crowd-sensing data is generally coarse-grained as a result of irregular spatial-temporal distribution of vehicles and the alternating data-offering intervals, authors introduce a method called Spatial-Temporal Correlation (*STC*) to overcome the coarseness problem. *STC* generates a fully covered distribution of travel speeds for road sections in time and space domains with regard to crowd-sensing data.

Operating of road networks becomes costly, time consuming, and environmentally unfriendly (with the increasing urban pollution) in case of an accident or event that leads to anomalies causing traffic congestion. Furthermore, traffic patterns indicate important information about how people live in the city. For instance, people are attracted and gathered in one location when there is a popular sporting event, which may increase the traffic activity and lead to congestions or protests may cause road closures, etc. Authors in [115] introduce a technique for identifying and describing the traffic anomalies, which might happen due to various reasons such as traffic accidents, traffic controls, celebrations, protests, disasters, etc. Proposed technique takes advantage of human mobility data (e.g., GPS trajectories of vehicles) and social media data. To this end, GPS trajectory is used to identify the significant routing changes, and the subgraph of the road network where the anomaly is detected is used to fetch the social media information in the perimeter to define the anomaly. Authors classified the usage of the proposed system under two category: (i) user oriented, which detects the anomalies around the user, and (ii) city planner and traffic controller oriented, which help city planners and traffic controllers to monitor and visually analyze the anomalies.

2.3.2 Public Transportation Management

Transportation systems, particularly the bus system, are effected severely from traffic congestion in rush hours. Long waiting durations for a bus due to unforeseen events (including traffic congestion) force people to select another transportation means for travelling to their destination. An approximate estimation on the whereabouts of bus and when it will arrive the specific bus stop can alleviate the waiting time, enhance the performance of bus riders, and thus attract progressively more people to prefer bus transportation. Moreover, transit agencies can profit from offering real-time traffic and arrival time prediction. One of the ways that transit agencies use to track the vehicles to offer these services is installing GPS-enabled in-vehicle devices to their vehicles. However, due to well-known disadvantages of GPS-based devices on energy consumption, requirement of sky visibility to approximate the location better, deployment and maintenance costs that are not profitable for small-scale transit agencies and so on, building a GPS-dependent tracking and arrival time prediction in real world is not the best option available.

The paper [177] introduces two energy-efficient tracking and prediction of arrival time in urban areas frameworks for buses: (i) *WiLocator* and (ii) *WiLocator(p)*. While *WiLocator* offers real-time tracking and prediction of the arrival time of a bus in urban environment settings, *WiLocator(p)* expands the capabilities of *WiLocator* method to track a bus by calculating the dominated road sub-segment for each *WiFi* Access Point (*AP*) with regard to the crowd-sensed historical data. *WiLocator(p)* minimizes the computational delay, and enhances the positioning accuracy while providing an offline learning phase when compared with *WiLocator*.

Authors in [48] study on a simulation environment based on an urban parking case study, which is capable of analyzing the behavior of the crowd and evaluating the critical user with regard to certain crowd-sensing scenarios. The use case scenario adopted by the authors in this work is parking assistance for drivers in a busy urban environment. The simulator enables drivers to pick a geographic region, and then – based on the selected region – it computes and build a model of certain crowd generated events (such as leaving or occupying a parking spot), and eventually simulates the anticipated behavior of the crowd of drivers, and evaluate the gains (such as duration of finding an available parking spot). Moreover, parking related events are sensed and share with other drivers by the smart citizens. In addition, these events can be displayed real-time on a city map and integrated into a navigation software to assist others to reduce parking related issues such as duration of searching a vacant parking location.

Vehicles operating on fossil fuel creates environmental pollutions as the demand keep increasing to transportation systems. Electric vehicles (*EVs*) can potentially resolve the problems caused by conventional vehicles. However, increasing usage of *EVs* will require and constitute complex interactions between transportation, power, and information systems, especially when regulating charging behavior of *EVs*. Therefore, analyzing the further demands due to *EV* charging/discharging and coming up with a strategy to mitigate the relevant impact of these behavior on the grid is of great importance. The study in [168]

introduces an optimal strategy established on real-time crowd-sensing for *EV* to select a route when navigating to the charging station. Authors implemented a road velocity matrix acquisition and restoration algorithm with the help of crowd-sensing. Furthermore, queue theory is adopted to address the duration of waiting at charging stations. The goal of proposed solution is to decrease the travel time of *EV* users and charging or total cost based on the time of use price mechanism, which depends on different technical constraints such as selection of path, travel time, capacity of battery, and constraints related with charging or discharging, etc.

2.3.3 Emergency Management

Quality of the roads has a direct impact on the quality of driving, which may lead to the traffic accidents when it is too low. The condition of road surfaces is an important sign when assessing the quality of roads. Traditional parameters to be considered in order to determine the quality of road surfaces are potholes, bumps and slipperiness. Surface condition of a road not only causes bad driving experience that may lead to an incident but also damages the car and ages it quicker.

Authors in [20] take advantage of mobile sensors planted in either vehicles or *RSUs* that are connected to cloud servers. Anomalies that vehicles experience (e.g., hitting a pothole) while moving on the road are collected and reported to centralized cloud servers to be further analyzed. *RSUs* along with the bases stations assist to carry data to the cloud to be analyzed and to offer recommendations. Real-time data processing is necessary to provide timely recommendation based on the condition of the road surface. Another study [53] aims to detect potholes and mark them on map. Embedded sensors in the smartphones of drivers are the key enablers for this work. Authors have developed an application called "Road Condition Monitoring App", which drivers install on their smartphones. Since the developed application is depending on crowd-sourcing and participatory sensing, it requests access grant to a number of sensors to operate effectively such as *GPS*, accelerometer, magnetometer, compass, and connectivity on device of the user. Alerts related with the road condition is then sent to the user based on the crowd-sensed data, which makes the innate behavior of application ubiquitous as the user of application does not necessarily need to know how the application operates by just using the sensor data to detect the road condition. Moreover, the user of application does not need to mark any potholes on the device. Designing an application based on the principles of crowd-sourcing and participatory sensing has remarkable benefits over the other approaches. For instance, no additional deployment cost is required as a result of usage of existing devices for the sensor platform.

Not only the road surface condition, or improved driving habits are enhancing the quality of driving and thus alleviating risk of traffic accidents, but also the constructions (e.g., bridge) that enable vehicles to travel and promote a flowing traffic. Evaluating the condition of bridges conventionally relies on visual inspections, which are conducted irregularly and

effects eventually timely structural analysis and assessments reports that maintenance plans depend on. Observation of the operational behavior of a bridge must be done frequently to keep track of its condition adequately and to prevent any unforeseen rehabilitation. To this end, authors in [99] use smartphone accelerometers sensors that are capable of providing aggregated sensor data – if provided through crowd-sourcing – about structural health monitoring. Smartphone data acquired from vehicles that are passing over a bridge are capable of identifying various modal frequencies of the bridge, which makes the process of maintaining the structural consistency of the bridge more efficient and helps making well-informed decision to prevent possible catastrophic events such as unexpected collisions.

2.3.4 Information Management

Vehicular networks are the key enablers when developing an application in the context of smart transportation. Vehicles are enabled to behave as mobile sensors to acquire miscellaneous data depending on the context of the application, which in turn facilitates to offer different services such as traffic management, safety, smart navigation, environment monitoring, pollution measurements, and urban surveillance. A new way of communication management is inevitable to acquire, share, transmit, and process the data collected from vehicles. Since the vehicles are mobile, the network formed by the vehicles changes dynamically in a way that could create isolated nodes, changing quality links, and node densities. However, this extremely mutable and heterogeneous topology poses a problem for conventional network protocols, which are normally created for homogeneous use cases.

Anycasting is an appropriate methodology for vehicular sensing applications as it offers the information to those who show an interest in the information. The paper [16] introduces an "adaptive Anycasting solution for Vehicular Environments" (*AVE*), which is a protocol to deliver messages adaptively. *AVE* incorporates geographical and topological information to adjust briskly its behavior to the condition of the network. Proposed *AVE* approach has been assessed in an urban simulation environment under realistic settings. Moreover, rational propagation models are adopted, and urban layouts from the database of OpenStreetMap are integrated to accomplish almost real-life-like road layout when combined with a meticulous building distribution.

WiFi *APs*, mesh routers, wireless sensors, and other wireless routes established throughout the road are capable of operating as an *RSU*. Furthermore, these *RSUs* can cater infrastructural supports for wireless access and dissemination of data in cyber-transportation systems. Authors in [163] propose two considerable applications in urban vehicular network for promoting the development of cyber-transportation systems: (i) crowd-sensing application to collect information from *RSU* resources, and (ii) data dissemination application for vehicular network through crowd-sensed results and location aware service. An hybrid routing solution – Routing for data Dissemination by Crowd-Sensing (*RDCS*) – is designed for inter-vehicle, vehicle-to-roadside, and inter-roadside data dissemination to overcome the

data dissemination issues in urban hybrid networks. Vehicles and *RSUs* are the communication nodes in the proposed hybrid routing solution. Moreover, implemented crowd-sensing module is capable of sensing physical world online via nearby *RSUs*, current coordinates, and speed by crowd-vehicles, and afterwards proceeding with the offline process of these aggregated data in the crowd-server. To facilitate routing decision to disseminate data, crowd-sensed results are used as input by the cyber elements in vehicles.

In the context of *VANET*, crowd-sensing of the environment is one of the key-enabling methodology for many applications. It empowers a group of vehicles to sense the environment with their perception sensors, and share the sensed data with other vehicles to stay always aware of whats happening in their visible and invisible perimeters. However, security gaps in crowd-sensing mechanism may lead to attack towards *V2X* networks, which can have a destructive effect on the road safety. The paper [23] classifies potential security threats, and defines various *V2X*-specific attacks that can be performed easily and may decrease the performance of crowd-sensing mechanism in *V2X* network. The evaluation of the attacks are based on the likelihood of weakening the road safety such as the platooning disruption (*PD*) attack, which focuses on the critical platoon signaling (that enables a group self-driving vehicles to travel safely on the road), and the perception data falsification (*PDF*) attack, which transmits fake sensing data to cause traffic accidents.

2.4 Data Analytics for Smart Transportation Application and Services

”Smart” transportation applications are made capable of predicting/forecasting events through data analytics methodologies. We categorize data analytics into two separate sub-sections: (i) Machine and Deep Learning, and (ii) Signal Processing. In this section, we will investigate enabling methodologies under each section.

2.4.1 Machine and Deep Learning

Machine and/or deep learning techniques are broadly used in the context of smart transportation applications. There exist various ways of integrating a sense of machine intelligence into an application such as algorithms, tools, and frameworks. Therefore, we will divide this sub-section into the following sub-sub-sections: Machine and Deep Learning (i) Methodologies and Algorithms, and (ii) Tools and Frameworks. Furthermore, we will investigate applications incorporating machine and/or deep learning methodologies through algorithms, tools and/or frameworks under the aforementioned services we have proposed in the previous Section 2.2.

Methodologies and Algorithms

Traffic Management: Applications and services that incorporate machine and/or deep learning algorithms and help manage traffic to improve the way of vehicles to move from a point to another efficiently are categorized under this category. Such applications and services includes but not limited to traffic flow prediction, traffic congestion prediction, route planning, and free parking sport locator.

One of the most common hassle drivers deal with everyday is searching an available parking location, which contributes a non-negligible source of atmospheric and acoustic pollution during the time of searching along with causing stress to drivers and wasting their time. Developing systems that serves real-time information about the vacancy of on-street parking spots in a city not only would help escalate the livability standards of the city but also increase the sustainability of the city. Combining the data acquired by variety of network types can help tackle the vacant parking sport detection problem. To this end, authors integrates [135] the data collected by small-scale sensor networks and widespread CCTV camera networks, which makes the creation of smart traffic management systems – that are able to offer real-time information to the users about the vacancy of available parking lots, to decrease the duration for seeking a parking space, and to reduce eventually the traffic density along with its direct adverse effect of air pollution – possible. This work involves a pattern recognition system that depends on a high performance classifiers, k-nearest neighbors (kNN) and Support Vector Machines (SVM). With kNN , availability of a specific parking spot will be decided based on the status of other visually similar parking spots stored in the training database. kNN – measuring the similarity of two parking sports with Euclidean distance – is able to achieve good performance in various recognition problems even though it has a lazy-learning approach and simple architecture. On the other hand, system learns to identify the available and occupied parking spots in order to determine the status of a newly discovered parking sport with SVM classification approach. The performance of both techniques will highly correlated with the tuning parameters. Therefore, authors have tested the proposed system with different number of neighbors for kNN – ranging from $k=1$ to $k=80$ with k symbolizing the number of neighbors – and the kernel type for SVM – linear, polynomial, quadratic, or radial basis function (RBF). Another paper focusing on parking location detection is [150], authors of which explains the feasibility of vacant parking spot detection with the usage of mobile phone magnetic sensors, magnetometer. Readings of magnetometer is analyzed with a supervised machine learning technique, SVM to facilitate identify the situation of the test-car. System acquires sensors readings from any car and SVM running in the server estimates the situation in the surrounding of the car by stating in the end that there is/are car(s) parked in the perimeter of the car (such as right/left, both side, or no cars). Furthermore, the system is capable of drawing a heat map of free parking spaces with the help of GPS data.

Precise and well-timed traffic flow information brings so many advantageous – including support for better travel decisions, lessen traffic congestion, decreasing the carbon foot-print left to earth by vehicles, and enhancing the efficiency of traffic operation – to the individual

travels, business sectors, and government agencies. Predicting the flow of traffic is highly correlated with the historical and real-time traffic data acquired from variety of sensors such as inductive loops, radars, cameras, mobile *GPS*, crowd-sourcing, social media, etc. Considering those listed (and not included) sources, authors in [94] offer a traffic flow prediction service based on deep-learning. A stacked auto-encoder (*SAE*) which learns generic traffic flow features is integrated into the proposed architecture and trained in a layer-wise greedy fashion. The modeling of the *SAE* is done by considering the spatial and temporal correlations.

Increasing traffic congestion in urban areas with the enhancing trend metropolization of the current century, smart route planning becomes an inevitable, essential service in the smart city development. A system to plan an individual trip that includes up-coming traffic hazards in routing is presented in [91]. A Spatio-Temporal Random Field established on a stream of sensor readings helps calculating the future traffic conditions. Moreover, in areas with low sensor presence, the proposed system is capable of estimating the traffic flow using a Gaussian Process Regression. Future traffic flow values are modeled based on their temporal patterns, correlations, and dependencies with the integration of *kNN* into the system. Furthermore, proposed system uses less memory, a specific issue introduced by *kNN* (which has to store all previous sensor readings for continuous traffic flow estimation).

With the advent in the area of data acquisition, storage, and processing, traffic routing applications that take advantage of forecasted data (e.g., traffic volume data, travel time data, and traffic condition information) helps saving time and cost of drivers on the road. Moreover, efficient management of traffic systems based on the forecasted information can be improved with the valuable contributions of transportation agencies. Estimating traffic conditions in a vehicular network depends on various parameters, one of which is travel speed of the vehicles. Regular historical traffic data plays a curial role when estimating the short-term future traffic flow. As the nature of acquired traffic flow data is high dimensional and non-linear, the candidate prediction algorithms that we are to integrate into traffic flow forecasting systems are expected to be equipped with advanced technologies and computational ability to handle the traffic flow data. The neural network [184] is one approach highly used to forecast traffic flow time series. In order to improve the computational performance and efficiency of conventional neural networks, there exists some improvements algorithms proven to be superior such as state-space neural network model [97] and Long Short-Term Memory Neural Network (*LSTM*) [148]. Moreover, the *ANN* created with the combination of fuzzy inference system and network structure – fuzzy-neural networks (*FNN*) – demonstrates good performance in traffic flow prediction accuracy. Optimization of parameters in an adaptive fuzzy rule-based system for traffic flow prediction in urban arterial networks offered in [43] is furthermore improved using a generic algorithm approach.

Mapping a high-dimensional data (e.g., traffic flow) into a high-dimensional feature space with the help of *SVM*'s non-linear relationship and linear regression capabilities. The support vector regression (*SRV*) model has been extensively employed in travel time [162] and speed [15] prediction application. Optimization of parameters in *SVM* to improve its efficiency, least squares support vector machines (*LS-SVMs*) [178] is applied in several actual

applications. With the integration of supervised and unsupervised machine learning techniques in [146], authors propose a new learning architecture to improve the learning ability of *FNN* model. While *k-means* method is applied to separate input samples into different clusters, Gaussian fuzzy membership function is employed to each cluster for measuring the degree of membership of samples to the cluster centers. Moreover, a trigonometric regression function tailors the periodic patterns derived from the traffic flow data. The prediction accuracy of multi-step ahead forecasting can be enhanced with the *FNN* model that takes periodic features into account.

Information of the present and past traffic congestion of a road network comes in handy when forecasting the future status of the roads in the road network. The paper [134] introduces a system that is capable of predicting the densely populated roads by using the present and past traffic congestion data. Furthermore, the proposed system offers alternative paths based on the provided source and destination. A densely congested path on the route map is assigned with a weight to signify the level of congestion. The weight values are produced by applying Density-based Clustering *DBSCAN* to the fuzzified data.

Public Transportation Management: The people using public transportation services and applications are the center of this section as efficient management of public transportation systems starts with a good understanding of the movement pattern of the resident living in a city such as their transportation mode and time. Authors in [47] consider both transportation and vehicular mode classification by adopting three machine learning algorithms: Decision Tree (*DT*), *kNN*, and *SVM*. While transportation modes investigated in this work are classified into 5 categories: still, walk, run, bike, and vehicle, vehicular modes are classified into 6 categories: motorcycle, car, bus, metro, train, and high speed rail. The data fed into machine learning algorithms for classification are acquired from three types of low-power-consumption sensors: accelerometer, magnetometer, and gyroscope. The features are obtained from the time series of these sensor measurements by accommodating them into the time and frequency domain.

In [142], authors cope with the short-term passenger flow prediction problems to enhance the prediction performance in the subways system. Authors introduce a novel method which integrates Wavelet-*SVM* [175]. By merging the complementary benefits of Wavelet and *SVM*, the weak points of those methods are reduced and overcome. There exist three important stages in Wavelet-*SVM* forecasting approach: (i) decomposition of passenger flow into different low and high frequency series by Wavelet, (ii) application of *SVM* method for learning and predicting the corresponding low and high frequency series, and (iii) reconstruction of the diverse predicted sequences by Wavelet.

Public transportation methods are used regularly by some individuals and used occasionally at specific times or purposes by the others. While general mobility of individuals can be inferred from their public transport travel, activity patterns of different individuals can be obtained from their smart card transaction which contributes various levels of information about individuals' activities. Only a small portion of an activity lasted for a period time

can be revealed from the fare transactions history of passengers using public transport occasionally. The study in [57] detects passengers whose activity pattern can be derived almost entirely from smart card data. The number of days and the spread of days between the first and last day that cards are observed are two variables that are used along with $k - means$ algorithm to cluster each individuals level of public transport usage. 3 user clusters are revealed at the end of this process: a group of non-recurrent users, a group of occasional users, and a group of frequent users.

Leveraging usage of public transport facilitate to solve the problem of meeting travel demand of users, and thus alleviate the issue of serious traffic congestion due to lack of capacity of current road-network. However, a driving force to allure people toward to public transport is necessary. One important factor many people consider when choosing their way of public transport is the duration of operational delay. In [140], authors presents a long-term analytics model, which examines the historical bus delay patterns of arrival time delay at time points and travel time delay for all route segments. $k - means$ method is employed in this work to provide users and city planners typical delay information – grouped based on time of the day –, and other features (e.g., weather). Another study, [92], implements a modified kNN method in order to analyze the bus arrival time *BAT* using historical bus *GPS* data. The modified kNN merges both cluster and principal component analysis and it relies on the processed *GPS* data. Validation of the proposed kNN method is performed with the empirical analysis of data collected from a bus. The results demonstrate that the modified version of kNN makes less prediction errors when compared with the other techniques, such as *ANN*, and it is more powerful in the *BAT* prediction.

Many transportation systems, especially public transportation systems, take advantage of the convenience of contact-less, electronic payment methods to eliminate the hassle of paying with cash. Automated Fare Collection (*AFC*) is one system used widely in the public transportation networks (e.g., bus, metro, and tram). *AFC* not only eases the way passengers pay the fair of usage by their smart cards provided by the operator company, but also provides additional data related with passengers’ trip to the operator company. In [102], authors outlines two types of transaction collected through smart cards: (i) monetary transactions as a result of card-holder adding credit or renewing travel pass, and (ii) journey transactions as a consequence of passengers entering stations, boarding buses, etc. The initial purpose of *AFC* systems is to automate and handle different billing operations presented in the fare collection operation. However, the collection of data formed by those transactions above (specifically journey transactions) creates a remarkable opportunity to infer and extract valuable (e.g., group behavior, frequent patterns) information that be used for various purposes such as performance evaluations, transit planning, etc. To this end, authors explain the necessity of employing advanced data analysis techniques (e.g., clustering, classification).

DBSCAN is used for individual trip chains to extract recurrent travel patterns of each passengers [96]. Furthermore, passengers are grouped by $k-means++$ with regard to regularity of travels, which depends upon four descriptive features of the passengers: (i) number of travel days, (ii) number of similar first boarding times, (iii) number of similar route sequences, and (iv) number of stop sequences. Comparatively, authors in [81] apply $k-means$

to distinguish infrequent passengers from frequent passengers based on the number of trip chains, and adopts *DBSCAN* to further divide the latter group based on boarding and alighting time and location regularity. In addition, the paper [21] also uses *DBSCAN* to portion passengers into four distinct groups based on their habitual travel times and the origins and destinations of their trips.

Group behaviors are analyzed by applying Hierarchical Agglomerative Clustering (*HAC*) and *k-means* into smart card transactions data. In [85], authors adopt *HAC* technique to cluster weekday profiles of a passenger (trip counts over five time bins within the day) for revealing various travel behaviors such as typical commuters, evening-only travel, etc. This capability, in turn, facilitates to build user-tailored transport information services. Moreover, the authors also provide several predictive models to offer personalized travel time estimates with regard to travel history of a passenger.

Clustering method is also applied in [104] to examine individual travel regularity. The trips of a passenger are collected into a daily profile to express a registered boarding pass for each time bin (a given hour). *k-means* method is used to determine the clusters of similar days that are relative to boarding times.

Another work that takes advantage of the transactions provided through smart cards is [34] which studies station congestion patterns. The station data consists of individual trips. Each observation in a station data accommodates the difference between entries and exits for a given station over 2-minutes intervals. Stations are re-clustered depending on their usage by employing Dynamic Time Warping (*DTW*) and hierarchical clustering techniques. Furthermore, station crowdedness is forecasted with three classification techniques that are Historic Value, Historic Mean, and Historic Trend.

Emergency Management: Accidents on the roads are inevitable and one of the most life-threatening dangers for travelers – not to forget to mention their potential to cause great damages. Therefore, a smart transportation system should assure the safety of travelers through emergency management services and application. Forecasting the future traffic accidents in advance is an effective solution to offer in smart transportation systems. By providing useful information to drivers, dangers on the road can totally be bypassed or the damage can be reduced by responding quickly. Traffic accidents can be predicted by applying classification analysis – a data mining procedures that required sufficient data for constructing a learning model [117]. There exists various application employing different techniques to predict traffic accidents, such as [95] uses video image from cameras on the road, and [172] analyzes numerous traffic data in order to prevent accident before happening or reduce impact to minimum in case of an accident. Traffic accidents happening on an highway are related with the short-term turbulence of traffic flow. In [95], authors analyzes real-time traffic data with *kNN* to identify the potential accidents. To this end, they initially determined the traffic accident precursors and their calculation time slice duration and then classify the traffic patterns. The results indicate that *kNN* approach outperforms *C-means* clustering approach.

Furthermore, *ANNs* are also extensively used in traffic safety management applications. Particularly, *BP* neural network is one of the common traffic safety forecasting methods. On the other hand, the forecasts by *BP* neural network will be influenced due to its innate drawbacks (e.g., local optimization solution). Another approach that is commonly used in prediction problems is *SVM*, which facilitates the forecasting procedure by requiring small, non-linear data samples. However, selection of the training parameters greatly affects the performance *SVM* method as inappropriate parameter settings causes poor forecasting results. The paper [50] proposes a novel approach by integrating particle swarm optimization with *SVM* for traffic safety forecasting. The method out of this merge is called *PSO-SVM*. Authors first and foremost analyzed the factors influencing traffic safety, and then established the forecasting model by employing *PSO-SVM* with regard to influencing factors.

The networks formed by vehicles such as Vehicular Ad Hoc Networks (*VANETs*) paves the way of message broadcasting among vehicles about road condition and emergent events (e.g., traffic accidents and congestions nearby the vehicles). While those networks make security and safety services and applications for smart transportation plausible, there could be a potential chance of having security breaches in the system that allow inside attackers to initiate false information attacks by injecting fake emergency messages to misinform other vehicles, and constitutes a great danger to road safety. Therefore, not only creating emergency management services and applications is crucial, but protecting the system from the attackers is also very important. Authors in [182] offer a false information detection scheme which depends upon information fusion technology where vehicles gather information from various sources to understand traffic condition. To this end, Basic Belief Assignments (*BBAs*) are exchanged between vehicles. However, bogus *BBAs* can be created and injected by multiple adversaries to support the false emergency message. To prevent this, *BBAs* received by a vehicle must be examined. A distance metric can be adopted for the quantification of dissimilarity rate between two *BBAs*. A density-based outlier detection algorithm such as *DBSCAN* can later be applied to eliminate the *BBAs* deviated largely from the others so that authors can use the remaining *BBAs* in the second layer fusion.

Information Management: Applications and services offered to the users in *VANETs* requires Internet connectivity. Detecting the activities of those in *VANETs* is becoming a challenging task with the rapid increase in the number of users. Furthermore, this challenging task becomes even more complicated as the mobility and density of the nodes in a *VANET* varies quickly. The information (e.g., speed of the nodes, position of the nodes, direction of motion of the nodes) related with a specific area is shared by nodes with the other nodes while they are traveling so that vehicles are aware of their surrounding to take precautions in case of an emergency on the road. One of the drawbacks of *VANETs* is the security issues due to the high mobility of the nodes. During the information sharing with other nodes, it is plausible that information of the node that shares its data can be compromised. Therefore, a smart transportation service/application should ensure the security and stability of the network in order to avoid intrusions.

The article [82] uses a novel Collaborative Learning Automata (*CLA*) based algorithm to

ensure the detection of an intrusion in *VANETs*. The algorithm accepts three parameters as an input: (i) density, (ii) mobility, and (iii) direction of motion of the vehicles. Based on these parameters, corresponding automation executes its action. Moreover, a Collaborative Trust Index (*CTI*) is calculated for automaton to perform the actions. New values can be added from *CTI* or already exiting values in *CTI* can be discarded with regard to the successful and unsuccessful operations such as penalty or reward received from the environment by the automaton. The action probability vector of the automaton that will be used in the next iteration is kept updated with these parameters. The solution converges to a specific value upon performing the action for a period of time.

Performance degradation in vehicular sensor networks (*VSNs*) due to an increase in the quantity of the vehicles on road and traffic jams leads to delay when receiving the emergency alerts. To overcome this problem in *VSNs* installed in dense urban regions, authors in [83] introduce a *CLA*-based routing algorithm to transmit information to the designated destination with minimum delay and maximum throughput. By integrating the learning automata (*LA*) in the access points (*APs*) in the network, *APs* are made capable of learning from their pas experience, which eventually help them make routing decisions swiftly. The whole region is divided into various clusters. An optimized path is then chosen by applying *CLA* with the following input parameters: vehicle density, distance from the nearest service unit, and delay.

Decreasing the traffic collisions at intersections is possible in *VANETs* with the warning messages sent to the vehicles. A performance improvement in the *VANETs* is necessary to ensure delivery of the messages – especially safety messages – to the target destination. An effective way to reduce packet loss and delay and to improve the reliability of *VANETs* is to use data congestion control. The paper [144] offers a centralized and localized data congestion control strategy for managing data congestion with the help of *RSUs* at intersections. Three units are integrated into the proposed strategy to provide congestion detection, message clustering, and data congestion control. Data congestion in the channels is identified by measuring the channel usage level. An unsupervised machine learning algorithm, *k-means*, is applied to cluster the messages collected and filtered. *K-means* method groups the messages based on three certain criteria: (i) message size, (ii) validity of messages, and (iii) type of messages. Moreover, the data congestion unit selects proper values of transmission range and rate, contention window size, and arbitration interframe spacing for each cluster. Finally, communication collisions are decreased with the communication parameters transmitted to the vehicles (that are stopped before the red traffic lights) by *RSUs* at the intersections.

Tools and Frameworks

Traffic Management: Under this section, we will analyze traffic management services and applications integrating various tools and frameworks in their platform. An experimental study, [150], validates the feasibility of using mobile phone magnetic sensors for spotting available parking locations. A supervised machine learning algorithm, *SVM*, is used for classification. Authors particularly has integrated *SVM* in their project to perform Support

Vector Classification for constructing a multiclass model following the one-vs-one scheme developed initially by [61]. In this work, *SVM* implementation in scikit-learn library [119] is adopted.

Another car parking solution is proposed in [183], authors of which takes two smart car parking solution based on real-time car parking information into account. Three feature sets are constructed based on the data analyzed and key parameters selected thereafter analyzing the data. Prediction model for the parking availability is derived from those feature sets. Additionally, three non-parametric algorithms are adopted to create parking occupancy rate (*OCCR*) model: (i) Regression Tree (*RT*), (ii) Neural Network (*NN*), and (iii) Support Vector Regression (*SVR*). Python programming language is used to implement the aforementioned algorithms. While *RT* and *SVR* algorithms are implemented using scikit-learn library [119], *NN* is implemented using PyBrain [132]. Different datasets from two cities (Melbourne and San Francisco) are used to compare the performance of feature sets and model combinations. Their results revealed that the regression tree with a feature set that consists of the history of observations, time of day, and day of week produces the best performance. Moreover, the proposed system can be realized as (i) a mobile application to facilitate trip planning for users by finding available parking spots, and/or (ii) a car navigation system to assist drivers when selecting parking locations with regard to their destination.

