

**FUSÃO NO ESPAÇO DE DADOS VEICULARES:
UMA ABORDAGEM PARA MOBILIDADE
INTELIGENTE**

PAULO HENRIQUE LOPES RETTORE

**FUSÃO NO ESPAÇO DE DADOS VEICULARES:
UMA ABORDAGEM PARA MOBILIDADE
INTELIGENTE**

Tese apresentada ao Programa de Pós-Graduação em Ciência da Computação do Instituto de Ciências Exatas da Universidade Federal de Minas Gerais como requisito parcial para a obtenção do grau de Doutor em Ciência da Computação.

ORIENTADOR: ANTONIO ALFREDO FERREIRA LOUREIRO
COORIENTADOR: LEANDRO A. VILLAS, JOÃO GUILHERME M. DE
MENEZES

Belo Horizonte
Março de 2019

PAULO HENRIQUE LOPES RETTORE

**FUSION ON VEHICULAR DATA SPACE: AN
APPROACH TO SMART MOBILITY**

Thesis presented to the Graduate Program in Computer Science of the Federal University of Minas Gerais in partial fulfillment of the requirements for the degree of Doctor in Computer Science.

ADVISOR: ANTONIO ALFREDO FERREIRA LOUREIRO
CO-ADVISOR: LEANDRO A. VILLAS, JOÃO GUILHERME M. DE
MENEZES

Belo Horizonte

March 2019

© 2019, Paulo Henrique Lopes Rettore
Todos os direitos reservados

Ficha catalográfica elaborada pela Biblioteca do ICEx - UFMG

Rettore, Paulo Henrique Lopes

R237f Fusion on vehicular data space: an approach to smart mobility / Paulo Henrique Lopes Rettore—
Belo Horizonte, 2019.
xxviii, 214 p.: il.; 29 cm.

Tese (doutorado) - Universidade Federal de Minas Gerais – Departamento de Ciência da Computação.

Orientador: Antonio Alfredo Ferreira Loureiro
Coorientador: João Maia de Menezes
Coorientador: Leandro Aparecido Villas

1. Computação – Teses. 2. Sistemas de transporte inteligentes 3. Fusão de dados heterogêneos. 4. Mobilidade inteligente. I. Orientador. II. Coorientador. II. Coorientador. III. Título.

CDU 519.6*65(043)



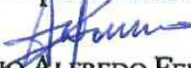
UNIVERSIDADE FEDERAL DE MINAS GERAIS
INSTITUTO DE CIÊNCIAS EXATAS
PROGRAMA DE PÓS-GRADUAÇÃO EM CIÊNCIA DA COMPUTAÇÃO

FOLHA DE APROVAÇÃO


Fusion on Vehicular Data Space: An Approach to Smart Mobility

PAULO HENRIQUE LOPES RETTORE

Tese defendida e aprovada pela banca examinadora constituída pelos Senhores:

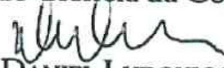

PROF. ANTONIO ALFREDO FERREIRA LOUREIRO - Orientador
Departamento de Ciência da Computação - UFMG

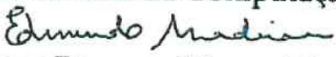

PROF. JOÃO GUILHERME MAIA DE MENEZES - Coorientador
Departamento de Ciência da Computação - UFMG


PROF. LEANDRO APARECIDO VILLAS - Coorientador
Departamento de Sistemas de Computação - UNICAMP


PROF. ANDRÉ LUIZ LINS DE AQUINO
Instituto de Computação - UFAL


PROF. CLODOVEU AUGUSTO DAVIS JÚNIOR
Departamento de Ciência da Computação - UFMG


PROF. DANIEL LUDOVICO GUIDONI
Departamento de Ciência da Computação - UFSJ


PROF. EDMUNDO ROBERTO MAURO MADEIRA
Instituto de Computação - UNICAMP

Belo Horizonte, 15 de Março de 2019.

Acknowledgments

I would like to thank the research agencies, CAPES/CNPq for funding this Ph.D. Also my co-workers and advisers. Especially my wife, daughter, mother and deceased father which was hoping to live to see myself concluding this work.

Along these four and half years, I have learned how to push myself forward, even though I had to fight my fears and limitations. Each person who passed through my life in this process contributed, somehow, to make me stronger than ever. At the end, I felt I did my best. The conditions I lived through allowed me to do what I did, not less, not more. I summarized my trajectory in Figure 1. Even though I lived hard personal and professional moments, I am grateful to experience these things and became a person which I am today.

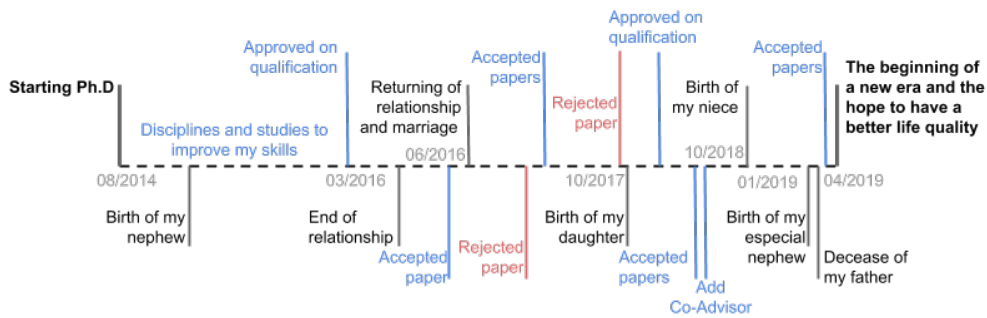


Figure 1: My Ph.D. time line.

“To my beloved wife and my little princess who was born in September 2017.”
(Paulo Rettore)

Resumo

Esta tese tem como objetivo o uso de diversas fontes de dados para promover a melhora da mobilidade atual nas cidades. No entanto, um desafio substancial surge quando combinamos várias fontes de dados, aumentando os problemas de cobertura espaço-temporal que afetam o desenvolvimento de soluções para Sistemas de Transporte Inteligentes – Intelligent Transportation System (ITS), especificamente Mobilidade Inteligente – Smart Mobility (SM). Nesse sentido, investigamos soluções para melhorar a qualidade dos dados do sistema de transporte, fornecendo aplicações e serviços, permitindo que a fusão entre Dados Intra-Veiculares – Intra-Vehicular Data (IVD) e Dados Extra-Veiculares – Extra-Vehicular Data (EVD) melhore a qualidade do transporte e mobilidade. Projetamos uma plataforma de fusão de dados heterogêneos para SM, com o objetivo de analisar os dados do sistema de transporte, introduzido como Espaço de Dados Veiculares – Vehicular Data Space (VDS), considerando seus aspectos espaço-temporais e identificar métodos e técnicas para a fusão desses dados. Foi criado o conceito VDS, que mapeia os dados disponíveis e usados pela comunidade para desenvolver soluções para ITS. Depois disso, desenvolvemos um conjunto de abordagens para fundir vários conjuntos de dados em benefício do ITS e SM. Inicialmente, realizamos estudos com o objetivo de fundir IVD economizando combustível, reduzindo as emissões de gases e garantindo a segurança no compartilhamento de carros em Redes Veiculares – Ad-hoc Networks (VANETs). Além disso, fundindo EVD desenvolvemos um modelo baseado em dados de mídia social, para enriquecer as informações atuais de trânsito, oferecendo mais opções para as pessoas se locomoverem na cidade. Finalmente, desenvolvemos uma abordagem para fundir Dados Intra-Extra-Veiculares – Intra-Extra-Vehicular Data (IEVD), permitindo

melhorar a qualidade dos dados de tráfego e enriquecer a atual cobertura do dados.

Abstract

Urban mobility aspects have become a challenge with the constant growth of the global population. As a consequence of such increase, more data has become available, which allows new information technologies to improve the mobility systems, especially the transportation system. Thus, a possible strategy to handle these issues is to employ an Intelligent Transportation System (ITS). However, the development of new applications and services for the ITS environment, improving the mobility, depending on the availability of vast amounts of data, despite its currently slow availability. This thesis aims to explore data from a vast number of sources from the ITS context to provide directions to improve mobility in cities. However, a substantial challenge emerges when we combine multiple data sources, increasing the data aspects as spatiotemporal coverage, which affects the development of Smart Mobility (SM) solutions. In this sense, we investigate solutions to improve the data quality of transportation systems, providing applications and services, enabling Intra-Vehicle Data (IVD) and Extra-Vehicle Data (EVD) fusion to enrich the raw data. We design a heterogeneous data fusion platform to SM, aiming to fuse those data considering their aspects, highlighting the most relevant methods and techniques to achieve the application goals. We introduce the concept of Vehicular Data Space (VDS), which maps the data available and used by the research community to design solutions for ITS. After that, we develop a set of approaches to fuse various datasets in benefit of SM. Initially, we conducted studies to fuse IVD to save fuel, reduce emissions and ensure the security of car-sharing in Vehicular *Ad-hoc* Network (VANET). Moreover, using the fusion of EVD, we developed a model, based on social media data to enrich the current traffic information, offering more options to people to move in a city. Finally, we

developed an approach to fuse Intra and Extra-Vehicle Data (IEVD), allowing to enhance the road traffic data quality and enriches the current spatiotemporal data coverage.

List of Figures

1	My Ph.D. time line.	ix
1.1	The data cycle on the transportation system.	3
1.2	Design of fusion on VDS.	5
2.1	The big picture of Vehicular Data Space and its respective state of data cycle in the VDS.	12
2.2	Vehicular data space provided in the urban area.	14
2.3	Data provided by infrastructure.	15
2.4	Data provided by government entities.	16
2.5	Data provided by media.	19
2.6	Taxonomy of vehicular data space based on the point of view of the source.	20
2.7	(a) Most used data source in VDS. (b) An overview of data acquisition based on its granularity and financial costs.	39
2.8	Applications based on vehicular data.	39
2.9	Overview of application groups based on their granularity, financial costs and data availability.	56
3.1	Comparison Between Vehicle Speed and GPS Speed Collected Every Second and Every Minute.	71
3.2	Comparison of GPS Speed and Incomplete GPS Speed Data.	73
3.3	Difference Between Vehicle Speed and GPS Speed (a) and Correlation Between Sensors Data in a Vehicle (b).	75
3.4	Disparateness Between Revolution per Minute and Carbon Dioxide. . .	77

4.1	Correlation between sensors.	84
4.2	Vehicle sensor data behavior along the trace.	84
4.3	Correlation between vehicle speed and RPM.	87
4.4	Vehicle's speed and RPM relation in a time series.	87
4.5	Design of fusion on VDS for gear virtual sensor.	89
4.6	Virtual sensor design scheme.	91
4.7	Example of calculated virtual sensors.	95
4.8	Distribution of acceleration values (a) and Route between areas delimited using the geofencing technique (b).	96
4.9	Speed and RPM relationship defined by gears.	97
4.10	Accelerometer readings on trips (a) and Cumulative precision of driver identification (b).	98
4.11	Smartphone accelerometer sensor with thresholds to determine driver behavior.	101
4.12	Instant precision of driver identification.	103
4.13	Design of fusion on VDS for vehicular virtual sensors.	105
4.14	Correlation between vehicle's speed and RPM after clustering.	110
4.15	Speed and fuel consumption relationship for different gears of Vehicle 1.	111
4.16	Gear frequency at a given speed.	112
4.17	Gear frequency at a given speed.	113
4.18	Pairs of drivers sharing data.	116
4.19	Design of fusion on VDS for eco-driving.	118
4.20	Identification of a legitimate/illegitimate driver.	122
4.21	The most representative variables of the dataset.	125
4.22	Accuracy vs. number of features using different data treatment techniques.	128
4.23	Classifier results when treating driver 10 as a legitimate and suspect driver.	130
4.24	The spread of infected vehicles in VANET scenarios.	132
4.25	Design of fusion on VDS for driver authentication.	134
5.1	The design of Road Data Enrichment (RoDE).	137
5.2	Route sentiment based on the tweets text analysis	140

5.3	Tweets frequency and Here Jam Factor time series.	144
5.4	Twitter MAPS (T-MAPS) modeling process.	146
5.5	Route recommendation similarity between T-MAPS and Google Direc- tions (dots represent the mean).	148
5.6	Route sentiment based on the tweets text analysis	149
5.7	The Area' Tags (AT) of each region of the path.	150
5.8	The spatial coverage by data sources used.	153
5.9	Hour of an incident by data source and the intersection of them.	154
5.10	Spatial incident coverage per data layer.	155
5.11	Design of Twitter Incident (T-Incident).	155
5.12	Spatiotemporal grouping based on a radius of 0.01 km ((a) and (b)) and 0.5 km ((c) and (d)).	159
5.13	The learning curve of a given kernel and spatiotemporal grouping.	164
5.14	Classification results based on different kernels and evaluation metrics.	165
5.15	Design of fusion on VDS for RoDE.	168
6.1	Spatiotemporal analysis of vehicular traces.	175
6.2	Traffic data analysis in Monchengladbach.	177
6.3	Design of Traffic Data Enrichment Sensor (TraDES).	178
6.4	Metrics per set of features.	186
6.5	Learning curve of RFE algorithm.	187
6.6	Evaluation of trips and street segments between the raw data and the fused data.	188
6.7	Traffic map coverage between the raw data and fused data in Monchengladbach.	189
6.8	Evaluation of trips and street segments between the raw data and the fused data.	190
6.9	Design of fusion on VDS for TraDES.	191

List of Tables

2.1	Data from a vehicle and additional devices embedded in it.	17
2.2	Summarizing of data source in vehicular data space taxonomy.	37
2.3	Class of data from VDS based on a given application group.	40
2.4	Vehicular data space focus on safety applications.	45
2.5	Vehicular data space focus on eco-driving applications.	48
2.6	Vehicular data space focus on traffic monitoring and management applications.	51
2.7	Vehicular data space focus on general purpose applications.	53
2.8	Availability of Vehicular data space.	54
3.1	Most used classes of machine learning algorithms by the ITS applications.	67
3.2	On-Board Diagnostic (OBD) Signaling Protocols	69
3.3	Sensors Collected from OBD and Smartphone	70
4.1	Data acquisition characteristics	86
4.2	Data acquisition characteristics	93
4.3	Engine Control Unit (ECU) data, smartphone and virtual sensors . . .	108
4.4	Characteristics of data collected	108
4.5	Evaluation of gear recommendation system	114
5.1	Data acquired from different data sources.	152
5.2	Number of tweets for each spatiotemporal grouping model.	156
5.3	Relevant features based on radius of 0.01 km.	160
6.1	Features from vehicles and roads.	173

6.2	Data acquired from different data sources.	174
6.3	Set of features resulted by each selection technique.	183
6.4	Data to feed the learning-based model.	184

List of Acronyms

- ADAS** Advanced Driver Assistant Systems. 9, 118, 121–123, 126, 134
- ANN** Artificial Neural Networks. 8, 176
- CAN** Controlled Area Network. 21–26, 172
- DAS** Driver Assistant Systems. 25, 46
- DVS** Device as Vehicular Sensor. 20
- ECU** Engine Control Unit. xviii, 9, 17, 20, 21, 23, 25, 29, 37, 38, 42, 57, 58, 68, 69, 71, 73, 106, 108, 172
- EDAS** Ecological Driving Assistance System. 25, 27
- EVD** Extra-Vehicle Data. xiii, 5, 6, 40, 56, 59, 66, 68, 169, 192, 193, 195
- EVS** Extra-Vehicular Sensor. 31, 66
- IEVD** Intra and Extra-Vehicle Data. xiv, 5, 6, 192, 193
- IMU** Inertial Measurement Unit. 23, 26, 27, 29, 57
- InfraVS** Infrastructure as Vehicular Sensor. 31, 33
- IoT** Internet of Things. 195
- IoV** Internet of Vehicle. 28