There exist some challenges when estimating the traffic flows, one of which is the sharp nonlinearities due to transitions between free flow, breakdown, recovery, and congestion. A deep learning architecture is capable of detecting these nonlinear spatio-temporal effects. Therefore, a deep learning model for predicting traffic flows is proposed in [123]. While the first layer of the deep learning network determines spatio-temporal relations among predictors, the rest of the layers build non-linear relations. Furthermore, the Stochastic Gradient Descent (*SGD*) technique is employed using *H2O* package [62] to discover an optimal network structure and weights. Moreover, similar techniques can also be found in Python’s Theano [19] or TensorFlow framework [3]. Authors present their methodology by predicting traffic flows during two special events where there is sharp traffic flow changes occurring unexpectedly and very suddenly. Proposed deep learning method provides accurate short-term traffic flow predictions.

Public Transportation Management: Observing the travel mode choice with regard to behavior of travellers can be very useful when analyzing the travel demand, planning the transportation schedule, and managing the traffic. Transportation agencies are enabled by speculating the travel mode distribution to develop proper strategies that mitigate travel time of users, traffic congestion, and air pollution. A Convolutional Neural Network (*CNN*) architecture is introduced in [42] to determine the transportation mode(s) of the trip of individuals based on the information of raw *GPS* trajectories. The proposed system is capable of categorizing the transportation mode into five distinct groups: (i) walk, (ii) bike, (iii) bus, (iv) driving, and (v) train. Data structure used in this work contains four kinematic features, which are speed, acceleration, jerk, and bearing rate. Those features together are

sufficient to define the motion characteristics of individuals. An effective *CNN* architecture that is capable of achieving the state-of-the-art accuracy on the *GPS* dataset is designed for the pre-processed feature dataset. While Python programming language is used to process the data, Keras [39] (a Python-based deep learning library) using the TensorFlow [3] backend with the *CPU* support only is used to implement the *CNN* architecture.

One approach to address the increasing congestion problems is to develop and apply a distributed traffic optimization system. However, such system should be capable of routinely handling the visual information from a scene in real-time as traffic is a dynamic environment where vehicles pass through intersections and sudden appearance of pedestrians on the scene is quite usual. Since, those type of systems analyze the visual of a scene, an image recognition technique is what's needed. *CNNs* are extensively adopted in the systems where image recognition is required. *CNNs* can be applied for miscellaneous tasks in traffic applications such as detecting a vehicle, capturing the velocity of a vehicle, and calculating the total wait time of vehicle, which can be handful information to have when developing a traffic flow optimization system. However, computational complexity of *CNN* introduce heavy load on the system. Therefore, the sparse representation of *CNN* – a simplified version – is an efficient way to shrink the computational burden. In [90], authors study the effect and the potential of sparse representations of *CNNs* in terms of run-time with different implementations. Moreover, a code-generator creates optimized and ready-to-deploy code upon the completion of pre-processing. This work demonstrates that the proposed solution on two different sparse *CNNs* is faster than the performance of the *CPU*-mode TensorFlow and is energy efficient than the *GPU*-mode TensorFlow.

On-demand ride service platforms provides relevant information for empty cars (e.g., an uber, or a taxi that looks for a customer) to encourage them to move from over-supply regions to over-demand regions through the demand forecasting of short-term passenger. However, there exist spatial, temporal, and exogenous dependencies, which makes short-term passenger demand forecasting challenging. Authors in [79] studies a novel *DL* technique – the fusion convolutional long short-term memory network (*FCL-Net*) – to solve the challenge introduced by the aforementioned three dependencies. Authors developed an end-to-end learning architecture. The proposed *DL* model contains multiple convolutional long short-term memory (*LTSM*) layers, standard *LTSM* layers, and convolutional layers. The spatio-temporal characteristics and relationships among explanatory variable are better detected with the fusion of convolutional techniques and the *LSTM* network. Moreover, explanatory variables is ranked with regard to their importance by using random forest (*RF*) algorithm. In this work, scikit-learn [119], TensorFlow [3], Keras [39], and XGBoost [36] are adopted to compare the models.

Emergency Management: Numerous research activities are recently being conducted to alleviate the drivers through in-vehicle systems to chance their driving behavior that could potentially cause fatal or non-fatal traffic accidents. The main purpose of these systems generally is to avoid an accident with an upcoming vehicle or pedestrian through the built-in collision-avoidance module, which sends warning messages to drivers. The recent studies,

such as [129], carried out with these systems shows encouraging and significant positive effects. By applying real world location analytics on traffic accident hotspots, in-vehicle warning messages can be produced to test the in-vehicle warning system proposed in [129]. Moreover, the proposed in-vehicle warning system uses the data acquired by connected vehicles to determine the dangerous locations. The data is collected from the Swiss Road Authority (*FEDRO*) and it contains more than 266,000 traffic accidents for a period of five years. Furthermore, the data is classified into 20 different combinations of two categories: (i) road type such as "motorway" and "main road", and (ii) infrastructure such as "roundabout" and "intersection". To capture the satellite imagery of the location of each accident are obtained from Google Maps by providing random accident samples taken from the categories in a balanced way. Finally, TensorFlow [3] Inception-v3 is used to produce the high-level feature vectors for the labeled images.

With the advent of high technological driving aids to ensure the safety on the roads, there exists plethora of other in-vehicle systems one of which is on-time vehicle braking system in hazardous road traffic situations. The study in [122] proposes a system which is capable of forecasting the emergency situation with regard to the motion pattern of the accelerator pedal. Data pattern obtained through the lifting motion of the accelerator pedal is used to classify the emergency breaking situations. Eventually, the system is trained to decrease the time delay between the following two events: (i) the driver lifts their foot from the accelerator pedal, and (ii) the driver energetically presses the brake pedal to stop the vehicle in dangerous and emergency braking situations. The time delay introduce in varying meters of traveled distance depending on the speed of the vehicle. For instance, a higher travel distance to stop the vehicle is expected for vehicle moving at high speed. Reducing the delay not only would help decreasing or totally eliminating the risks of traffic accidents but also enhance road safety. The proposed system adopts a machine learning approach when classifying the emergency braking situation on gradient boosting. Scikit-learn [119] library is used to implement the algorithms described in their work.

Autonomous vehicles is another smart transportation application that gains popularity. One of the fundamental requirements to integrate in autonomous vehicles is to make use of vehicles following other vehicles in the perimeter safely and taking proper actions to avoid rear-end collisions and accidents. Drivers understand the behavior and intention of the other vehicles moving ahead with rear-lights signals. Thus, it is crucial for an autonomous vehicle to adopt rear-lights signal recognition, which is a computer vision problem incorporates processing of color images and analyzing brightness (i.e., blob technique). In [158], authors introduce a two-stage approach that is capable of recognizing rear-lights from a sole image instantly and robustly. As the traffic environment is dynamic, the proposed system conforms the standards for analyzing the real-time data. Rear appearance images of vehicle are used as the training data. A multi-layer perception neural network is adopted to learn "Brake Lights Patterns" (*BLP*) from a large database storing rear appearance images of vehicles. Furthermore, deep classifier will group the given image of a vehicle under two categories: "brake" and "normal". Integrating multi-layer lidar (IBEO Lux fusion system) and a camera, a vehicle can be identified instantly and robustly. Moreover, in order to accelerate the vehicle

identification and to enhance the robustness, different methods are investigated such as road segmentation and a novel vanishing point region of interest (*ROI*). In this work, an open source deep learning library – Caffe [152] – is adopted to integrate already available deep convolutional neural networks, such as *VGG-16* [137] or *AlexNet* [151]. *AlexNet* model (an eight-layer neural network) is used in their system to obtain real-time system.

Traffic accidents not only causes potential injuries and damages but also creates road congestion. Therefore detecting incidents in advance decrease the likelihood of congestion on the road. One way of detecting incidents is to take advantage of machine learning methods to analyze video feeds at fixed intersections. Authors in [147] study on a system that is capable of distinguishing the congestions caused by high density of traffic (normal traffic jam) and the congestions caused by a traffic accident. To this end, the proposed system analyze the latent traffic data from readily available sources such as smart phones with supervised machine learning algorithms – Logistic Regression, a Bagging classifier of logistic regression, and an adaBoost classifier. These algorithms are adopted from scikit-learn [119] library. A multi-Agent System (*MAS*) is used to simulate traffic by modelling behavior of independent vehicles on the roadway. With the traffic data generated in the simulation environment, supervised learning classifiers are trained to identify the normal congestion versus congestion as a result of an accident for any given time step of a simulation.

Information Management: As the services and applications under smart transportation category can possibly be used by many, experiencing low information delivery efficiency and high levels of redundant traffic is unavoidable. Data can be shared between users with various methods (e.g., via peer-to-peer (*P2P*) networks, content delivery (*CDN*) networks) to minimize the impact of the aforementioned issues. However, enhancing network capacity by allocating traffic or other resources efficiently is still an issue to be resolved. The network systems are enabled to comply with the changing network environments via gathering the traffic data and implementing the smart statistical analysis and multi-dimensional modeling. Therefore, the network system can accommodate the allocation of network resources or settings, which eventually will enhance both the network service and users’ experience (and thus Quality of Experience (*QoE*)).

A fine-grain network traffic prediction algorithm – Temporal Poisson traffic prediction (*TP2*) is introduced in [174] for managing the traffic in a smart way. The proposed *TP2* algorithm is capable of modeling properly the non-linearity and non-stationary in Information Centric Network (*ICN*). *TP2* is also capable of predicting the traffic demands for various contents in the next time windows. Furthermore, with the help of *TP2*, a collaborative cache resource placement algorithm is introduced for allocating the cache resources effectively. A simulation environment based on *NDNSIM* is used to measure the performance of the proposed caching strategy. As a result of intelligent network management, cache resources are utilized efficiently, which, in turn, escalate the user experience of the end clients. Intelligent resource management in this work is implemented by using Theano [19], Scipy, and Scikit-learn [119] modules.

Providing a better networking infrastructure is a necessary to build a functioning smart transportation system where information flows consistently from one node to another. During the exchange, vehicles send sensitive data (generally without encrypting) such as their location, speed, and heading. Anybody within the perimeter of wireless coverage range can track vehicles and their drivers by obtaining the sensitive data via intervening the network illegally. There are many works in the literature proposing new privacy protection schemes. In this work, authors compare the power of different privacy metrics or analyzes, which constitutes a great importance for the accurate measurement of privacy to assess new privacy protections. The paper [154] proposes a new criterion – monotonicity – for the strength of privacy metrics such as low privacy for strong adversaries, and high privacy for weak adversaries. Furthermore, proposed method is applied to 32 privacy metrics studies in the literature along with the ones that have not been adopted by vehicular privacy before. Proposed technique is originally introduced in [153] and it consists of four core components: (i) user behavior, (ii) adversary behavior, (iii) criteria for metric strength, and (iv) privacy metrics. Those four components were implemented using Python programming language and the packages developed in Python such as numpy, scipy, scikit-learn [119], and mpi4py.

One of the techniques to acquire information from wide-spectrum of sources in intelligent transportation systems is Participatory Sensing (*PS*). It enables modern sensing-capable devices to collect information practically from anywhere, at anytime, about anything. Information is generated by the users of *PS*. They contribute sensed data with variety of enabled-devices available to them such as smartphones, smart-vehicles, wearable sensor vests and armbands. As information shared through personal devices, it is sensitive disclosing the location, daily routines, and social relations of users. Moreover, it can be used to generate a profile for a user, who will in the end stop using those services to stop information leakage. Therefore, powerful privacy protection along with encouragement mechanisms must be developed to facilitate mass user participation. Authors introduce a framework – SHIELD – in the article [55] to overcome these challenges. SHIELD is a new data verification framework that integrates both evidence handling (Dempster-Shafer Theory) and data-mining methods to detect and eliminate erroneous/malicious user contributions. The proposed architecture trains a collection of classifiers such as *RF*, naive Bayes (*NB*) classifier, and nearest neighbor (*NN*) classifier. Those classifiers are implemented with scikit-learn [119] library.

For intelligent transportation applications/services stable and robust wireless ad hoc communication is necessary. Faults and shortages due to communication channel may lead to dangerous situations in vehicle safety and automation applications. For instance, a collision detection system without proper network communication infrastructure may delay or fail the delivery of warning messages, which may cause accidents. No communication network failures are considered by many systems. Communication problems caused by partial network faults are not deeply considered. However, partial network faults (which are usually hard to detect, and therefore ignored in mobile ad hoc network applications/services) are critical and they must be addressed to ensure durability and quality of the communication channel. The paper [174] proposes a method to identify failure in mobile ad hoc networks. To this end, the proposed technique correlates the acquired data to demonstrate the abil-

ity to communicate with a probabilistic model created after investigating the acquired data from a fleet of vehicles. As a result of this comparison, a measure signifying the current functionality state of the antenna is outputted. Moreover, this measure also reveals the possible performance degradation. The system is built by employing Theano [19], Scipy, and scikit-learn [119] modules. Furthermore, the system can detect the current state of the individual radio systems to flag them for repair or replacement if necessary.

2.4.2 Signal Processing

Signal processing is another technique that can be used to analyze the behavior of a sensing phenomenon (e.g., image processing methods to analyze the traffic camera feeds for detecting road condition such as traffic congestions, accidents, etc.). In this sub-section, we will investigate various signal processing algorithms in the context of smart transportation services that we have described in Section 2.2.

Methodologies and Algorithms

Traffic Management: Traffic light junctions are locations where road congestions are likely to take place. Decreasing the waiting time spent while waiting for the traffic light to turn into green prevents and/or minimize the chance of traffic jam occurrences. The cycles and the phases of the traffic lights are two key aspects that need to be considered when controlling the traffic lights. A traffic light cycle denotes to the series in the traffic light signals: 1. green, 2. yellow, and 3. red. Each item in the series turns on and off in that order cyclically. Almost every traffic light currently operates by following fixed cycles, which in turn makes the duration of green light phase unsuitable and unbalanced for dynamically varying traffic conditions. Therefore, duration and phase of the traffic lights must be calculated by considering the actual traffic flows or the actual queue lengths in order to eliminate congestions in the traffic light junctions. citecollotta2015novel proposes a system to dynamically manage the traffic lights in an isolated intersection where incoming vehicle movements are not influenced negatively by upstream traffic lights. Propose system aims to decrease the average queue waiting times of the vehicles. To this end, they built the system by incorporating an IEEE 802.15.4 WSN to monitor traffic with one fuzzy logic controller for each traffic light phase, which enables system to detect the green time duration of each phase. A model is developed in Simulink/Matlab to simulate the behavior of traffic light controller when selecting the phase and managing the duration of green light. Furthermore, a prototype board is built established on the model developed in Simulink/Matlab. The processing unit of the board is a microchip called PIC24FJ256GB108 micro-controller, which combines the control features of a micro-controller together with the processing and throughput efficiency of a digital signal processor.

Traffic control management applications such as traffic signal controller, warning systems in case of lane departure, automatic vehicle accident detection, and automatic traffic density

estimation started relying on automatic vehicle detection systems. In [22], authors introduce a system aiming to detect and count the number of vehicles from video feeds captured from a fixed cameras over roads such *CCTV* cameras planted near traffic intersections/junctions. Number of vehicles in a particular location and time is a useful input that facilitates to predict the future status of the traffic. Gaussian mixture model (*GMM*) based object detection method is adopted to accomplish the task. This simple approach offers a non-predictive regional tracking and a counting of tracked objects by establishing simple rules. The frames are extracted from the video data obtained from the stationary traffic cameras. *GMM* separates the moving object from the background in each frame by keeping track of identified objects within a particular region, and count them later on. Blob detection is applied on a binary image to calculate a detected region with its centroid and the area of the blob. The *GMM* provides the pixels detected as foreground. Furthermore, in order to improve the performance of detection, some pixel manipulations methods are applied such as de-noise and filter the binary image from deteriorated blobs and noisy pixels and fill the closed contours in the binary image.

Traffic congestions due to various reasons such as insufficient road capacity, unrestrained demand, delays exposed by traffic lights affects mostly mega cities. Among the various causes of traffic, insufficient road capacity and uncontrolled demand are two reasons complementing to each other. Whereas, delay related with the traffic light is hard-coded and usually is not self-adapting itself depending on the current status of the traffic. Authors introduces a technique that leverages the live video feed from the cameras at the traffic intersections to compute real-time traffic density by adopting the video and image processing methods in [74]. Furthermore, the proposed system also enables changing the traffic lights depending on the vehicle density on road, which eventually decreases the traffic congestions and therefore cut down the number of accidents related with traffic congestion. Recorded video feeds are transmitted to the servers where video and image processing methods are applied to detect the vehicle density on every side of the road to switch the traffic lights based on the congestion level of the road. Proposed system is developed using video and image processing toolbox of Matlab and C++ compiler to produce algorithmic results.

Public Transportation Management: Detecting moving vehicles is a key challenge and must be addressed to provide an efficient transportation management. There are plethora of techniques introduced in the literature to detect the moving vehicle such as background subtraction [35], inter-frame difference [30], and mixed Gaussian background modeling [139]. In the study [93], authors introduce a motion-foreground-detection technique that depends on three-frame difference and mixed Gaussian model and is capable of extracting the foreground moving vehicles. A training classifier encompassing improved *Adaboost* algorithm and *Haar*-like feature is developed to eliminate undesired vehicles upon detecting the moving vehicles with the proposed methodology. Moreover, the proposed system takes advantage of *Canny* operator to locate the characteristics of a bus, which eventually differentiates the bus and coach in their design. That makes it plausible to locate the windows and compute the color pixel ratio of vehicle window area, and to compare later on with the pre-set threshold. In

the end, correct identification of public transport vehicles is accomplished.

Transportation systems constitutes by their nature spatial and temporal complexity problems. With increasing number of vehicles moving on the road, traditional information processing techniques and equipment are insufficient to deliver properly operating smart transportation application and services. New ideas are necessary to blueprint, design, and eventually implement state of the art modeling and control systems. When coping with control problems, involvement of measurement and signal processing tasks are inevitable. Therefore, to handle issues that came into existence from the ever-increasing complexity of today's systems, experts concentrate exclusively on modeling and signal processing methodologies and application of non-classical, *AI*, and soft computing based techniques. In the paper [120], authors demonstrate current issues with the measurement and signal processing techniques and how this issues may lead to new dimensions and re-assessment of these techniques if necessary. Conventional methodologies are unsuccessful in many cases when generating handy, valuable solutions. Particularly, measurement and signal processing problems are significantly complex due to involvement of a broad range of different disciplines and requirement for a multiple components and methods. This work presents a short overview of miscellaneous imprecise computational methods and debates on their applicability to handle complex modeling, signal processing, and control problems.

Emergency Management: Two main causes of accidents on the road are driving 1. carelessly and 2. in a hurry. It is possible to prevent accidents prior to happen with the early detection of dangerous driving conditions and warning drivers in such cases. Roadways that are equipped with the cameras and speed detectors can monitor and detect the drivers exceeding the speed limit allowed in that part of the road. This approach is fundamental, how it is primitive posing many limitations such as undetected drivers who usually accelerate to a speed exceeding the permitted limit outside of speed detector zone but slowdown in the range of a speed detector zone. Authors of the study [171] introduce an architecture for driver assistant system by applying image processing methods. A camera installed on the front window of a vehicle is used to identify road lanes markings and based on this information locate the position of the vehicle on the road. Hough Transform (*HT*) is modified to enhance the computational efficiency, and is adopted in the proposed architecture to determine the lanes automatically in real-time.

Pedestrian mortality rate at night due to traffic accidents constitutes a significant portion when compared with the pedestrian mortality rate during day-light. In [121], authors propose an automotive night vision system that is capable of identifying automatically pedestrians from a vehicle. The proposed system integrates the passive solutions such as systems adopting thermal vision, and active solutions such as systems using the near infrared camera. Systems with the former solution offers larger range of detection. On the other hand, systems with the active solution operates in a closer range, therefore provides more natural images. The software adopted to detect and classify objects takes advantage of modern digital signal processing methods such as connected component labeling (*CeL*), histogram of oriented gradients (*HOG*), and *SVM*.

Each year, accidents caused by tiredness of drivers are unprecedented. Vehicles that are controlled by tired drivers forms a serious thread for other drivers sharing the same road. Driver' fatigue does not only affect negatively the capability of drivers to understand what's happening around of drivers, but also affect negatively their mood. Some studies has been conducted to identify and quantify fatigue by using the measurement of physiology variables i.e., electroencephalogram (*EEG*), electrooculogram (*EOG*), and electromyogram (*EMG*). The paper [173] proposes a method to distinguish automatically the normal state from fatigue state with the given *EEG*, *EOG*, and *EMG* signals. Moreover, more complex features are developed to differentiate the experience drivers from the rest by computing the Lempel-Ziv complexity (*LZC*) of *EEG* approximate entropy. Three entropy-based features: (i) the wavelet entropy (*WE*), (ii) the peak-to-peak value of approximate entropy (*PP-ApEn*), and (iii) the peak-to-peak value of sample entropy (*PP-SampEn*) are derived from the signals acquired. Furthermore, authors introduce *WE* in a sliding window (*WES*), *PP-ApEn* in a sliding window (*PP-ApEnS*), and *PP-SampEn* in a sliding window (*PP-SampEnS*) to analyze driver fatigue in real-time. An *ANN* classifier is applied to real-time computation outcomes of these three sliding window values in order to determine the level of driver fatigue.

Information Management: Data transmission in vehicular networks between *V2X* is provided via antennas. Performance of wireless communication system can be made better with smart antennas by installing multiple antennas on one or both ends of the communications link. Smart antennas use extensively signal processing technologies. What enables data transmission between center-, traveler-, roadside-, and vehicle- type *ITS* subsystems possible is numerous wireless communication standards and technologies. *V2I* links relies on short-range communication technologies such as *RFID*, wireless *LAN*, and *UWB*. The range of short-range communication technologies, however, can be enhanced with mobile ad-hoc networks. Wide area mobile networks are used to establish the communication link between infrastructure and the roadside, vehicle, or traveler. The operating environment where wireless communication link is established poses various challenges. Furthermore, variation of data types transmitted through the communication channels indicate design constraints for the systems relying on wireless communication technologies. For instance, some applications are delay tolerant but they require no transmission errors such as web-surfing, e-mail. Whereas, some applications can allow transmission errors to some extend but not delays. Moreover, high data rate is needed when transferring large data files and video, and higher priority when using the network capacity is required in emergency response applications.

Authors in [69] categorize the smart antennas based on the number of transmitters and number of receivers in the wireless link into three: (i) Single-Input Multiple Output (*SIMO*), (ii) Multiple-Input Single-Output (*MISO*), and (iii) Multiple-Input Multiple-Output (*MIMO*). Authors describes the superiorities of a smart antennas over the conventional antennas when coping with the challenges described above in wireless communication links. Smart antennas is capable of increasing the range through array gain, enhancing the performance through diversity gain, increasing the capacity and quality of the wireless link through interference cancellation. Moreover, data rate can be increased through higher

Signal to Noise plus Interference Ratio ($SINR$). Additionally, authors introduce in [105] a simulation environment for *VANETs* that accommodates the smart antennas with dynamic radiation patterns. In order to enhance the performance of applications in *VANETs*, beam forming based on geocase messages is adopted in this work.

IEEE 802.11p standard is developed by *IEEE* to add wireless access in vehicular environments (*WAVE*). Bandwidth necessary for applications that can potentially be used in densely populated areas like in urban city, multiband CR-enabled *802.11p* waveform generation is recommended. In [149], authors simulate *IEEE 802.11p* waveforms in Matlab and AWR Virtual system simulator (*VSS*). The generated waveforms are, then, plotted with cancellation plot and Bit Error Rate (BER) graph to evaluate the performance of the waveforms. Moreover, the propose simulation also involves the multiband transmission by spectrum sensing to facilitate the generation of cognitive radio-enabled *802.11p* waveform.

Table 2.1: Overview of smart transportation services and referenced applications. Services are s1: traffic management, s2: public transportation management, s3: emergency management, and s4: information management.

Service	Data Analytics Methodologies		Tools and Frameworks	Vehicular Sensing
	Machine/Deep Learning	Signal Processing		
s1	Smart parking: <i>SVM</i> [135, 150], Traffic flow prediction: <i>kNN</i> [94], <i>ANN</i> [184], State-space Neural Network Model [97], <i>LTSM</i> [148], and <i>ANN</i> with <i>FNN</i> [43], Smart driving: <i>SRV</i> [162, 15] and <i>LS-SVM</i> [178]	Traffic control management: <i>GMM</i> [22] Traffic congestion detection: <i>Matlab</i> Image Processing Toolbox [74]	Smart parking: scikit-learn and PyBrain library [150, 183], Traffic flow prediction: <i>H2O</i> package [123]	Traffic accident prediction [155], Traffic condition evaluation [156], Traffic anomaly analysis [115]
s2	Travel mode detection: <i>DT</i> , <i>kNN</i> , and <i>SVM</i> [47], Passenger flow prediction: Wavelet- <i>SVM</i> [142], Passenger activity detection: <i>k-means</i> [57], Bus arrival prediction: <i>k-means</i> [140, 92], <i>AFC</i> data analysis: <i>DBSCAN</i> , <i>k-means</i> , <i>HAC</i> , and Hierarchical Clustering Techniques [96, 81, 21, 85, 104, 34]	Moving vehicle detection [35, 30, 139, 93], Efficient computation for transportation systems [120]	Travel mode detection: <i>Keras</i> library and <i>TensorFlow</i> framework [39], Traffic optimization: <i>TensorFlow</i> [90], Smart routing: <i>scikit-learn</i> , <i>TensorFlow</i> , <i>Keras</i> , and <i>XGBoost</i> [79].	Bus arrival prediction [177], Smart parking [48], Smart routing [168]
s3	Traffic accident prediction: <i>kNN</i> [117, 95, 172, 95], Traffic safety forecasting: <i>SVM</i> [50], False information detection: <i>BBAs</i> [182]	Driver assistance system: Hough Transform [171], Automotive night vision system: <i>CeL</i> , <i>HOG</i> [121], Drivers fatigue detection [173]	Collison avoidance: <i>TensorFlow</i> [129] and <i>Caffe</i> [158], On-time vehicle breaking system: <i>scikit-learn</i> [122], Classification of traffic congestions: <i>scikit-learn</i> [147]	Road surface analysis [20], Pothole detection [53], Bridge condition analysis [99]
s4	Intrusion detection: <i>CLA</i> [82], Smart data transmission: <i>CLA</i> [83], Managing data congestion: <i>k-means</i> [144]	Smart antenna [69, 105], Waveform simulation [149]	Network status prediction: <i>Theano</i> , <i>Scipy</i> , and <i>scikit-learn</i> [174], Vehicular privacy: <i>numpy</i> , <i>scipy</i> , and <i>scikit-learn</i> [154, 153], Privacy protection: <i>scikit-learn</i> [55], Communication failure identification: <i>Theano</i> , <i>Scipy</i> , and <i>scikit-learn</i> [174]	Smart message delivery [16], Smart data acquisition and dissemination [163], Security treat classification [23]

Chapter 3

Public Transport Tracking via Vehicular Crowd-Sensing: Proposed Supervised Methodology

3.1 Introduction

3.1.1 Problem Description

As the connected vehicles have become a reality, vehicular sensing has appeared as a viable solution to assist in several services –in addition to infotainment services– in smart cities such as traffic management, forensics, and air pollution control [131]. Wide usage of smartphones and the CAN-Bus interfaces in vehicles can be consolidated into a semi-dedicated sensing platform so that the smartphones can read the data through the interface and transform the data into a usable format by vehicular services [7].

Wide usage of smartphones and the CAN-Bus interfaces in vehicles can be consolidated into a semi-dedicated sensing platform so that the smartphones can read the data through the interface and transform the data into a usable format by vehicular services [7]. However, localization and tracking of vehicles is a crucial problem in a smart environments [26, 29, 6, 27, 113, 28, 114, 5]. For instance, under hostile settings where GPS signal quality is not of high quality, a smart localization approach has to be used by connected vehicles.

3.1.2 Proposed Solution

We propose a participatory GPS-less solution for the localization of public transportation vehicles moving over static routes [78]. Our proposed framework relies on two primary building blocks: *(i)* Multi-dimensional crowd-sensed data that is acquired from the built-in sensors of the smartphones and/or tablets carried by the passengers, *(ii)* Online prediction engine that

couples probabilistic location regions using a machine learning-based classification. Our proposed solution targets to obtain crowd-sensed data from multiple built-in sensors to estimate the location of a public transportation vehicle within the bounds of a region. Moreover, as we focus on vehicles moving on static routes, while eliminating GPS usage, by estimating a bounded region as the location of the vehicle, we protect location privacy of the participants who provide their sensor readings.

Through simulations, we show that the location of a bus can be estimated by the proposed framework with a 90% accuracy. The quality of our accuracy results – when compared with the accuracy results of *GPS* – are relative to the environment where public transportation vehicles operate. In urban areas (e.g., downtown of a city) with full of obstacles (e.g., high-rise buildings) that disable *GPS* signals to reach the *GPS*-enabled devices, the localization accuracy of the public transport vehicle is going to be lower compared to rural areas where localization accuracy of *GPS* may reach up to 100%. Since public transportation services are mostly active in urban environments, our proposed framework shows near *GPS*-like, or even better than *GPS* – with the enhancements we will explain in Chapter 5 – performance in terms of localization accuracy. Furthermore, we also show that instead of aggregating participatory sensed GPS data, which would have been possible, acquiring ambient and kinematic sensor data can lead to some 38–46% savings in the battery energy of participating devices. Besides the benefits of the proposed framework in terms of prediction accuracy and energy savings, avoiding exact GPS data in the sensory acquisition stage also ensures location privacy of the participants.

3.2 Proposed System Overview

Our proposed framework is illustrated in Fig. 3.1, which consists of four main subsystems: **1)** Data acquisition, **2)** Data aggregation, **3)** Probabilistic prediction, and **4)** Analytics. Acquired data is stored on a remote server, on which analytics and probabilistic prediction algorithms run. Table 3.1 presents the notation used in the description of the proposed framework. We explain the components of our proposed framework in detail in the Section 3.3. Moreover, Figure 3.2 illustrates the sequence of events that takes place during localization of vehicles.