ITS Intelligent Transportation System. xiii, xviii, 2–8, 10, 11, 13, 19, 36–38, 41, 49, 50, 59–63, 65–68, 77, 78, 117, 118, 133, 135, 154, 167, 169–171, 176, 178, 185, 187, 189, 192, 193

IVC Intra-Vehicular Communication. 9

IVD Intra-Vehicle Data. xiii, 5, 6, 40, 56, 59, 66, 68, 134, 169, 192, 193, 195

IVN Intra-Vehicular Network. 9

IVS Intra-Vehicular Sensor. 20, 21, 33, 59, 66, 78

IVSN Intra-Vehicle Sensor Network. 9

IVWSN Intra-Vehicle Wireless Sensor Network. 9

KNN k-Nearest Neighbors. 161, 164, 183, 184

LBSM Location-Based Social Media. 7, 8, 18, 19, 35, 49, 135–145, 151, 152, 154, 157, 161–163, 166, 167, 174, 193, 194

MANETs Mobile *Ad-hoc* Networks. 12

MLP Multi-Layer Perceptron. 184–186

MVS Media as Vehicular Sensor. 31, 34, 135

NLP Natural Language Processing. 138, 151, 157, 162

NN Neural Network. 185

OBD On-Board Diagnostic. xviii, 17, 21–29, 43, 69, 70, 73, 74, 78, 79, 81, 82, 88, 91, 93, 98, 102–104, 106, 108, 109, 117–120, 123, 133, 171, 172, 178, 191, 194

PAYD Pay-As-You-Drive. 26, 30, 43

PHYD Pay-How-You-Drive. 43

PSN Participatory Sensor Networks. 13

QVS Questionnaire as Vehicular Sensor. 31, 32

RF Random Forest Classifier. 161, 164, 184

RoDE Road Data Enrichment. xvi, xvii, 6–8, 136, 137, 139, 150, 151, 166–168, 193, 194

RPM Revolutions Per Minute. 20, 22–24, 26, 52, 66

SDL Smart Device Link. 55

SM Smart Mobility. xiii, 3, 4, 6, 77, 78, 89, 104, 117, 133, 135, 169, 192, 193, 195

SVM Support Vector Machine. 161, 164, 184

T-Incident Twitter Incident. xvii, 8, 137, 150–152, 155, 162, 163, 165, 166, 194, 195

T-MAPS Twitter MAPS. xvii, 7, 8, 136, 137, 139, 140, 145–150, 194

TraDES Traffic Data Enrichment Sensor. xvii, 6, 8, 169–171, 176, 178, 180, 182, 185–189, 191, 193, 195

V2I Vehicle-to-Infrastructure. 17

VANET Vehicular *Ad-hoc* Network. xiii, xvi, 12, 13, 55, 78, 92, 115–119, 122, 129, 131–133, 194

VD Vehicular Data. 11

VDS Vehicular Data Space. xiii, xv–xvii, 4–7, 10–13, 17, 19, 31, 38, 39, 44, 56, 59, 61, 63–67, 88, 89, 104, 105, 117, 118, 133, 134, 167–169, 191–193

VDSource Vehicular Data Source. 11, 20, 38, 59, 64, 66, 68

VS Virtual Sensor. 118, 119, 123, 124, 133

VSD Vehicular Sensor Data. 28

VSN Vehicular Sensor Network. 12, 30, 34, 52

VSocN Vehicular Social Networks. 28, 49

List of Algorithms

1	Spatiotemporal LBSM Data Grouping	158
2	Spatiotemporal Traffic Data Grouping	179

Contents

Acknowledgments	ix
Resumo	xi
Abstract	xiii
List of Figures	xv
List of Tables	xviii
List of Acronyms	xx
List of Algorithms	xxiii
1 Introduction	1
1.1 Motivation	1
1.2 Objectives	4
1.3 Contributions	6
1.4 Outline	7
2 Vehicular Data Space	9
2.1 Vehicular Data Space	10
2.2 Entities of the Vehicular Data Space	12
2.2.1 Infrastructure	14
2.2.2 Transit Authority	15
2.2.3 Vehicle	16

2.2.4	Publicity	17
2.2.5	Media	18
2.3	Taxonomy of Vehicular Data Source	20
2.3.1	Intra-Vehicular Sensor	20
2.3.2	Extra-Vehicular Sensor	31
2.3.3	Considerations	36
2.4	Potential Applications	39
2.4.1	Safety	41
2.4.2	Eco-Driving	46
2.4.3	Traffic Monitoring and Management	49
2.4.4	General Purpose	52
2.4.5	Infotainment	54
2.4.6	Data Availability	55
2.4.7	Overview	57
2.5	Chapter Remarks	59
3	Heterogeneous Data Fusion	61
3.1	Contextualization	61
3.2	Data Preparation	63
3.3	Data Processing	65
3.4	Vehicular Sensor Data Fusion	68
3.4.1	Vehicular Data	68
3.4.2	Heterogeneous Data	69
3.4.3	Problems of Heterogeneous Data Fusion: Case Study	70
3.5	Chapter Remarks	77
4	Intra-Vehicular Data Fusion	78
4.1	Vehicular Sensor Data: Characterization and Relationships	78
4.1.1	Related Work	79
4.1.2	Characteristics of Vehicular Data	81
4.1.3	Case Study	85
4.1.4	Results	86
4.1.5	Section Remarks	88

4.2	Vehicular Virtual Sensor	89
4.2.1	Related Work	91
4.2.2	Data Acquisition	92
4.2.3	Operating Vehicular Data	94
4.2.4	Mathematical Operators	94
4.2.5	Section Remarks	102
4.3	A Method of Eco-driving	104
4.3.1	Related Work	106
4.3.2	Data Acquisition	107
4.3.3	Data Preparation	109
4.3.4	Gear Sensor	109
4.3.5	Efficient Gear Change Service	110
4.3.6	Results	114
4.3.7	Collaborative Recommendation Service	115
4.3.8	Section Remarks	117
4.4	Driver Authentication in VANET	117
4.4.1	Related Work	119
4.4.2	Extra Factor for Driver Authentication	121
4.4.3	Data Acquisition	123
4.4.4	Data Preparation	124
4.4.5	Identification of Drivers and Suspects	125
4.4.6	Suspicious Vehicles in VANETs	131
4.4.7	Section Remarks	133
4.5	Chapter Remarks	134
5	Extra-Vehicular Data Fusion	135
5.1	Enriching Road Data Based on Social Media	135
5.2	Related Work	137
5.3	RoDE: Route Service	139
5.3.1	Data Acquisition	139
5.3.2	What We Have Learned From The Data Aspects	141
5.3.3	Twitter as a traffic sensor	144
5.3.4	T-MAPS Modeling Process	145

5.3.5	A Case Study	146
5.3.6	Route Description Services	148
5.3.7	Discussion	150
5.4	RoDE: Incident Service	150
5.4.1	Data Acquisition	151
5.4.2	Incident Data Fusion	152
5.4.3	T-Incident Design Architecture	154
5.4.4	Evaluation	162
5.4.5	Discussion	166
5.5	Chapter Remarks	167
6	Intra-Extra-Vehicular Data Fusion	169
6.1	Introduction	169
6.2	Related Work	170
6.3	Data Acquisition	171
6.3.1	Data Characterization	174
6.4	TraDES' Design	176
6.4.1	Input and Output Data	178
6.4.2	Data Preparation	178
6.4.3	Learning-based Model	184
6.5	Evaluation	185
6.6	Chapter Remarks	189
7	Final Remarks	192
7.1	Conclusions	192
7.2	Future Work	193
7.3	Comments on Publications	195
7.3.1	Contributions from the Thesis	195
7.3.2	Other Publications	197
	Bibliography	198

Chapter 1

Introduction

Over the years, cities have required new improvements in their transportation systems. In that way, initiatives to enhance road traffic efficiency, safety and people's mobility become important challenges to advance transportation systems, paving the way to Smart Cities. Considering the need of transportation systems data to develop smart solutions, we face the problem of poor data quality currently available and its aspects such as imperfection, inconsistencies, spatiotemporal gaps (incompleteness), outliers, unstructured data, non-standardized data acquisition and others. Applications and services for transportation systems need to use a vast range of data sources to deal with those aspects. In this thesis, we aim to provide a set of applications and services to improve the current transportation systems, through the use of methods and technique to apply heterogeneous data fusion.

This chapter is organized as follows. Section 1.1 motivates the current research. Section 1.2 presents the objectives of this thesis. Section 1.3 presents the main contributions conducted in this investigation. Finally, Section 1.4 outlines the following chapters.

1.1 Motivation

In general, medium and large cities have significant issues related to transportation and traffic because people are in constant need of quicker and safer mobility

modes. The number of fatalities and injuries on the road have achieved an alarming number. Globally, 1.3 million people die every year and up to 50 million suffer severe injuries. These facts have a direct impact on the economy of nations, leading to costs in the order of about 2% to 5% of the Gross Domestic Product (GDP) in many countries [Bank, 2017]. It is also reported that traffic congestion results in critical economic and environmental costs. In 2011, 498 U.S. urban areas were evaluated regarding the impact of congestion. It was found that about USD 121 billion was wasted on fuel consumption and more than 25 billion kg of CO₂ was emitted. Those values were USD 24 billion and 4,53 billion in 1982, respectively. In 2014, 471 U.S. urban areas were observed, and the costs related to wasted fuel consumption due to congestion reached USD 160 billion [Schrank et al., 2012, 2015].

Over the years, governments and car manufacturers launched initiatives to improve road traffic efficiency, safety and people's mobility. They have been working on various aspects of Intelligent Transportation Systems (ITSs), which aim to improve decision-making by leveraging the availability of information and communication technologies to provide applications and services to boost the transportation systems. Some initiatives are described in [Agency, 2017; Commission, 2017; of Transportation, 2017b; Council, 2017; of Transportation, 2017a; Board, 2017; Thyssenkrupp, 2017; ClickutilityTeam, 2017]. Mike [2013] discussed the considerable growth of on-board informatics inside vehicles. Currently, each vehicle has an average of 60-100 embedded sensors, and these numbers can go up to as much as 200 sensors per vehicle in 2020. Moreover, according to Machina Research [Machina Research and Telefonica, 2013], in 2020, about 90% of new cars will feature an Internet-integrated, while it was about 10% in 2013.

We also have sensors on road infrastructure such as inductive loop traffic detectors, monitoring cameras, radars, traffic lights, and weather sensors have increased in number and quality (accuracy) in the transportation systems. Besides, the use of media in the transportation scenario has also increased, once these sources may report incidents, traffic conditions, number of fatalities, and road conditions.

Based on these various data related to the transportation systems, a relevant research challenge emerges aiming to answer *how those data can be used to improve*

people's life quality in large cities, especially regarding mobility and traffic?

In this direction, Smart Mobility (SM) plays a crucial role regarding technological solutions to answer that research question. SM aims to integrate ITS considering people's mobility with a focus on green initiatives (e.g., electric vehicles and bikes) and reduced emissions, leading to better access to public transport and integration of different transportation modes. However, the development of new applications and services to ITS depends on the availability of vast amounts of data, despite its current slow availability. In fact, many data sources become a gold coin in the development of new solutions, tools and businesses. Nevertheless, to study and develop solutions to SM, we first need to deeply comprehend the data cycle from the transportation systems. In other words, solutions to improve the current transportation systems depend on advances at each stage of the data cycle. Figure 1.1 shows the data cycle of the transportation system and a short description.

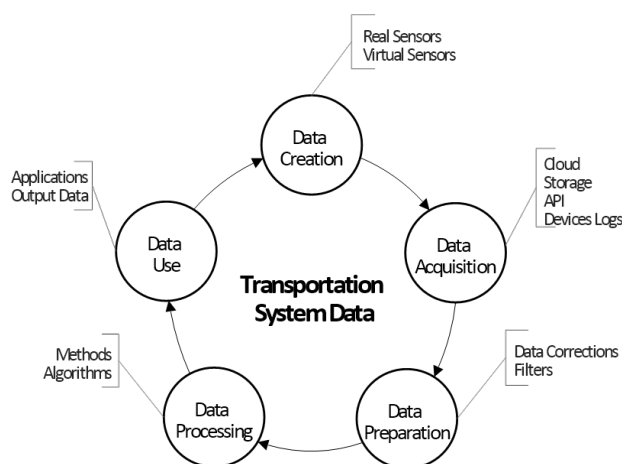


Figure 1.1: The data cycle on the transportation system.

The data cycle begins with *Data Creation*. Data can come from real sensors responsible for measuring the environment or virtual sensors. In this stage, a problem that arises when using real sensor data to monitor and control entities is the data reliability, which includes availability and data quality. A solution to monitor and improve physical sensors, or temporarily replace them, is the use of virtual sensor. This type of sensor may combine data from other sensors, correct or filter failures, apply adequate methods and algorithms considering a given problem

domain, and take the resulting data to applications or input it to a new cycle. *Data Acquisition* represents its availability to the community and its spatiotemporal coverage which constitutes limitations to develop general and broad solutions. Also, there are issues related to the data storage and the data structure, which become relevant in an ITSs, due to the need of big data analysis. The *Data Preparation*, in general, represents the most critical stage of any study in ITS, since it is responsible for establishing the data to develop solutions in a given scenario. The *Data Processing* transforms the treated data into valuable and more informative data to be used in the next stage. *Data Use* is the last stage, which provides the application to users, or outputs the data to start a new data cycle.

We identified challenges and open issues of each data stage. But also, we noticed a lack on both the availability of data and on the data quality. Then, we aim to answer the following question: *"How to deal with the lack of both the availability of data and data quality from the transportation scenario and propose solutions to improve people's life quality in large cities, especially regarding mobility and traffic?"*

Our hypothesis is that *"Through the use of heterogeneous data fusion we can improve the data quality, providing methods and applications to achieve SM"*. In this sense, we focused on two main stages, which are *Data Preparation* and *Data Processing*. These two stages may deal with solutions to improve the data quality of transportation systems. The integration of multiple data sources becomes an essential process to provide consistent, accurate and useful information to applications in ITS. Such a process is called Data Fusion and constitutes a challenging task especially when considering heterogeneous data and their spatiotemporal aspects [Khaleghi et al., 2013b].

1.2 Objectives

The overall goal of this thesis is to provide a set of methods and applications to achieve SM through the use of heterogeneous data fusion. Figure 1.2 depicts the refinement of this goal by showing the design of our fusion process considering an ITS. We consider the concept of a Vehicular Data Space (VDS) as the input data to the whole process. The VDS covers all data related to the ITS environment.

Based on that, all data created or acquired is used as input to feed the fusion stage according to three types of combination. The Intra-Vehicle Data (IVD) only uses the data provided by vehicles. The Extra-Vehicle Data (EVD) focuses on fusing data surrounding vehicles, while the Intra and Extra-Vehicle Data (IEVD) aims to combine both types of data. The output of these three types of data fusion approaches are applications and services to improve current mobility, or they can be used as input data for a new data fusion cycle.