3.3 Proposed Methodologies

3.3.1 Data Acquisition

Data acquisition mainly involves the recruitment of passengers for participatory sensing of ambient and kinematic data. Here, a possible challenge might be the quantification of the trustworthiness of the acquired data [124]. However, by making a fundamental assumption,

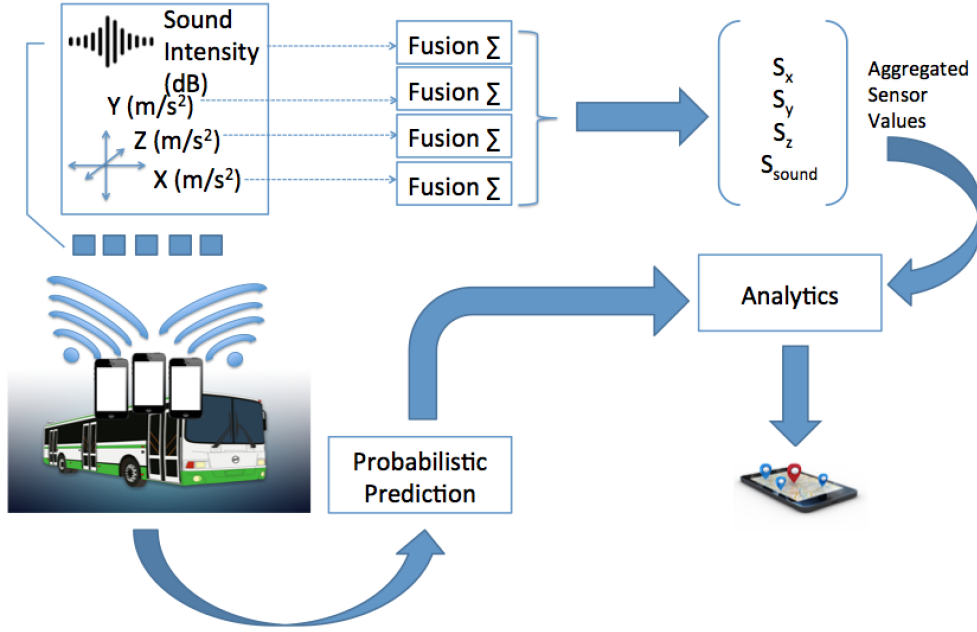


Figure 3.1: GPS-less localization architecture based on supervised machine learning.

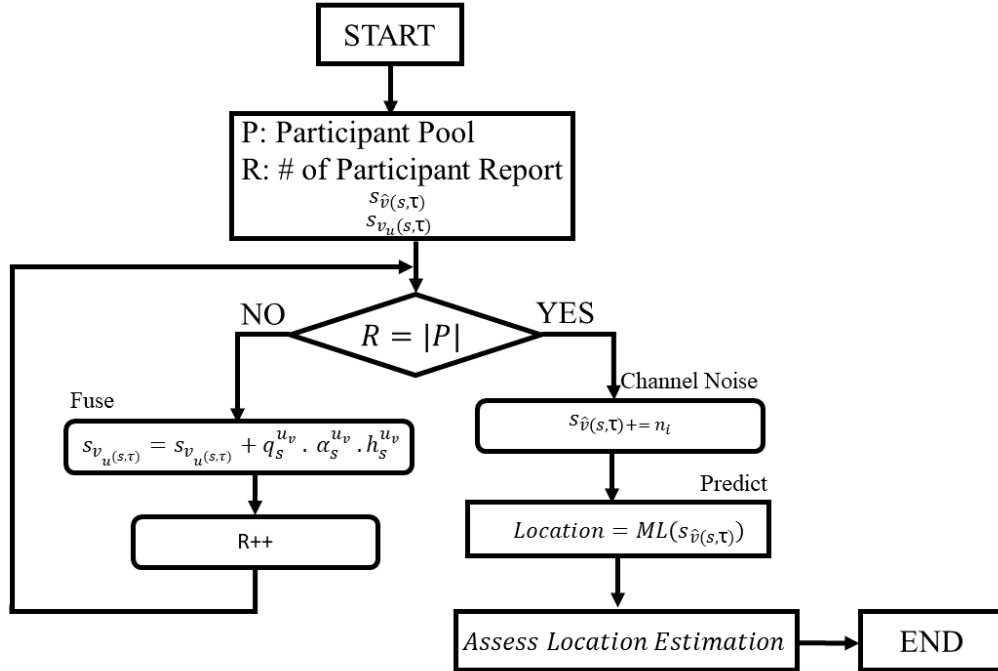


Figure 3.2: A generic flowchart for *GPS*-less vehicle localization.

Table 3.1: Notations for GPS-less localization architecture based on supervised machine learning. The variables and constants are grouped in three categories: The first category (top section of the table) denotes the objects, i.e., sensors, sensory detection zones and coordinates, second category includes the variables regarding aggregated sensed values, the third category is related to the variables regarding communications, and the fourth category (bottom section of the table) is related to participant recruitment scenarios.

Parameter	Value
k	Index of a window of ℓ landmarks
v	Index of a vehicle
τ	Time (discrete) of interest
r_τ^v	The sensory detection zone (SDZ) of vehicle v at time τ
z_i	SDZ i
z_{ix}	x coordinate of the SDZ i
z_{iy}	y coordinate of the SDZ i
z_{iz}	z coordinate of the SDZ i
ℓ	# of possible landmarks that a vehicle can be present
$\mathfrak{T}_k^v$	k^{th} sliding window for set of possible landmarks for vehicle v
Δ_τ	Time between two predictions
ν_τ^v	Velocity of vehicle v at time τ
ν_τ^v	Differential velocity vehicle v during Δ_τ
$\hat{\nu}_\tau$	Aggregated velocity at time τ
$\hat{\nu}_{\tau-1}$	Aggregated velocity at time $\tau - 1$
$\Phi_{\mathfrak{T}_k}$	Centre point of the SDZ of the k^{th} window
$\Phi_{\mathfrak{T}_k,x}$	X coord. of the centre point of the SDZs of k^{th} window
$\Phi_{\mathfrak{T}_k,y}$	Y coord. of the centre point of the SDZs of k^{th} window
$\Phi_{\mathfrak{T}_k,z}$	Z coord. of the centre point of the SDZs of k^{th} window
$\mathfrak{D}(a, b)$	Euclidean distance between point a and point b
$S_{\hat{v}(s,\tau)}$	Aggregated value of sensor s in vehicle v at time τ
$S_{v_u(s,\tau)}$	Sensor s reading reported by user u in vehicle v . time τ
q_s^{uv}	Noise factor of sensor s of user u in vehicle v
α_s^{uv}	Amplification of sensed data s of user u in vehicle v
h_s^{uv}	Fading coeff. of sensor reading s of user u in vehicle v
n_c	Channel noise experienced by the aggregator
$\eta - GPS$	Heterogeneous participant recruitment with GPS
$\eta - nGPS$	Heterogeneous participant recruitment without GPS
$\alpha - GPS$	Active participant recruitment with GPS
$\alpha - nGPS$	Active participant recruitment without GPS

we model each participant/passenger with rational and cooperative behavior which results in acquired data that does not interfere with the value of the aggregated data beyond possible noise factors.

3.3.2 Data Aggregation

Data aggregation is a fusion center which runs a distributed estimation technique over the acquired data from the sensors. As mentioned above, no adversary behavior is assumed in the data aggregation subsystem but possibly noisy readings and communication channel at each participant that would distort the individual signal transmitted to the aggregator. To this end, the aggregator in this framework simply adopts the estimation technique in [18] with the assumption that transmission of every sensor reading ($S_{v_u(s,\tau)}$) is prone to fading effect of $h_s^{u_v}$, and amplified by $\alpha_s^{u_v}$ whereas the aggregator experiences a channel noise with normalized variance n_c . Given these, the aggregated value of the built-in sensors of U participants in vehicle v ($S_{\hat{v}(s,\tau)}$) is formulated as shown in Eq. 3.1.

$$S_{\hat{v}(s,\tau)} = \sum_{u=1}^U (S_{v_u(s,\tau)} + q_s^{u_v}) \alpha_s^{u_v} h_s^{u_v} + n_c. \quad (3.1)$$

3.3.3 Probabilistic Prediction

Probabilistic prediction mainly involves estimating the location of the vehicle within a window of ℓ landmarks, i.e., **Sensory Detection Zone (SDZ)**. The concept of window of SDZ is illustrated in Fig. 3.3, which accelerates the classification process in the analytics module. It is not possible to estimate the exact location of a vehicle without GPS; however given the probability of a vehicle's presence in a window of ℓ stationary regions (i.e., in this example, $\ell = 3$), probabilistic deduction can be made for the next possible ℓ stationary regions that the vehicle can be present. In the figure, at $\tau = 1$, the vehicle is predicted to be in a stationary region denoted by $\mathfrak{T}_1 = \{r_1, r_2, r_3\}$. If the probability of the vehicle's presence in this window is one, the vehicle is expected to be in $\mathfrak{T}_2 = \{r_2, r_3, r_4\}$ at $\tau = 2$. Given that the vehicle was predicted to be in a landmark of the timeslot $TS_{k' < k}$ at $t = \tau - \Delta_\tau$, the probability of vehicle v being located in the landmarks of the timeslot T_k at time $t = \tau$ can be calculated as a conditional probability as formulated in Eq. 3.2.

$$\begin{aligned} & Pr(r_\tau^v \in \mathfrak{T}_k^v | r_{\tau-\Delta_t}^v \in \mathfrak{T}_{k' < k}^v) \\ &= \frac{Pr(r_\tau^v \in \mathfrak{T}_k^v \cap r_{\tau-\Delta_t}^v \in \mathfrak{T}_{k' < k}^v)}{Pr(r_{\tau-\Delta_\tau}^v \in \mathfrak{T}_{k' < k}^v)}. \end{aligned} \quad (3.2)$$

Due to the discrete nature of these two occurrences, the probability formulation can be transformed to the simpler form in Eq. 3.3.

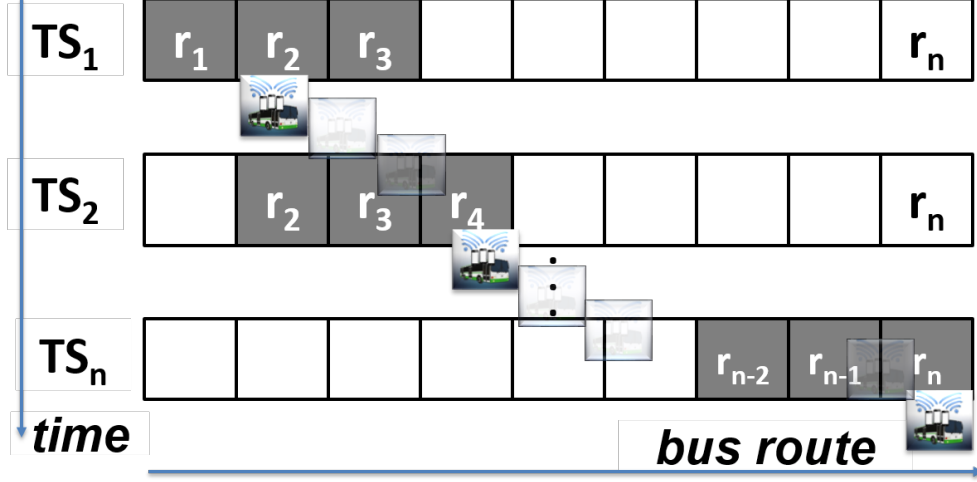


Figure 3.3: The concept of sliding window of stationary regions.

$$\begin{aligned}
 & Pr(r_\tau^v \in \mathfrak{T}_k^v | r_{\tau-\Delta_t}^v \in \mathfrak{T}_{k'<k}^v) \\
 &= \frac{Pr(r_\tau \in \mathfrak{T}_k) Pr(r_{\tau-\Delta_\tau} \in \mathfrak{T}_{k'<k})}{Pr(r_{\tau-\Delta_\tau} \in \mathfrak{T}_{k'<k})}.
 \end{aligned} \tag{3.3}$$

The probability of a vehicle being in a SDZ of the window $\mathfrak{T}_k$ can be calculated as the ratio of the total distance that can be covered from the center point of the SDZs of the previous window to the center point of the SDZs of the k^{th} window ($\mathfrak{D}(\Phi_{\mathfrak{T}_{k-1}}, \Phi_{\mathfrak{T}_k})$) as formulated in Eq. 3.4. Here it is worth noting that the differential aggregated value of the velocity of the vehicle during the interval of Δ_τ is used to estimate the distance that can be covered. The center point of window $\mathfrak{T}_k$ is denoted by a vector of xyz coordinates, and is calculated by obtaining the mean of the x, y, and z coordinates of each SDZ as shown in Eq. 3.5.

$$\begin{aligned}
 & Pr(r_\tau \in \mathfrak{T}_k) \\
 &= \frac{\sqrt{\mathfrak{D}(\Phi_{\mathfrak{T}_{k-1}}, \Phi_{\mathfrak{T}_k})^2 - \left(\frac{\hat{v}_\tau - \hat{v}_{\tau-1}}{\Delta_\tau}\right)^2}}{\mathfrak{D}(\Phi_{\mathfrak{T}_{k-1}}, \Phi_{\mathfrak{T}_k})}.
 \end{aligned} \tag{3.4}$$

$$\begin{aligned}
 & \Phi_{\mathfrak{T}_k} = [\Phi_{\mathfrak{T}_k, x}, \Phi_{\mathfrak{T}_k, y}, \Phi_{\mathfrak{T}_k, z}] \\
 &= \frac{1}{\ell} \left[\sum_{i=1}^{\ell} z_{i_x}, \sum_{i=1}^{\ell} z_{i_y}, \sum_{i=1}^{\ell} z_{i_z} | z_i \in \mathfrak{T}_k \right].
 \end{aligned} \tag{3.5}$$

3.3.4 Analytics

Analytics module takes the output of probabilistic estimation of potential landmarks (i.e., SDZs) that the vehicle is expected to be at, and runs (and keeps continuously training) a machine learning-based classifier to estimate the location among the ℓ landmarks.

As mentioned earlier, localizing public transportation vehicles by GPS can be costly in terms of energy, as well as GPS coverage issues. If GPS signals from multiple mobile devices are used to predict the current location on earth, the energy consumption on the participating devices becomes a critical factor as most of those devices run on batteries. Thus, we propose utilizing built-in sensors in smart mobile devices to acquire data, and integrate the sensed data with the proposed framework in Fig. 3.1 that consists of the four components explained above. The output of the analytics module of the proposed framework is expected to provide an estimated region (i.e., approximate location) of a public transportation vehicle. To this end, while sensor data acquisition application collects data from passengers inside of a public bus, probabilistic prediction module provides an input to the analytics module with a window of predicted stationary regions throughout the static route, which accelerates the classification procedure in the analytics module. The analytics module runs a machine learning algorithm.

Analytics and probabilistic prediction modules run simultaneously. After the initial simulation day, as the analytics module classifies/trains aggregated sensor data, probabilistic prediction module starts making predictions on already classified sensor data with the next simulation day's sensor data.

To clarify the scope of this paper, it is worth mentioning that in this paper, we use Support Vector Machines (SVM) for the purpose of analyzing and making predictions on the acquired data; however feasibility analysis of various machine learning algorithms on the proposed framework is included in the ongoing extensions of this work. Besides, here, sensors that are used to collect data in order to replace GPS during data acquisition process are limited to accelerometer and microphone, whereas various combination of different sensors is also being investigated by our ongoing research agenda.

In the next section, we will introduce two performance evaluation schemes: (i) *GPS-less* localization accuracy, and (ii) *GPS-less* localization energy consumption. Furthermore, we will investigate the energy consumption of proposed architecture under different participant profile (e.g., active, inactive participant profile), and different sensor distribution.

3.4 Performance Evaluation

3.4.1 Data Acquisition and Simulation Settings

To evaluate our proposed framework, we have collected real data by using Google's Science Journal application [56] for five consecutive days on a static route under a linear velocity

varying between 30-40 kmph towards 15 landmarks (i.e., stops) that are located at least 500 meters apart (available at [109]). The built-in sensors used in data collection are mainly accelerometer (movement along x, y, z axes in terms of meters) and the microphone which senses the sound pressure level in dB. Probability distribution of the collected data has been fed into the CrowdSenSim [49] simulator to generate sensed data by each type of sensor throughout the virtual simulation duration. CrowdSenSim is a discrete-event simulator which enables the simulation of a large-scale crowd-sensing activity in an urban scenario in order to generate innovative solutions in data collection, task assignment, monitoring, and resource management. Simulation settings are summarized in Table 3.2.

Table 3.2: General settings

Parameter	Value
Number of vehicles	50
Channel noise variance at the aggregator(n_c)	1.0
Average velocity [kmph]	{30, 40}
Average stop time [s]	[30, 60]
Minimum inter-station distance	0.5 km
Simulation duration	5 days
Propagation environment	Rayleigh fading
Noise model in sensor readings	AWGN
Classifier	SVM

We have simulated 50 public buses that move over 17 different lines (possibly with overlaps) each of which consists of 15 stops. To simulate bus stops, we have randomly selected 15 random points from the street map in the simulator, which are in form of (latitude, longitude, altitude). Those points are picked according to following criteria: Each consecutive points are ensured to be at least 500 meters apart from each other. There should not be any colliding points. Once the stops (i.e., GPS coordinates of the landmarks) have been determined, they are sorted. As the public transportation vehicles move along static routes, simulating the mobility model is straightforward. Initially each bus is placed on the initial point of its corresponding line. Bus moves forward to its next stop with a constant randomly chosen velocity. After arriving the next stop, the bus waits between 30 seconds to 60 seconds. This waiting time is also determined randomly on every time a vehicle arrives at a stop. When a bus reaches the final destination, it returns back to its initial position.

In the simulation of the data acquisition module, each passenger’s sensor reading is subject to Additive White Gaussian Noise (AWGN). The analytics module uses Support Vector Machine (SVM) with polynomial Kernel type and a Kernel degree of 15 under the termination criteria of either 100 iterations or < 0.01 difference between two consecutive results. In the simulation results, every point represents the average of ten runs with a confidence level of 95%.

GPS-less Localization Accuracy

Accuracy is calculated as the ratio of the correct predictions to total predictions throughout the total simulation duration (T). A correct prediction denotes the case where the actual GPS coordinates of a public vehicle is within the estimated stational region (r_τ^v) as formulated in Eq. 5.6.

$$\text{Accuracy} = \frac{\sum_{\tau}^T (r_\tau^v \in TS_\tau \wedge r_\tau^{lv} = r_\tau^v)}{T}. \quad (3.6)$$

We investigate the impact of the following parameters on the accuracy: Window size that is the number of possible landmarks that a vehicle can be present (ℓ), route length, and the noise factor at the built-in sensor reading as listed in Table 3.3. All tests have been conducted under varying number of passengers that participate in the crowd-sensing campaign throughout the route.

Table 3.3: Accuracy settings

Parameter	Value
Number of participants	{5, 10, 15, 20, 25, 30}
Stational region window size (ℓ)	{3, 4, 5}
Static route length	{15, 20, 30}
Sensing noise (q_s^{uv})	{1.0, 1.5, 2.0, 2.5, 3.0}

GPS-less Localization Energy Consumption

We investigate the energy consumption under two different schemes:

1) *Active vs Inactive Participant Profile Scheme*: We have compared energy consumption of GPS with the energy consumption of our proposed GPS-less localization framework. To this end, we categorize the passengers/participants into two categories, namely *active* and *inactive*. Activity denotes a coefficient regarding interaction frequency of a participant with their mobile device. Energy consumption of an active participant (E) in our simulation is the total energy used by previously determined social media applications (E_{apps}) [11] plus the energy consumption of GPS (E_{GPS}) or data acquisition application ($E_{mic} + E_{acc}$) as formulated in Eq. 3.7.

$$E = \begin{cases} E_{apps} + E_{mic} + E_{acc} & \eta - nGPS \vee \alpha - nGPS \\ E_{apps} + E_{GPS} & \eta - GPS \vee \alpha - GPS \end{cases}. \quad (3.7)$$

Under this scenario, participants are recruited according to following four recruitment scenarios: (i) $\eta - GPS$, (ii) $\eta - nGPS$, (iii) $\alpha - GPS$, and (iv) $\alpha - nGPS$. While Heterogeneous participant recruitment (η) acquires data from both active and inactive participants, active participant recruitment (α) acquires data from only active participants.

After evaluating the accuracy results (that we will present in Section 3.4.2), we have determined the simulation settings listed in Table 3.2 that yield the worst and the best accuracy results. While simulation settings with static route length of 15, window size of 5, and buses containing 20 participants performs the best, simulation settings with static route length of 15, window size of 3, and buses containing 30 participants performs the worst. Energy consumptions are measured with regard to heterogeneous and active participant recruitment under the worst and the best simulation settings. The results are presented in Section 3.4.2.

2) *Identical vs Non-identical Sensor Array Pool (Combination) Scheme*: Heterogeneous participant recruitment scenario is adopted with identical or non-identical sensor distribution methods to visualize the energy consumption rates of various sensor combinations in a bus under the worst and the best simulation settings. In identical sensor array pool (i.e., combo) scheme, participants in a bus can use accelerometer, microphone, or both of them equally distributed among participants. For instance, a bus can acquire 30% microphone, 30% accelerometer, and 40% both microphone, and accelerometer data. On the other hand, sensors are distributed to participants in the bus with non-identical sensor array pool (i.e., combo) as follows:

- **Non-identical Sensor Array Pool - 1:** In a bus, 50% of passengers using only accelerometer; 20% of passengers using only microphone; and 30% of passengers using both accelerometer and microphone.
- **Non-identical Sensor Array Pool - 2:** In a bus, 20% of passengers using only accelerometer; 30% of passengers using only microphone; and 50% of passengers using both accelerometer and microphone.
- **Non-identical Sensor Array Pool - 3:** In a bus, 30% of passengers using only accelerometer; 50% of passengers using only microphone; and 20% of passengers using both accelerometer and microphone.

Energy consumptions are measured with regard to heterogeneous participant recruitment technique along with the sensor combination types presented above under the worst and the best simulation settings. The results are presented in Section 3.4.2.

3.4.2 Numerical Results

GPS-less Localization Accuracy

Figure 3.4a illustrates the scenario where route length is set to 15. As seen in the figure, the higher the number of participants, the less the accuracy experienced. This is due to the added noise to the sensor readings as the number of participants (i.e., sensors) increases. On the other hand in each iteration, the larger the window size, the higher the GPS-less localization accuracy. The reason of this phenomenon is that higher number of clusters increases the probability of accurate estimation of the location within the stational regions. As setting $\ell = 5$ results in the best accuracy in the previous figure, we set the window size ℓ to 5 in the second experiment, and vary route length.

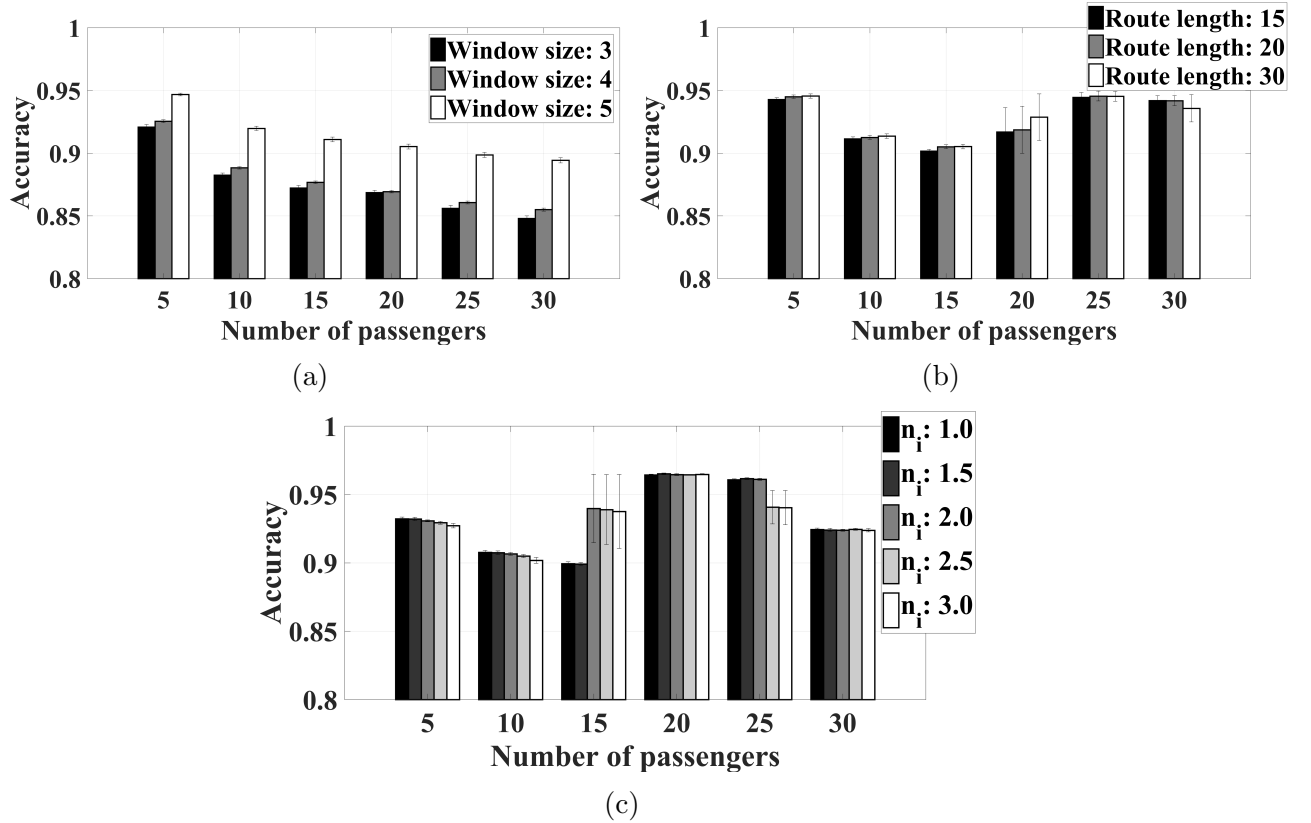


Figure 3.4: Accuracy performance under three varying simulation settings (a) The effect of window size on accuracy (b) The effect of route length on accuracy (c) The effect of Gaussian noise on accuracy.

As it is illustrated on Fig. 3.4b, longer routes improve the accuracy of GPS-less localization. The best results can be achieved when the number of participants is limited to five and the route length is set to 30 stops. As the window size is large, the noise impacting sensor readings can be low when the number of participants is small, and when the route

is long. As the route gets longer, and once the prediction of the next window can be done accurately, classification problem can be done more effectively.

To see the effect of AWGN on the accuracy, the third experiment is conducted, the results of which are presented in Fig. 3.4c where n_i is the variance factor of Gaussian noise. In our simulations, it is the Gaussian distribution of standard deviation of 100 subsequent sensed values of sensor type i . In order to conduct this experiment, standard deviation of 100 subsequent sensors is multiplied with a varying coefficient, n_i . Initial value of n_i is to 1. In every test case, it is increased by 0.5 until 3.0. Here, route length and window size have been set to 5 and 15, respectively as that combination resulted in the best accuracy in the previous results. As seen in Fig. 3.4c, increasing n_i has a slight effect on the accuracy in most of the test cases except the iterations with 15 and 25 passengers. Nevertheless, the best accuracy has been recorded with 20 passengers.

Based on these results, the following specifications can be made for GPS-less localization in smart public transportation. The overall best accuracy value of 97% can be reached with the following parameters set: $r = 5$, route length = 15, number of participating passengers = 20, and $n_i = 1.0$ -3.0. Conversely, the worst overall accuracy has been recorded as 86% with the following specifications: $r = 3$, route length = 15, number of participating passengers = 30, and $n_i = 1.0$.

GPS-less Localization Energy Consumption

1) *Active vs Inactive Participant Profile Scheme*: Figs. 3.5.a-b compare the energy consumption of all four recruitment scenarios per vehicle under best and worst cases in terms of GPS-less localization accuracy. Best case denotes the settings (i.e., route length, number of participants, window length and noise characterization) that results in the best accuracy value whereas the worst case denotes the opposite. As the inactive participant ratio increases in heterogeneous recruitment scenarios, overall energy consumption of mobile device decreases by 38%-46%.

In 3.5.c-d, multi-dimensional comparison of energy, accuracy and activity is presented under best and worst case settings for localization accuracy of $\eta - nGPS$. As seen in both sub-figures, increased user activity results in reduced energy consumption due to localization, and better localization accuracy under both settings.

2) *Identical vs Non-identical Sensor Array Pool (Combination) Scheme*: Experiments for varying sensor combinations are conducted under $\eta - nGPS$. 3.6 illustrates the energy consumptions of $\eta - nGPS$ with four different sensor combinations. While 3.6a displays the energy consumption results under the best simulation settings, 3.6b displays the energy consumption results under the worst simulation settings. The worst performing sensor combination is non-identical sensor array pool - 2 under both simulation settings, whereas the best performing sensor combination is non-identical sensor array pool - 3. Therefore, a brisk increase in the number of passengers only using both sensors also increases the total energy consumption.