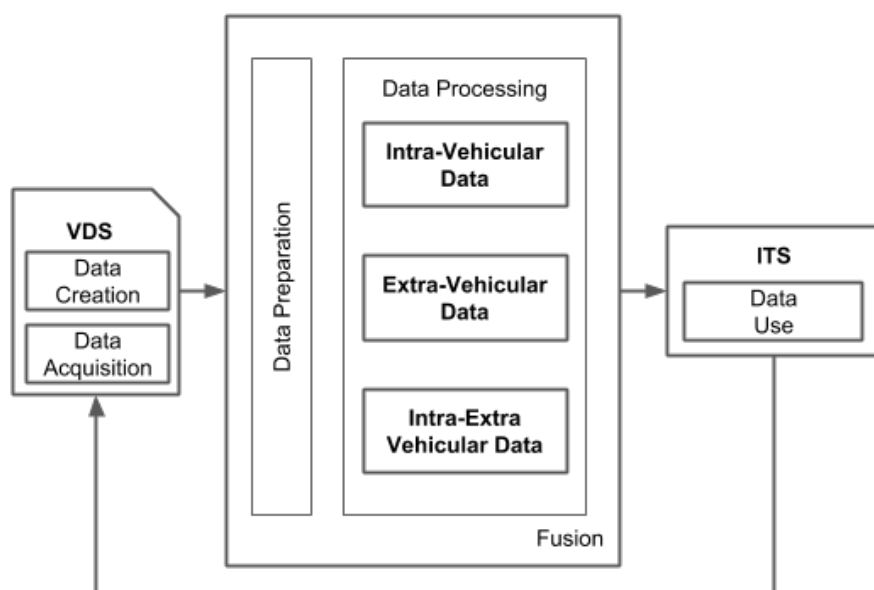


Figure 1.2: Design of fusion on VDS.

The fusion process depends on the data availability and data preparation, which aims to deal with data issues. Nevertheless, the most critical data issue that may affect the development of efficient solutions for ITS is related to data incompleteness. In other words, when combining multiple types of data, there is an increase in the spatiotemporal coverage issues that negatively affect the development of ITS approaches. When all data sources from the VDS, such as vehicles and their surrounding environment, are observed at the same time and space, we can notice that not all of them present the same spatiotemporal coverage. Thus, we argue that new methods to fuse the VDS are required to allow the analysis of the same event from different data perspectives. This allows us to enrich information related to VDS.

1.3 Contributions

This thesis investigates solutions to improve the data quality for transportation systems, thus enabling IVD and EVD fusion to provide the conception of new applications and services in all fields, particularly, to improve overall mobility. Hence, we propose a heterogeneous data fusion platform for SM, aiming to analyze each data type from the VDS, considering the data aspects and its spatiotemporal coverage, in order to improve the current transportation system scenario.

The contributions of this thesis are the results of a literature review and a temporal and spatial data fusion using the same or other data sources available for VDS. In that direction, we use mathematical methods, geostatistics, and machine learning techniques in the following contributions:

- A vast literature review to provide the concept of VDS.
- A methodology to develop applications and services for ITS, specifically SM, based on the transportation system data cycle.
- Intra-Vehicle Data (IVD) Fusion: Techniques to perform Intra-Vehicle Data (IVD) fusion applied to eco-driving methods to reduce fuel consumption, emissions and vehicle maintenance. An extra-authentication factor based on driver identification, and also a virtual gear sensor.
- Extra-Vehicle Data (EVD) Fusion: Techniques to combine the user's viewpoint and road data to enrich the current transportation system data. We propose Road Data Enrichment (RoDE), a framework that fuses heterogeneous data sources to enhance ITS' services, such as routing and event detection.
- Intra and Extra-Vehicle Data (IEVD) Fusion: Techniques to fill the road spatiotemporal data gaps, using vehicular trace and road data, improving road data quality and route suggestion. We propose Traffic Data Enrichment Sensor (TraDES), a low-cost traffic sensor for ITS based on heterogeneous data fusion.

1.4 Outline

In the following, we present the thesis organization and provide a brief summary of each chapter.

Chapter 2 examines the most remarkable studies of the last five years, which describe services and applications for Intelligent Transportation System (ITS)s, however with a focus on the data employed by them. We introduce the concept of Vehicular Data Space (VDS), which is then used to describe the vehicular scenario considering the data perspective. Moreover, we outline a taxonomy, according to the data source; and categorize the applications according to the data used.

Chapter 3 discusses the data fusion aspects of VDS. We highlight several issues in the transportation data that must be treated before the fusion process. Moreover, we conduct an exploratory analysis over real vehicle data to identify data issues (e.g., imperfection, correlation, inconsistencies, among others) found in our data set. We also point out some fundamental aspects concerning ITS, heterogeneous data fusion, challenges and opportunities in this area.

Chapter 4 focuses on Intra-Vehicular Data Fusion and the issues related to data heterogeneity, correlation and characterization. We also present the design of a vehicular virtual sensor that allows the development and evaluation of eco-driving based on a gear virtual sensor. Our methodology gives the driver recommendations of the best gear by considering speed and torque, thus saving fuel and reducing CO₂ emissions. Besides, we design the virtual sensor to identify the driver, treating it as an extra authentication factor to local services and vehicular networks. This virtual sensor is also used to determine a suspicious driver, promoting the discussion on the impacts of these drivers during the data dissemination process in a vehicular network.

Chapter 5 discusses the Extra-Vehicular Data Fusion. We propose RoDE, a framework that fuses heterogeneous data sources to enhance ITS' services, such as routing and event detection. We present RoDE through two services: (i) Route service, and (ii) Incident service. For the first one, we present the Twitter MAPS (T-MAPS), a low-cost spatiotemporal model to improve the description of traffic conditions through Location-Based Social Media (LBSM) data. As a case study, we explain how T-MAPS is able to enhance routing and trajectory description

using tweets. We compare T-MAPS routes with Google maps routes. Moreover, we present three route description services over T-MAPS: Route Sentiment (RS), Route Information (RI), and Area' Tags (AT) aiming to enhance the route information. For the second service, we present the Twitter Incident (T-Incident), a low-cost learning-based road incident detection and enrichment approach built using heterogeneous data fusion. We use a learning-based model to identify patterns on social media data which may describe a class of events, aiming to detect its types. The methodology results to detect events achieved scores above 90% in *F1 score*, *Recall* and *Precision* metrics, thus allowing incident detection and its description as RoDE' services. Besides, the event description service allows us to better understand the LBSM user's viewpoint, regarding the transit events and points of interest.

Chapter 6 presents the proposal of Intra-Vehicular and Extra-Vehicular data fusion to provide novel applications and services to improve smart mobility. We propose TraDES, a low-cost traffic sensor for ITS based on heterogeneous data fusion. TraDES aims at fusing data from vehicular traces with road traffic data to enrich current spatiotemporal traffic data. In that direction, we propose a robust methodology to spatially and temporally group these different data sources, producing a vehicular trace with its respective traffic conditions, which is then given as input to a learning-based model based on Artificial Neural Networks (ANN). Hence, TraDES is an enriched traffic sensor that is able to sense (detect) traffic conditions using a scalable and low-cost approach and increase the spatiotemporal traffic data coverage.

Chapter 7 presents the conclusions and future work of this thesis, and also the publications obtained during the doctorate.

Chapter 2

Vehicular Data Space

Given the importance of sensors to a vehicle's operation, new vehicular models embed many high-quality sensors [Faezipour et al., 2012] to get more reliable and diverse information about themselves. In that way, Advanced Driver Assistant Systems (ADAS) offer a means to enhance, among other things, the driver's safety and comfort [Bengler et al., 2014]. In the last years, the development of vehicular sensors had a significant increase. As a consequence, the number of connecting cables inside the vehicle has also increased, resulting in an additional 50 kg to the vehicle mass, besides the increase of the final vehicle cost, and the difficulty of installing and maintaining all systems working properly [Qu et al., 2010]. For that reason, an Intra-Vehicle Sensor Network (IVSN)¹ may need to rely on wireless communication for its operation. Thus, the Intra-Vehicle Wireless Sensor Network (IVWSN) is a research topic in the field of vehicular sensor communication.

An important issue here is how to have a wireless connection among sensors and the Engine Control Unit (ECU). This sensor network usually has some particular characteristics, such sensors are stationary and are only one hop away to the ECU, and have no energy constraint. In spite of these characteristics, there are some challenges related to the efficient use of wireless channels, such as latency, reliability, security and interference issues in a dense urban scenario. In particular, we are interested in challenges and opportunities related to the whole data space

¹Also mentioned as Intra-Vehicular Communication (IVC) and Intra-Vehicular Network (IVN)

that influences or is influenced by vehicles. How those sensors communicate, wired or wireless, to provide useful data is not the focus of our study. For more details, see [Tonguz et al., 2006, 2007; Ahmed et al., 2007; Tsai et al., 2007; de Francisco et al., 2009; Lu et al., 2014a; Reis et al., 2017], and [Tuohy et al., 2015] for a broad comprehension of Intra-Vehicle Networks.

The development of new applications and services for ITS depends on the availability of different data sources, what it is not the current case. In fact, many data sources may play a central role in the development of new solutions, tools and businesses. In the literature, there are some studies describing the main features and properties of Intelligent Transportation System (ITS) applications [Qu et al., 2010; Engelbrecht et al., 2015; Abdelhamid et al., 2015]. In this chapter, we survey recent proposals describing services and applications for ITS, but with a focus on the data employed by them. We introduce the concept of Vehicular Data Space (VDS), which is then used to describe the vehicular scenario from the perspective of data. Moreover, we outline a taxonomy and applications based on that concept, and we end with the challenges and open issues based on the data cycle on the VDS.

The rest of the chapter is organized as follows. In Section 2.1, we introduce the VDS concept and discuss the methodology used to identify relevant studies in the literature. In Section 2.2, we present an example of the VDS environment and its respective entities and data. In Section 2.3, we present a taxonomy of the vehicular data space from the perspective of data sources, and analyze existing solutions. In Section 2.4, we discuss some potential applications in VDS, focusing on the data point of view. Finally, in Section 2.5, we conclude the survey with some possible future directions.

2.1 Vehicular Data Space

Given the importance of data to ITS, this work looks at the ITS field using the perspective of data. For that, we categorize existing literature research according to the data sources employed by them. The aim here is to consider different data aspects, such as availability, spatiotemporal correlations, acquisition challenges, frequent used data types and their applicability, and heterogeneous data fusion

issues. Therefore, our goal is to present the vast ITS field according to the vehicular data context.

For that, we introduce the concept of a VDS, which covers the various aspects regarding data to provide a descriptive view of the transportation scenario, however, differently from the approach presented in [Qu et al., 2010]. Here, we assume that a VDS encompasses both the data sources and the data produced by them. Hence, we conduct a literature review focusing on the concepts of Vehicular Data Source (VDSsource), Section 2.3, and Vehicular Data (VD), Section 2.4. Besides, we created the data cycle for VDS, aiming to show stages which may serve as a guideline to propose new solutions to the ITS scenario and allow a whole comprehension of the VDS. Figure 2.1 summarizes the subsets of VDS and the five stages of the proposed data cycle, which span from the data creation to their use. Each subset in the VDS can be briefly described as follows:

- Vehicular Data Source (VDSsource)
 - *Data Creation*: The process of sensing environment variables through real or virtual sensors.

- Vehicular Data (VD)
 - *Data Acquisition*: The process of making these data available through device logs, storage, cloud or even APIs.
 - *Data Preparation*: The filters or corrections applied to the data so it can be processed.
 - *Data Processing*: The methods and algorithms applied to the data according to its properties and desired use.
 - *Data Use*: The proposed use (e.g., applications) which may power other data cycles or applications.

Based on that, the VDSsource deals with the *Data Creation*, whereas the VD covers the rest of the data cycle, i.e., *Data Acquisition*, *Data Preparation*, *Data Processing* and *Data Use*, allowing the developing of services and applications for ITSs. As mentioned, it is out of our scope to provide a deep discussion about each

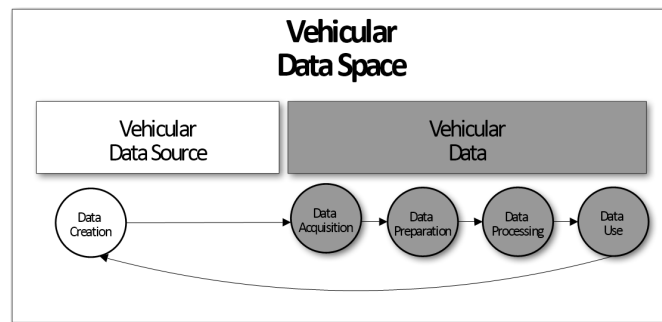


Figure 2.1: The big picture of Vehicular Data Space and its respective state of data cycle in the VDS.

of these steps in Section 2.4 (application section), except the *Data Use*, which discusses how the data may be used, disregarding its acquisition and processing.

2.2 Entities of the Vehicular Data Space

Vehicular *Ad-hoc* Networks (VANETs) are a derivation of Mobile *Ad-hoc* Networks (MANETs), in which vehicles are equipped with computing, sensing and communication capabilities [Laberteaux and P., 2008; Hartenstein and Laberteaux, 2009; Karagiannis et al., 2011]. Moreover, VANET possess characteristics that are specific to the vehicular environment, such as vehicles are expected to move in well-defined patterns and concentrate in high-density urban regions, and vehicles have a more predictable mobility model. Built on top of VANET, the Vehicular Sensor Network (VSN) [Lee and Gerla, 2010; Jeong and Oh, 2016] is a powerful sensing platform that provides the capability for collecting, computing and sharing sensor data. A vehicle contains various types of highly reliable sensors and almost eliminates the energy constraints of traditional MANETs, due to its rechargeable battery. Moreover, vehicles can leverage the communication capabilities already deployed in urban areas, such as cellular and wireless networks.

The perception of the surrounding environment is paramount for provisioning many services in VANET. Physical sensors play an important role in control systems, as they provide data on operational states and malfunctions of monitored entities. Vehicular control systems are among those that depend on sensor data to actuate on their components to provide a safe and enjoyable driving experience.

Traffic control systems also depend on sensor data to measure the vehicle flow, traffic lights coordination, and delays. Weather monitoring systems rely on sensors for predicting storms. Moreover, Participatory Sensor Networks (PSN) also play a relevant role in monitoring and control systems in a wide scope. News and Social Media can act as a virtual sensor wherever there is a lack of physical sensors. For instance, an accident report can be filled out by Social Media users in areas with no road sensors infrastructure. Moreover, people's feelings who pass near an incident cannot be perceived by physical sensors.

Many studies in VANET focus on the communication issues for ITS and their associated challenges. For instance, assume an accident between two vehicles. Most studies are interested in knowing how this event can be disseminated through a road to alert other drivers and the road administrators, i.e., how to efficiently broadcast the emergency event. On the other hand, here we focus on the data. In other words, both vehicles are constantly producing data. Therefore, how can such data be used to improve an accident avoidance system? Furthermore, how can the road historical data be analyzed to reduce the risks of an accident?

We consider as a VDS all data used to provide a descriptive view of a vehicular scenario, such as intra-vehicle data, traffic flow data, traffic incidents data, infotainment and others. Notice that the data may be produced by intra-vehicle sensors, smart devices or even social media, for instance. The first step before proposing solutions for ITS is to understand the data and its sources, such as the entities responsible for acquiring and, in some cases, providing data access to the community.

We show an example of the VDS and its respective entities, which produce data in an urban area, in Figure 2.2. Figure 2.2 shows an example of the VDS and its respective entities, which produce data in an urban area. In the following, we describe some of data sources shown in this figure, grouping them in Infrastructure, Transit Authority, Vehicle, Publicity and Media. We highlight that the concept of data in our context may be related to the raw data or also data in a given context, i.e., a piece of information.

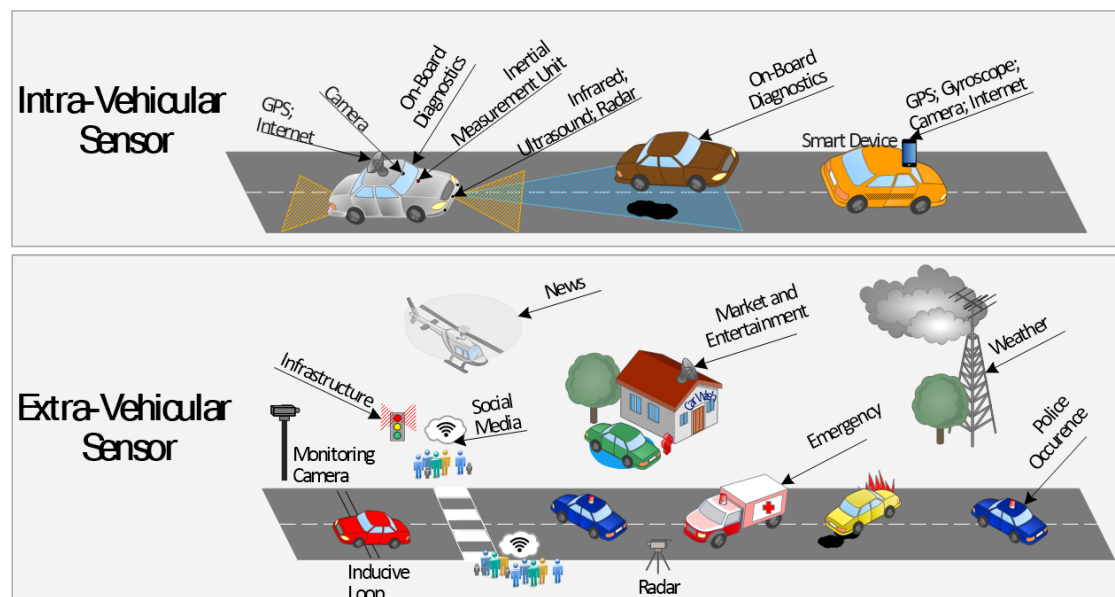


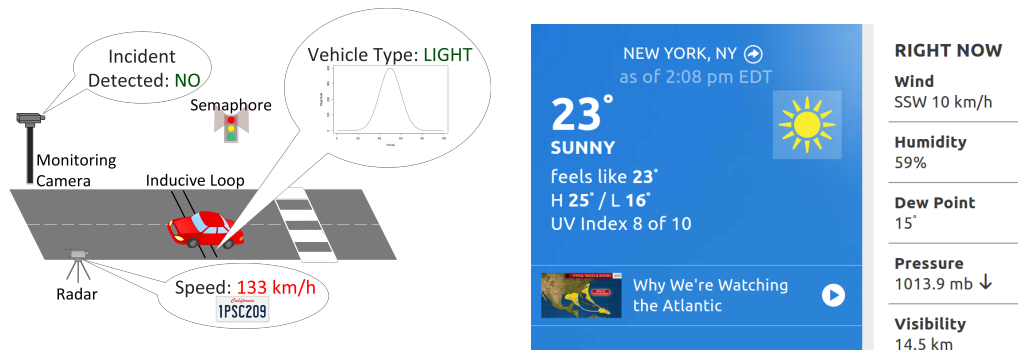
Figure 2.2: Vehicular data space provided in the urban area.

2.2.1 Infrastructure

Infrastructure data address a range of sensors, such as vehicle detection loops, called inductive loop traffic detectors, monitoring cameras, radars, traffic lights, and weather sensors. Inductive loops are based on a wired electromagnetic communication (see the black lines on the roads in Figure 2.2). They are installed on the pavement and can detect a vehicle passing at a certain point and its speed. Inductive loops have also been used to classify types of vehicles, based on their signatures [Jeng and Chu, 2015].

Similarly, however, with a higher deployment cost, monitoring cameras or radars can also be used to detect the speed of a vehicle or its type. Moreover, cameras have also been used to detect and prevent accidents, and to broadcast notifications to authorities. A preventive situation can be illustrated by an animal that crosses a road, and, then, the authorities are promptly notified about it, so they can take actions to avoid future accidents. Cameras can also record the vehicle's license plate when traffic rules are broken (e.g., the red car crossing a red traffic light in Figure 2.2). The combination of inductive loops, traffic lights, cameras and radars produces a virtual sensor that allows traffic agencies to apply

the governing legislation and eventually issuing traffic tickets. Figure 2.3a shows each data source just mentioned.



(a) Data provided by the inductive loop, monitoring cameras and radar.

(b) Data provided by a weather station in New York City [Weather, 2017].

Figure 2.3: Data provided by infrastructure.

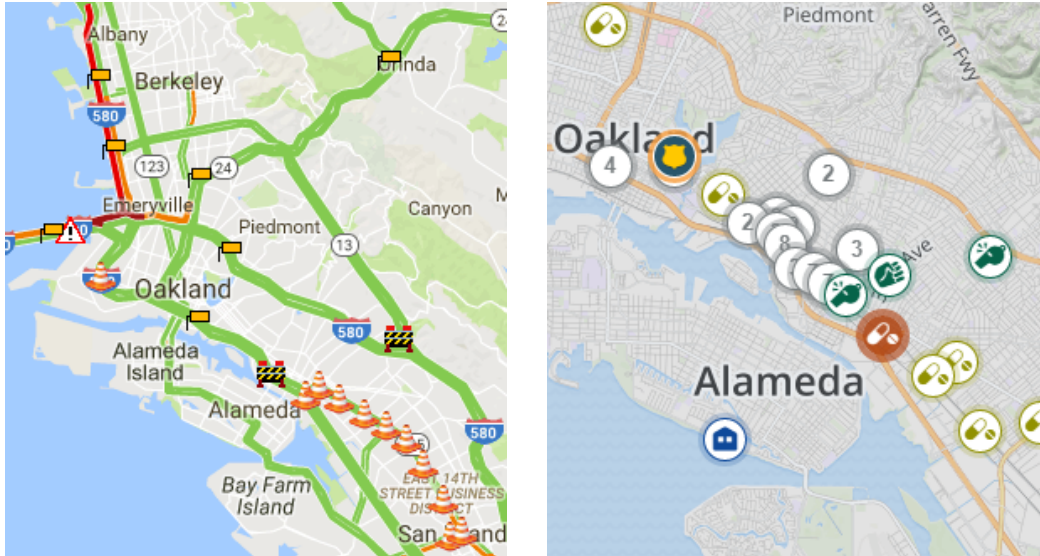
The road infrastructure needs to work together to prevent traffic jams and high traffic flow. For instance, a traffic light can be based on static time intervals, or adapt its behavior according to the perceived traffic conditions. The data traffic may be collected as a result of wired or wireless communication with other traffic lights, inductive loops and radars. Other data provided by the infrastructure are weather stations, which provide, in real time and for a certain area, data about temperature, pressure, wind speed, dew point, humidity, and also prediction data on the chances of precipitation. Figure 2.3b shows an example of a New York weather station².

2.2.2 Transit Authority

Government entities play an essential role in the transportation system since they help decision makers and overall people to better understand the mobility behavior in a city. Most countries possess agencies that provide traffic-related data, such as statistics about traffic jams, accidents, road state, police occurrences, medical occurrences, fatalities, and injuries on the road, and mobility patterns. Such data may be used by different stakeholders to make informed decisions. For instance, in possession of data about fatalities and injuries on a specific road, drivers

²<https://weather.com>

can change their actions and drive more carefully. As an example of government data, Figure 2.4a shows the traffic alerts provided by the U.S. Department of Transportation (DoT) in the state of California, aiming to show blocked roads, incidents, traffic intensity and alerts to road users.



(a) Road conditions, incidents and traffic level provided by the U.S. Department of Transportation

(b) Data provided by the government and available through a Crime Reports platform [Merritt, 2017].

Figure 2.4: Data provided by government entities.

Figure 2.4b shows another type of data provided by Police Departments. Using an online platform, Socrata [Merritt, 2017] makes government data available to citizens. Crime Reports show a variety of crimes, such as disorder, vehicle thefts, property crime, robbery, sexual offense and drugs. Such data allow users to better understand a particular area. Notice that the data provided by these entities may not be the raw data. Some treatment may be introduced to offer a more detailed scenario. Despite this, we still consider them as data.

2.2.3 Vehicle

An important data source in a VANET scenario is the vehicle itself. Vehicles have sensors to collect data about speed, acceleration, movement, luminosity, location, the presence of people or obstacles, external and internal temperatures, and

Table 2.1: Data from a vehicle and additional devices embedded in it.

Vehicular Sensor Data									
<i>From Additional Devices</i>				<i>From Engine Control Unit</i>					
Time	Obstacles Detection	Video Record	Road Condition	Throttle Position	Tire Pressure	Fuel Level	Torque	Engine RPM	Acceleration
Location	3-axis Acceleration	Audio Record	Atmospheric Pressure	Steering Wheel Angle	Battery Voltage	Intake Air Temp	Fuel Flow	Speed	Light
GPS Speed	Altitude	Ambient Air Temp		Fuel Consumption	Gear	Trip Distance	Engine Coolant Temp	CO ₂	Air Conditioner Temp

current structural state, which can provide information to alert the driver about events about the vehicle. Moreover, sensors may be used to control the operation of vehicles. For instance, data provided by the luminosity sensors can control the automatic functioning of the lights, turning them on during the night. Furthermore, proximity sensors can help drivers to keep a safe distance from neighboring vehicles, avoiding collisions. These sensors play an important role in autonomous vehicles. Table 2.1 presents some data that can be acquired directly from the ECU of vehicles or additional devices embedded in vehicles.

Sensors embedded in a vehicle can also be used to detect many events in the surrounding environment during the vehicle's trajectories. Using the On-Board Diagnostic (OBD), data collected from sensors can be used to monitor the traffic and events around the city. For instance, the vehicle's GPS data can support a traffic monitoring service, alerting about traffic jams. In another scenario, combining data from both accelerometer and GPS, it is possible to monitor the presence of holes on the roads.

2.2.4 Publicity

The VDS also contains data provided by market and entertainment companies. These data aim to offer personalized products, services or comfort applications to the drivers. Figure 2.2 shows a simple example of a market on the road, where a Car Wash company tries to sell its services to vehicles that will pass in front of its location, using a Vehicle-to-Infrastructure (V2I) infrastructure. Based on that same idea, a car maintenance company can offer services to the driver since the vehicle sends data about its state and eventual malfunction to the car manufac-

turer.

A variety of applications can be developed to provide entertainment to the passengers of a vehicle, based on information about them and their vehicles. For instance, their smartphones carry a personal user data and applications which become useful through the dashboard display and multimedia kit inside the cars. This allows a better involvement between passengers and the environment around them. There are private companies with initiatives, focusing on connecting costumers with their cars, growing the comfort and the client satisfaction. For instance, the General Motors developed OnStar³, Audi offers Audi Connect⁴, Apple developed CarPlay⁵, Google developed Android Auto⁶, and Toyota and BMW have also an infrastructure for their users, Toyota Touch 2⁷ and BMW ConnectedDrive⁸, respectively.

2.2.5 Media

The growth and popularity of the Internet implied the increase of media in reporting the conditions of transportation. The incidents, traffic conditions, the number of fatalities, road conditions, the events in a given location and so on become the goal of many types of media, i.e., social media, news blogs, newspapers, map navigation and transit insights, radios, and TVs. Constituting, a relevant way to disseminate and provide information to the better comprehension of the transportation system. Even though the data provided by media can be subjective and biased, those data can provide information difficult to obtain with other data sources.

The use of social media is a novel possibility to obtain information about the traffic and road conditions, or report events to other drives. These are particular Location-Based Social Media (LBSM) apps, which enable mobile users to act as mobile sensors, monitoring the environment, weather, urban mobility and traffic

³<https://www.onstar.com/us/en/home.html>

⁴<https://www.audiusa.com/help/audi-connect>

⁵<https://www.apple.com/ios/carplay/>

⁶<https://www.android.com/auto/>

⁷<https://www.toyota-europe.com/world-of-toyota/articles-news-events/2016/toyota-touch-2>

⁸<http://www.bmwusa.com/standard/content/innovations/bmwconnecteddrive/connecteddrive.aspx>

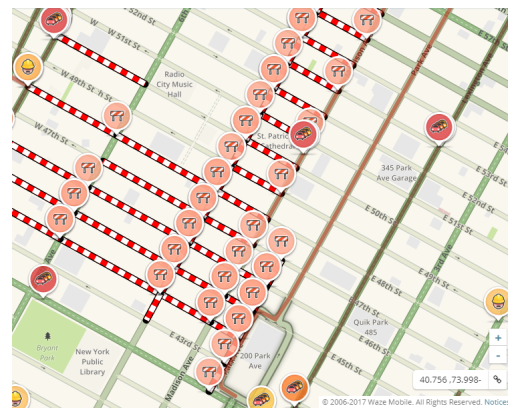
conditions. The main feature of this data type is the real-time information of the sensed events. Typically, users retrieve the accurate data about the traffic conditions. Another important feature is its large coverage, since all users connected to the network can access these data with no restrictions. Figure 2.5 shows examples of types of media used in the VDS in benefit of applications to the ITSs. Figure 2.5a shows textual data provided by reports of the user from the Twitter Platform⁹, whereas Figure 2.5b displays visual data provided by a combination of users' reports of the Waze¹⁰ app, allowing other users to have a better overview of the traffic conditions.

Bridge closed in #Sayreville on Rt-35 SB between 1st St and CR-689, stopped traffic back to Lorraine Ave #traffic bit.ly/11xKLzq

11:13 AM - 9 Sep 2017 from Sayreville, NJ



(a) Data provided by the LBSM Twitter [Twitter, 2006], reporting the traffic occurrence in NY City.



(b) Data provided by Waze map [Waze, 2006] in NY City.

Figure 2.5: Data provided by media.

A different way to obtain data of VDS comes from radio stations created to disseminate information about the road state. For instance, there are radio stations focused on broadcasting information about the road conditions like a road blockade, accident and animals on the road. These pieces of information are obtained from drivers' notifications and road employee observations.

⁹<https://twitter.com>

¹⁰<https://www.waze.com/>

2.3 Taxonomy of Vehicular Data Source

As suggested by previous section, we categorize the Vehicular Data Source (VD-Source) into two main groups named Intra-Vehicular Sensors (Section 2.3.1) and Extra-Vehicular Sensors (Section 2.3.2), as shown in Figure 2.6. Afterwards, we discuss each leaf of the taxonomy tree and present an overview.

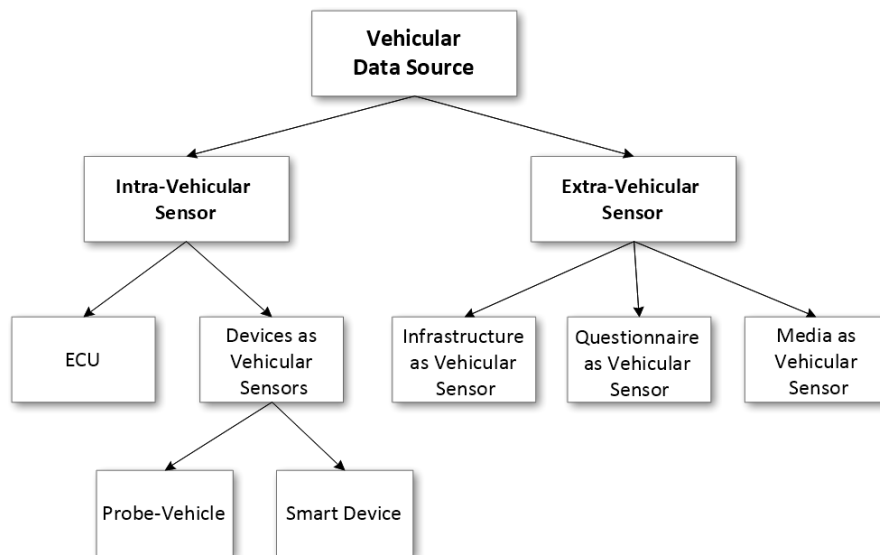


Figure 2.6: Taxonomy of vehicular data space based on the point of view of the source.