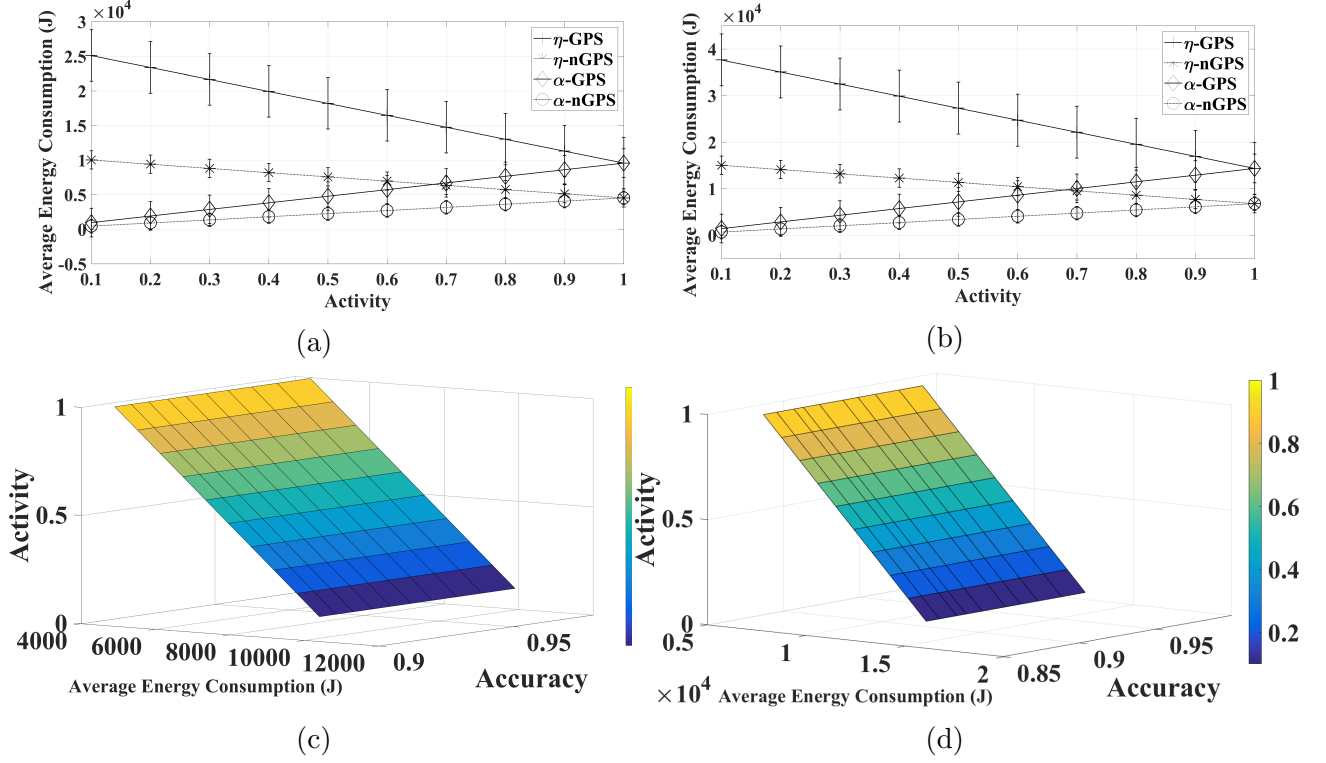


Figure 3.5: Energy performance per vehicle under different scenarios and recruitment schemes (a) Best case energy consumption comparison of $\eta - GPS$, $\eta - nGPS$, $\alpha - GPS$ and $\alpha - nGPS$ (b) Worst case energy consumption comparison of $\eta - GPS$, $\eta - nGPS$, $\alpha - GPS$ and $\alpha - nGPS$ (c) Best case of $\eta - nGPS$ energy vs accuracy vs activity (d) Worst case of $\eta - nGPS$ energy vs accuracy vs activity.

In Fig. 3.7 and 3.8 multi-dimensional comparison of energy, accuracy, and activity of identical and non-identical sensor array pools are presented respectively. While Fig. 3.7a presents the multi-dimensional results under the best simulation settings, 3.7b presents the multi-dimensional results under the worst simulation settings. In all experiments illustrated in Fig. 3.7 and 3.8, higher activity rates lead to high accuracies and higher energy consumptions. We have observed from Fig. 3.8a and 3.8b, the location accuracy can go under 60% with high number of participants using accelerometer, and less number of participants using microphone. Moreover, experiment results show from Fig. 3.8c, and 3.8e yields high location accuracies while consuming less energy compared to Fig. 3.8d, and 3.8f. While overall less energy consumption with the highest performance in terms of localization accuracy is achieved by identical sensor array pool under the best simulation settings, high energy consumption with the least localization accuracy is achieved with the non-identical sensor array pool - 1.

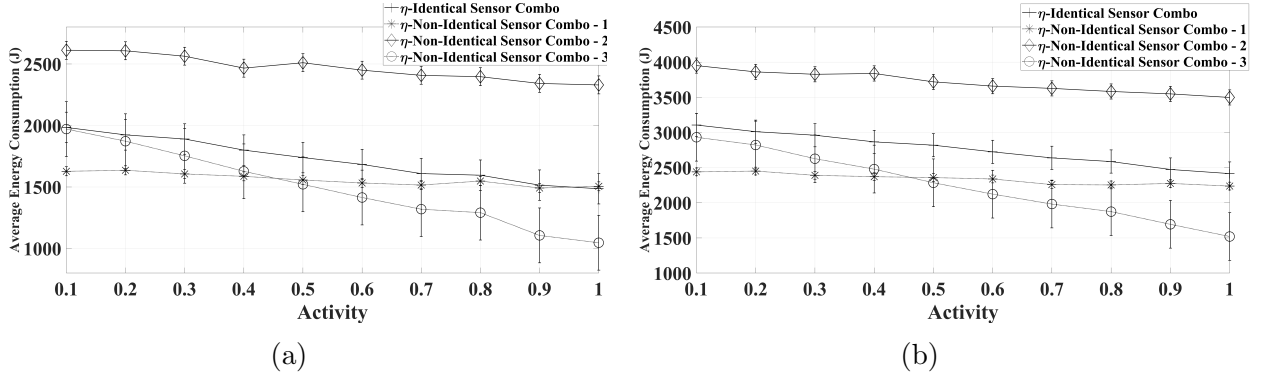


Figure 3.6: Energy performance per vehicle with four different sensor array pool schemes under best and worst simulation settings respectively (a) Best case energy consumptions of identical sensor array pool (i.e., combo), non-identical sensor array pool (i.e., combo) - 1, 2, and 3 (b) Worst case energy consumptions of identical sensor array pool, non-identical sensor array pool - 1, 2, and 3.

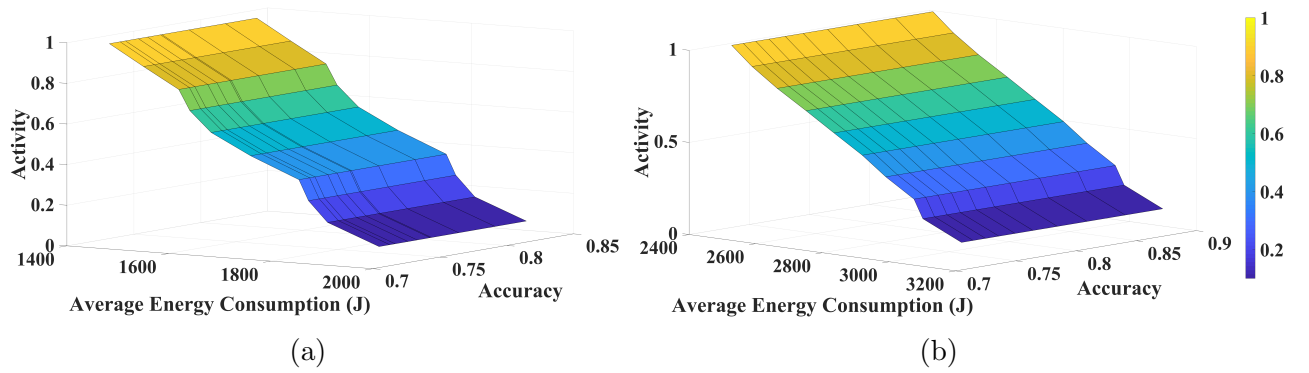


Figure 3.7: Energy (per vehicle) and accuracy performance with identical sensor array pool scheme under best and worst simulation settings (a) Best case energy consumption vs accuracy vs activity (b) Worst case energy consumption vs accuracy vs activity.

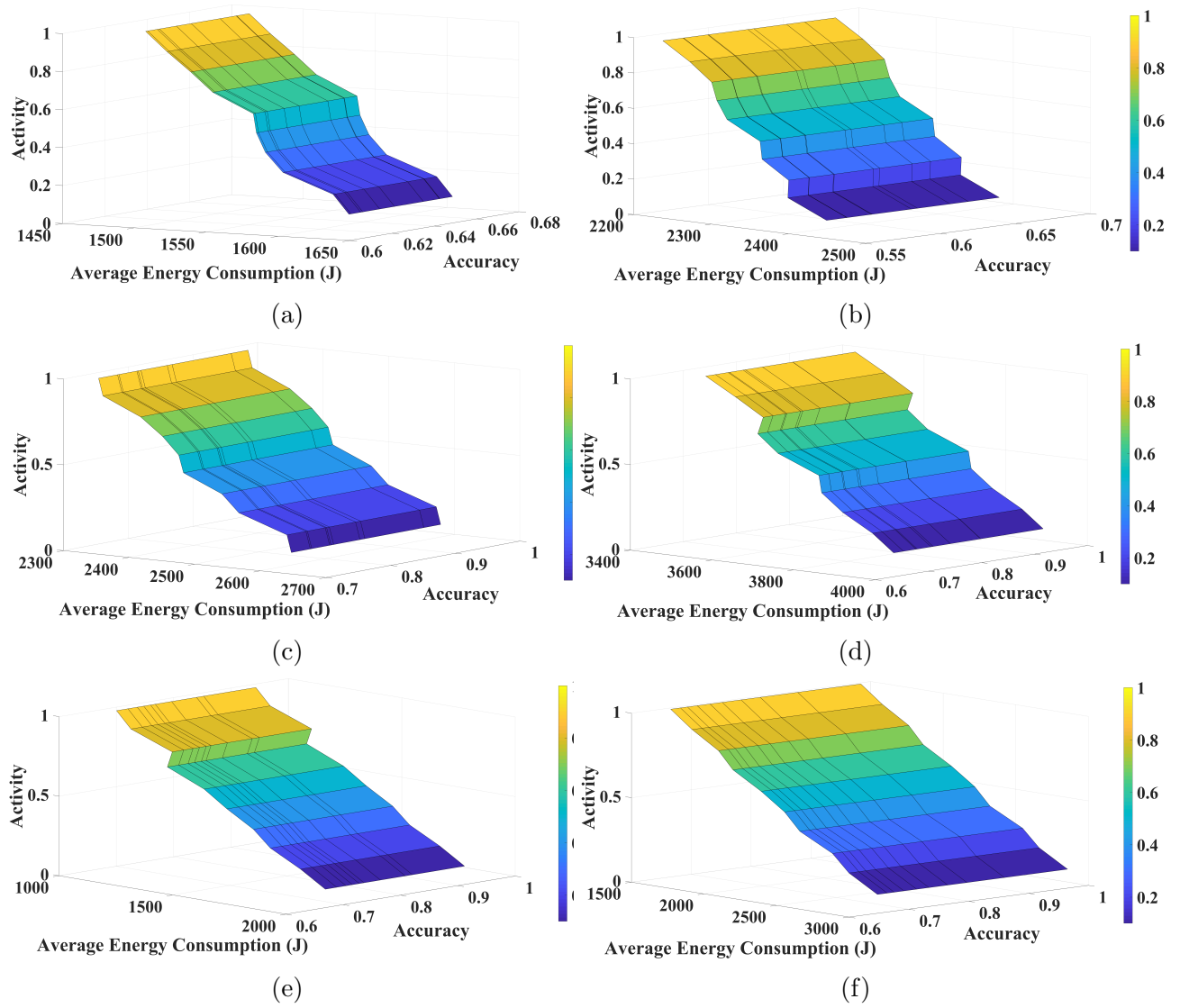


Figure 3.8: Energy (per vehicle) and accuracy performance with three different non-identical sensor array pool (i.e., combo) schemes (non-identical sensor array pool - 1, 2, and 3) under best and worst simulation settings. Best case energy consumption vs accuracy vs activity with (a) non-identical sensor array pool - 1 (c) non-identical sensor array pool - 2 (e) non-identical sensor array pool - 3. Worst case energy consumption vs accuracy vs activity with (b) non-identical sensor array pool - 1 (d) non-identical sensor array pool - 2 (f) non-identical sensor array pool - 3.

3.5 Concluding Remarks

Smart transportation and vehicular sensing are crucial components of smart city services. Localization of public transport vehicles by reducing the usage of GPS sensors can be achieved by covering the location data through other type of sensors that are already available in a non-dedicated manner in a smart infrastructure. Based on this motivation, we have proposed a crowd-sensing-based framework to locate public transportation vehicles in a smart city setting without using GPS. Moreover, we also evaluated the performance of the proposed framework in terms of energy consumption with different availability of the built-in sensors to investigate the effect of sensor availability on accuracy and energy performance. The proposed framework consists of data acquisition, sensory data aggregation, probabilistic prediction and analytics modules. Accelerometer and microphones on mobile devices have been used to acquire sensed data and replace GPS. Data collected through a real application have been fed into simulation environment where 50 bus lines have been simulated with each having 15 bus stops. We adopted SVM in the analytics module, and tested the proposed framework by varying the probabilistically derived location region, route length of buses and the characterization of the noise factor introduced from the mobile devices. We have achieved an overall accuracy of above 90%, and 80% in the best and worst cases, respectively. We have also investigated the energy consumption of our best and worst cases performing simulation settings with active and passive participant profiles. Our results showed that GPS-based localization results in draining up to 46% higher battery power than the overall energy consumption of our proposed GPS-less localization solution. Particularly, our proposed identical sensor array pool yields the highest localization accuracy with the lowest energy consumption. On the other hand, non-identical sensor array pool - 1 yields the lowest localization accuracy with the highest energy consumption.

In the next chapter, we introduce an unsupervised machine learning method and replace the proposed supervised machine learning method integrated in analytics module. The performance of the framework in terms of localization accuracy is assessed under different availability scenarios of built-in sensors.

Chapter 4

Public Transport Tracking via Vehicular Crowd-Sensing: Proposed Unsupervised Methodology

4.1 Introduction

4.1.1 Problem Description

Public transportation is among the services that call for digital and intelligent solutions in smart cities. However, implementing even a basic feature to provide extra functionalities can be costly in terms of sensor deployment into vehicles and implementing the communication and processing infrastructure. For instance, in order to develop a service to track vehicles moving on static routes (e.g., buses, streetcars, or trains) and providing real-time or near real-time feedback about their distances to a certain station, all public transportation vehicles need to be equipped with GPS devices unless they have not yet been equipped with one.

Previously, we studied the emulation of GPS sensors under a case study for localizing public buses in a smart city setting [78] with crowd-sensing approach to overcome the problem of costly sensor deployment. To emulate a sensor, we applied an ML method on sensory data read in a non-dedicated manner. We used a *supervised machine learning*-based clustering to predict the location of a vehicle on a route based on the acquired data. Although, we achieved over 90% of accuracy when estimating the location of vehicles, the proposed framework had to work within limited location windows of landmarks over the paths to achieve that accuracy. Thus, a probabilistic module would estimate a sensor detection zone (SDZ) that a vehicles is expected to be present; and the ML algorithm would run on that limited window of landmarks (i.e., SDZ) to estimate the precise location.

4.1.2 Proposed Solution

Our aim is to eliminate the SDZ concept and propose a holistic and plug-and-play framework to unveil the potential of vehicular crowd-sensing in the localization of public transportation vehicles that move on static paths [77]. Furthermore, we also aim to respect privacy of sensing data providers by running the estimation for discrete stational landmarks as opposed to continuous GPS coordinates. To this end, we propose a framework that comprises of two main building blocks: (i) crowd-sensed multi-dimensional data gathered from multiple built-in sensors of smart-devices of participants, (ii) an unsupervised machine learning module – that integrates k-means algorithm – to estimate the approximate location of a vehicle. The results of our simulations show that the location of a vehicle can be predicted by the proposed framework with an accuracy up to 95% without not dropping lower than 80% in the presence of low quality data. Moreover, training of *k-means* algorithm took remarkably less time than the training of *SVM*. Based on our observations, the training of *SVM* took between 30 minutes to 30 hours to be completed, while the training of *k-means* has been completed between 5 minutes to 1 hour.

Furthermore, as a result of exploiting the VCS concept, capital and operational costs are significantly reduced allowing the proposed framework to leading to *easy to deploy* and *flexible to integrate* benefits for the proposed framework. Moreover, high energy savings of participants devices and preservation of participants locations (i.e., privacy respect) are shown to be possible through the elimination of GPS sensors.

4.2 Proposed System Overview

A minimalist illustration of our proposed architecture is presented in Figure 4.1 with a single bus line moving on its route, and acquiring accelerometer and microphone data from participants, which comprises of three primary building blocks: (i) Data Acquisition, (ii) Data aggregation, and (iii) Data analytics. A cloud platform stores the sensory data acquired by participants in the vehicles. Data analytics and predictions on the data accumulated and aggregated in the cloud run as time elapses. To help the readership, we also present the notation used in the paper as follows: Table 4.1 accommodates the constants, Table 4.2 represents the parameters used in the *k-means* formulation, Table 4.3 is associated with the communication parameters, and Table 4.4 describes the participant recruitment scenarios. We explain the components of our proposed framework in detail in the Section 4.3.

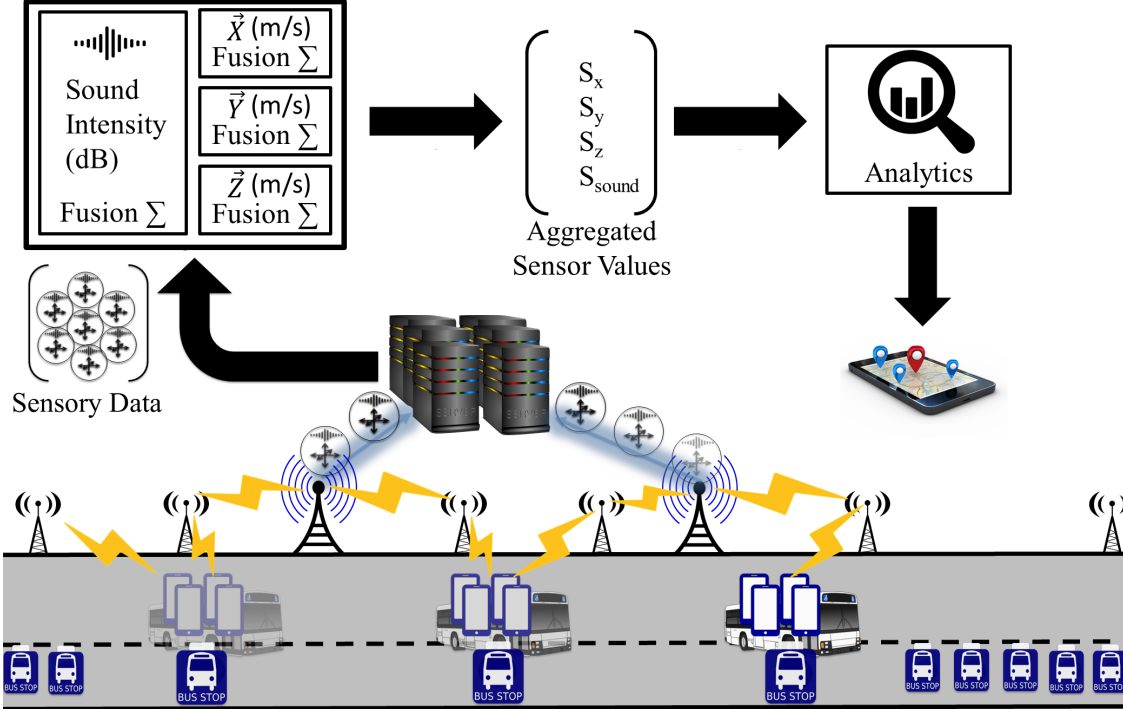


Figure 4.1: GPS-less localization architecture based on unsupervised machine learning.

4.3 Proposed Methodologies

4.3.1 Data Acquisition

Sensory data are acquired from the participants located in each bus. Buses move on their static routes with a constant velocity stopping at pre-defined stations/stops for a certain amount of time for boarding or dropping-off passengers. A subset of passengers are recruited as participants to crowd-sense data and communicate it to the cloud platform for analysis.

Acquired data from participants in vehicles are transmitted to platform by employing the principles of Vehicular Ad Hoc Networks (*VANETs*). Collected data at *RSUs* later will be conveyed to the platform. The *RSUs* are placed at strategically pre-determined locations so that provided vehicular communication sustains effectively and reliably [46]. In the data acquisition module, participants may send their aggregated sensory data to the nearest *RSU* within the coverage area of participant via cellular network interfaces. They may also transmit their sensory data via WiFi to a gateway deployed in the vehicle, which will convey the collected sensory data from to the *RSU* through WLAN. While sensory data flows from vehicle to the *RSU*, there is no data transmission among vehicles in our scenario. Besides these, as proposed in [101] vehicle-to-vehicle-to-infrastructure (*V2V2I*), which can leverage *V2V* communications.

Since sensory data is conveyed to its destination wirelessly, the loss rate of the data is

higher compared to wired communication links [45]. Typical factors affecting the lost rate of wireless communications are: **(i)** the gap/distance between the source and receiver, **(ii)** communication noise, and **(iii)** collision rate. Therefore, in order to ensure desired transmission rates and throughput, the communication plane of the proposed architecture needs to be complemented by error recovery techniques as studied in [45]. Here, a possible challenge might be the quantification of the trustworthiness of the acquired data [124]. However, by making a fundamental assumption, we model each participant/passenger with rational and cooperative behavior which results in acquired data that does not interfere with the value of the aggregated data beyond possible noise factors.

Table 4.1: The constants that are used in formulas such as number of vehicles and clusters, sensor array, and so on.

Parameter	Value
V	# of vehicles
v	Index of a vehicle
K	# of clusters
k	Index of cluster
P	# of participants in a vehicle
p	Index of participant
R	Randomly selected sensory data array
u_v^p	Participant p in vehicle v
π_k	Size of cluster k
c_k	δ -dimensional centroid of cluster k
θ	δ -dimensional sensor array
d	Day d
τ	Time (discrete) of sensing in a day
ℓ	# of possible landmarks
$\gamma(a, b)$	Binary function returns 1 if $a = b$
$\mathfrak{D}(a, b)$	The closest Euclidean distance between point a and the points in array b
$\min \mathfrak{D}(a, b)$	The smallest Euclidean distances between the points in array a and b
$\mathfrak{A}(a)$	Sum of elements of array a

Table 4.2: The parameters that are used to formulate the $k - means$ algorithm

Parameter	Value
$f(\theta, d, \tau)$	Total of squared errors of sensor value θ with respect to all centroids
$S^{\theta,d,\tau}$	Acquired sensory data on day d at time τ by the platform
$r^{\theta,d,\tau}$	Randomly selected sensory data from acquired sensory data on day d at time τ by the platform
S_k^θ	Sensory data in cluster k
$\epsilon_{S_k^\theta}^d$	Label of sensory data, S_k^θ , in cluster k on day d

Table 4.3: The parameters that are associated with the communication channel

Parameter	Value
$O_{u_p^v}^{\theta,d,\tau}$	Aggregated value of sensor θ of participant p in vehicle v on day d at time τ
$S_{u_v^p}^{\theta,d,\tau}$	Reading of sensor θ reported by participant p in vehicle v on day d at time τ
$q_\theta^{u_v^p}$	Noise factor of sensor θ of participant p in vehicle v
$\alpha_\theta^{u_v^p}$	Amplification of sensed data θ of participant p in vehicle v
$h_\theta^{u_v^p}$	Fading coefficient of sensor reading θ of participant p in vehicle v
n_a	Channel noise experienced by the aggregator

Table 4.4: The descriptions for participant recruitment scenarios

Parameter	Value
$\eta - rec$	Even participant recruitment
$\alpha - rec$	Uneven participant recruitment
$\eta - rec.i$	Even participant recruitment with sensor usage distribution scenario- i
$\alpha - rec.i$	Uneven participant recruitment with sensor usage distribution scenario- i

4.3.2 Data Aggregation

Data aggregation is an intermediate step acquisition and analysis of sensory data. Data aggregation involves processing the acquired sensory data and presenting the processed data in a compressed format. In our proposed architecture, input of the data analytics component is the compressed data generated by the data aggregation module.

In the proposed architecture, data is gathered and sent by participants carrying battery

dependent and resource-limited devices (such as smartphones or tablets) equipped with non-dedicated sensors to the cloud in order to be stored and processed. Multiple participants transmit their sensory data concurrently, which overloads the network and may cause network congestion if not handled properly. Aggregation of data is a solution to overcome the following problems: **(i)** battery constraint of devices that are used by participants for sensing tasks, **(ii)** network overload that results from sending huge sensed data simultaneously by many participants, and **(iii)** transmission costs due to the size of sensory data.

As we aim to formulate theoretical foundations of our proposed framework by considering practical and realistic assumptions, it is worth briefly studying the state-of-the-art in data aggregation to address the above mentioned challenges. These challenges in data aggregation can be addressed by various techniques. For instance, the studies in [118, 70, 12] address the battery limitation of devices and offer data aggregation based solution while the authors in [145] analyze different data aggregation solutions to solve network bandwidth issues that could possibly emerge due to enormous amount of participants transmitting sensory data. Decrease in the utilization of network resources helps to reduce energy consumption on participant devices.

A greedy aggregation approach evaluated in [70] reduces energy consumption by increasing the amount of path sharing with the adjustment of aggregation points. Cooperative-Input-Multiple-Output (CMIMO) combined with data aggregation has proven to drop the energy cost down by decreasing the amount of data in transit [118]. A fuzzy logic-based technique has been employed in [145] to solve network bandwidth issues that could possibly emerge due to enormous amount of participants transmitting sensory data. Decrease in the use of network helps to reduce energy consumption on participants' devices.

There is direct relation between the amount of data in transit and transmission cost. Hence by reducing the volume of sensory data, it might be possible to reduce the cost of transmission. This can be achieved by eliminating the correlated information in the sensed data. In our proposed framework, recruited passengers in a vehicle are in close proximity to each other. Therefore, it is expected that there exists a spatial correlation between the sensory data that are received from the same origin. Sensory data could possibly experience noise while being recorded by individual participants and while being transmitted through communication channel towards the central cloud system. We consider similar effects as in [18] in the aggregation of sensory data. Eq. 4.1 formulates the aggregated values of built-in sensors of participant p in vehicle v ($O_{u_v^p}^{\theta,d,\tau}$). In the equation, fading effect ($h_{\theta}^{u_v^p}$) amplified by $\alpha_{\theta}^{u_v^p}$ represents the noise, channel noise with normalized variance of n_a represents the distortion in the sensor reading ($S_{u_v^p}^{\theta,d,\tau}$) during transmission.

$$S^{\theta,d,\tau} = \sum_{v=1}^V \sum_{p=1, p'=u_v^p}^P (S_{u_v^p}^{\theta,d,\tau} + q_{\theta}^{u_v^p}) \alpha_{\theta}^{u_v^p} h_{\theta}^{u_v^p} + n_a. \quad (4.1)$$

4.3.3 Data Analytics

Data analytics is the last component in the proposed architecture where aggregated sensory data is used to estimate the location of the vehicle among the ℓ landmarks (i.e., stops / stations). In the analytics module, an unsupervised machine learning algorithm runs continuously in order to cluster the sensor readings into *landmark whereabouts* / *stational regions*.

While Figure 4.1 illustrates the high level architecture, Figure 4.2 presents the complete crowd-sensing process in our architecture. Passengers in each vehicle are considered as participants that are equipped with non-dedicated sensors. Participants record data for d days with their non-dedicated sensors and transmit them to the cloud to be processed and analyzed. Sensory values from all of the participants in a vehicle throughout a day are aggregated before being sent to the cloud. In the figure, $S_{v(\theta,\tau)}^d$ is the aggregated values of the non-dedicated sensor θ from vehicle v_m traveling from l_1 to l_n at time τ on day d .

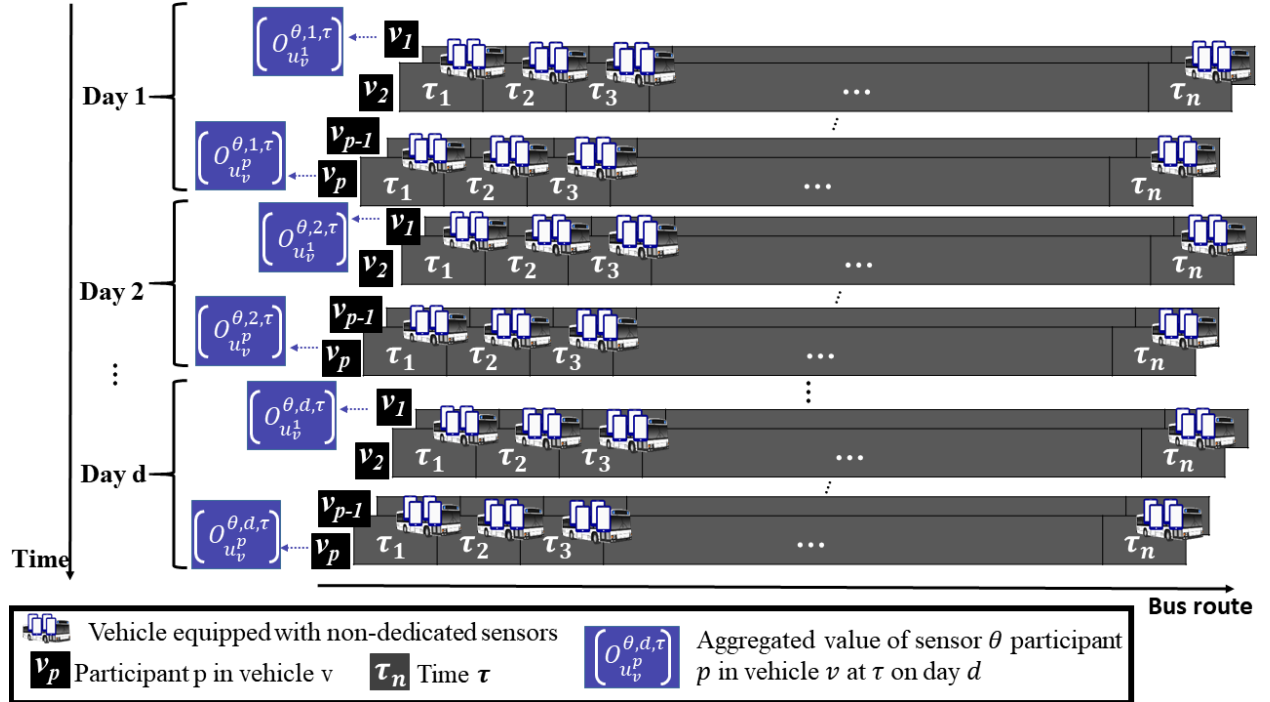


Figure 4.2: Application of the crowd-sensing concept to a public transportation network containing m vehicles each of which moves on a static route of n stops / stations.