2.3.1 Intra-Vehicular Sensor

Intra-Vehicular Sensor (IVS) corresponds to the subset of sensors that describe the main interactions between a vehicle and its driver, passengers or its surrounding environment, from the perspective of the vehicle itself. In other words, IVS represents all sensors embedded in a vehicle or on-board that measure the vehicle state, the drivers' behavior or the environment conditions in its surrounding.

IVS may collect data from the ECU, such as engine load, engine coolant temperature, engine Revolutions Per Minute (RPM), vehicle speed, throttle position, and others. Moreover, IVS may also collect data provided by devices on-board of a vehicle. These devices are classified as Device as Vehicular Sensor (DVS). We

further categorize these devices into *Probe-Vehicle*, where a set of precise sensors are used to monitor a particular event, and *Smart Device*, where devices, such as smartphones, tablets and other pieces of hardware act as data sources. In the following, we group the proposals according to the type of IVS they employ to collect data.

2.3.1.1 Engine Control Unit

Given the importance of sensors to a vehicle's operation, new models embed many high-quality sensors to get more reliable and diverse information about themselves. All data produced by sensors in a vehicle are delivered to its ECU through an internal network, named Controlled Area Network (CAN), which is accessible through the vehicle's OBD port. A useful analogy is to suppose that the OBD is the language that we use to speak about a vehicle's state, as informed by the ECU, using a communication device (CAN).

The OBD system was first introduced to regulate emissions. However, it is now used for a variety of applications. There are different signaling protocols to transmit internal sensor data to external devices through a universal port. Such a universal port is present in all cars produced since 1996 in the U.S. and Europe. There are Parameter IDs (PIDs) to access sensor information using the OBD, which identify individual sensors. Some PIDs are defined by regulatory entities and are publicly accessible. However, manufacturers may include other sensors' data under specific and undisclosed PIDs.

The 52 North Initiative for Geospatial Open Source Software [Bröring et al., 2015] proposed a platform named EnviroCar for collecting geographic data and vehicles' sensors. The EnviroCar is an open platform for Citizen Science projects, which aims to provide sustainable mobility, traffic planning and share the findings from the industry when collecting and analyzing car data. Using an OBD adapter into a car, they collected a variety of sensor data and uploaded it to the Web. The system consists of the EnviroCar app and the EnviroCar server. Bröring et al. [2015] described the spatiotemporal RESTful Web Service interface and the designed data model. Since 2015, there are over 500,000 measurement data points collected and these numbers are continuously growing. Reininger et al.

[2015] described a prototype to provide vehicular data access through a website. Using an OBD port and a smartphone, they provided data, such as speed, RPM, fuel consumption, coordinates, and altitude, for later post-processing and analysis. They also described a sandboxing mechanism that prevents malicious attacks from other programs on the smartphone.

Ly et al. [2013] showed the potential of using inertial sensors to distinguish drivers. They concluded that the acceleration feature does not play a significant role in such process, contrarily to the braking and turning features. As an experimental test-bed, they employed a LISA-X (probe-vehicle) to acquire all their data. This experimental vehicle was outfitted with a variety of sensors and vision system. They used signals from a CAN, such as an engine speed, brake pressure, acceleration, pedal pressure, vehicle speed and angular rotation to recognize the vehicle maneuvers represented by three types of events: braking, acceleration, and turning. D'Agostino et al. [2015] proposed a classification method for identifying driving events using short-scale driving patterns. For that, they relied on data provided by CAN and GPS.

Carmona et al. [2015] proposed a novel tool to analyze the driver's behavior and identify aggressive behavior in real time. For that, they relied on a variety of data, such as brake usage frequency, throttle usage, engine RPM, speed, and steering angle. Such data were retrieved using a Raspberry Pi device connected to the CAN through an OBD port. Kumtepe et al. [2016] developed a solution to detect the driver's aggressiveness in a vehicle using visual information and in-vehicle sensor data acquired from the CAN, such as vehicle speed and engine rotation (RPM). They could detect aggressive driving behavior with a success rate of over 93%.

Johnson and Trivedi [2011] showed that sensors available on smartphones can detect movement with a similar quality to a vehicle CAN bus, allowing the recognition and recording of driver's actions. However, Paefgen et al. [2012] showed that such quality depends on the smartphone positions and the type of event being identified. AbuAli [2015] collected data from vehicular sensors using an OBD port to detect the driver's behavior, road artifacts and accidents. To address these issues, it was used the vehicle speed, throttle position, RPM and coordinates to track the vehicle's location. That work showed that the proposed

system can detect road artifacts with a success rate of about 84%.

Zhang et al. [2016] developed a driver's identification model using sensors available both on mobile phones and vehicles, in which data was collected through an OBD port. They evaluated three vehicles in two different environments, a controlled and a naturalistic. Considering only the vehicular sensors, such as acceleration pedal position D, throttle position manifold, absolute throttle position B, relative throttle position, acceleration pedal position E, engine RPM and torque, the classification model obtained a 30.36% accuracy in the controlled environment with 14 drivers whereas in the naturalistic environment with two drivers per vehicle it obtained an 85.83% accuracy. Satzoda and Trivedi [2015] proposed techniques to extract semantic information from raw data provided by vehicles in order to minimize the effort needed for data reduction in Naturalistic Data Studies (NDS). They applied fusion techniques to data from a forward-looking camera, vehicle's speed from a CAN bus, and Inertial Measurement Unit (IMU) and GPS as well. As result, they extracted a set of 23 pieces of semantic information about the location and position of the vehicle on the lane, its speed, the traffic density and the road curvature.

Corcoba Magaña and Muñoz Organero [2016] proposed a solution to reduce the impact of traffic events on fuel consumption. For that, they first developed a system to detect traffic incidents based on the rolling resistance coefficient, the road slope angle and the vehicles speeds. Next, they found an optimal deceleration by anticipating traffic incidents, improving fuel consumption by up to 13.47%. Through an OBD port, they obtained the vehicle speed, acceleration, engine speed and the fuel consumption. Meseguer et al. [2013] developed a smartphone app aiming to characterize the road type as well as the aggressiveness of each driver. For this purpose, they relied on data, such as speed, acceleration, and RPM acquired from the CAN. As result, they achieved an accuracy of 98% when attempting to characterize road types and 77% when characterizing the driving style. Similarly, Hong et al. [2014] developed a platform to model aggressive driving styles based on data from smart devices and ECU. From a smartphone, they used GPS location and 3-axis acceleration. From the IMU, they employed the number of turns and acceleration, whereas from the vehicle they used the speed, engine RPM and throttle position. In addition, they employed the Manchester Driving Behavior

Questionnaire (DBQ) to complement the characterization of the driving style. As a result, using all three data sources, their prediction achieved 90.5% accuracy, while the questionnaire data achieved 81%.

Hallac et al. [2016] developed a method for predicting the identity of drivers based on in-vehicle sensor data collected from a CAN. In particular, they used the steering wheel angle, steering wheel velocity, vehicle speed, brake pedal position and gas pedal position. The results achieved an accuracy of about 76.9% for a two-driver classification and 50.1% for a five-driver classification. Martinez et al. [2016] proposed a non-intrusive method for identifying impostor drivers. They relied on a dataset Abut et al. [2007] that allowed access to a variety of sensor data. However, a reduced set of variables from the CAN was used, such as RPM, brake pedal and throttle position. As result, they achieved an identification rate greater than 80% for every evaluated group category.

Riener and Reder [2014] conducted a study aiming to show that traffic safety and efficiency improve when competent drivers support the not so competent ones by sharing the road and driving data. The data acquisition was made using the OpenXC Platform [OpenXC, 2012] and a smartphone. They used the steering wheel angle, torque, RPM, vehicle speed, throttle position, fuel consumption, gear position, GPS and 3-axis acceleration. They developed a social driving app that provides recommendations about how to drive on a given track based on experiences shared by other drivers. Rettore et al. [2018a] explored the driver's identification as an extra authentication factor to local services and vehicular networks. In this respect, they developed a virtual sensor to determine the driver's identity (legitimate or suspect), with a precision above 98%, using embedded sensor data such as vehicle speed, fuel flow, gear, engine load, throttle position, emissions and RPM.

We also developed a virtual gear sensor for manual transmission cars, which allows to relate each gear with the fuel consumption. They proposed a methodology to recommend the best gears according to current speed and torque. Using such methodology, they were able to reduce the fuel consumption and the CO₂ emissions by approximately 29% and 21%, respectively. They collected data from vehicle sensors, such as engine load, engine RPM, fuel flow, throttle position, trip distance and CO₂ through an OBD port. Ruty et al. [2013] conducted a study

to show the impact of eco-driving training in a municipal fleet. They used the CarChip [CarChip, 2013] technology to acquire data from the CAN and evaluate their proposal. The results showed an average decrease of engine idling between 4% and 10%, and an average reduction of emissions of 1.7 kg of CO₂ per vehicle per day. One year later, Ruttly et al. [2014a] assessed the value of vehicle monitoring technology (VMT) and eco-driver training to reduce emissions and fuel. They showed the results of eco-driving training in a fleet of vehicles at the ski resort operation in Ontario, Canada. The fleet reduced 14% of their average daily speed, 55% of abrupt deceleration, 44% of hard accelerations, and 2% of idling time. Finally, they achieved a decrease of 8% in fuel costs and CO₂ emissions.

Similarly, Ayyildiz et al. [2017] developed an advanced telematics platform to compare the driving style before and after eco-driving training. They acquired data from an OBD port, such as vehicle speed, fuel consumption, emissions and GPS location using a smartphone. The study presented a reduction of 5.5% in fuel consumption for heavy vehicles, while light vehicles did not show significant variations. Brace et al. [2013] proposed an onboard Driver Assistant Systems (DAS), which encourages to improve the driver's driving style. Specifically, the system aims to decrease fuel consumption by reducing the rates of acceleration and early gear changes. For that, they employed data from the vehicle ECU. The used data include vehicle speed, throttle position, engine speed, engine load, engine fueling demand and engine coolant temperature for a total of 39,300 km of collected trip data. They showed fuel savings of up to 12% and an average fuel savings of about 7.6%. Zhao et al. [2016] proposed and evaluated the Dynamic Traffic Signal Timing Optimization Strategy (DTSTOS), aiming to reduce the total fuel consumption and traffic delays in a road intersection. Using the VISSIM traffic simulator [Group, 1992], they obtained data, such as vehicle speed and fuel consumption.

Araújo et al. [2012] proposed a smartphone app to help drivers to change their behavior and, consequently, reduce the fuel consumption. For that, they used the vehicle state data acquired from the CAN bus, through an OBD and the smartphone sensors. They relied on data, such as vehicle speed, acceleration, altitude, GPS, throttle position, instant fuel consumption and the engine rotations. Andrieu and Pierre [2012] developed an efficient Ecological Driving Assistance

System (EDAS) aiming to detect eco-driving behavior and provide drivers with recommendations to help them to reduce the fuel consumption and preserve their safety. They used the CAN and OBD to monitor driving parameters, for instance, vehicle speed, RPM, fuel, brake pedal and throttle position. They showed that it is possible to reduce fuel consumption just by following simple rules of eco-driving. After applying those rules, the average fuel consumption, the speed, and the time spent above the legal speed limit reduced approximately 12.5%, 5.8% and 30%, respectively.

Paefgen [2013] conducted a study aiming to determine the risk of an accident according to collected vehicular sensory data. Focusing on the automobile insurance market and aiming to introduce adaptive insurance tariffs, known as Pay-As-You-Drive (PAYD), the author used a dataset of location trajectories and vehicle's speed data from an OBD port to develop an algorithm to reconstruct trajectories when GPS data were missing. The result was a business model for insurance telematics offerings.

2.3.1.2 Probe-Vehicle

A Probe-Vehicle is a vehicle specifically designed for collecting traffic data, road data, driver data and other types of data in real-time. Its main feature is the high quality of sensors embedded in it. For that reason, many public and private initiatives use that kind of vehicle to measure the quality of roads, weather and driver's behavior. In the following, we analyze studies that employed probe-vehicles to achieve their goal.

Mednis et al. [2012] designed an embedded device (CarMote) that focus on monitoring road surface and weather. They used a microphone, accelerometer, temperature and humidity sensors to create a detailed map of the road quality and meteorology. Ly et al. [2013] collected sensor data from the front side radar, front/rear camera, lateral (Left/Right) and longitudinal (Forward/Backward) acceleration and Yaw Angular Velocity sensors to describe three types of events: braking, acceleration and turning. Satzoda and Trivedi [2015] associated inertial data from the IMU, GPS and camera with the vehicle speed obtained from its CAN bus. Beyond the in-vehicle data used by D'Agostino et al. [2015], they also

used a camera, aiming to record the trips and label the main events while en-route.

Guo and Fang [2013] conducted a study aiming to identify features associated with dangerous driving. Using demographic, personality and driving characteristic data, they predicted who the high-risk drivers are. The authors used the first large-scale study conducted in the United States in 2006, the 100-Car Naturalistic Driving Study (NDS), to develop their methodology and application. The vehicles were instrumented with a set of sensors, such as five camera views around the vehicle, GPS, speedometer, three-dimension accelerometer, radar, and others. The data were collected continuously for 12 months with approximately 43,000 hours and 2 million vehicle miles. The results associated the driver's age, personality and critical incident rate with the risk of crashes and near-crash events. They also showed that approximately 6% of drivers are high-risk drivers, 12% are moderate-risk while 84% are low-risk.

Elhenawy et al. [2015] introduced a new predictor for driver's aggressiveness and demonstrated that this measure enhances the modeling of driver stop/run behavior. They also developed a model that can be used by traffic signal controllers to predict the driver's stop/run decisions. The vehicles were equipped with a Differential Global Positioning System (DGPS) unit, a longitudinal accelerometer, acceleration and brake pedal position, and, in some cases, cameras as well. Carmona et al. [2015] also used in their analysis of the driver's behavior a DGPS, which is composed of a base station that provides improved location accuracy in real-time. They also used an IMU, which has embedded accelerometers and gyroscopes.

Relying on visual information, Kumtepe et al. [2016] developed a method to detect the driver's aggressiveness by detecting lane deviation and collision time. Andrieu and Pierre [2012] employed a GPS, front car camera and a fuel flow meter to develop an efficient EDAS. They also used a specific fuel flow hardware aiming to validate the fuel consumption provided by an OBD port. In this direction, Honda Sensing [Honda, 2015] is an example of a practical solution currently available for their customers. Since 2015, Honda embeds in its cars a suite of safety and driver-assistive technologies such as Collision Mitigation Braking, Road Departure Mitigation, Adaptive Cruise Control, Lane Keeping Assist, Traffic Sign Recognition and Auto High-Beam Headlights.

2.3.1.3 Smart Device

Similarly to probe-vehicles, a smart device also collects and stores traffic data, road data and driver data in real time, however using a low-cost device to sense the environment around and inside the vehicle. In other words, we consider a smart device as a non-intrusive kind of sensor inside the vehicle and not embedded in it. Consider, for instance, smartphones, tablets or a hardware working as data sources inside a vehicle. In the following, we analyze proposals that rely on smart devices as Vehicular Sensor Data (VSD).

Aloul et al. [2015] presented a smartphone app to detect and report car accidents automatically. They used accelerometer and GPS data to determine the severity of an accident and, if necessary, inform its location to the rescue personnel. Fox et al. [2015] designed a crowdsourcing pothole detection scheme using real-world data collected from a smart device with sensors, such as GPS, vehicle speed, the three-axis acceleration and data from the mobility simulator CarSim [Corporation, 2010]. They simulated an environment with 500 vehicles and were able to detect 99.6% of the potholes. In a real-world scenario, their approach could detect 88.9% of the potholes.