As mentioned earlier in data acquisition section, multiple participants in a vehicle are recruited to crowd-sense data through their smart phones or tablets. Sensory data acquired by any participant in any day is transmitted to the platform upon aggregation (*Aggregated data*: $O_{u_v^p}^{\theta,d,\tau}$) as illustrated in Figure 4.2. Aggregated sensory data is fused in the fusion center of the cloud platform and is fed into the analytics module.

In the analytics module of the cloud, an unsupervised learning method, namely *k-means*, is used to analyze and approximate the location of a vehicle in the vehicular network. The ML algorithm, *k-means* [166] aims at partitioning the given data into pre-determined number (k) of clusters without leaving any unclustered data as an outlier (i.e., noise).

Aggregated sensory data from a participant p in a vehicle v at time τ of the day d ($O_{u_v^p}^{\theta,d,\tau}$) is given as sensory data input to *k-means*. Sensory data is clustered in such a way that all the values in the same cluster show similarity to the other points in the same cluster. Indeed, this implies that data points in different clusters are expected to be as dissimilar as possible. Clusters are defined by their centroids that are initially determined using generally one of the two methods: (i) Forgy or (ii) Random Partition. While the former method randomly selects k sensory values from $S^{\theta,d,\tau}$ and assigns these as the initial centroids of the clusters, the latter assigns all sensory values to random clusters and then re-calculates location of the centroids along with the sensory values in clusters until the locations of centroids are stabilized. In this paper, we use a modified randomized initialization technique introduced in [14] called *k-means++* for our initialization phase. *K-means++* initialization method selects the centers one by one in a controlled manner, where currently selected centroids randomly alter the choice of the next centroid.

Initialization process starts with an empty set, R , which stores randomly selected centroids c_k for each cluster k . *K-means++* initialization technique selects randomly a sensory value, $r^{\theta,d,\tau}$ from $S^{\theta,d,\tau}$ to be the cluster of the first centroid c_1 (See Eq. 4.2). Remaining centroids are chosen based on the following probabilistic metric formulated in Eq. 4.3; i.e., the closest Euclidean distance of a random sensory value, $r^{\theta,d,\tau}$, to the random sensory values in set R over the sum of minimum Euclidean distances between every sensory value in $S^{\theta,d,\tau}$ and R . $r^{\theta,d,\tau}$ values yielding higher probabilities will be added to the random centroid set, R .

$$R \Leftarrow \emptyset \cup r^{\theta,d,\tau}. \quad (4.2)$$

$$c_k = \left[\frac{\mathfrak{D}(r^{\theta,d,\tau}, R)}{\mathfrak{A}(\min \mathfrak{D}(S^{\theta,d,\tau}, R))} \right] \cup R, \quad \forall r^{\theta,d,\tau} \in S^{\theta,d,\tau} \forall r^{\theta,d,\tau} \notin R, \forall k \in K. \quad (4.3)$$

After initialization of centroids, each sensory value is placed to the cluster that is represented by a centroid that is closest to that sensory value among all other centroids. This similarity /closeness is calculated by using a distance-based approach. *K-means* uses euclidean distance method for this purpose as it is formulated by the Eq. 4.5. This processes is also called labeling as each cluster is represented with a label and sensory data assigned to a cluster is classified with its corresponding label.

$$\min f(\theta, d, \tau), \quad (4.4)$$

such that

$$f(\theta, d, \tau) = \sum_{k=1}^K (S^{\theta, d, \tau} - c_k)^2. \quad (4.5)$$

The objective of Eq. 4.5 is to minimize the squared error function, $f(\theta, d, \tau)$ as stated with Eq. 4.4 by placing sensory values to a cluster k , as formulated in Eq. 4.6.

$$\begin{aligned} \{S_k^\theta : (S_k^\theta - c_k)^2 \leq (S_k^\theta - c_j)^2\}, \\ \forall k, j | k \in K, j \in K, k \neq j. \end{aligned} \quad (4.6)$$

Centroids of clusters are updated following the completion of placing all the sensory data into randomly selected clusters. New centroid of cluster k , c_k , is calculated by the mean of sensory data in that cluster as formulated by Eq. 4.7.

$$c_k = \frac{1}{\pi_k} \sum_{j=1}^{\pi_k} S_j^\theta. \quad (4.7)$$

Euclidean distances between new centroids and the sensory data are re-computed using Eq. 4.5 based on the criterion formulated by Eq. 4.6. After re-labeling every sensory value in each cluster, centroids are re-calculated according to Eq. 4.7 once again. This process continues until total squared errors of sensed value ($f(\theta, d, \tau)$) converges to a local optimum for every θ , which denotes that locations of the centroids have become steady. It is worth noting that $k - means$ algorithm reaches a local optimum within a polynomial time complexity as shown in the Eq. 4.8 for the quantity of $\pi_k \delta$ dimensional sensory data in K clusters, whereas reaching global optimum by $k - means$ is an $NP - complete$ problem [166].

$$\begin{aligned} O(\pi_k \delta K), \\ \forall k \in K. \end{aligned} \quad (4.8)$$

The output of $k - means$ algorithm is the labeled version of the given sensory data, $S^{\theta, d, \tau}$. A cluster and the sensory values that belong to the same cluster are represented with the

same label. Under normal circumstances (rational and dependable participants with non-maliciously functioning non-dedicated sensors), aggregated sensory data of a participant in a vehicle, $O_{u_p}^{\theta, d, \tau}$, will have similar patterns throughout the journey on the same route everyday. Therefore the acquired sensory data of a participant on day d will be correlated with those acquired on the next day, $d + 1$. Sensory data received from the same stops/stations are expected to be placed within the same cluster, and thus marked with the same label. Setting the number of stops / stations equal to the number of clusters equal is needed to identify the location of transmitted sensory data. Grouping data in different clusters, even though they are not transmitted from those origins implies possible deviation in sensor value that can be caused by various factors, such as readings under unusual conditions, malfunctioning of sensor(s), or untrustworthy participant(s).

By comparing the labels of sensory data on day d and day $d + 1$, an approximation about the vehicle’s whereabouts can be made. If the labels of sensory data of a participant on day d at time τ are equal to the labels of sensory data of the same participant on the next day, $(d + 1)$, at time τ , it is interpreted as follows: There is a consistency between day d and $d + 1$, which will strengthen the estimation about vehicle’s location.

4.4 Performance Evaluation

4.4.1 Data Acquisition and Simulation Settings

In our performance evaluation tests, we simulate a vehicular crowdsensing network with varying physical conditions whereas the data used in the simulation environment are generated through real data [110] that was collected on our personal devices in a private vehicle following a bus route in downtown Ottawa by using the Science Journal mobile app of Google [56]. The app is capable of detecting the available sensors in a device once installed and launched. Out of all the sensors available, we have selected the following common sensors: **(i)** accelerometer, and **(ii)** microphone. The reason behind the selection of those sensors is that they can be found in the majority of nowadays’ smart-devices.

To emulate heterogeneity in the crowd-sensing environment, we have collected sensory data through 5 different device types (an iPhone 6s Plus, a Nexus 7, an iPhone 7, an iPad Pro, a Samsung Galaxy Tab) and 2 different operating systems (iOS and Android).

We have maintained the same route of 15 stops / stations that are placed at least 500 meters away from each other, and in each consecutive day, velocity of a vehicle has been varied between 30-40 kmph. We have collected 14 days worth of data, out of which 7 days of data have been used for conducting the experiments due to exclusion of the data containing irregular patterns caused by coarse deviation in non-qualified days. That has led us run the simulation for 7 days.

CrowdSenSim [49] (the same simulator that is used in Chapter 3) is used as the simulation platform for experiments, and it is introduced required classes and methods to simulate a ve-

Table 4.5: Global simulation settings

Parameter	Value
Number of vehicles	50
Antenna type of participants' smart devices	3G
Average velocity [kmph]	{30, 40}
Average stop time [s]	[30, 60]
Minimum inter-station distance	0.5 km
Simulation Duration	7 days
Propagation environment	Rayleigh fading
Noise model in sensor readings	AWGN
Clustering Method	K-means

hicular crowd-sensing environment. In our scenario, each sensor value is recorded over static paths; hence we aim to place each sensory data point in the most suitable cluster. Therefore, K-means is selected as our preferred clustering algorithm to determine the approximated location of the public vehicle from which those values have been received. When participants transmit their sensory values, the data could experience distortion to some extent due to communication channel noise. Hence the sensory data that are sensed by participants' mobile devices would lead to slight variations compared to the data accumulated in the cloud. In our simulation, we emulate this scenario by applying Additive White Gaussian Noise (AWGN) to each sensor value collected by the participants.

For simulation scenarios, random points in the form of latitude, altitude, and longitude from the street map of the simulation are selected as the stops / stations (i.e., landmarks) to create the static routes for the public transportation vehicles. Number of landmarks is set to 15 in most of the simulations; however we also vary the value of this parameter to investigate the impact of the route length on the performance of our proposal. Stops / stations of different routes may overlap with each other. The stops / stations are placed by at least 500 meters apart from each other while velocity and waiting time of any vehicle is set to a random value in [30, 40] kmph and [30, 60] s, respectively. Duration of the simulation is set to 7 days. 50 public vehicles that operate on 17 different lines are simulated. So far, we have presented the global simulation settings, which are summarized in Table 4.5.

Besides the global simulation settings, Table 4.6 lists the settings for customized experiments to investigate the best possible combination of the route length, cluster size and sensing noise under varying number of participants. The research findings of these investigations are further discussed in the Section 4.4.2.

Upon finding the configuration that would result in the best performance, we further investigate the impact of sensor availability on the participating devices when participant recruitment is pursued based on "nearly" even or non-even distribution. Even and non-even participant recruitments are denoted by $\eta - rec.$ and $\alpha - rec.$, respectively. To this end,

Table 4.6: Simulation settings for feasibility analysis

Parameter	Value
Number of participants	{5, 10, 15, 20, 25, 30}
Static route length	{15, 20, 30}
Cluster size	{10, 15, 20}
Sensing noise (q_s^{uv})	{1.0, 1.5, 2.0, 2.5, 3.0}

Table 4.7: Simulation settings for $\eta - rec.$ and $\alpha - rec.$

Parameter	Value
Number of participants	{3, 6, 9, 12, 15, 18, 21, 24, 27, 30}
Static route length	{10, 15}
Cluster size	{10, 15}
Sensor distributions for $\eta - rec.1$	{30%, 30%, 40%}
Sensor distributions for $\eta - rec.2$	{30%, 40%, 30%}
Sensor distributions for $\eta - rec.3$	{40%, 30%, 30%}
Sensor distributions for $\alpha - rec.1$	{50%, 20%, 30%}
Sensor distributions for $\alpha - rec.2$	{20%, 30%, 50%}
Sensor distributions for $\alpha - rec.3$	{30%, 50%, 20%}
Sensing noise (q_s^{uv})	2.5

availability of sensors are set to different values under each scenario. For instance, under the uneven recruitment scenario $\alpha - rec.1$, sensor distribution is set to {50%, 20%, 30%} denoting that 50% of the participants in a bus record only accelerometer data, 20% of the participants in a bus record only decibel data of microphone, and 30% of the participants in a bus record both accelerometer data and decibel values of microphone.

In each experiment, we keep a correlation between the route length and cluster size, i.e., if route length is 10, cluster size will also be set to 10. By doing so, we eliminate the SDZ concept in the previous work [78] as each stop / station over the route becomes a stand alone SDZ. It is worth noting that in the figures, each run simulates seven days, and the results in the figures present the average of five runs.

4.4.2 Numerical Results

We run three types of experiments: **(i)** Feasibility study for the combination of the route length, cluster size, sensing noise, and number of participants, **(ii)** $\eta - rec.$ analysis to examine the impact of uniform sensor availability on participating devices, **(iii)** $\alpha - rec.$ analysis to examine the impact of non-uniform availability of sensors on participating devices. All

experiments are conducted in a GPS-less manner.

The ratio of the correct predictions to total predictions that were made from the beginning of time till the end of a simulation time (T) denotes the accuracy as formulated in Eq. 4.9. If a given pair of accelerometer and microphone sensed by user in a vehicle at a stop/station is placed in the correct cluster where all the other sensed values from the very same landmark are placed, then the prediction made for this pair of values is correct. In the equation γ denotes a binary function which returns one if the label of sensory data in cluster k S_k^θ on day d and $d + 1$ remains the same.

$$\text{Accuracy} = \frac{\sum_{k=1}^K \sum_{d=1}^{D-1} \gamma(\epsilon_{S_k^\theta}^d, \epsilon_{S_k^\theta}^{d+1})}{K(D-1)}. \quad (4.9)$$

Feasibility analysis

We aim to determine the optimal values set for route length, cluster size, sensor noise and, number of passengers to conduct the participant recruitment scenarios under the best combination of these values. We run individual experiments for each parameter by varying the number of passengers from 5 to 30.

As depicted in all sub-figures displayed in Fig. 4.3, increasing number of passengers (i.e., participant) has an affirmative effect on the accuracy. Furthermore, as the cluster size of k-means algorithm increases from 10 to 15 and 20, accuracy also increases except the case where the number of passengers is 30, in which we see a slight drop-down from 10 to 20 cluster size. The reason behind this behavior is that it becomes easier to place participants in less number of clusters. The highest accuracy, approximately 93%, is achieved under the cluster size of 10, which is immediately followed by the cluster size of 15 with approximately 92% of accuracy where the number of passengers is 30.

Similar accuracy pattern as in the case of cluster size can be observed in Fig. 4.3b for route length. As the route gets longer, accuracy improves with the increasing number of passengers. However shorter route length tends to perform better, when the number of passenger reaches to 30. Since increasing number of passengers and route length increases the data generated, the clustering process, in turn, get facilitated better for k-means algorithm. Thus, the highest accuracy, approximately 92%, has been reached with the route length of 15 stops, and a participant population of 30 passengers.

Sensing noise, n_a , exhibits instable behavior under varying number of passengers (i.e., 5 and 10), however more stable behavior is achieved with the increasing number of passengers afterwards as depicted in Fig. 3.4c. The highest accuracy, approximately 92%, has been reached with $n_a = 2.5$ and $n_a = 3.0$ with a participant pool of 30 passengers per vehicle.

After analyzing these results, we conclude that the optimal combination of cluster size (π_k), route length (ℓ), sensing noise (n_a), and number of participants per vehicle (P) to lead

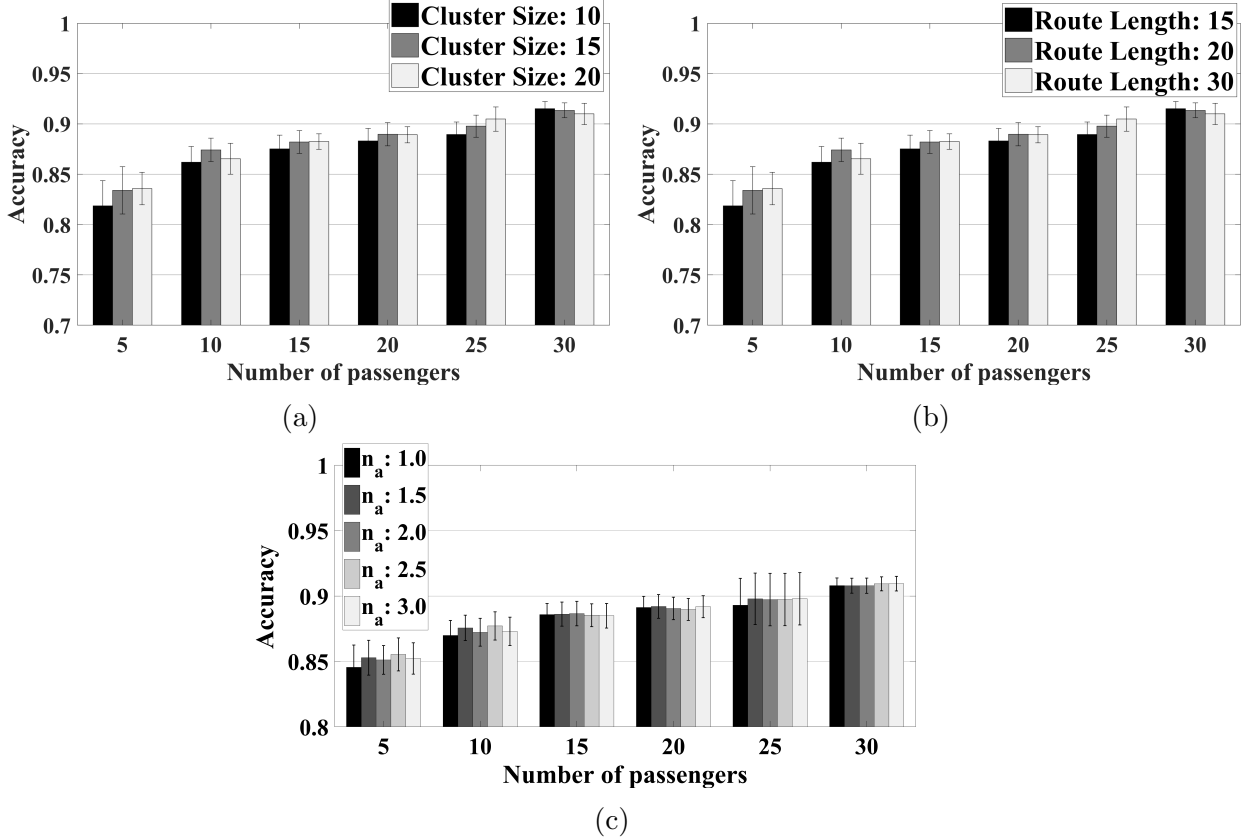


Figure 4.3: Accuracy performance under three varying simulation settings (a) The effect of cluster size on accuracy (b) The effect of route length on accuracy (c) The effect of sensing noise (n_a) on accuracy.

to higher accuracies can be constructed as follows: $\{\pi_k = 10, \ell = 15, n_a = 2.5, 3.0, P = 30\}$.

Localization accuracy under 'nearly even' availability of sensors ($\eta - rec.$)

In the simulations to test 'nearly even' availability of sensors ($\eta - rec.$), number of participants per vehicle is varied from 10% to 100% of its optimal value. A comparison between two different settings has been illustrated in the Figure 4.4: under cluster size (route length) of 10 and 15. As expected, the lower the number of clusters (i.e., number of landmarks/stations), the better the accuracy.

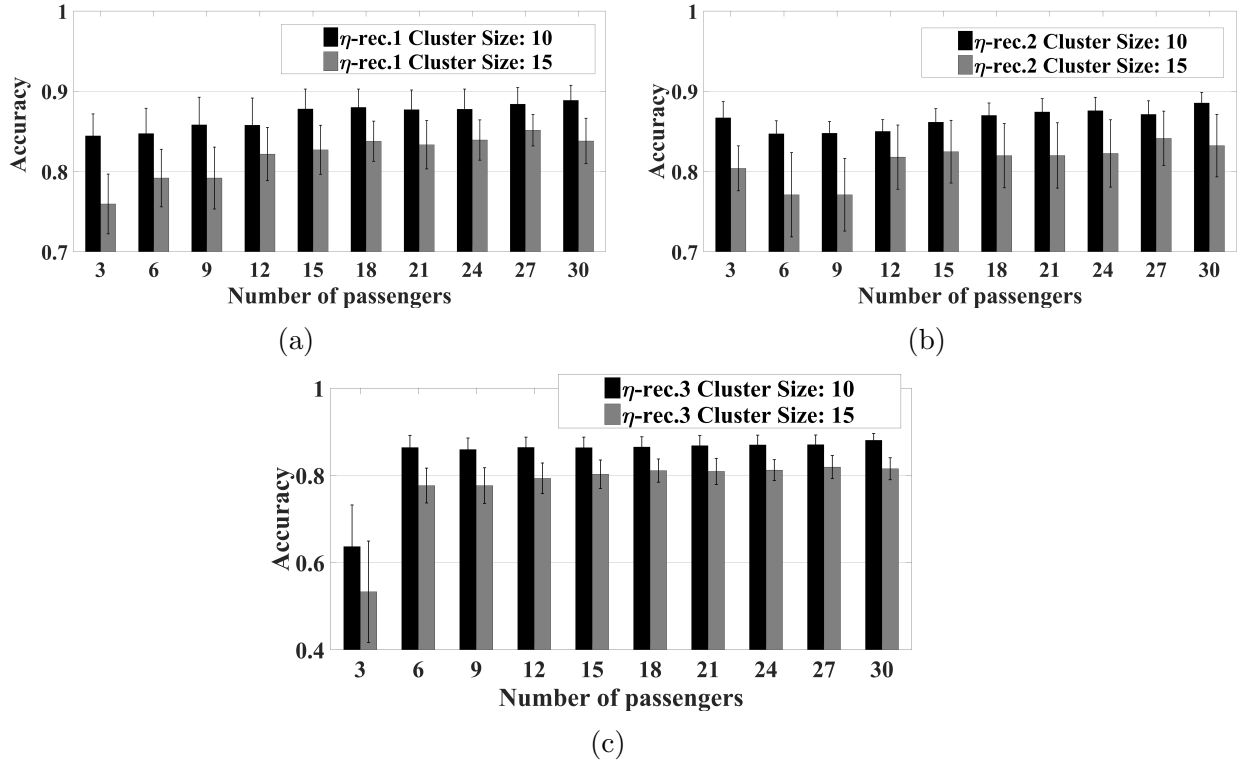


Figure 4.4: The effect of nearly even participant recruitment on localization accuracy under different sensor usage distribution scenarios and cluster sizes: (a) scenario-1 ($\eta - rec.1$) (b) scenario-2 ($\eta - rec.2$) (c) scenario-3 ($\eta - rec.3$).

Fig. 4.4a illustrates the impact of nearly even participant recruitment scenario on accuracy under 30% accelerometer, 30% microphone, and 40% both accelerometer and microphone availability. As the participant pool grows, accuracy under both cases increase. Furthermore, gradual increase in participant pool leads to narrower confidence intervals for localization accuracy. The highest accuracies under both route length (i.e., cluster size) settings were measured as 88% and 84%.

Nearly even recruitment scenario under the availability of 30% accelerometer, 40% microphone, and 30% both accelerometer and microphone with the cluster size (i.e., number

of stations / stops) of 15 has shown extreme low accuracy values as illustrated in Fig. 4.4b. The highest accuracies of both test cases are 88% and 84% with large confidence intervals.

Localization accuracies in Fig. 4.4c reveal overall consistencies throughout the entire simulation except the initial case where the participant pool is significantly small, i.e., 3 passengers per vehicle. Confidence interval has been also consistently measured minimal except the initial case. The highest accuracies under both route length settings are recorded as 89% and 81%.

Localization accuracy under uneven availability of sensors ($\alpha - rec.$)

Fig 4.5 encompasses of three sub-figures that illustrates the impact of uneven participant recruitment scenario on localization accuracy.

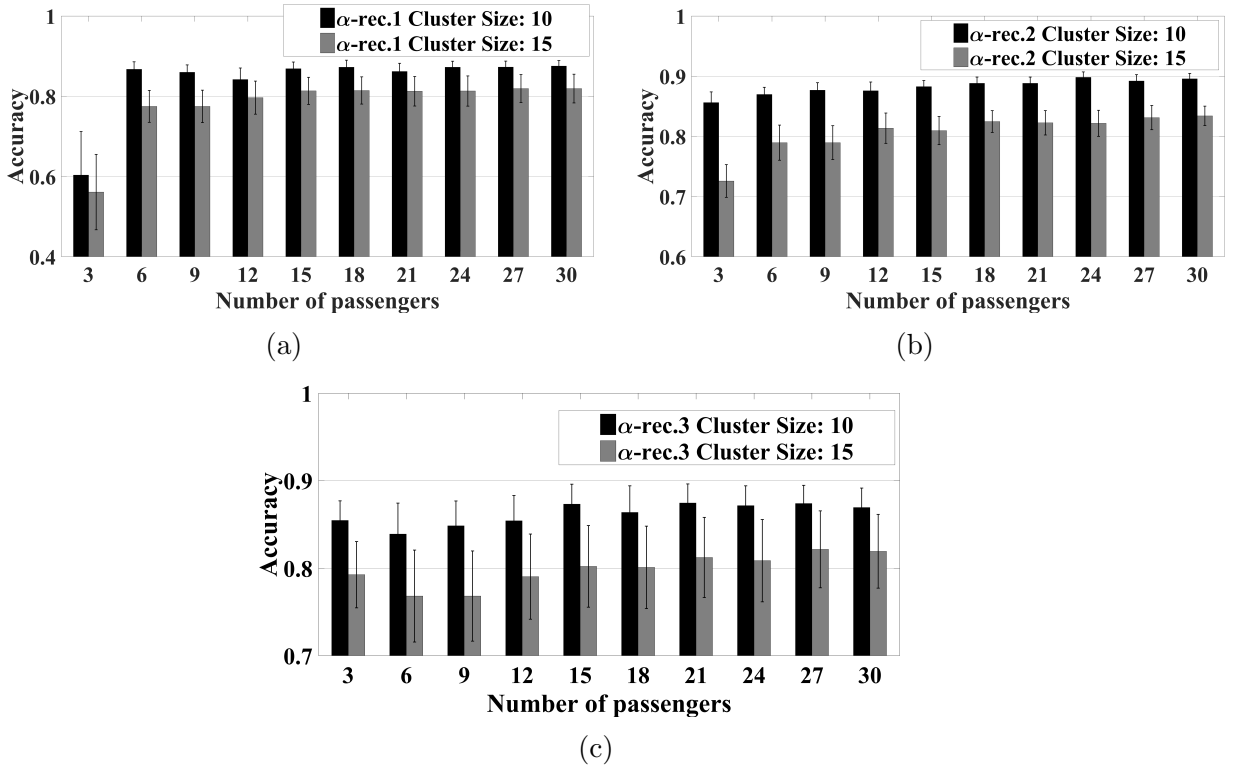


Figure 4.5: The effect of uneven participant recruitment on localization accuracy under different sensor usage distribution scenarios and cluster sizes: (a) scenario-1 ($\alpha - rec.1$) (b) scenario-2 ($\alpha - rec.2$) (c) scenario-3 ($\alpha - rec.3$).

Fig. 4.5a illustrates the impact of uneven participant recruitment scenario on localization accuracy under the availability of 50% accelerometer, 20% microphone, and 30% both accelerometer and microphone. Accuracy exhibits a sudden improvement after the first experiment (i.e., 3 participants per vehicle), and keeps a steady behavior for the rest of the

participant pool sizes. The highest accuracies achieved for both route length cases are 89% and 81%.

Fig. 4.5b presents a smooth increase in localization accuracy as the number of passengers increases. The highest accuracies recorded for both route length cases (i.e., 10 stops/stations and 15 stops/stations) are 90% and 83%. Accuracies in Fig. 4.5c reveal a remittent pattern. The highest accuracies were calculated as 88% and 82% with considerably high confidence interval ranges. The reason of large confidence intervals is due to the low availability of accelerometers and the low availability of both sensor types together. The accuracies start to increase as the participant pool grows; thus, the sound level data combined with the limited movement data can provide useful information only if the population is high. On the other hand, reliability of acquired data is low as one can tell from the large confidence intervals. When all three figures are considered, it can be concluded that the accelerometer provides the highest contribution to the accuracy whereas at least 30% availability of both sensors combined with high availability of accelerometers can guarantee high localization accuracies with reasonable confidence intervals.

We have shown with these extensive even and uneven participant recruitment scenarios that accelerometers in smart mobile devices make the highest contribution to the value of the aggregated sensory data whereas at least 30% of the participants are required to have both accelerometer and microphone available to be used opportunistically in order to achieve localization accuracies at the order of 90%.

4.5 Concluding Remarks

Smart transportation and vehicular networks are integral components of smart cities. In this article, we have proposed a vehicular crowdsensing-based effective solution to the problem of GPS-less discovery of the whereabouts of public transportation vehicles that move on static routes in a vehicular network setting. Despite its ubiquity, GPS is a power hungry and location dependent sensor. We have proposed to empower smart mobile devices of passengers in order to acquire non-dedicated sensory data in an opportunistic manner. In our proposed framework, upon being aggregated, crowd-solicited sensory data undergoes an unsupervised machine learning component to match aggregated data with one of the clusters, which corresponds to a landmark (i.e., stop / station) over a static route. We have evaluated the performance of our proposed framework through simulations that have been by the data plane of the proposed framework. The data plane is designed as a container for sensory data that was collected in real time by heterogeneous sensing environment consisting of smart mobile devices with various specifications. The control plane of the proposed architecture consists of data aggregation and analytics modules. While data aggregation module fuses vehicular network data, analytics module applies $k - means$ algorithm on the fused data to cluster them with regard to the origin of location. We have achieved an overall accuracy varying between 80% - 90% with all user requirement scenarios. Furthermore, through extensive simulations, we have pursued a feasibility study to find out the best availability

combination of built-in/non-dedicated sensors in smart mobile devices. We have shown that accelerometers in smart mobile devices make the highest contribution to the value of the aggregated sensory data whereas at least 30% of the participants are required to have both accelerometer and microphone available to be used opportunistically in order to achieve localization accuracies at the order of 90%.

In the next chapter, we introduce a new dependability modules with the proposed framework in order to assess the quality of data and evaluate the reliability of participants. Furthermore, we are also planning to address elimination of malicious participants in the presence of adversaries in the participant pool.

Chapter 5

Ensuring Trustworthiness of Crowd-Sensed Vehicle Tracking Service

5.1 Introduction

5.1.1 Problem Description

Processing and analysis of the data, as well as its presentation is a crucial factor. In the analysis of vehicular sensing data for Intelligent Transportation Systems (*ITS*)s, machine learning (*ML*) algorithms has attracted many researchers [71, 143] whereas deep learning (*DL*) methods are also gaining popularity under specific contexts [38, 73]. However, it is certain that untrustworthy/malicious data would lead to deterioration in data analytics performance. Distributed and heterogeneous nature of a crowd-sensing system introduces the trustworthiness concerns due to either or combination of the following reasons: (*i*) malfunctioning sensors, (*ii*) malicious/untrustworthy participants. Research on reliability and dependability of the acquired data aims to prevent the consequential misbehaviour of the platform, and proposes upfront actions against unreliable / untrustworthy sensed data [141, 13, 59].