Goncalves et al. [2014] designed a platform to acquire data about the traffic condition and to drive performance using a smartphone GPS. Han et al. [2014] developed the SenSpeed, an accurate vehicle speed estimation system, to address an unavailable GPS signal or inaccurate data in urban environments. The authors relied on smartphone sensors, such as gyroscope and accelerometer to sense turns, stops and crossing irregular road surfaces. The results show that the real-time speed estimation error is 2.1 km/h, while the offline speed estimation error is 1.21 km/h, using the vehicle speed through the OBD as ground truth in their experiments. Ning et al. [2017] conducted a study to detect traffic anomalies based on the analysis of trajectory data in Vehicular Social Networks (VSocN). Furthermore, they introduced a taxonomy for VSocN applications. The VSocN is an integration of social networks and the concept of the Internet of Vehicles (IoVs).

Chu et al. [2014] designed a solution to distinguish driver and passengers based on accelerometer and gyroscope data of a smartphone. The Driver Detection System (DDS) focus on identifying micro-activities that can be discriminated

using a popular and low-cost device. The results show an accuracy of up to 85% to determine who is the driver and the passenger. Aiming to identify the user's driving style, Vaiana et al. [2014] used acceleration data (longitudinal and lateral) from a smartphone GPS. Kaplan et al. [2015] reviewed and categorized techniques found in the literature for detecting driver drowsiness and distraction. They provided insights on techniques used for driver inattention monitoring and the recent solutions that use smart devices, such as smartphones and wearables.

Johnson and Trivedi [2011] developed an inexpensive way to detect and recognize driving events and driving styles based on a smartphone. They created a MIROAD system that uses Dynamic Time Warping (DTW) and a smartphone equipped with a gyroscope, magnetometer, accelerometer, GPS and video recording capability to detect, recognize and record actions without external processing. The results proved that the MIROAD was able to recognize the U-turn 77% of the time. Similarly, however broader, Engelbrecht et al. [2014] used accelerometer and gyroscope of a smartphone to recognize driving maneuvers. They validated the approach with an extra device equipped with a dedicated GPS and IMU. Hong et al. [2014] created a model to identify an aggressive driving style. When using the smartphone and ECU data, they achieved an accuracy of 81%, while using only the smartphone the accuracy was of about 66.7%.

Fazeen et al. [2012] also used smartphone sensors (three-axis accelerometer and GPS) to evaluate a vehicle's condition, such as gear shifts and road conditions (bumps, potholes, rough road, uneven road, and smooth road) and also various driver behavior. Paefgen et al. [2012] conducted a study to evaluate driver behavior based on critical driving events and capture driver variability under real-world conditions. They compared the results of using only a smartphone and its inertial sensors to a commercial sensor unit [Technology, 1999] connected directly to the vehicle's OBD port. Castignani et al. [2015] analyzed the capability of smartphone sensors to identify driving maneuvers and classify them as calm and aggressive. For such purpose, they developed the SenseFleet application. They used GPS and motion sensors from the smart device and also the weather and time of day to give them context information. They showed that SenseFleet can provide accurate detection of driving risks.

Yuan et al. [2016] proposed the AC-Sense, an adaptive and comprehensive

scheme for data acquisition in VSNs aiming to increase the quality of vehicular sensing. They used real taxi GPS trajectories and air quality data from Beijing. They combined these datasets to determine the capacity of taxis to sense the air quality. The results showed that the scheme can increase the sensing efficiency and maintain the data quality. Pan et al. [2013] also used real taxi GPS trajectories to detect and describe traffic anomalies. Wang et al. [2017] used vehicular trajectories with the location, heading and speed information to estimate the urban traffic congestion and detect anomalies on the road.

Bergasa et al. [2014] developed a smartphone app to detect the safety level while driving. The app, DriveSafe, was developed for iPhone and aimed to detect inattentive driving behaviors, alerting the drivers about unsafe behaviors. To achieve that goal, the authors relied on computer vision and pattern recognition techniques and data from the rear camera of the smartphone, microphone, inertial sensors and GPS. They also presented a general architecture of DriveSafe and evaluated its performance in a testbed using data from 12 participants (9 males and 3 females). Each participant carried out two types of tests (aggressive and normal). The tests involved 20 minutes of trips during different days and times. DriveSafe was able to detect an inattentive driver behavior with an overall precision of about 92%. They also compared DriveSafe to the commercial AXA Drive app [AXA, 2013] and obtained better results.

Ma et al. [2017] proposed the DrivingSense, which uses noise and other types of data provided by smartphone sensors to identify dangerous behaviors, such as speeding, irregular driving direction change and abnormal speed control. DrivingSense was able to detect events like driving direction changes and abnormal speed with a precision of 93.95% and 90.54%, respectively. Saiprasert et al. [2017] also proposed algorithms to detect and classify driving events based on smartphone sensors, such as GPS and accelerometer.

Corcoba Magaña and Muñoz Organero [2016] used location and road slope data obtained using a smart device to determine the risk of an accident based on the location of trajectories. Paefgen [2013] focused on the automobile insurance market to introduce an adaptive insurance tariff known as PAYD. Bröring et al. [2015] developed an app (EnviroCar) for Android smartphones to collect the location of vehicles and upload it to the Web.

Zhang et al. [2016] developed a model to classify dangerous drivers using only smartphone sensors like accelerometer, gyroscope and GPS. The classification model obtained an accuracy of about 79.88% in a controlled environment and 80.00% in a naturalistic environment. Araújo et al. [2012] developed an application to assess the driving behavior and reduce the fuel consumption. For that, besides in-vehicle sensors, they also used an accelerometer and GPS from a smartphone to acquire acceleration, altitude and location data.

Some studies [Reininger et al., 2015; Meseguer et al., 2013; Ayyildiz et al., 2017] rely solely on the smartphone GPS to develop an app to help drivers improve their driving behavior. AbuAli [2015] used GPS data to track the vehicle's location and store it on the Web. Rutty et al. [2013] and Rutty et al. [2014a] also used a GPS provided by CarChip(CarChip [2013]. Riener and Reder [2014] developed a social driving app aiming to improve the driving efficiency by providing recommendations about how to drive on a given track. Besides using in-vehicle data, they also relied on smartphone GPS and 3-axis acceleration data. Zuchao Wang et al. [2013] developed a system for visually analyzing urban traffic congestion. They used GPS trajectories and speed data from taxis in Beijing to design a model to extract and derive traffic jam information in a realistic road network. The process consists of an efficient data filtering step based on spatiotemporal aspects, size and network topology to create a graph structure and its visualizations.

2.3.2 Extra-Vehicular Sensor

The Extra-Vehicular Sensor (EVS) concepts of VDS corresponds to the subset of real and virtual sensors that seek to describe the driver's behavior and the environment around the vehicle by a variety of sources individually or fused. In that way, we categorize studies that use Questionnaire as Vehicular Sensor (QVS), Infrastructure as Vehicular Sensor (InfraVS) and Media as Vehicular Sensor (MVS), to provide data such as a descriptive driver's style, traffic behavior, weather conditions, and statistics related to drivers, gender, number of accidents, injuries, fatalities and others. In the following, we analyze each category and the related work.

2.3.2.1 Questionnaire as Vehicular Sensor

QVS can be considered the first way to sense the driver's perception of the road condition, accidents, distractions, behavior, expertise, feelings, gender and social aspects. Despite the high cost of applying a questionnaire, it gives a very detailed information about the context evaluated. There are studies that use a very known questionnaire of the Psychology to evaluate the aforementioned issues.

Driving involves a variety of skills including cognitive aspects such as attention and perception, but also emotional, motivational and social interaction. In that direction, the way in which a person performs this activity is described as the driving style. Moreover, it is well known that the driving style can lead to an inattentive and distracted direction, representing a significant issue to the road safety. There are different ways of understanding the driving style of a person or group. A wide solution adopted by psychologists is a questionnaire. There are diverse measurement instruments designed for this purpose, as the Driving Behavior Questionnaire [Parker et al., 1995], Driving Behavior Inventory [Glendon et al., 1993], Driving Style Questionnaire [French et al., 1993] and Driving Expectancy Questionnaire [Deery and Love, 1996].

Beanland et al. [2013] conducted a study to identify driver distraction and inattention in serious crashes, based on the Australian National Crash In-depth Study (ANCIS). The participation in ANCIS was voluntarily and represents a person who was admitted to a hospital for getting involved in an accident. The authors indicated that the most severe injury accidents involve driver's inattention. Despite the variety of observed inattention and distraction events, most of them are possible to prevent. The development of interventions to the driving style depends on studies about the driving behavior and personality traits. In that direction, Poó and Ledesma [2013] used the Zuckerman-Kuhlman Personality Questionnaire [Zuckerman, 2002] to assess the relationships among driving styles and personality traits, and their variation by gender and age. As result, they showed a more comprehensive understanding of personality traits and driving style relationships. Hong et al. [2014] obtained 81% accuracy in their method to determine the aggressiveness of driving style, using Manchester Driving Behavior Questionnaire (DBQ).

van Huysduynen et al. [2015] validated the different factors of Multidimensional Driving Style Inventory (MDSI) [Taubman-Ben-Ari et al., 2004], aiming to know if the questionnaire can measure driving styles. Also, they grouped the factor analysis in angry driving, anxious driving, dissociative driving, distress-reduction driving and careful driving style. Sagberg et al. [2015] conducted a vast literature review, aiming to understand the multidimensionality and complexity of driving styles. They found evidence that sociocultural factors, gender, age, driving experience, personality, cognitive style, group and organization values, and culture can determinate the driving style. The authors also observed the correlation between self-report instruments and observed behavior methods. Finally, but not limited to, they proposed a framework for predictions about how driving styles are established and modified, creating a base to test future empirical studies.

Truxillo et al. [2016] developed a study to compare the effectiveness of the supervisor training and the use of eco-driving educational materials to reduce the fuel consumption. They collected data through a survey, containing the attitudes, knowledge and behavior of a driver before using eco-driving educational materials. After that, they disseminated the material to those participating organizations and the second and third surveys were sent after two and four months, respectively. As part of their results, they found that both groups increased the eco-driving behavior, suggesting that the support for efficient driving behavior can change the fuel consumption.

2.3.2.2 Infrastructure as Vehicular Sensor

The infrastructure can also tell about the vehicle's state, traffic condition, weather and driver's behavior. The essential difference compared to the IVS way is that the InfraVS also can provide information about the group and not only the vehicle individually. Although, this kind of vehicular sensor shows information at different granularity compared to the IVS. The infrastructure gives an external and global view of the environment, in this case, the transportation view. In the following, we describe approaches that use infrastructure data to develop or evaluate, somehow, the proposed applications.

Aoude et al. [2011] developed algorithms for estimating the driver's behavior

at road intersections. They used a set of devices that provide data for the further analyses, as GPS to record the current time of each vehicle, four radars which identified the vehicles, their speed, range and lateral position, four cameras, and a phase-sniffer to record the traffic light signal phase. The authors introduced two classes of algorithms that can classify drivers as compliant or violating. Finally, their approach was validated using naturalistic intersection data, collected through the U.S. Department of Transportation Cooperative Intersection Collision Avoidance System for Violations (CICAS-V)

Castignani et al. [2015], in contrast to the current solutions, used contextual information, weather condition [Map, 2017], in their application SenseFleet, aiming to better describe the driving behavior. Yuan et al. [2016] used the air quality data in Beijing to create the AC-Sense, an adaptive and comprehensive scheme for data acquisition in VSNs. Wang et al. [2017] proposed a traffic congestion detection based on GPS trajectories, social media and infrastructure data (e.g., weather), and showed that it could affect traffic conditions, leading to complementary traffic information.

Lu et al. [2014a] discussed the challenges and review the state-of-the-art about wireless solutions for vehicle communication among different entities, as vehicle-to-sensor, vehicle-to-vehicle, vehicle-to-Internet and vehicle-to-infrastructure. Using VISSIM traffic simulator [Group, 1992], Zhao et al. [2016] proposed and evaluated the Dynamic Traffic Signal Timing Optimization Strategy (DTSTOS), aiming to reduce the total fuel consumption and traffic delays in a road intersection, based on the vehicle speed, fuel consumption and traffic light timing control.

2.3.2.3 Media as Vehicular Sensor

Nowadays, with the growing and popularity of the Internet, the use of media to report the transportation conditions has increased. Thus, issues as incidents, traffic conditions, fatalities, road condition and events in a given location become the goal of different media platforms. We consider MVS as any kind of media (e.g., social media, blogs, news, map tools with transit insights, and government reports) that disseminate information to better contribute to transportation comprehen-

sion. The highlight is the social media data with the potential to be used as a real-time traffic data source. In the following, we describe approaches that use some sort of media data to develop or evaluate the proposed applications.

Pan et al. [2013] proposed a method to detect and describe traffic anomalies based on GPS from vehicles' trajectories and social media data. The system provides real-time alerts when anomalies are detected, including the associated features and an event description based on social media. They used a GPS trajectory dataset of taxis to detect anomalies and the Twitter to provide details of these events. As result, the system detected 86.7% of the incidents reported to the transportation authority, whereas the baseline reported only 46.7%. Santos et al. [2018] argued that LBSM feeds may offer a new layer to improve traffic and transit comprehension. They presented the Twitter MAPS (T-MAPS), a low-cost spatiotemporal model to improve the description of traffic conditions through tweets. The authors developed three route description services based on natural language analyses, aiming to enhance the route information.

Gu et al. [2016] explored the posts from the Twitter platform to extract traffic incident information, which is a low-cost solution compared to existing data sources. In that way, the authors developed a methodology to data acquisition, processing and filtering. They validated the Twitter-based incidents using data from RCRS (Road Condition Report System) incident, 911 Call For Service (CFS) incident, and HERE travel time (a part of the National Performance Management Research Data Set). That study pointed out the significance of traffic incident reported by Influential Users (IU) and individual users, frequency of reports on weekends and weekdays, and also during the day, and the volume of information from the center of a city and outside it. As conclusion, they demonstrated the potential of social media data to enrich the incident reporting sources.

In the same way, but using different social media as a data source, Septiana et al. [2016] used text mining system about RSS feed Facebook E100 aiming to categorize road conditions into six types: floods, traffic jams, congested roads, road damage, accidents and landslides. They showed an accuracy of 92% in the road condition monitoring. Shekhar et al. [2016] focused on the vehicular traffic monitoring using more than one social media, instead of traditional traffic sensors and satellite information which can be quite expensive. Using a Natural Language

Processing (NLP) technique, they examined Twitter and Facebook posts to address traffic problems at a specific location and time interval. Besides, they looked for the causes of recurrent traffic congestion, and noticed that the obtained results were consistent when compared to the HERE Driver+, since more information was added to the context analysis.

Wang et al. [2017] proposed a framework to integrate GPS trajectories data and social media data, aiming to compute urban traffic congestion more precisely. Using vehicular trajectories with location, heading and speed, social events from Twitter, road features, Point of Interest (POI), and weather information, they estimated the urban traffic congestion and also detected anomalies on the road. Sinha et al. [2017] discussed the management of urban infrastructure based on insights from public data, which was used to categorize and visualize the urban public transportation issues. Their holistic framework considered the public transportation agency data, social media as Twitter and Facebook posts, and web portals. Their goal was to help governments and common citizens to have a whole visualization and understanding of transportation in a city. Kurkcü et al. [2017] proposed to fuse data from the Transportation Operations Coordinating Committee (TRANSCOM) and Twitter posts to allow real-time, inexpensive and geographical coverage. Using Twitter and Sina Weibo, Lau [2017] presented an approach to extract and analyze traffic information to enhance ITSs.