5.1.2 Proposed Solution

While the proposed architectures in our previous studies [78, 77] yield location approximations above 80% on the average, in order to achieve higher accuracy and dependability levels, root causes of misleading estimations need to be further investigated. One of the major drawbacks of a vehicular crowd-sensing-based localization framework is the presumption that all participating sensors are fully trustworthy. This presumption unintentionally

leads to the contribution of untrustworthy data (in other words, unreliable sensor readings) to the platform, which drastically effects the performance of the data analytics module (namely the performance of analytics algorithm that is being used) in terms of estimated location accuracy. Therefore, to achieve remarkable improvement in the accuracy and dependability of vehicular crowd-sensing-based localization, participant trustworthiness and reliability need to be taken into account in the design of sensor recruitment policies. A participant’s trustworthiness will determine the amount of sensor data that will be accepted into the platform from that participant. Therefore, there is a direct relationship between a trustworthy/untrustworthy participant and reliable/unreliable sensor readings qualified to be placed into the trustworthy sensor reading pool – a process we will discover in detail in the Section 5.2 and 5.3.

5.2 Proposed System Overview

Vehicular crowd-sensing-based localization framework adopts the components that have been introduced in [77]. As illustrated in Fig. 5.1, our contribution to the original crowd-sensing-based localization is the introduction of trustworthiness assessment components. Proposed architecture encompasses three components: **I.** Data acquisition, **II.** Data aggregation, and **III.** Analytics and Prediction. Trustworthiness of the participants are assessed in the analytics module, which is highlighted by a filter. Three types of databases are maintained: **a.** SR-DB that stores the collected sensory data from the participants, **b.** PT-DB that stores trustworthiness scores of the participants, and **c.** VT-DB that stores trustworthiness scores of the vehicles. Multiple participants cooperate in the acquisition of sensing data that is offloaded to cloud-based server and storage units for aggregation, analysis and archival. Furthermore, Figure 5.2 illustrates the sequence of events that take place during *GPS*-less vehicle localization with trustworthy sensor data.

As mentioned in the motivation, it is not realistic to assume that all participants contribute to the crowd-sensed data reliably since any crowd-sensing campaign is vulnerable to sensor malfunctioning or malicious participants. To cope with this phenomenon, we present three recruitment policies where the first one serves as a benchmark to the second and third schemes, which –to the best of our knowledge– are proposed for the first time in this context. These recruitment policies are as follows: **(i)** Non-restricted recruitment (*NRR*), **(ii)** Reliability-driven naive recruitment (*RDNR*), and **(iii)** Reliability-driven exclusive recruitment (*RDER*). To provide an easy access to readers, we present the notation used in the paper as follows: Table 5.1 encompasses list of constants, Table 5.2 represents the parameters used in the recruitment policies, Table 5.3 accommodates the auxiliary functions used in the equations, and Table 5.4 describes the activity profiles of the participants. We present the details of the three sensor recruitment policies that operate in the presence of reliability / trustworthiness breaches in Section 5.3

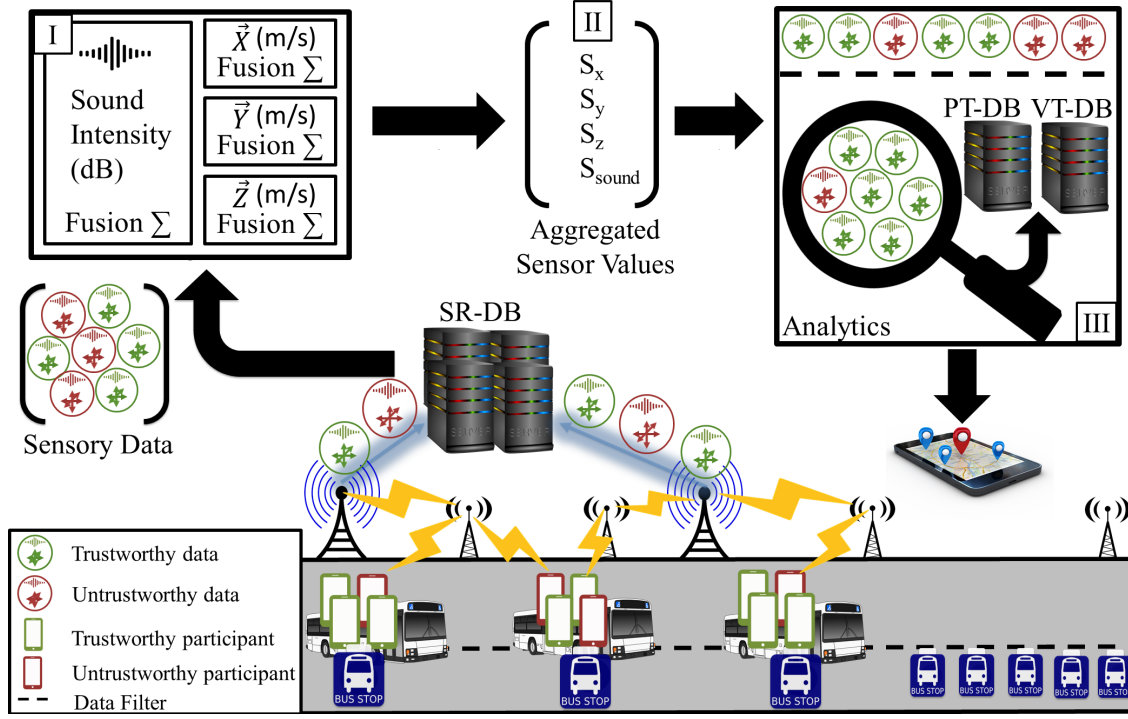


Figure 5.1: Trustworthiness-driven *GPS*-free localization architecture.

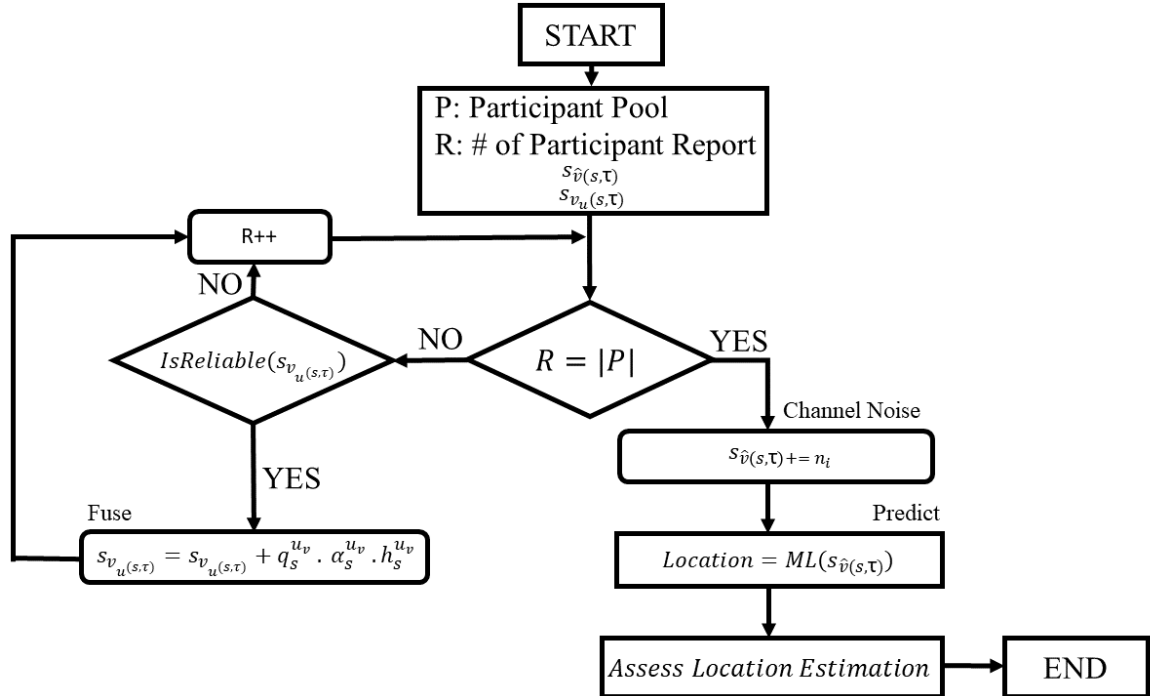


Figure 5.2: A generic flowchart for *RDNR* and *RDER*. *RDNR* and *RDER* have their own reliability assessment methods explain in detail in Section 5.3.2 and 5.3.3 respectively.

Table 5.1: List of constants that are used in the formulas and throughout the Chapter 5

Parameter	Value
V	# of vehicles
v	Index of a vehicle
K	# of clusters
k	Index of cluster
P	# of participants in a vehicle
p	Index of participant
θ	δ -dimensional sensor readings
D	# of days
d	Day d
T	Total duration of the simulation
τ	Time (discrete) of sensing in a day
ℓ	# of possible landmarks

5.3 Proposed Methodologies

5.3.1 Non-Restricted Recruitment

This recruitment strategy relies on the sensor readings of participants without applying any trustworthiness assessment for the participants. It is worth noting that *NRR* has been adopted as the benchmark solution from previous work in [78, 77]. Thus, in the localization process for public transportation vehicles, the entire sensory data is passed to the analytics module without any elimination. Sensor readings transmitted by malfunctioning sensors or sensory data that has been distorted should be expected to degrade accuracy of localization of vehicles moving on static paths. Under the hypothetic assumption that the entire data is trustworthy, the analytics module will only need to call an effective clustering algorithm (e.g., k-Means) to cluster sensory data into clusters that correspond to anchor coordinates, which in the case of public transportation vehicles, correspond to stops.

5.3.2 Reliability-driven Naive Recruitment

RDNR is the first trustworthiness-driven data acquisition regime for vehicular crowd-sensing-based localization. The *naive* property of RDNR is because of that the decision for the recruitment of a participating sensor is made upon a comparison between the reliability assessment (i.e., trust score, $S_{u_v^p}^{\theta, d, \tau}$) of the sensor and the value of a random variable. In

Table 5.2: The parameters that are used to formulate the user recruitment scenarios

Parameter	Value
u_v^p	Participant p in vehicle v
$S_{u_v^p}^{\theta,d,\tau}$	Reading of sensor θ reported by participant p in vehicle v on day d at time τ
$\mathfrak{T}_{u_v^p}^{\theta,d,\tau}$	Trustworthy sensory data of participant p in vehicle v on day d at time τ .
$\epsilon_{u_v^p}^{k,\theta,d}$	Label of sensory data (in cluster k) of participant p in vehicle v on day d
$\mathfrak{A}_{u_v^p}^{k,\theta,d}$	Label of sensory data in SR-DB (in cluster k) of participant p in vehicle v on day d
$\mathfrak{B}_{u_v^p}^{k,\theta,d}$	Binary trust score of sensory data (in cluster k) of participant p in vehicle v on day d
$\mathcal{E}_{u_v^p}^{k,d}$	Trust score of participant p in vehicle v on day d for cluster k
$\mathcal{V}_{u_v^p}^{k,d}$	Threshold trust score of participant p in vehicle v on day d for cluster k
$\mathcal{L}^{k,\theta,d}$	Labels of the trustworthy sensory data on day d
$\mathcal{J}_v^{k,d}$	“Rolling” trust score of vehicle v on day d on cluster k

Table 5.3: List of auxiliary functions that are used in the equations

Parameter	Value
$\gamma(a, b)$	Binary function returns 1 if $a = b$
$\mathfrak{R}(a, b)$	Random value generator between a and b (inclusively)
$\mathfrak{f}(a, b)$	Sums the values in array a until the index b .

Table 5.4: Description for the participant activity profiles

Parameter	Value
η -participants	Active + inactive participant profiles
α -participants	Active participant profiles

RDNR, the platform collects data from every participant to create an initial Sensor Reading Database (*SR-DB*). The database contains cluster information of the sensory data of the participants in each vehicle. *SR-DB* plays a crucial role when assessing the trustworthiness

of the sensory data of a participant.

While participants feed sensory data into remote storage system, data analytics module evaluates them and creates the Data Trustworthiness Matrix (DTM) which consists of only binary values. Each row of DTM corresponds to a participant p on vehicle v on day d at time τ and each column of DTM corresponds to a landmark/stop (i.e., cluster) for vehicle v . A tuple $[row, column]$ in DTM stores a binary trust score that corresponds to the sensor reading θ in cluster k for participant p on vehicle v on day d ($\mathfrak{B}_{u_v^p}^{k,\theta,d}$).

When participant p in vehicle v passes through the same landmark, data analytics module compares the data stored in $SR-DB$ ($\mathfrak{A}_{u_v^p}^{k,\theta,d}$) with the cluster of sensor reading that participant has just fed into the platform ($\epsilon_{u_v^p}^{k,\theta,d}$). If they match, binary value “1” denoting that the data is trustworthy is recorded into corresponding $[row, column]$ cell in DTM . Otherwise, the value of “0” in the cell will denote that the data is untrustworthy as it is formulated with the Eq. 5.1.

$$\mathfrak{B}_{u_v^p}^{k,\theta,d} = \gamma(\epsilon_{u_v^p}^{k,\theta,d}, \mathfrak{A}_{u_v^p}^{k,\theta,d}),$$

$$\forall d \in D, \forall k \in K, \forall v \in V, \forall p \in P. \quad (5.1)$$

Additionally, Participant Trustworthiness Matrix (PTM) is generated using DTM to evaluate the level of trustworthiness of each participant. Trustworthiness values (i.e., scores) of each index in PTM is calculated by using Eq. 5.2. As new clusters/landmarks (i.e., indices) are added to DTM , PTM updates itself accordingly. Value of an index in PTM ($\mathcal{E}_{u_v^p}^{k,d}$) is calculated by adding consecutively the values of indices (i.e., landmarks or clusters) DTM until a certain cluster and dividing the sum by the index of that cluster. As a numerical example to solidify this explanation, one may assume DTM values consisting of 5 landmarks as $[1 \ 0 \ 0 \ 0 \ 1]$. Then the corresponding PTM will be $[1 \ (1+0)/2 \ \dots \ (1+0+0+0+1)/5]$ or $[1 \ 0.5 \ \dots \ 0.2]$.

This method enables better understanding of the trustworthiness of individual participants. Whenever a participant transmits sensory data that is not consistent with the landmarks, their DTM entries for those landmarks remain 0. Sensory data labeled as untrustworthy will decrease the aggregated sum at the landmark, which eventually will reduce gradually trustworthiness of participant at the landmark (or cluster) $\mathcal{E}_{u_v^p}^{k,d}$.

$$\mathcal{E}_{u_v^p}^{k,d} = \frac{\mathfrak{f}(\mathfrak{B}_{u_v^p}^{k,\theta,d}, k)}{k},$$

$$\forall d \in D, \forall k \in K, \forall v \in V, \forall p \in P. \quad (5.2)$$

A random value generated between 0 and 1 is inclusively set as a threshold ($\mathcal{V}_{u_v^p}^{k,d}$) in $RDNR$. The threshold value is compared against the values in PTM ($\mathcal{E}_{u_v^p}^{k,d}$). If the threshold

is greater than trustworthiness score of the sensor reading ($\mathcal{E}_{u_v^p}^{k,d}$), then sensory data $S_{u_v^p}^{\theta,d,\tau}$ is deemed as untrustworthy by the data analytics module and will not be considered when determining the approximate location of the vehicle by using Eq. 5.3. Otherwise, the analytics module gathers the trustworthy sensor readings ($\mathcal{E}_{u_v^p}^{k,d}$) and feeds them as an input to the clustering module (e.g., k -means clustering), which produces labels for the trustworthy data ($\mathcal{L}^{k,\theta,d}$) to be used in the GPS-less public transport vehicle localization process.

$$\begin{aligned} \mathcal{V}_{u_v^p}^{k,d} &= \mathfrak{R}(0, 1), \\ \mathfrak{T}_{u_v^p}^{\theta,d,\tau} &= \begin{cases} \mathcal{E}_{u_v^p}^{k,d-1} \geq \mathcal{V}_{u_v^p}^{k,d} & S_{u_v^p}^{\theta,d,\tau} \\ \mathcal{E}_{u_v^p}^{k,d-1} < \mathcal{V}_{u_v^p}^{k,d} & null \end{cases}, \\ \forall d \in D, \forall \tau \in T, \forall k \in K, \forall v \in V, \forall p \in P. \end{aligned} \quad (5.3)$$

5.3.3 Reliability-driven Exclusive Recruitment

Another technique we have developed when assessing the trustworthiness of the participant is *RDER*, which principally adopts the idea of comparing participants trustworthiness scores in each cluster k on day d ($\mathcal{E}_{u_v^p}^{k,d}$) with the corresponding average trustworthiness score of the vehicle in cluster k on day d (denoted by $\mathcal{J}_v^{k,d}$). Vehicle Trustworthiness Matrix (*VTM*) contains the average trustworthiness score for each vehicle v in cluster k on day d $\mathcal{J}_v^{k,d}$, which determines a threshold for the vehicle when assessing the trustworthiness of a particular sensory data point.

While the rows of *VTM* denote vehicle v on route r , the columns of *VTM* represent the landmarks of route r . Thus, the tuple $[row, column]$ denotes the trustworthiness of a particular vehicle on a given landmark, which is calculated by Eq. 5.4. Averages of binary values in *DTM* that corresponds to the related vehicle for each landmark are computed. Each average value corresponds to trustworthiness score for vehicle v in cluster k on day d ($\mathcal{J}_v^{k,d}$) and represents a $[row, column]$ pair in *VTM*.

$$\begin{aligned} \mathcal{J}_v^{k,d} &= \frac{\sum_{d'=1}^d \sum_{p=1}^P \mathcal{E}_{u_v^p}^{k,d'}}{PD}, \\ \forall k \in K, \forall v \in V, \forall d \in D. \end{aligned} \quad (5.4)$$

Participant trustworthiness score ($\mathcal{E}_{u_v^p}^{k,d}$) in *PTM* is compared to the corresponding $\mathcal{J}_v^{k,d}$. If the score of participants exceeds the score of the vehicle, sensory data $S_{u_v^p}^{\theta,d,\tau}$ is considered to be trustworthy by the analytics module and included in the evaluation dataset that will be used for estimating the location of the public transportation vehicle. Otherwise, it is deemed

to be illegitimate by the analytics module for location estimation process. We adopt the $k - means$ method to cluster the dataset containing the trustworthy sensor readings ($\mathfrak{Z}_{u_v^p}^{\theta,d,\tau}$) and produce an output of labels ($\mathfrak{L}^{k,\theta,d}$) for the dataset that will be used when approximating the location of the vehicles.

$$\mathfrak{Z}_{u_v^p}^{\theta,d,\tau} = \begin{cases} S_{u_v^p}^{\theta,d,\tau} & \mathcal{E}_{u_v^p}^{k,d-1} \geq \mathcal{J}_v^{k,d-1} \\ null & \mathcal{E}_{u_v^p}^{k,d-1} < \mathcal{J}_v^{k,d-1} \end{cases},$$

$$\forall d \in D, \forall \tau \in T, \forall k \in K, \forall v \in V, \forall p \in P. \quad (5.5)$$

5.4 Performance Evaluation

In this section, we present series of experiment results that evaluate and compare the performance of *NRR*, *RDNR*, and *RDER* in terms of location estimation accuracy. Moreover, we also present the average recruitment probability of *RDNR*, and *RDER*. Finally, we show the energy efficiency of *RDER* approach in terms of η -participants and α -participants profiles.

5.4.1 Simulation Settings

We simulate a vehicular crowd-sensing environment where participants in public transportation buses transmit their sensory data to the cloud storage unit via cellular connection. Crowdsensim [49] is used to simulate crowd-sensing events for localization. Sensory readings in the simulation are generated from real data. In order to reflect a similar behavior, real dataset has been collected under a situation closely alike to the one set in the simulation environment (available at [111] and [108]). Real data consists of two sensors: (i) accelerometer, and (ii) microphone collected by Googles Science Journal Application [56]. Therefore, generated sensor readings are also made up of those sensors. Sensor readings experience value degradation to some extent. Hence, values on the storage unit are noisy readings, which are generated by Additive White Gaussian Noise (*AWGN*) in the simulation.

Table 5.5 lists the details of our simulation settings. There exists 50 vehicles that carry certain number of participants varying between 3 and 18, which we have kept the same within one experiment but increased gradually to check the effect of increasing number of participant on accuracy. Every vehicle belongs to a certain route, each of which contains three vehicles.

Vehicles move on a static route length for each experiment, which has a value of either 10 or 15. Size of a cluster for $k - means$ algorithm is directly related with the value of static route length. Thus, if the static route length is set to 10 for one experiment, cluster size will also be set to 10. The reason behind this is to group the sensor readings from a landmark into the corresponding cluster (i.e., landmark). For instance, if the route length is set to a bigger

Table 5.5: Simulation settings

Parameter	Value
Number of vehicles	50
Average velocity [kmph]	{30, 40}
Average stop time [s]	[30, 60]
Minimum inter-station distance	0.5 km
Simulation Duration	7 days
Propagation environment	Rayleigh fading
Noise model in sensor readings	AWGN
Clustering Method	K-means
Number of participants	{3, 6, 9, 12, 15, 18}
Static route length	{10, 15}
Cluster size	{10, 15}
Sensing noise (q_s^{uv})	{1.0, 2.0, 3.0}

number than cluster size, there will be sensor readings grouped into wrong clusters due to insufficient amount of clusters. We have analyzed the effect of varying number of participants and route length on accuracy. Moreover, we have examined the effect of different sensing noise associated with each vehicle on accuracy.

5.4.2 Numerical Results

Advantage of Reliability-aware Solution over Reliability-unaware Solution

Before comparing the proposed three sensor recruitment policies, a preliminary study to illustrate the advantage of a very simple reliability-aware sensor recruitment policy is compared in Figure 5.3 with the reliability unaware sensor recruitment policy. We have determined two parameters when evaluating the performance of *GPS*-less localization accuracy: (i) route length and (ii) sensing noise factor (q_θ^{uv}). In each experiment, number of participants in each vehicle is increased gradually from 3 to 18 with a step size of 3 participants in every iteration.

The effect of route length on accuracy can be observed in Fig. 5.3a. The route length, l , is first set to 10, then to 15, and finally to 20 in every experiment with a static number of participants. We observe that the localization accuracies for both recruitment (reliability-aware, and reliability-unaware) policies are higher as the route length gets shorter. On the other hand, we observe slight increase in the accuracy as the population of participants in each vehicle gets higher.

The change of sensing noise on accuracy is illustrated in Fig 5.3b. We compare the

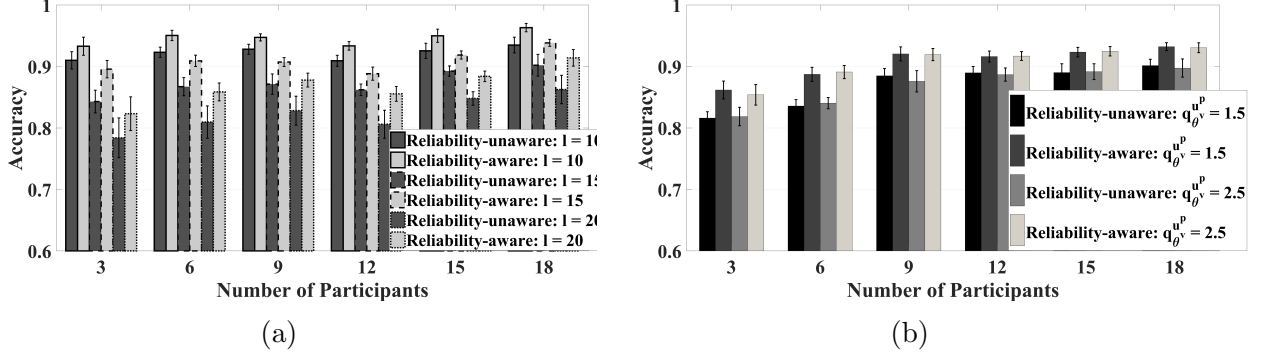


Figure 5.3: Accuracy performance of reliability-aware and reliability-unaware sensor recruitment policies under two varying parameters: (a) The effect of route length on accuracy (b) The effect of noise factor, $q_{\theta}^{u_p}$, on accuracy.

accuracies of two recruitment policies with two different sensing noise: $q_{\theta}^{u_p} = 1.5$ and $q_{\theta}^{u_p} = 2.5$ under a fixed number of participants in each experiment. While sensing noise change has no remarkable impact on the accuracy of the two recruitment policies – except the case with reliability unaware where there is a subtle increase and decrease from one iteration to another –, it has an impact on the confidence intervals of the accuracies. As the sensing noise factor gets higher, confidence interval gets broader because high sensing noise changes the actual sensor reading more, which eventually causes fluctuations in the accuracy.

Benchmark

In section 5.1.2, we explained the details of three approaches we have integrated into data analytics module. We have chosen *NRR* as the benchmark solution among the others since we have employed this method in our previous works and we aim to achieve remarkable performance improvements with the other two techniques: *RDNR* and *RDER*. Furthermore, we also highly anticipate to see that *RDER* would outperform the other two in terms of accuracy. Since *RDNR* evaluates trustworthiness of participants sensor readings in a totally random manner, we also expect to see fluctuations on the accuracies of consecutive experiments. On the other hand, *RDER* is expected to provide steady accuracy results between consecutive experiments as it is based on average trustworthiness analysis of participants and vehicles.

Performance analysis of the proposed framework has been conducted by running set of simulation experiments which settings are described in Section 5.4.1. Furthermore, we classify the experiments into two sub-categories based on their objectives: (i) *GPS*-less localization accuracy with trustworthiness in mind and (ii) *GPS*-less energy consumption. Accuracy is the ratio of the correct predictions to total predictions that were made during the run time of the simulation (T), which is formulated in Eq. 5.6. A prediction in *k-means* method is considered to be correct when a trustworthy pair of sensor reading (accelerometer

and microphone) from a participant at a landmark is placed in the correct cluster where other sensor readings transmitted from the same vehicle and landmark are placed.

$$\text{Accuracy} = \frac{\sum_{k=1}^K \sum_{d=1}^{D-1} \gamma(\mathfrak{L}^{k,\theta,d}, \mathfrak{L}^{k,\theta,d+1})}{K(D-1)}. \quad (5.6)$$

GPS-less localization accuracy with trustworthiness in mind

We conduct a series of experiments by varying the number of participants, route length/cluster size, and sensing noise for each recruitment policy (*NRR*, *RDNR*, and *RDER*).

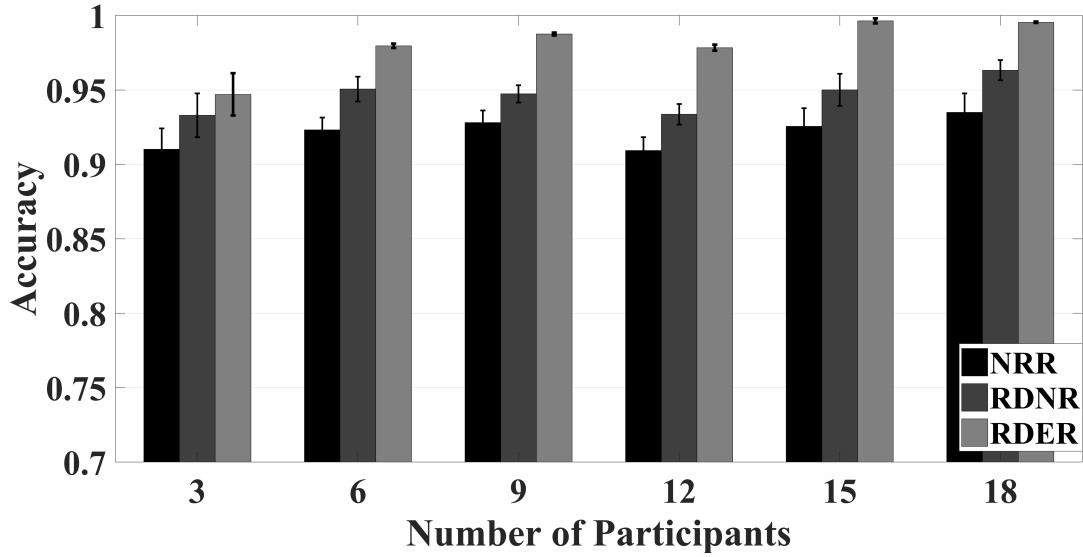


Figure 5.4: The effect of $l=10$ on accuracy with increasing number of participants.

Fig. 5.4 presents accuracy results where route length (therefore cluster size) l is set to 10. While route length is stabilized by a static number of landmarks, volume of the participants in vehicles keep increasing gradually starting from 3 up until 18 with a step size of 3 in every iteration. This increase in the participant count translates into an affirmative effect on the accuracy. Moreover, *RDER* has the highest performing recruitment schema among the other two having an accuracy at the worst case (number of participants = 3) around 95% and 99% at the best case (number of participants = 18). Furthermore, *NRR* and *RDNR* not only outperformed by *RDER* but they also exhibit higher data inconsistencies as it can be observed from their large confidence intervals.

For the next series of experiments (varying sensing noise values (q_s^{uv})) route length and cluster size values are assigned to 15. q_s^{uv} values that we have selected are: 1.0, 2.0, and 3.0,

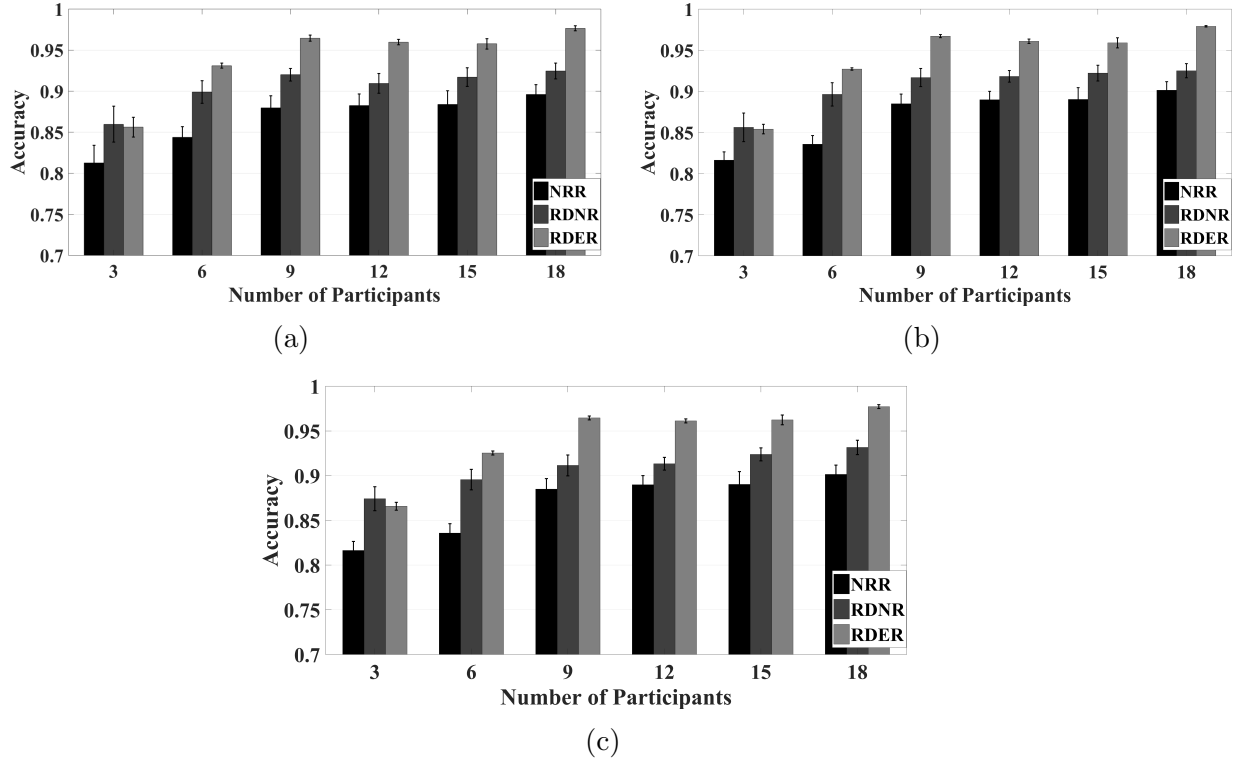


Figure 5.5: The effect of sensing noise (q_s^{uv}) on accuracy (a) $q_s^{uv} = 1.0$ (b) $q_s^{uv} = 2.0$ (c) $q_s^{uv} = 3.0$.

where the first value has the lowest impact on the sensory data degradation stored on the platform, and the last value has the highest impact. We can observe from the Figure 5.5 that they have similar (but not the same) accuracy values with increasing number of participants. However, confidence interval of *RDER* is getting larger with the increasing q_s^{uv} , which means noisy sensory data is increasing the chance of observing sensor reading variances in the evaluation dataset. *NRR* and *RDNR* recruitment schemes exhibit larger confidence intervals in all sensing noise experiments due to containing larger amount of untrustworthy data in their evaluation dataset than *RDER*. We can, furthermore, see *RDNR* outperform *RDER* when the number of participants is 3.

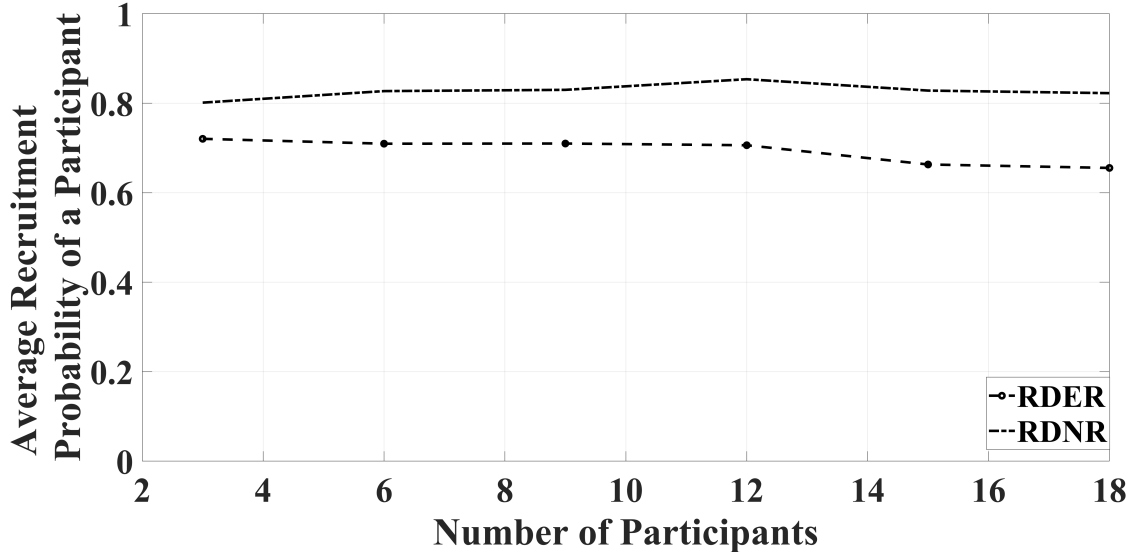


Figure 5.6: Comparison of recruitment probabilities of *RDNR* and *RDER* participant recruitment scenarios.

Fig. 5.6 illustrates the average recruitment probability of a participant, which denotes to the ratio of the number of trustworthy sensor data to total sensor data that has been transmitted to the platform. We compare the ratios of *RDER* and *RDNR* recruitment scenarios. It is obvious in the figure that *RDER* exhibits a lower recruitment ratio than *RDNR*, which means *RDER* is eliminating remarkably more untrustworthy data than the *RDER*. Therefore, *RDER* outperforms the other two scenarios (*NRR* and *RDNR*).

GPS-less energy consumption

Our energy experiments involve two different participant profiles: (i) α -participants profile and (ii) η -participants profile. The former denotes only participants who use their smart devices frequently (for web browsing and social media, etc.) while participating in the sensing task whereas the latter consists of participants from both profiles (meaning some participants

who use their devices actively (i.e., active participants) along with the rest who do not use their phone but contribute to the sensing phenomena (i.e., inactive participants)).

That being said, Fig. 5.7 consist of two sub-plots the first of which illustrates the effect of increasing number of participants that belong to α -participant profile and the second of which plots the effect of increasing ratio of active participants to inactive participants. In the first plot, number of participants is gradually increased from 3 to 18. It is obvious that the energy consumption and location accuracies are increasing as the number of participants increases. On the other hand, in the second plot, activity ratio is gradually increased meaning that number of α -participant profile has been increased and number of inactive participants has been decreased at the same ratio, which eventually increases energy consumption.

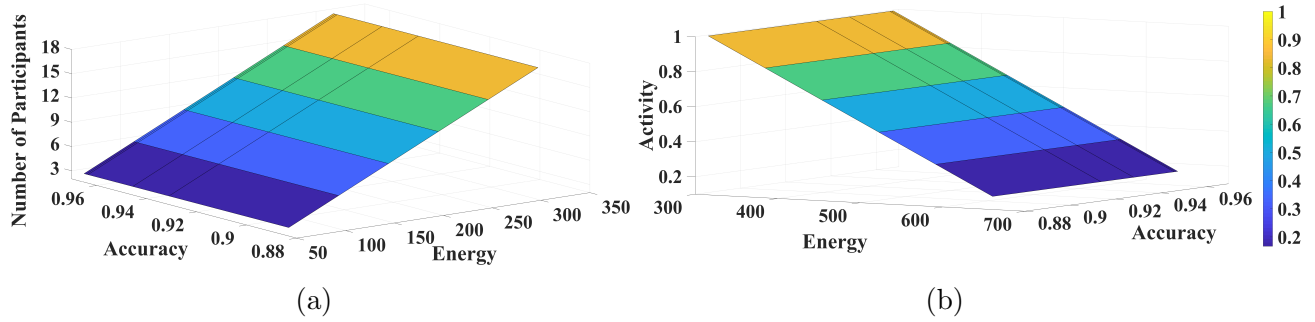


Figure 5.7: Energy performance per vehicle under *RDER* participant recruitment scenario with different participant profiles (a) α -participants profile (b) η -participants profile.

5.5 Concluding Remarks

Public transportation methods are integral part of the cities and used by many every day. Smart cities require a solid foundation of smart transportation. In our previous studies, we have already developed a *GPS*-less localization technique by employing both supervised [78] and unsupervised [77] machine learning methods on crowd-sensed sensory data. As participants that feed their sensor readings into the proposed platform could be untrustworthy due to various reasons such as malfunctioning of built-in device sensors of participants, in this chapter, we have initially analyzed the advantage of reliability-aware technique over reliability-unaware technique and then proposed different participant recruitment scenarios to eliminate the untrustworthy sensor readings and analyze their performance in terms of localization accuracy. These recruitment scenarios are Non-Restricted Recruitment (*NRR*) as the benchmark, Reliability-Driven Naive Recruitment (*RDNR*), and Reliability-Driven Exclusive Recruitment (*RDER*). Moreover, we have selected *NRR* as our benchmark solution to compare it with the *RDNR* and *RDER* as it totally allows untrustworthy data from participants. We have shown that *RDER* introduces the highest accuracy among the other strategies with a localization accuracy up to 98% in the best case ($l = 10$ and number of par-

ticipants = 18) and approximately 85% in the worst case ($l = 15$ and number of participants = 3).

Chapter 6

Conclusions, Open Issues, and Future Directions

As the urban population grows rapidly, the challenge to both supply basic resources, and ensure overall economic, social, and environmental sustainability becomes harder. Access to sufficient fresh water, and cleaner energy, the ability to travel efficiently from one point to another, providing safety, and security are some of the services that a modern, smart city should offer to its citizens. Smart transportation and vehicular sensing are integral elements of smart city services. Recent advents on wireless communication technologies (e.g., improvements on *LTE*, on-going development of *5G*, and novel infrastructure on vehicular communication such as *VANETs*) enables smart transportation services to be built on robust network, and therefore minimizes the delay and package lost rates due to dynamic nature of transportation environment. Improvements on cloud and fog computing, accelerates the speed data is stored and processed. Furthermore, data analytic methodologies such machine learning, deep learning, and signal processing pave the way for building intelligent system that is capable of predicting and proposing solutions for the possible future events.

We have categorized the smart transportation services into 4 distinct groups as follows: (i) traffic management, (ii) public transportation management, (iii) emergency management, and (iv) information management. A briefly summary for the functionality of each service, traffic management service: provides applications that aim mainly to minimize traffic congestion; public transportation management service offers applications that generally focuses on improving the efficiency of public transportation vehicles; emergency management service encompasses applications that target to provide safer roads by alleviating the traffic accident rates; and information management service proposes applications that aim to provide efficient delivery and secure storage of the transportation data (by taking privacy measures). In this thesis, we propose *GPS*-less public transport tracking framework via vehicular crowd-sensing to eliminate the possible problems introduced by *GPS* (low location accuracy performance in certain environments such as indoor, urban areas; high energy consumption; and privacy violation of individuals due to direct location sharing). Furthermore, we extend and improve this idea by introducing various methodologies such

as different machine intelligence techniques, crowd recruitment scenarios, and assessment of the trustworthiness of the crowd-sensed data.

In [78], we introduced supervised machine learning approach – specifically *SVM* – to approximate the location of public transport. The proposed architecture consists of four modules: 1) data acquisition, 2) crowd-sensed non-dedicated sensor data aggregation, 3) probabilistic prediction, and 4) data analytics. We have used following non-dedicated sensors: accelerometer and microphone to acquire data from participants in public transport vehicles. To validate the performance of proposed solution, we have set-up a simulation environment with 50 bus lines, each of which has 15 bus stops. Furthermore, each bus contains a static number of participants that contributes sensor data to the platform. Data acquired through a real application (which has only one vehicle) have been fed into the simulation environment (which has 3 buses per lines except the last line having 2 bus). Therefore, extensive analysis with a large scale application made possible with the simulation environment. *SVM* has been applied to the acquired sensor readings. To conduct the performance analysis on the proposed framework, we changed the probabilistically derived location region, route length of buses, and the noise factor induced by the network channel. Through our performance analysis, we have determined the best and worst performing combinations of those metrics. We have obtained an overall accuracy rate above 90% and 80% in the best and worst cases respectively. Moreover, energy consumptions under best and worst case performing simulation settings have been investigated with two different participant profiles: active and inactive. As an extension to this work, we have also proposed varying sensor combinations (such as identical sensor array pool, non-identical sensor array pool - 1, etc.) scenarios and computed both location accuracies and energy consumptions under the best and worst simulation settings. While proposed identical sensor array pool yields the highest location accuracy and the lowest energy consumption, non-identical sensor array pool - 1 offers the lowest location accuracy with the highest energy consumption among the rest.

We have proposed an unsupervised machine learning method – *k-means* – in [77], which essentially has the same architectural design in [78]: a data plane to acquire the crowd-sensed data, and a control plane to aggregate and analyze the crowd-sensed data. *K-means* in the analytics module clusters the crowd-sensed readings and matches a sensor reading transmitted from a vehicle with one of the clusters while data aggregation module fuses vehicular network data. Performance of the proposed unsupervised approach has been tested through simulations, which use sensor readings acquired in real time by heterogeneous sensing environment encompassing smart mobile devices having diverse specifications. Different participant recruitment scenarios with varying ratio of sensors (e.g., α -rec and η -rec) and cluster sizes have been adopted to find out the best performing participant recruitment scenario and cluster size. Overall location accuracy of the proposed architecture is between approximately 80% - 90% with all participant recruitment scenarios. Moreover, we have proven that at least 30% of participants' mobile devices should have both accelerometer and microphone available to be used to obtain localization accuracies approximately around, or more than 90%.

Since crowd-sensing recruits participants who contribute their crowd-sensed data into the

proposed platform, there is always a risk of accommodating untrustworthy sensor readings, which eventually degrades the performance of locating the public transport vehicles with higher accuracies. Untrustworthy sensor readings could be introduced into platform due to various reasons such as malfunctioning of built-in device sensors of participants, and malicious participants (who feed untrustworthy data into the platform deliberately). In [75], we proposed a *GPS*-less vehicle tracking system via vehicular crowd-sensing that employs trustworthiness-aware recruitment technique together with the trustworthiness-unaware recruitment (architectures proposed in [78, 77]) technique. While the highest localization accuracy is obtained with trustworthiness-aware scheme is approximately 90% when l (route length) = 20 and number of participants = 18, the highest achieved with the same simulation settings is 93%. Furthermore, we have extended this work with an additional trustworthiness-aware recruitment scheme and compared the performance of three participant recruitment schemes (Non-Restricted Recruitment *NRR*, Reliability-Driven Nave Recruitment *RDNR*, and Reliability-Driven Exclusive Recruitment *RDER*). While *RDER* yields the best localization accuracy among the other participant recruitment schemes up to 98% in the best case scenario with l (route length) = 10 and number of participants = 18, it achieves approximately 85% of localization accuracy in the worst case with $l = 15$ and number of participants = 3.

We have presented three main *GPS*-less public transport vehicle tracking architectures in thesis and summarized their architectural designs and performance in terms of location accuracy and energy consumptions. In our simulation environment, we have assumed that traffic is flowing without any traffic congestion, or any situation that forces bus to stop except bus stops (such as traffic lights). Therefore, emulating real-life traffic environment in order to test our proposed architectures and to analyze the effect of various traffic flow behaviors on localization accuracy is still an open issue to solve. Another open issue is to simulating the effects of signal quality degradation of communication channel on localization accuracy with the frequently changing environment in an urban location. For instance, signal quality of mobile phones might alleviate when buses pass through downtown where there exists signal blocking obstacles (e.g., high-rise buildings). Therefore, varying transmission channel quality throughout to simulation to understand its effects on localization accuracy is still an open issue to solve. Moreover, trustworthiness assessment mechanism falsely accepts the crowd-sensed provided from vehicles where all participants contribute unreliable data to the platform. Our experiments are based on the fact that there exists at least one trustworthy participant in each bus.

We are currently focusing on extending this work to eliminate the open issues described above. Adopting trustworthiness assessment module in our supervised architecture and analyze its behavior in this context is also in our agenda. Moreover, our simulation environment (*CrowdSensim* [49]) currently supports three built-in maps among which we are currently using Luxembourg. With the new release of *CrowdSensim*, custom maps can be added with the coordinates extracted from Open Street Map [63]. Therefore, another future goal is to integrate Ottawa’s public bus transport map into our simulation environment.

References

- [1] A dynamically reconfigurable architecture for emergency and disaster management in ITS, author=Niar, Smail and Yurdakul, Arda and Unsal, Osman and Tugcu, Tuna and Yuceturk, Aziz, booktitle=Proc. International Conference on Connected Vehicles and Expo (ICCVE), pages=479–484, year=2014, organization=IEEE.
- [2] United States Department of Transportation. <https://www.transportation.gov/>. Accessed: 2018-06-16.
- [3] Martin Abadi, Paul Barham, Jianmin Chen, Zhifeng Chen, Andy Davis, Jeffrey Dean, Matthieu Devin, Sanjay Ghemawat, Geoffrey Irving, Michael Isard, et al. TensorFlow: a system for large-scale machine learning. In *Proc. OSDI*, volume 16, pages 265–283, 2016.
- [4] Khadige Abboud, Hassan Aboubakr Omar, and Weihua Zhuang. Interworking of DSRC and cellular network technologies for V2X communications: a survey. *IEEE Transactions on Vehicular Technology*, 65(12):9457–9470, 2016.
- [5] K. Abrougui, A. Boukerche, and R. W. N. Pazzi. Design and evaluation of context-aware and location-based service discovery protocols for vehicular networks. *IEEE Transactions on Intelligent Transportation Systems*, 12(3):717–735, 2011.
- [6] O. Abumansoor and A. Boukerche. A secure cooperative approach for nonlinear-of-sight location verification in VANET. *IEEE Transactions on Vehicular Technology*, 61(1):275–285, 2012.
- [7] Najah Abu Ali, Mervat AbuElkhair, and Salah Bouktif. Utilizing VIN for improved vehicular sensing. In *Wireless Communications and Networking Conference (WCNC)*, pages 1–6. IEEE, 2016.
- [8] Moayad Aloqaily, Burak Kantarci, and Hussein T Mouftah. An auction-driven multi-objective provisioning framework in a vehicular cloud. In *Proc. IEEE Global Communications Conference (GLOBECOM)*, pages 1–6, 2015.
- [9] Moayad Aloqaily, Burak Kantarci, and Hussein T Mouftah. Multiagent/multiobjective interaction game system for service provisioning in vehicular cloud. *IEEE Access*, 4:3153–3168, 2016.

- [10] Sheng-hai An, Byung-Hyug Lee, and Dong-Ryeol Shin. A survey of intelligent transportation systems. In *Proc. International Conference on Computational Intelligence, Communication Systems and Networks*, pages 332–337. IEEE, 2011.
- [11] Fazel Anjomshoa, Moayad Aloqaily, Burak Kantarci, Melike Erol-Kantarci, and Stephanie Schuckers. Social behaviometrics for personalized devices in the internet of things era. *IEEE Access*, 5:12199–12213, 2017.
- [12] Fazel Anjomshoa and Burak Kantarci. SOBER-MCS: sociability-oriented and battery efficient recruitment for mobile crowdsensing. *MDPI Sensors*, 18(5):1593, 2018.
- [13] Muhammad Arshad, Zahid Ullah, Naveed Ahmad, Muhammad Khalid, Haithiam Criuckshank, and Yue Cao. A survey of local/cooperative-based malicious information detection techniques in VANETs. *Springer EURASIP Journal on Wireless Communications and Networking*, 2018(1):62, 2018.
- [14] David Arthur and Sergei Vassilvitskii. K-means++: the advantages of careful seeding. In *Proc. ACM-SIAM Symposium on Discrete Algorithms*, pages 1027–1035. Society for Industrial and Applied Mathematics, 2007.
- [15] Muhammad Tayyab Asif, Justin Dauwels, Chong Yang Goh, Ali Oran, Esmail Fathi, Muye Xu, Menoth Mohan Dhanya, Nikola Mitrovic, and Patrick Jaillet. Spatiotemporal patterns in large-scale traffic speed prediction. *IEEE Transactions on Intelligent Transportation Systems*, 15(2):794–804, 2014.
- [16] Miguel Baguena, Carlos T Calafate, Juan-Carlos Cano, and Pietro Manzoni. An adaptive anycasting solution for crowd sensing in vehicular environments. *IEEE Transactions on Industrial Electronics*, 62(12):7911–7919, 2015.
- [17] Fan Bai and Donald K Grimm. Method and apparatus for augmenting smartphone-centric in-car infotainment system using vehicle Wi-Fi/DSRC, April 28 2015. US Patent 9,021,049, Google Patents.
- [18] Mahesh K Banavar, Cihan Tepedelenlioglu, and Andreas Spanias. Estimation over fading channels with limited feedback using distributed sensing. *IEEE Transactions on Signal Processing*, 58(1):414–425, 2010.
- [19] Frédéric Bastien, Pascal Lamblin, Razvan Pascanu, James Bergstra, Ian Goodfellow, Arnaud Bergeron, Nicolas Bouchard, David Warde-Farley, and Yoshua Bengio. Theano: new features and speed improvements. *ArXiv Preprint ArXiv:1211.5590*, 2012.
- [20] Sultan Basudan, Xiaodong Lin, and Karthik Sankaranarayanan. A privacy-preserving vehicular crowdsensing-based road surface condition monitoring system using fog computing. *IEEE Internet of Things*, 4(3):772–782, 2017.

- [21] Ashish Bhaskar, Edward Chung, et al. Passenger segmentation using smart card data. *IEEE Transactions on Intelligent Transportation Systems*, 16(3):1537–1548, 2015.
- [22] Prem Kumar Bhaskar and Suet-Peng Yong. Image processing based vehicle detection and tracking method. In *Proc. International Conference on Computer and Information Sciences (ICCOINS)*, pages 1–5. IEEE, 2014.
- [23] Kaigui Bian, Gaoxiang Zhang, and Lingyang Song. Toward secure crowdsensing in vehicle-to-everything networks. *IEEE Network*, 32(2):126–131, 2018.
- [24] R Blasco, H Do, S Shalmashi, S Sorrentino, and Y Zang. 3GPP LTE enhancements for V2V and comparison to IEEE 802.11 p. In *ITS European Congress*, pages 6–9. ERTICO, 2016.
- [25] Tadilo Endeshaw Bogale and Long Bao Le. Massive MIMO and mmWave for 5G wireless HetNet: potential benefits and challenges. *IEEE Vehicular Technology Magazine*, 11(1):64–75, 2016.
- [26] A. Boukerche, H. A. B. F. Oliveira, E. F. Nakamura, and A. A. F. Loureiro. Localization systems for wireless sensor networks. *IEEE Wireless Communications*, 14(6):6–12, 2007.
- [27] A. Boukerche, H. A. B. F. Oliveira, E. F. Nakamura, and A. A. F. Loureiro. DV-Loc: a scalable localization protocol using voronoi diagrams for wireless sensor networks. *IEEE Wireless Communications*, 16(2):50–55, 2009.
- [28] A. Boukerche, C. Rezende, and R. W. Pazzi. Improving neighbor localization in vehicular ad hoc networks to avoid overhead from periodic messages. In *Proc. IEEE Global Telecommunications Conference (GLOBECOM)*, pages 1–6, 2009.
- [29] Azzedine Boukerche, Horacio ABF Oliveira, Eduardo F Nakamura, and Antonio AF Loureiro. Vehicular ad hoc networks: a new challenge for localization-based systems. *Elsevier Computer communications*, 31(12):2838–2849, 2008.
- [30] Gary R Bradski. Real time face and object tracking as a component of a perceptual user interface. In *Proc. Workshop on Applications of Computer Vision (WACV)*, pages 214–219. IEEE, 1998.
- [31] Jeffrey A Burke, Deborah Estrin, Mark Hansen, Andrew Parker, Nithya Ramanathan, Sasank Reddy, and Mani B Srivastava. Participatory sensing. *UCLA Sensor for Embedded Network Sensing*, 2006.
- [32] Andrew T Campbell, Shane B Eisenman, Nicholas D Lane, Emiliano Miluzzo, and Ronald A Peterson. People-centric urban sensing. In *Proc. International Workshop on Wireless Internet*, page 18. ACM, 2006.

- [33] Giuseppe Cardone, Andrea Cirri, Antonio Corradi, Luca Foschini, Raffaele Ianniello, and Rebecca Montanari. Crowdsensing in urban areas for city-scale mass gathering management: geofencing and activity recognition. *IEEE Sensors*, 14(12):4185–4195, 2014.
- [34] Irina Ceapa, Chris Smith, and Licia Capra. Avoiding the crowds: understanding tube station congestion patterns from trip data. In *Proc. International Workshop on Urban Computing*, pages 134–141. ACM, 2012.
- [35] Lei CHEN and Bei-ji ZOU. New algorithm for detecting moving object based on adaptive background subtraction and symmetrical differencing [J]. *CNKI Application Research of Computers*, 2:053, 2008.
- [36] Tianqi Chen and Carlos Guestrin. Xgboost: a scalable tree boosting system. In *Proc. International Conference on Knowledge Discovery and Data Mining*, pages 785–794. ACM, 2016.
- [37] Xiang Cheng, Chen Chen, Wuxiong Zhang, and Yang Yang. 5G-enabled cooperative intelligent vehicular (5GenCIV) framework: when benz meets marconi. *IEEE Intelligent Systems*, 32(3):53–59, 2017.
- [38] Zheyuan Cheng, Mo-Yuen Chow, Daebong Jung, and Jinyong Jeon. A big data based deep learning approach for vehicle speed prediction. In *Proc. International Symposium on Industrial Electronics*, pages 389–394. IEEE, 2017.
- [39] François Chollet et al. Keras. <https://github.com/keras-team/keras>. Accessed: 2018-06-18.
- [40] IEEE Computer Society LAN MAN Standards Committee et al. Wireless lan medium access control (MAC) and physical layer (PHY) specifications. *IEEE Standard 802.11-1997*, 1997.
- [41] Alejandro Correa, Guillem Boquet, Antoni Morell, and Jose Lopez Vicario. Autonomous car parking system through a cooperative vehicular positioning network. *MDPI Sensors*, 17(4):848, 2017.
- [42] Sina Dabiri and Kevin Heaslip. Inferring transportation modes from gps trajectories using a convolutional neural network. *Elsevier Transportation Research Part C: Emerging Technologies*, 86:360–371, 2018.
- [43] Loukas Dimitriou, Theodore Tsekeris, and Antony Stathopoulos. Adaptive hybrid fuzzy rule-based system approach for modeling and predicting urban traffic flow. *Elsevier Transportation Research Part C: Emerging Technologies*, 16(5):554–573, 2008.
- [44] Mahmoud Hashem Eiza, Qiang Ni, and Qi Shi. Secure and privacy-aware cloud-assisted video reporting service in 5G-enabled vehicular networks. *IEEE Transactions on Vehicular Technology*, 65(10):7868–7881, 2016.

- [45] Elias Eze, Sijing Zhang, Enjie Liu, Shehu Muhammad, and Joy Eze. Achieving reliable communication in vehicular ad-hoc networks (VANETs): a survey. In *Proc. International Conference on Automation and Computing (ICAC)*, pages 1–6. IEEE, 2017.
- [46] Elias C Eze, Sijing Zhang, and Enjie Liu. Vehicular ad hoc networks (VANETs): current state, challenges, potentials and way forward. In *Proc. International Conference on Automation and Computing (ICAC)*, pages 176–181. IEEE, 2014.
- [47] Shih-Hau Fang, Hao-Hsiang Liao, Yu-Xiang Fei, Kai-Hsiang Chen, Jen-Wei Huang, Yu-Ding Lu, and Yu Tsao. Transportation modes classification using sensors on smart-phones. *MDPI Sensors*, 16(8):1324, 2016.
- [48] Karoly Farkas and Imre Lendak. Simulation environment for investigating crowd-sensing based urban parking. In *Proc. International Conference on Models and Technologies for Intelligent Transportation Systems (MT-ITS)*, pages 320–327. IEEE, 2015.
- [49] C. Fiandrino, A. Capponi, G. Cacciatore, D. Kliazovich, U. Sorger, P. Bouvry, B. Kantarci, F. Granelli, and S. Giordano. CrowdSenSim: a simulation platform for mobile crowdsensing in realistic urban environments. *IEEE Access*, 5:3490–3503, 2017.
- [50] Ren Gang and Zhou Zhuping. Traffic safety forecasting method by particle swarm optimization and support vector machine. *Elsevier Expert Systems with Applications*, 38(8):10420–10424, 2011.
- [51] Raghu K Ganti, Fan Ye, and Hui Lei. Mobile crowdsensing: current state and future challenges. *IEEE Communications Magazine*, 49(11), 2011.
- [52] Xiaohu Ge, Hui Cheng, Guoqiang Mao, Yang Yang, and Song Tu. Vehicular communications for 5G cooperative small-cell networks. *IEEE Transactions on Vehicular Technology*, 65(10):7882–7894, 2016.
- [53] Avik Ghose, Provat Biswas, Chirabrata Bhaumik, Monika Sharma, Arpan Pal, and Abhinav Jha. Road condition monitoring and alert application: Using in-vehicle smart-phone as internet-connected sensor. In *Proc. International Conference on Pervasive Computing and Communications Workshops (PERCOM Workshops)*, pages 489–491. IEEE, 2012.
- [54] Arindam Ghosh, Vishnu Vardhan Paranthaman, Glenford Mapp, Orhan Gemikonakli, and Jonathan Loo. Enabling seamless V2I communications: toward developing cooperative automotive applications in VANET systems. *IEEE Communications Magazine*, 53(12):80–86, 2015.
- [55] Stylianos Gisdakis, Thanassis Giannetsos, and Panos Papadimitratos. SHIELD: a data verification framework for participatory sensing systems. In *Proc. Conference on Security & Privacy in Wireless and Mobile Networks*, page 16. ACM, 2015.