2.3.3 Considerations

As previously mentioned, in this section, we discussed the studies considering the Vehicular Data Space. Table 2.2 summarizes recent proposals and their respective categories based on our taxonomy.

This area provides some initial and exciting results that can lead to new research challenges, when considering the data aspects and their applicability. It is interesting to observe that there are studies in Vehicular Data Source (VDS) that considered a different number of data sources in their proposals. In particular, for one data source we have the following: Engine Control Unit; probe-vehicles; smart devices; infrastructure; questionnaire, and some sort of media. Considering the intersections of data sources, there are studies that used simultaneously two

Table 2.2: Summarizing of data source in vehicular data space taxonomy.

Papers	Vehicular Data Space: A Source Point of View				
	Intra-Vehicular Sensor			Extra-Vehicular Sensor	
	ECU	Probe-Vehicle	Smart Device	Infrastructure	Questionnaire Media
[Hallac et al., 2016; Martinez et al., 2016]					
[Rettore et al., 2017, 2018a]	✓				
[Brace et al., 2013]					
[Mednis et al., 2012; Guo and Fang, 2013]					
[Elhenawy et al., 2015]		✓			
[Zuchao Wang et al., 2013; Fazeen et al., 2012]					
[Goncalves et al., 2014; Engelbrecht et al., 2014]					
[Chu et al., 2014; Vaiana et al., 2014]					
[Han et al., 2014; Bergasa et al., 2014]			✓		
[Aloul et al., 2015; Fox et al., 2015]					
[Kaplan et al., 2015; Ma et al., 2017]					
[Ning et al., 2017; Saiprasert et al., 2017]					
[Ly et al., 2013; Satzoda and Trivedi, 2015]					
[Andrieu and Pierre, 2012; D'Agostino et al., 2015]	✓	✓			
[Carmona et al., 2015; Kuntepe et al., 2016]					
[Johnson and Trivedi, 2011; Araújo et al., 2012]					
[Paefgen et al., 2012; Meseguer et al., 2013]					
[Paefgen, 2013; Riener and Reder, 2014]					
[Bröring et al., 2015; Reiningger et al., 2015]	✓		✓		
[Rutty et al., 2013, 2014a]					
[AbuAli, 2015; Zhang et al., 2016]					
[Corcoba Magaña and Muñoz Organero, 2016; Ayyildiz et al., 2017]				✓	
[Aoude et al., 2011]					
[Poó and Ledesma, 2013; Beanland et al., 2013]					
[van Huysduynen et al., 2015; Sagberg et al., 2015]					✓
[Truxillo et al., 2016]					
[Hong et al., 2014]	✓		✓		✓
[Castignani et al., 2015; Yuan et al., 2016]			✓	✓	
[Lu et al., 2014a]	✓		✓	✓	
[Zhao et al., 2016]	✓		✓	✓	
[Wang et al., 2017]			✓	✓	✓
[Rettore et al., 2019]				✓	✓
[Gu et al., 2016; Septiana et al., 2016]					
[Shekhar et al., 2016; Sinha et al., 2017]					
[Kurkcü et al., 2017; Lau, 2017]					✓
[Santos et al., 2018]					

data sources: Engine Control Unit and probe-vehicles; Engine Control Unit and smart devices; Engine Control Unit and infrastructure; and smart devices and infrastructure. For three data sources, we have: Engine Control Unit, probe-vehicles and questionnaires; Engine Control Unit, smart devices and infrastructure; and smart devices, infrastructure and media.

Additionally, we can quantify the use of each data source in the studies above. Figure 2.7a shows the percentage of the use of each vehicular data source. Smart Device (typically smartphones) and ECU represent approximately two-thirds of all data sources employed in the development of applications and methods for ITSs. Smartphones are being designed with more and more sensors capable of sensing different physical variables, which explain their large use as a data source. An

ECU also allows to sense the environment with high-quality sensors and assess the driver's behavior.

Next comes the Probe-Vehicle data source. In this case, only active research groups and companies use this data source due to its high cost to equip the vehicle and design solutions based on the embedded technologies.

The three least used data sources are Media, Questionnaire and Infrastructure. The use of media as a data source to the ITS has increased in the last years, and, probably, we can expect a stronger presence in the future. Media has the power to overcome the limitations of the data coverage provided by all other data sources mentioned in this study. Moreover, media can also offer the transportation view through the lens of users, companies and governments. Questionnaires report the behavior of a group and depend on the sample, and, thus, cannot be generalized. We noticed that the investigations about ITS do not use too much this data source such as media and its variations. Finally, but not less important, the infrastructure has taken its initial steps to be a data source to the VDS. The reasons are the low incentive, security and privacy issues to make the data available to the community.

While these issues keep untreated, we have to live with a short range of data, conducting studies only in large cities, which know the importance of having data available to investigate new applications and services to their citizens. Figure 2.7b shows the relationship between *Costs* to develop and use of each VDSource and its respective *Granularity* and *Scalability*¹¹. *Cost* represents the value to use a data source, *Granularity* how much descriptive the data source can be, and *Scalability* the capacity of acquiring large amounts of data from different agents.

The questionnaire is one of the cheapest ways to acquire vehicular data. However, their responses may not completely correlate with real-world events. On the other hand, the use of infrastructure as vehicular data is more scalable given its capacity to sense a variety of agents¹² in the transportation system. However, it typically involves high financial costs and a management solution for the transportation system. An example of a low-cost and scalable solution to acquire vehicular data is the use of social media as a vehicular data source. Its

¹¹It means the capacity to provide amounts of data from a variety of agents.

¹²For instance, people, vehicles and companies.

broad use allows a wide information dissemination about road conditions, accidents and other events.

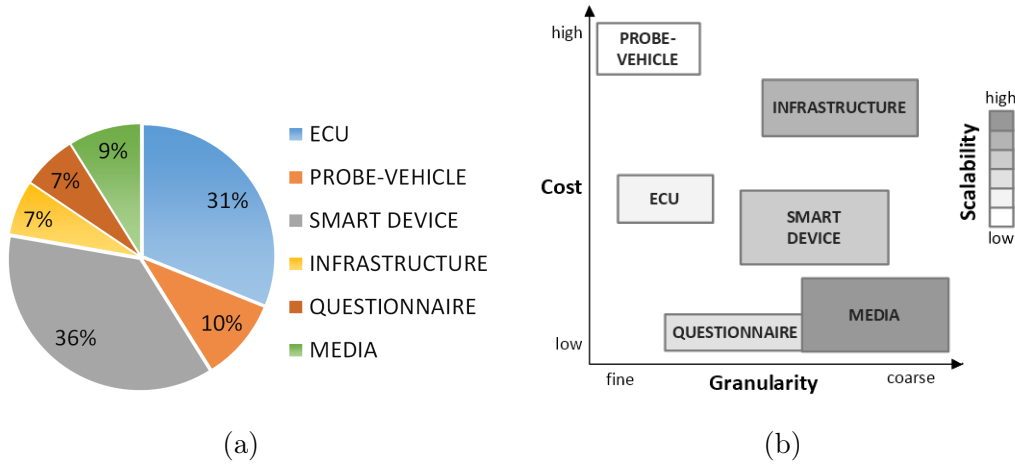


Figure 2.7: (a) Most used data source in VDS. (b) An overview of data acquisition based on its granularity and financial costs.

2.4 Potential Applications

Many are the applications designed for the vehicular environment, with different functions and goals. In this section, we categorize these applications based on the taxonomy described in Section 2.3. Figure 2.8 depicts the main applications based on vehicular data related to safety, eco-driving, traffic monitoring and management, infotainment, and also general purpose.

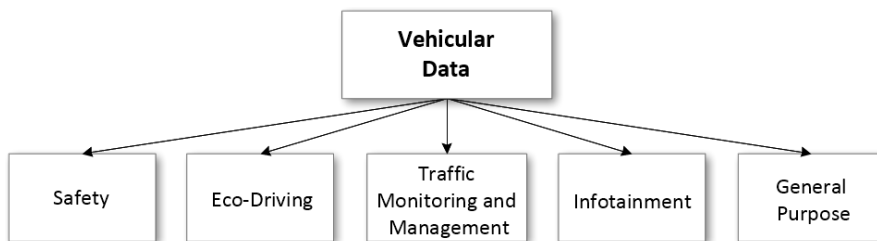


Figure 2.8: Applications based on vehicular data.

To present an overview of the applications, we summarized them in data classes using the VDS. We grouped the investigations into two data categories:

Intra-Vehicle Data (IVD) and Extra-Vehicle Data (EVD). Table 2.3 describes the groups of applications mentioned before. We noticed that 64% and 16% of them only used Intra-Vehicle Data (IVD) and Extra-Vehicle Data (EVD) to develop their applications, respectively, whereas 20% dealt with both groups. This clearly shows some interesting opportunities to explore the EVD and the fusion between IVD and EVD. For the rest of this section, we discussed the data of each category used by a given investigation. Furthermore, we highlighted the data availability which most of those group of applications utilized, and overview the whole section at the end.

Table 2.3: Class of data from VDS based on a given application group.

Application Group	Goals	Authors	Vehicular Data Space	
			Intra-Vehicle Data	Extra-Vehicle Data
Traffic Monitoring and Management	Event Detection (Incidents, Potholes, Traffic)	[Mednis et al., 2012; Pan et al., 2013; Zhao et al., 2016] [Wang et al., 2017]	✓	✓
		[Zuchao Wang et al., 2013; Goncalves et al., 2014; Han et al., 2014]	✓	
		[Gu et al., 2016; Septiana et al., 2016; Shekhar et al., 2016] [Kurken et al., 2017; Lau, 2017; Sinha et al., 2017] [Santos et al., 2018]		✓
Safety	Driver Style/ Behavior	[Aoude et al., 2011; Hong et al., 2014; Castignani et al., 2015] [Angkititrakul et al., 2009; Johnson and Trivedi, 2011; Paefgen et al., 2012] [Paefgen et al., 2012; Fazeen et al., 2012; Meseguer et al., 2013] [Ly et al., 2013; Guo and Fang, 2013; Engelbrecht et al., 2014] [Vaiana et al., 2014; Chu et al., 2014; Bergasa et al., 2014] [Elhenawy et al., 2015; Carmona et al., 2015; Martinez et al., 2016] [Kumtepe et al., 2016; Zhang et al., 2016; Hallac et al., 2016] [Ma et al., 2017; Saiprasert et al., 2017; Rettore et al., 2018a]	✓	✓
		[Beanland et al., 2013; Poó and Ledesma, 2013; van Huysduynen et al., 2015] [Sagberg et al., 2015]		✓
		[Aloul et al., 2015; Fox et al., 2015; D'Agostino et al., 2015] [Meseguer et al., 2013; Riener and Reder, 2014; AbuAli, 2015] [Ning et al., 2017]	✓	
		[CarChip, 2013; Technology, 1999; Paefgen et al., 2013]	✓	
		[Araújo et al., 2012; CGI, 2014]	✓	✓
		[Andrieu and Pierre, 2012; Brace et al., 2013; Meseguer et al., 2013] [Rutty et al., 2013, 2014a; Ayyildiz et al., 2017] [Rettore et al., 2017] [Truxillo et al., 2016]	✓	✓
		[Corcoba Magaña and Muñoz Organero, 2016; Zhao et al., 2016]	✓	✓
		[Riener and Reder, 2014]	✓	
		[Bröring et al., 2015]	✓	
		[Reiminger et al., 2015; Yuan et al., 2016] [Angkititrakul et al., 2009; Bergasa et al., 2014; Bröring et al., 2015] [OpenXC, 2012; MirrorLink, 2017; Ford, 2010] [Magister54, 2015]	✓	✓
	✓			
General Purpose	Data Acquisition, Data Available, Developers			

Intra-Vehicle Data = Location, Speed, RPM, Acceleration, Brake Pedal, Engine Load, Throttle Position, Gear, Fuel, Emissions, Engine Temp, Turning, Radar, Video/Audio, Light;
 Extra-Vehicle Data = Altitude/Atmospheric Pressure, wind speed/humidity/temperature, and traffic light/inductive loop; sociocultural factors, gender/age, driving experience, personality, and cognitive style; Social Media, News, and Government data;

2.4.1 Safety

There are many ways to increase the safety on the roads. The advance of technology has allowed investments on vehicles and roads to achieve this goal. Some studies support the necessity of improvements to decrease the number of road accidents. Most accidents could be avoided if the driver received a warning half a second before the moment of collision. In that way, studies to improve the recognition of driver's style have emerged, aiming to better understand the driver's behavior. In the safety category, we considered applications that propose to identify driver's patterns (e.g., style, behavior), offer customized insurance services, and improve the car security.

Driving analysis is a topic of interest due to the increase of the safety issue in vehicles. In 2015, the U.S. Department of Transportation showed the number of deaths in motor vehicle crashes, which is above 35 thousand people [Administration, 2016]. They also argued that alcohol, speeding, lack of safety belt use and other problematic driver's behaviors contribute to the death in vehicle crashes. The driver's behaviors vary considerably depending on age, gender, drugs consumption, types of used roads, distracted driving attitudes [Schroeder et al., 2013], and other factors. For these reasons, the study of driver's style has emerged, aiming to increase driving safety and, as consequently, reduce deaths in traffic. Engelbrecht et al. [2015] analyzed the use of smartphones to support a variety of ITS applications in a safety field as the driver's behavior, and road condition monitoring. Kaplan et al. [2015] also conducted a review to detect driver's drowsiness and distraction.

Considering as input data acceleration, braking and turning collected from the accelerometer sensor of a smartphone, once inside the vehicle, it is possible to sense the vehicle longitudinal and lateral acceleration. Then, thresholds on these measurements can detect different maneuvers. In that way, if we apply thresholds on the z-axis (representing acceleration and brakes), we can obtain rules to define the driver's style, aiming to identify sharp peaks that indicate aggressive increases of speed or hard braking. Additionally, analyzing thresholds on the x-axis acceleration, it is possible to detect excessive speed in left or right turns.

Several studies have focused on driving style and driving maneuvers recognition [Ly et al., 2013; Carmona et al., 2015; Kumtepe et al., 2016; Johnson and Trivedi, 2011; Zhang et al., 2016; Meseguer et al., 2013; Hallac et al., 2016; Martinez et al., 2016; Riener and Reder, 2014; Rettore et al., 2018a; Vaiana et al., 2014; Engelbrecht et al., 2014; Fazeen et al., 2012; Castignani et al., 2015; Bergasa et al., 2014; Saiprasert et al., 2017]. Some of these studies identify who the driver is whereas others classify the driver's behavior as aggressive or normal, and driving maneuvers. Ma et al. [2017] discussed the influence of noise provided by smartphone sensors, to identify dangerous behaviors. Satzoda and Trivedi [2015] extracted semantic information from raw data provided by the vehicle. D'Agostino et al. [2015] and AbuAli [2015] proposed a classification method for driving events recognition, using short-scale driving patterns. Fox et al. [2015] designed a pothole detection scheme using a real-world data and simulator.

In the same way, Aoude et al. [2011] developed algorithms for estimating the driver's behavior at road intersections. Wang et al. [2014] presented a survey of a wide range of mathematical identification and modeling methods of driver's behavior. Guo and Fang [2013] conducted a study aiming to identify factors associated with individual driver's risk and also predict the high-risk drivers, based on demographic data, driver's personality, and driving characteristics. Elhenawy et al. [2015] presented a model that can be integrated with in-vehicle safety systems to predict driver's stop/run behavior and then taking actions to avoid collisions. Chu et al. [2014] developed a smartphone app that focuses on determining if its user is a passenger or a driver. Using different approach, Beanland et al. [2013]; Poó and Ledesma [2013]; van Huysduynen et al. [2015], and Sagberg et al. [2015] used questionnaires from the literature to understand the multidimensionality and complexity of the driving styles concept. Hong et al. [2014] developed a platform, aiming to model the aggressiveness of the driving style, based on different data sources as smart devices, ECU and questionnaire.