- [56] Google. Science Journal. <https://makingscience.withgoogle.com/science-journal>. Accessed: 2018-06-18.
- [57] Gabriel Goulet-Langlois, Haris N Koutsopoulos, and Jinhua Zhao. Inferring patterns in the multi-week activity sequences of public transport users. *Elsevier Transportation Research Part C: Emerging Technologies*, 64:1–16, 2016.
- [58] IEEE 802.11 Working Group et al. IEEE standard for information technology–telecommunications and information exchange between systems–local and metropolitan area networks–specific requirements–part 11: wireless LAN medium access control (MAC) and physical layer (PHY) specifications amendment 6: wireless access in vehicular environments. *IEEE Std*, 802(11), 2010.
- [59] Jyoti Grover, Nitesh Kumar Prajapati, Vijay Laxmi, and Manoj Singh Gaur. Machine learning approach for multiple misbehavior detection in VANET. In *Proc. International Conference on Advances in Computing and Communications*, pages 644–653. Springer, 2011.
- [60] Maxime Guériau, Romain Billot, Nour-Eddin El Faouzi, Julien Monteil, Frédéric Armetta, and Salima Hassas. How to assess the benefits of connected vehicles? a simulation framework for the design of cooperative traffic management strategies. *Elsevier Transportation Research Part C: Emerging Technologies*, 67:266–279, 2016.
- [61] Isabelle Guyon, B Boser, and Vladimir Vapnik. Automatic capacity tuning of very large VC-dimension classifiers. In *Proc. Advances in Neural Information Processing Systems*, pages 147–155. NIPS, 1993.
- [62] H2O.ai. H2o. <https://github.com/h2oai/h2o-3>. Accessed: 2018-06-18.
- [63] Mordechai Haklay and Patrick Weber. OpenStreetMap: user-generated street maps. *IEEE Pervasive Computing*, 7(4):12–18, 2008.
- [64] John Harding, Gregory Powell, Rebecca Yoon, Joshua Fikentscher, Charlene Doyle, Dana Sade, Mike Lukuc, Jim Simons, and Jing Wang. Vehicle-to-vehicle communications: readiness of V2V technology for application. Technical report, 2014.
- [65] John Harris, Mark Beach, Andrew Nix, and Paul Thomas. The potential of offloading and spectrum sharing for 5G vehicular infotainment. In *Proc. Vehicular Technology Conference*, pages 1–5. IEEE, 2016.
- [66] Hannes Hartenstein and LP Laberteaux. A tutorial survey on vehicular ad hoc networks. *IEEE Communications*, 46(6), 2008.
- [67] Mikko Heiskala, Jani-Pekka Jokinen, and Markku Tinnilä. Crowdsensing-based transportation services – an analysis from business model and sustainability viewpoints. *Elsevier Research in Transportation Business & Management*, 18:38–48, 2016.

- [68] Xianan Huang, Ding Zhao, and Huei Peng. Empirical study of dsrc performance based on safety pilot model deployment data. *IEEE Transactions on Intelligent Transportation Systems*, 18(10):2619–2628, 2017.
- [69] Xin Huang. Smart antennas for intelligent transportation systems. In *Proc. International Conference on Intelligent Transportation Systems Telecommunications*, pages 426–429. IEEE, 2006.
- [70] Chalermek Intanagonwiwat, Deborah Estrin, Ramesh Govindan, and John Heidemann. Impact of network density on data aggregation in wireless sensor networks. In *Proc. International Conference on Distributed Computing Systems*, pages 457–458. IEEE, 2002.
- [71] Arash Jahangiri and Hesham A Rakha. Applying machine learning techniques to transportation mode recognition using mobile phone sensor data. *Transactions on Intelligent Transportation Systems*, 16(5):2406–2417, 2015.
- [72] Jesus A Jimenez. Smart transportation systems. In *Smart Cities*, pages 123–133. Springer, 2018.
- [73] Jaeyoung Jung and Keemin Sohn. Deep-learning architecture to forecast destinations of bus passengers from entry-only smart-card data. *IET Intelligent Transport Systems*, 11(6):334–339, 2017.
- [74] Anurag Kanungo, Ayush Sharma, and Chetan Singla. Smart traffic lights switching and traffic density calculation using video processing. In *Proc. Recent Advances in Engineering and Computational Sciences (RAECS)*, pages 1–6. IEEE, 2014.
- [75] C. Kaptan, B. Kantarci, and A. Boukerche. On the dependability of vehicular crowd-sensing data. *IEEE Transaction on Vehicular Technology (TVT)*.
- [76] C. Kaptan, B. Kantarci, and A. Boukerche. Reliability-driven vehicular crowd-sensing: a case study for localization in public transportation. In *Proc. IEEE Global Communications Conference (GLOBECOM)*. IEEE, 2018 (Accepted).
- [77] C. Kaptan, B. Kantarci, and A. Boukerche. Vehicular crowd-sensing for gps-less tracking of public transportation vehicles. *IEEE Transaction on Vehicular Technology*, 2018 (Revision Requested).
- [78] C. Kaptan, B. Kantarci, T. Soyata, and A. Boukerche. Emulating smart city sensors using soft sensing and machine intelligence: a case study in public transportation. In *Proc. IEEE International Conference on Communications (ICC)*. IEEE, 2018.
- [79] Jintao Ke, Hongyu Zheng, Hai Yang, and Xiqun Michael Chen. Short-term forecasting of passenger demand under on-demand ride services: a spatio-temporal deep learning approach. *Elsevier Transportation Research Part C: Emerging Technologies*, 85:591–608, 2017.

- [80] Muhammad Awais Khan, Susana Sargento, and Miguel Luis. Data collection from smart-city sensors through large-scale urban vehicular networks. pages 1–6, 2017.
- [81] Le Minh Kieu, Ashish Bhaskar, and Edward Chung. Mining temporal and spatial travel regularities for transit planning. <https://eprints.qut.edu.au/62976/>. Accessed: 2018-06-20.
- [82] Neeraj Kumar and Naveen Chilamkurti. Collaborative trust aware intelligent intrusion detection in VANETs. *Elsevier Computers & Electrical Engineering*, 40(6):1981–1996, 2014.
- [83] Neeraj Kumar, Sudip Misra, and Mohammad S Obaidat. Collaborative learning automata-based routing for rescue operations in dense urban regions using vehicular sensor networks. *IEEE Systems Journal*, 9(3):1081–1090, 2015.
- [84] Nicholas D Lane, Shane B Eisenman, Mirco Musolesi, Emiliano Miluzzo, and Andrew T Campbell. Urban sensing systems: opportunistic or participatory? In *Proc. Workshop on Mobile Computing Systems and Applications*, pages 11–16. ACM, 2008.
- [85] Neal Lathia, Chris Smith, Jon Froehlich, and Licia Capra. Individuals among commuters: building personalized transport information services from fare collection systems. *Elsevier Pervasive and Mobile Computing*, 9(5):643–664, 2013.
- [86] Uichin Lee and Mario Gerla. A survey of urban vehicular sensing platforms. *Elsevier Computer Networks*, 54(4):527–544, 2010.
- [87] Uichin Lee, Eugenio Magistretti, Mario Gerla, Paolo Bellavista, and Antonio Corradi. Dissemination and harvesting of urban data using vehicular sensing platforms. *IEEE Transactions on Vehicular Technology*, 58(2):882–901, 2009.
- [88] Lun Li, Jiqiang Liu, Lichen Cheng, Shuo Qiu, Wei Wang, Xiangliang Zhang, and Zonghua Zhang. CreditCoin: A privacy-preserving blockchain-based incentive announcement network for communications of smart vehicles. *IEEE Transactions on Intelligent Transportation Systems*, 2018.
- [89] Wenmin Li, Qiaoyan Wen, Qi Su, and Zhengping Jin. An efficient and secure mobile payment protocol for restricted connectivity scenarios in vehicular ad hoc network. *Elsevier Computer Communications*, 35(2):188–195, 2012.
- [90] Zhizhou Li, Justin Eichel, Akshaya Mishra, Andrew Achkar, and Kshirasagar Naik. A CPU-based algorithm for traffic optimization based on sparse convolutional neural networks. In *Proc. Canadian Conference on Electrical and Computer Engineering (CCECE)*, pages 1–5. IEEE, 2017.
- [91] Thomas Liebig, Nico Piatkowski, Christian Bockermann, and Katharina Morik. Dynamic route planning with real-time traffic predictions. *Elsevier Information Systems*, 64:258–265, 2017.

- [92] Tao Liu, Jihui Ma, Wei Guan, Yue Song, and Hu Niu. Bus arrival time prediction based on the k-nearest neighbor method. In *Proc. International Joint Conference on Computational Sciences and Optimization (CSO)*, pages 480–483. IEEE, 2012.
- [93] X. Luo, H. Zhao, H. Ogai, and C. Zhu. A novel video detection design based on modified adaboost algorithm and HSV model. In *Proc. Advanced Information Technology, Electronic and Automation Control Conference (IAEAC)*, pages 2328–2331. IEEE, March 2017.
- [94] Yisheng Lv, Yanjie Duan, Wenwen Kang, Zhengxi Li, and Fei-Yue Wang. Traffic flow prediction with big data: a deep learning approach. *IEEE Transactions on Intelligent Transportation Systems*, 16(2):865–873, 2015.
- [95] Yisheng Lv, Shuming Tang, and Hongxia Zhao. Real-time highway traffic accident prediction based on the k-nearest neighbor method. In *Proc. International Conference on Measuring Technology and Mechatronics Automation (ICMTMA)*, volume 3, pages 547–550. IEEE, 2009.
- [96] Xiaolei Ma, Yao-Jan Wu, Yinhai Wang, Feng Chen, and Jianfeng Liu. Mining smart card data for transit riders’ travel patterns. *Elsevier Transportation Research Part C: Emerging Technologies*, 36:1–12, 2013.
- [97] Xiaolei Ma, Haiyang Yu, Yunpeng Wang, and Yinhai Wang. Large-scale transportation network congestion evolution prediction using deep learning theory. *Public Library of Science PloS One*, 10(3):e0119044, 2015.
- [98] Makoto Maekawa. Intelligent transportation systems solutions. *NEC Journal of Advanced Technology*, 1(3):194–198, 2004.
- [99] Thomas J Matarazzo, Paolo Santi, Shamim N Pakzad, Kristopher Carter, Carlo Ratti, Babak Moaveni, Chris Osgood, and Nigel Jacob. Crowdsensing framework for monitoring bridge vibrations using moving smartphones. *IEEE Proceedings*, 106(4):577–593, 2018.
- [100] Hamid Menouar, Ismail Guvenc, Kemal Akkaya, A Selcuk Uluagac, Abdullah Kadri, and Adem Tuncer. UAV-enabled intelligent transportation systems for the smart city: applications and challenges. *IEEE Communications Magazine*, 55(3):22–28, 2017.
- [101] Jeffrey Miller. Vehicle-to-vehicle-to-infrastructure (V2V2I) intelligent transportation system architecture. In *Proc. Intelligent Vehicles Symposium*, pages 715–720. IEEE, 2008.
- [102] K Mohamed, Etienne Côme, Latifa Oukhellou, and Michel Verleysen. Clustering smart card data for urban mobility analysis. *IEEE Transactions on Intelligent Transportation Systems*, 18(3):712–728, 2017.

- [103] Prashanth Mohan, Venkata N Padmanabhan, and Ramachandran Ramjee. Nericell: rich monitoring of road and traffic conditions using mobile smartphones. In *Proc. Conference on Embedded Network Sensor Systems*, pages 323–336. ACM, 2008.
- [104] Catherine Morency, Martin Trépanier, and Bruno Agard. Analysing the variability of transit users behaviour with smart card data. In *Proc. Intelligent Transportation Systems Conference (ITSC)*, pages 44–49. IEEE, 2006.
- [105] Steffen Moser, Simon Eckert, and Frank Slomka. An approach for the integration of smart antennas in the design and simulation of vehicular ad-hoc networks. In *Proc. International Conference on Future Generation Communication Technology (FGCT)*, pages 36–41. IEEE, 2012.
- [106] Min Mun, Sasank Reddy, Katie Shilton, Nathan Yau, Jeff Burke, Deborah Estrin, Mark Hansen, Eric Howard, Ruth West, and Peter Boda. PEIR, the personal environmental impact report, as a platform for participatory sensing systems research. In *Proc. International Conference on Mobile Systems, Applications, and Services*, pages 55–68. ACM, 2009.
- [107] IEEE 802.11 n Working Group et al. Draft amendment to standard for information technology-telecommunications and information exchange between systems-local and metropolitan networks-specific requirements-part 11: wireless LAN medium access control and physical layer specifications: enhancements for higher throughput. *IEEE P802. 11n/D1. 0*, 2006.
- [108] Nextcon Lab. Dataset of All Experiments for IEEE Global Communications Conference. <http://nextconlab.academy/GC/>. Accessed: 2018-06-18.
- [109] Nextcon Lab. Dataset of All Experiments for IEEE International Conference on Communication. <http://nextconlab.academy/ICC/>. Accessed: 2018-06-18.
- [110] Nextcon Lab. Dataset of All Experiments for IEEE Transactions on Vehicular Technology. <http://nextconlab.academy/TVT/>. Accessed: 2018-06-18.
- [111] Nextcon Lab. Dataset of All Experiments for Reliability-unaware and Reliability-aware Participant Recruitment Techniques. <http://nextconlab.academy/MSWIM/>. Accessed: 2018-06-18.
- [112] Jianbing Ni, Aiqing Zhang, Xiaodong Lin, and Xuemin Sherman Shen. Security, privacy, and fairness in fog-based vehicular crowdsensing. *IEEE Communications Magazine*, 55(6):146–152, 2017.
- [113] Horacio A.B.F. Oliveira, Azzedine Boukerche, Eduardo F. Nakamura, and Antonio A.F. Loureiro. Localization in time and space for wireless sensor networks: an efficient and lightweight algorithm. *Elsevier Performance Evaluation*, 66(3):209 – 222, 2009. Modeling and Analysis of Wireless Networks: Selected Papers from MSWiM 2007.

- [114] Horacio A.B.F. Oliveira, Eduardo F. Nakamura, Antonio A. F. Loureiro, and Azzedine Boukerche. Error analysis of localization systems for sensor networks. In *Proc. International Workshop on Geographic Information Systems, GIS '05*, pages 71–78. ACM, 2005.
- [115] Bei Pan, Yu Zheng, David Wilkie, and Cyrus Shahabi. Crowd sensing of traffic anomalies based on human mobility and social media. In *Proc. International Conference Advances in Geographic Information Systems*, pages 344–353. ACM, 2013.
- [116] Hyunsuk Park and Youngchan Kim. Model for filtering the outliers in DSRC travel time data on interrupted traffic flow sections. *Springer Journal of Civil Engineering*, pages 1–13, 2018.
- [117] Seong-hun Park, Sung-min Kim, and Young-guk Ha. Highway traffic accident prediction using VDS big data analysis. *Springer Supercomputing*, 72(7):2815–2831, 2016.
- [118] Nandini S Patil and PR Patil. Data aggregation in wireless sensor network. In *Proc. International Conference on Computational Intelligence and Computing Research*, volume 6. IEEE, 2010.
- [119] Fabian Pedregosa, Gael Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, et al. Scikit-learn: machine learning in python. *JMLR Journal of Machine Learning Research*, 12(Oct):2825–2830, 2011.
- [120] Tam&s Peter. Fuzzy and anytime signal processing approaches for supporting modeling and control of transportation systems. In *Proc. International Conference on Computational Cybernetics (ICCC)*, pages 339–344. IEEE, 2005.
- [121] Karol Piniarski, Pawel Pawlowski, and Adam Dabrowski. Pedestrian detection by video processing in automotive night vision system. In *Proc. Signal Processing: Algorithms, Architectures, Arrangements, and Applications (SPA)*, pages 104–109. IEEE, 2014.
- [122] Albert Podusenko, Vsevolod Nikulin, Ivan Tanev, and Katsunori Shimohara. Prediction of emergency braking based on the pattern of lifting motion of accelerator pedal. In *Proc. Annual Conference of the Society of Instrument and Control Engineers of Japan (SICE)*, pages 1622–1627. IEEE, 2017.
- [123] Nicholas G Polson and Vadim O Sokolov. Deep learning for short-term traffic flow prediction. *Elsevier Transportation Research Part C: Emerging Technologies*, 79:1–17, 2017.
- [124] Maryam Pouryazdan, Burak Kantarci, Tolga Soyata, Luca Foschini, and Houbing Song. Quantifying user reputation scores, data trustworthiness, and user incentives in mobile crowd-sensing. *IEEE Access*, 5:1382–1397, 2017.

- [125] Maryam Pouryazdan, Burak Kantarci, Tolga Soyata, and Houbing Song. Anchor-assisted and vote-based trustworthiness assurance in smart city crowdsensing. *IEEE Access*, 4:529–541, 2016.
- [126] Khaled Rabieh, Mohamed MEA Mahmoud, and Mohamed Younis. Privacy-preserving route reporting schemes for traffic management systems. *IEEE Transactions on Vehicular Technology*, 66(3):2703–2713, 2017.
- [127] Filippo Rebecchi, Marcelo Dias De Amorim, Vania Conan, Andrea Passarella, Raffaele Bruno, and Marco Conti. Data offloading techniques in cellular networks: a survey. *IEEE Communications Surveys & Tutorials*, 17(2):580–603, 2015.
- [128] Ju Ren, Yaoyue Zhang, Kuan Zhang, and Xuemin Shen. Exploiting mobile crowdsourcing for pervasive cloud services: challenges and solutions. *IEEE Communications Magazine*, 53(3):98–105, 2015.
- [129] Benjamin Ryder and Felix Wortmann. Autonomously detecting and classifying traffic accident hotspots. In *Proc. International Joint Conference on Pervasive and Ubiquitous Computing and International Symposium on Wearable Computers*, pages 365–370. ACM, 2017.
- [130] Jagruti Sahoo, Soumaya Cherkaoui, and Abdelhakim Hafid. A novel vehicular sensing framework for smart cities. In *Proc. Conference on Local Computer Networks*, pages 490–493. IEEE, 2014.
- [131] Jagruti Sahoo, Soumaya Cherkaoui, Abdelhakim Hafid, and Pratap Kumar Sahu. Dynamic hierarchical aggregation for vehicular sensing. *IEEE Transactions on Intelligent Transportation Systems*, 18(9):2539–2556, 2017.
- [132] Tom Schaul, Justin Bayer, Daan Wierstra, Yi Sun, Martin Felder, Frank Sehnke, Thomas RĆ1/4ckstieĆ, and JĆ1/4rgen Schmidhuber. PyBrain. *JMLR Journal of Machine Learning Research*, 11(Feb):743–746, 2010.
- [133] Ramon S Schwartz, Rafael RR Barbosa, Nirvana Meratnia, Geert Heijenk, and Hans Scholten. A directional data dissemination protocol for vehicular environments. *Elsevier Computer Communications*, 34(17):2057–2071, 2011.
- [134] E Vijay Sekar, J Anuradha, Anshita Arya, Balamurugan Balusamy, and Victor Chang. A framework for smart traffic management using hybrid clustering techniques. *Springer Cluster Computing*, pages 1–16, 2017.
- [135] Xavier Sevillano, Elena MÀrmol, and Virginia Fernandez-Arguedas. Towards smart traffic management systems: vacant on-street parking spot detection based on video analytics. In *Proc. International Conference on Information Fusion (FUSION)*, pages 1–8. IEEE, 2014.

- [136] Vishal Sharma, Ilsun You, Giovanni Pau, Mario Collotta, Jae Deok Lim, and Jeong Nyeo Kim. LoRaWAN-based energy-efficient surveillance by drones for intelligent transportation systems. *MDPI Energies*, 11(3):573, 2018.
- [137] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *ArXiv Preprint ArXiv:1409.1556*, 2014.
- [138] Mani Srivastava, Tarek Abdelzaher, and Boleslaw Szymanski. Human-centric sensing. *The Royal Society Phil. Trans. R. Soc. A*, 370(1958):176–197, 2012.
- [139] Chris Stauffer and W. Eric L. Grimson. Learning patterns of activity using real-time tracking. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 22(8):747–757, 2000.
- [140] Fangzhou Sun, Yao Pan, Jules White, and Abhishek Dubey. Real-time and predictive analytics for smart public transportation decision support system. In *Proc. International Conference on Smart Computing (SMARTCOMP)*, pages 1–8. IEEE, 2016.
- [141] Mingshun Sun, Ming Li, and Ryan Gerdes. A data trust framework for VANETs enabling false data detection and secure vehicle tracking. In *Proc. Communications and Network Security*, pages 1–9. IEEE, 2017.
- [142] Yuxing Sun, Biao Leng, and Wei Guan. A novel wavelet-SVM short-time passenger flow prediction in Beijing subway system. *Elsevier Neurocomputing*, 166:109–121, 2015.
- [143] Paula Ta-Shma, Adnan Akbar, Guy Gerson-Golan, Guy Hadash, Francois Carrez, and Klaus Moessner. An ingestion and analytics architecture for IoT applied to smart city use cases. *IEEE Internet of Things*, 5(2):765–774, 2018.
- [144] Nasrin Taherkhani and Samuel Pierre. Centralized and localized data congestion control strategy for vehicular ad hoc networks using a machine learning clustering algorithm. *IEEE Transactions on Intelligent Transportation Systems*, 17(11):3275–3285, 2016.
- [145] Irina Tal and Gabriel-Miro Muntean. Using fuzzy logic for data aggregation in vehicular networks. In *Proc. International Symposium on Distributed Simulation and Real Time Applications (DS-RT)*, pages 151–154. IEEE, 2012.
- [146] Jinjun Tang, Fang Liu, Yajie Zou, Weibin Zhang, and Yinhai Wang. An improved fuzzy neural network for traffic speed prediction considering periodic characteristic. *IEEE Transactions on Intelligent Transportation Systems*, 18(9):2340–2350, 2017.
- [147] Robert W Thomas and José M Vidal. Toward detecting accidents with already available passive traffic information. In *Proc. Computing and Communication Workshop and Conference (CCWC)*, pages 1–4. IEEE, 2017.

- [148] JWC Van Lint, SP Hoogendoorn, and Henk J van Zuylen. Accurate freeway travel time prediction with state-space neural networks under missing data. *Elsevier Transportation Research Part C: Emerging Technologies*, 13(5-6):347–369, 2005.
- [149] Ponnusamy Vijayakumar, Ravi Ranjan, and S Malarvizhi. Waveform generation and reception of IEEE 802.11p standard for CR-VANET application. In *Smart Computing and Informatics*, pages 583–591. Springer, 2018.
- [150] Félix Jesús Villanueva, David Villa, Maria José Santofimia, Jesus Barba, and Juan Carlos Lopez. Crowdsensing smart city parking monitoring. In *Proc. World Forum on Internet of Things (WF-IoT)*, pages 751–756. IEEE, 2015.
- [151] Berkley Vision. AlexNet. https://github.com/BVLC/caffe/tree/master/models/bvlc_alexnet. Accessed: 2018-06-18.
- [152] Berkley Vision. Caffe. <http://caffe.berkeleyvision.org/>. Accessed: 2018-06-18.
- [153] Isabel Wagner. Evaluating the strength of genomic privacy metrics. *ACM Transactions on Privacy and Security*, 20(1):2, 2017.
- [154] Isabel Wagner. Measuring privacy in vehicular networks. In *Proc. Conference on Local Computer Networks (LCN)*, pages 183–186. IEEE, 2017.
- [155] Jiafu Wan, Jianqi Liu, Zehui Shao, Athanasios V Vasilakos, Muhammad Imran, and Keliang Zhou. Mobile crowd sensing for traffic prediction in internet of vehicles. *MDPI Sensors*, 16(1):88, 2016.
- [156] Cheng Wang, Zhenzhen Zhang, Lu Shao, and MengChu Zhou. Estimating travel speed via sparse vehicular crowdsensing data. In *Proc. World Forum on Internet of Things (WF-IoT)*, pages 643–648. IEEE, 2016.
- [157] Hongjian Wang, Yanmin Zhu, and Qian Zhang. Compressive sensing based monitoring with vehicular networks. In *Proc. INFOCOM*, pages 2823–2831. IEEE, 2013.
- [158] Jian-Gang Wang, Lubing Zhou, Yu Pan, Serin Lee, Zhiwei Song, Boon Siew Han, and Vincensius Billy Saputra. Appearance-based brake-lights recognition using deep learning and vehicle detection. In *Intelligent Vehicles Symposium (IV)*, pages 815–820. IEEE, 2016.
- [159] Jingjing Wang, Chunxiao Jiang, Kai Zhang, Tony QS Quek, Yong Ren, and Lajos Hanzo. Vehicular sensing networks in a smart city: principles, technologies and applications. *IEEE Wireless Communications*, 2017.
- [160] Min Wang, Martin Winbjork, Zhang Zhang, Ricardo Blasco, Hieu Do, Stefano Sorrentino, Marco Belleschi, and Yunpeng Zang. Comparison of LTE and DSRC-based connectivity for intelligent transportation systems. In *Proc. Vehicular Technology Conference (VTC)*, pages 1–5. IEEE, 2017.

- [161] Shangguang Wang, Tao Lei, Lingyan Zhang, Ching-Hsien Hsu, and Fangchun Yang. Offloading mobile data traffic for qos-aware service provision in vehicular cyber-physical systems. *Elsevier Future Generation Computer Systems*, 61:118–127, 2016.
- [162] Chun-Hsin Wu, Jan-Ming Ho, and Der-Tsai Lee. Travel-time prediction with support vector regression. *IEEE Transactions on Intelligent Transportation Systems*, 5(4):276–281, 2004.
- [163] Di Wu, Yuan Zhang, Juan Luo, and Renfa Li. Efficient data dissemination by crowd-sensing in vehicular networks. In *Proc. International Symposium of Quality of Service (IWQoS)*, pages 314–319. IEEE, 2014.
- [164] Hansong Xu, Jie Lin, and Wei Yu. Smart transportation systems: architecture, enabling technologies, and open issues. In *Secure and Trustworthy Transportation Cyber-Physical Systems*, pages 23–49. Springer, 2017.
- [165] Qing Xu, Tony Mak, Jeff Ko, and Raja Sengupta. Vehicle-to-vehicle safety messaging in dsrc. In *Proc. International Workshop on Vehicular Ad Hoc Networks*, pages 19–28. ACM, 2004.
- [166] Rui Xu and Donald Wunsch. Survey of clustering algorithms. *IEEE Transactions on Neural Networks*, 16(3):645–678, 2005.
- [167] Hehua Yan, Qingsong Hua, Daqiang Zhang, Jiafu Wan, Seungmin Rho, and Houbing Song. Cloud-Assisted mobile crowdsensing for traffic congestion control. *Springer Mobile Networks and Applications*, 22(6):1212–1218, 2017.
- [168] Hongming Yang, Youjun Deng, Jing Qiu, Ming Li, Mingyong Lai, and Zhao Yang Dong. Electric vehicle route selection and charging navigation strategy based on crowdsensing. *IEEE Transactions on Industrial Informatics*, 13(5):2214–2226, 2017.
- [169] Kan Yang, Kuan Zhang, Ju Ren, and Xuemin Shen. Security and privacy in mobile crowdsourcing networks: challenges and opportunities. *IEEE Communications Magazine*, 53(8):75–81, 2015.
- [170] Ibrar Yaqoob, Ibrahim Abaker Targio Hashem, Yasir Mehmood, Abdullah Gani, Salimah Mokhtar, and Sghaier Guizani. Enabling communication technologies for smart cities. *IEEE Communications Magazine*, 55(1):112–120, 2017.
- [171] Shu-Chung Yi, Yeong-Chin Chen, and Ching-Haur Chang. A lane detection approach based on intelligent vision. *Elsevier Computers & Electrical Engineering*, 42:23–29, 2015.
- [172] Rujun Yu and Xiuqing Liu. Study on traffic accidents prediction model based on RBF neural network. In *Proc. International Conference on Information Engineering and Computer Science (ICIECS)*, pages 1–4. IEEE, 2010.

- [173] Chi Zhang, Hong Wang, and Rongrong Fu. Automated detection of driver fatigue based on entropy and complexity measures. *IEEE Transactions on Intelligent Transportation Systems*, 15(1):168–177, 2014.
- [174] Hengyang Zhang, Shixiang Zhu, Renchao Xie, Tao Huang, and Yunjie Liu. Intelligent resources management system design in information centric networking. *IEEE China Communications*, 14(8):105–123, 2017.
- [175] Li Zhang, Weida Zhou, and Licheng Jiao. Wavelet support vector machine. *IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics)*, 34(1):34–39, 2004.
- [176] Ning Zhang, Nan Cheng, Amila Tharaperiya Gamage, Kuan Zhang, Jon W Mark, and Xuemin Shen. Cloud assisted HetNets toward 5G wireless networks. *IEEE Communications Magazine*, 53(6):59–65, 2015.
- [177] Rui Zhang, Wenping Liu, Yufu Jia, Guoyin Jiang, Jing Xing, Hongbo Jiang, and Jiangchuan Liu. WiFi sensing-based real-time bus tracking and arrival time prediction in urban environments. *IEEE Sensors*, 18(11):4746–4760, 2018.
- [178] Yang Zhang and Yuncai Liu. Traffic forecasting using least squares support vector machines. *Taylor & Francis Transportmetrica*, 5(3):193–213, 2009.
- [179] Dong Zhao, Huadong Ma, Liang Liu, and Xiang-Yang Li. Opportunistic coverage for urban vehicular sensing. *Elsevier Computer Communications*, 60:71–85, 2015.
- [180] Jian Zhao, Chuan Wu, and Zongpeng Li. Cost minimization in multiple IaaS clouds: a double auction approach. *ArXiv Preprint ArXiv:1308.0841*, 2013.
- [181] Kan Zheng, Qiang Zheng, Periklis Chatzimisios, Wei Xiang, and Yiqing Zhou. Heterogeneous vehicular networking: a survey on architecture, challenges, and solutions. *IEEE Communications Surveys & Tutorials*, 17(4):2377–2396, 2015.
- [182] Qiusheng Zheng. Detecting bogus messages in vehicular ad-hoc networks: an information fusion approach. In *Proc. China Wireless Sensor Network Conference (CWSN)*, volume 812, page 191. Springer, 2018.
- [183] Yanxu Zheng, Sutharshan Rajasegarar, and Christopher Leckie. Parking availability prediction for sensor-enabled car parks in smart cities. In *Proc. International Conference on Intelligent Sensors, Sensor Networks and Information Processing (ISSNIP)*, pages 1–6. IEEE, 2015.
- [184] Yajie Zou, Xuedong Hua, Yanru Zhang, and Yinhai Wang. Hybrid short-term freeway speed prediction methods based on periodic analysis. *NRC Research Press Canadian Journal of Civil Engineering*, 42(8):570–582, 2015.