Another agent interested in issues related to the vehicle safety is the manufacturers. They pay attention to their vehicles' behavior to foresee problems, allowing them to offer their services in advance. Thus, in that class of application, the manufacturers use the vehicular sensor data to improve their technology to make their automobiles safety and comfortable. As safety applications, we have

other two classes as prevention and correction. A diagnostics application is included in the prevention class and provides information about the components malfunction, aiming to avoid further breakdowns or damages. The applications in the correction class is designed to protect the vehicle and its passengers. The airbag application is activated based on a sudden stop (in most cases), the wheel speed can be changed depending on the lack of traction, for instance.

Many approaches considered the high costs involved in evaluating and improving vehicular safety solutions. They allowed a low-cost way for companies and researchers to develop and test their solutions. As an example, CarSim [Corporation, 2010] or generally VehicleSim (VS) is a product conceived to provide a realistic view of the vehicle components (e.g., tires, suspension, and steering) in different environments. Many companies and researchers use it as a tool for kinematic and control simulation testing to improve their development process.

Other market solutions focused on fleet companies. For instance, the CarChip Connect [CarChip, 2013] is an easy-to-use fleet monitoring tool. CarChip is a small telematics device with GPS and accelerometer, which connects to the vehicle by the OBD-II port. This tool provides the vehicle location and real-time alerts to improve the safety and the productivity. This tool tracks and sends reports data to the cloud, allowing clients to manage their fleets. In the same way, Scope Technology [Technology, 1999] aims to provide end-to-end telematics products and services. Their solutions empower insurance providers, fleet operators and aftermarket service providers to implement their personalized services.

The possibility to sense the vehicle and detect the driver's behavior opened the opportunity to customize applications and services developed according to the client's needs. As an example of these approaches, there are applications for insurance companies aiming to offer personalized services to their customers. The concepts of PAYD or Pay-How-You-Drive (PHYD) promote a new vision of how to charge rates, based not on the range of risk as age, address and gender, but also considering the driver's behavior, i.e., aggressive or standard. The aim of these applications is to classify the drivers' behavior to describe a distinguished attitude and its respective degree of safety for themselves and all around them. Besides that, the ability to offer flexible insurance services promises a significant improvement in traffic safety, taking into account the incentive to customers to

drive safely.

Paefgen [2013] focused on evaluating an accident risk based on continuous measurement of vehicular sensor data in the context of adaptive insurance tariffs. That work of Telematics strategy for automobile insurers also pointed out the business implications of risk-adaptive insurance taxes. Showing the less applicability to the current market, but a promising perspective on the new market entrants. As an example of a market, AXA is an insurance company that focuses on protecting personal property (e.g., cars, homes) and liability (personal or professional). AXA Drive [AXA, 2013] gives the driver real insights and personalized tips to help them to improve their driving behavior. State Farm insurance company developed a smartphone app, Drive Safe & Save [StateFarm, 2017], aiming to offer to their clients the reduction of auto insurance based on safer driving. Besides the car insurance, another promising field is related to the Health insurance. It aims to provide fast medical assistance based on a smart device application that automatically detects serious vehicle crashes, also known as Real-Time Medical Response [Detch, 2017]. Aloul et al. [2015] also conducted a study in that way, with the development of a smartphone app to detect and report car accidents.

Section 2.3 reviewed the literature through the lens of VDS and its data sources. However, we can have new insights when we look at the data used to achieve specific goals. Thus, Table 2.4 classifies applications into three groups: (i) safety; (ii) application goals as driver style/behavior, event detection, and insurance, fleet monitoring, and aftermarket; and (iii) data used for these applications.

That table categorizes 38 applications as safety, of which 28 focused on the driver's style/behavior, 7 on event detection, and 7 on insurance, fleet monitoring and aftermarket.

Table 2.4: Vehicular data space focus on safety applications.

Application Group	Goals	Authors	Vehicular Data Space																				
			Location	Speed	RPM	Acceleration*	Brake Pedal	Engine Load	Throttle Position	Gear	Fuel	Emissions	Engine Temp*	Turning*	ATM*	Radar	Video/Audio	Light	Infrastructure	Questionnaire*	Media*	Car Features	
		[Ly et al., 2013]	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
		[Carmona et al., 2015]	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
		[Johnson and Trivedi, 2011; Bergassa et al., 2014; Ma et al., 2017]	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
		[Paefgen et al., 2012]	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
		[Zhang et al., 2016]	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
		[Hallac et al., 2016]	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
		[Guo and Fang, 2013]	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
		[Ellenawy et al., 2015]	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
		[Chur et al., 2014; Engelbrecht et al., 2014]	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
		[Vaiana et al., 2014]	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
		[Fazeen et al., 2012; Sairasert et al., 2017]	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
		[Paefgen et al., 2012]	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
		[Beakland et al., 2013; Poó and Ledesma, 2013; van Huysduynen et al., 2015; Sagberg et al., 2015]	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
		[Aonde et al., 2011]	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
		[Angkitrakul et al., 2009]	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
		[Martinez et al., 2016]	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
		[Kumtepe et al., 2016]	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
		[Mesguer et al., 2013]	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
		[Hong et al., 2014]	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
		[Casignani et al., 2015]	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
		[Rettore et al., 2018a]	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
			17	11	9	19	7	3	7	2	2	2	1	15	0	3	9	0	2	5	0	0	0
Safety	Count Result		28																				
		[Aloul et al., 2015]	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
		[Fox et al., 2015]	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
		[D'Agostino et al., 2015]	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
		[Mesguer et al., 2013]	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
		[Riener and Reder, 2014]	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
		[AbuAli, 2015]	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
		[Ning et al., 2017]	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
			7	5	3	5	2	2	1	1	1	1	1	1	1	0	2	0	2	0	0	0	0
	Count Result		7																				
		[CarChip, 2013; Technology, 1999; Paefgen et al., 2013]	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
			3	3	3	3	3	3	3	0	3	3	3	3	3	3	0	0	0	0	0	0	0
	Count Result		3																				

Acceleration = longitudinal/3-axis, Engine Temp = Engine Coolant Temp, Turning = Rotation Angle;
 ATM = Altitude/Atmospheric Pressure, Infrastructure = wind speed/humidity/temperature, and traffic light/inductive loop;
 Questionnaire = sociocultural factors, gender/age/driving experience, personality, and cognitive style;
 Media = Social Media, News, and Government;

2.4.2 Eco-Driving

Fuel consumption is a factor that varies according to the drivers' habits. Two different vehicles are expected to consume more or less fuel according to their engines' size. However, the same vehicle may behave differently depending on the person who is driving it. As an example, someone who drives a car aggressively and accelerates it more than another person who uses it more consciously is expected to consume more fuel. From both environmental and economic points of view, it is desirable that drivers interact with their vehicles in a way that is as fuel efficient as possible, which reduces costs with refueling and greenhouse gases emissions. Collecting vehicular fuel consumption and emission data can lead to applications that help drivers to optimize these aspects in their driving styles.

Different initiatives and studies [Corcoba Magaña and Muñoz Organero, 2016; Meseguer et al., 2013; Riener and Reder, 2014; Rettore et al., 2017; Ruddy et al., 2013; Ayyildiz et al., 2017; Araújo et al., 2012; Andrieu and Pierre, 2012; Truxillo et al., 2016] are investing specialized services for Eco-driving to encourage driving style improvements, in order to reduce fuel consumption. Eco-driving refers to behavior and techniques designed to reduce fuel consumption, which includes recommendations for a person's driving style, the way, and frequency they use a vehicle, its configuration, accessories and maintenance. Eco-driving is part of a comprehensive approach to reduce the transport sector's contribution to the greenhouse effect. Bröring et al. [2015] developed a solution to acquire vehicular data and made it available to the community.

Brace et al. [2013] proposed a DAS to reduce fuel consumption decreasing the rates of acceleration, and the early gear changes, demonstrating a fuel savings of up to 12%, and average fuel savings of 7.6%. The CGI Group Inc [CGI, 2014] conducted a study based on more than 3 million Scania Truck trips, across seven European countries. They compared the impact of eco-driving coaching for different fleets and countries. Moreover, they proposed an estimated effect of coaching (EEOC), which provides a realistic estimate of the fuel savings gained from eco-driving coaching. Zhao et al. [2016] proposed the dynamic traffic signal timing optimization strategy (DTSTOS), also aiming to reduce the vehicle fuel consumption in a road intersection.

Table 2.5 summarizes all applications reviewed in this section, grouping them in the following groups: (i) eco-driving application; (ii) application goals as driver style/behavior, event detection, and data acquisition; and (iii) data used for these applications. Thus, we categorized 14 applications as eco-driving, of which 10 focused on the driver's style/behavior, 3 on event detection, and 1 on data acquisition.

Table 2.5: Vehicular data space focus on eco-driving applications.

Application Group	Goals	Authors	Vehicular Data Space																				
		[Brace et al., 2013]	Location	Speed	RPM	Acceleration*	Brake Pedal	Engine Load	Throttle Position	Gear	Fuel	Emissions	Engine Temp*	Turning*	ATM*	Radar	Video/Audio	Light	Infrastructure	Questionnaire*	Media*	Car Features	
		[Rettore et al., 2017]	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
		[Truxillo et al., 2016]	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
Driver Style		[CGI, 2014]			✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
		[Araújo et al., 2012]	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
		[Mesguier et al., 2013]	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
		[Andrieu and Pierre, 2012]	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
		[Rutty et al., 2013, 2014a; Ayyildiz et al., 2017]	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
Count		10	6	8	6	4	1	2	3	2	8	5	1	0	1	0	0	0	1	1	0	4	
Result			✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
Event Detection (Incidents, Potholes, Traffic)		[Corcoba Magaña and Muñoz Organero, 2016]	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
		[Zhao et al., 2016]	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
		[Riener and Reder, 2014]	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
Count		3	2	3	2	1	1	1	1	1	3	1	1	1	1	0	1	1	1	0	0	1	
Result			✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
Data Acquisition		[Bröring et al., 2015]	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
Count		1	1	1	1	0	1	1	0	1	1	1	1	1	0	0	0	0	0	0	0	0	1
Result			✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓

Acceleration = longitudinal/3-axis, Engine Temp = Engine Coolant Temp, Turning = Rotation Angle;
ATM = Altitude/Atmospheric Pressure, Infrastructure = wind speed/humidity/temperature, and traffic light/inductive loop;
Questionnaire = sociocultural factors, gender/age, driving experience, personality, and cognitive style;
Media = Social Media, News, and Government;

2.4.3 Traffic Monitoring and Management

It is well known the issues related to transportation and traffic in large cities, such as time spent on traffic jams, and number of fatalities and injuries on the roads, which achieved an alarming scenario. These numbers prompted new initiatives from governments and private sectors to improve the road traffic efficiency and safety. Thus, an ITS becomes a way to find smart and low-cost solutions to improve decision-making and obtain rich traffic information. In this field, to acquire rich information about the traffic, we need to comprehend the environment such as weather condition, vehicle characteristics and the road condition as influencers to the driving style. Thus, we show some applications that are interested in the characterization of traffic and road conditions.

Goncalves et al. [2014] used a smartphone GPS to study and characterize traffic and road conditions. They built the Iris Geographic Information System (GIS)- based platform using the smartphone Android on a client side and a server side for collect data by store, pre/post processing, analyze and manage the traffic condition. Zuchao Wang et al. [2013] developed a system for visual analysis of urban traffic congestion, using only GPS trajectories. Han et al. [2014] developed the SenSpeed an accurate vehicle speed estimation system to urban environments. Ning et al. [2017] studied the traffic anomaly detection based on trajectory data analysis in VSocN. Using a public data, Gu et al. [2016] explored the Twitter platform, aiming to extract traffic incident through users posts, providing a low-cost solution to increase the road information.

Santos et al. [2018] argued that LBSM feeds may offer a new layer to improve traffic and transit comprehension. They presented the Twitter MAPS (T-MAPS) a low-cost spatiotemporal model to improve the description of traffic conditions through tweets.

Septiana et al. [2016] proposed the categorization of the road conditions, based on text mining of Facebook feeds. In the same way, Shekhar et al. [2016] focused on the vehicular traffic monitoring using Facebook and Twitter posts. Pan et al. [2013] also used social media data to enrich the anomalies detection based on GPS from vehicles trajectories. Sinha et al. [2017] and Lau [2017] presented some insights based on public data to enrich urban public transportation and the

ITS. Kurkcu et al. [2017] provided detailed information about incidents, based on agencies and social media data.

On the other hand, Mednis et al. [2012] proposed the CarMote, a dedicated hardware designed to monitor and create a detailed road map of the quality of the surface and weather. Zhao et al. [2016] proposed the DTSTOS, also aiming to reduce the traffic delays in a road intersection. Aquino et al. [2015] and Silva et al. [2019] proposed a characterization of vehicles velocities to identify traffic behaviors using information theory.

Table 2.6 summarizes these initiatives and studies into three groups: (i) traffic monitoring and management application; (ii) event detection as application goals; and (iii) data used for these applications. We categorized 14 applications focused on event detection (e.g., incidents, potholes and traffic condition).

Table 2.6: Vehicular data space focus on traffic monitoring and management applications.

Application Group	Goals	Authors	Vehicular Data Space																					
			Location	Speed	RPM	Acceleration*	Brake Pedal	Engine Load	Throttle Position	Gear	Fuel	Emissions	Engine Temp*	Turning*	ATM*	Radar	Video/Audio	Light	Infrastructure	Questionnaire*	Media*	Car Features		
Traffic Monitoring and Management	Event Detection (Incidents, Potholes, Traffic Condition)	[Mednis et al., 2012]	✓	✓	✓	✓	✓	✓						✓			✓			✓			✓	
		[Zuchao Wang et al., 2013; Goncalves et al., 2014]	✓	✓	✓	✓	✓	✓							✓			✓			✓			✓
		[Han et al., 2014]	✓	✓	✓	✓	✓	✓							✓			✓			✓			✓
		[Pan et al., 2013]	✓	✓	✓	✓	✓	✓							✓			✓			✓			✓
		[Wang et al., 2017]	✓	✓	✓	✓	✓	✓							✓			✓			✓			✓
		[Gu et al., 2016; Septiana et al., 2016; Shekhar et al., 2016; Simha et al., 2017]	✓	✓	✓	✓	✓	✓							✓			✓			✓			✓
		[Kunkeu et al., 2017; Lau, 2017]	✓	✓	✓	✓	✓	✓							✓			✓			✓			✓
		[Zhao et al., 2016]	✓	✓	✓	✓	✓	✓							✓			✓			✓			✓
		[Santos et al., 2018]	✓	✓	✓	✓	✓	✓							✓			✓			✓			✓
Count Result		14	5	3	0	2	0	0	0	0	1	0	0	1	0	0	1	0	2	1	0	2	1	9

Acceleration = longitudinal/3-axis, Engine Temp = Engine Coolant Temp, Turning = Rotation Angle;
ATM = Altitude/Atmospheric Pressure, Infrastructure = wind speed/humidity/temperature, and traffic light/inductive loop;
Questionnaire = socio-cultural factors, gender/age/driving experience, personality, and cognitive style;
Media = Social Media, News, and Government;

2.4.4 General Purpose

The general purpose category shows studies to develop solutions to data acquisition and its availability to the community. Table 2.7 summarizes the proposals in this category. For instance, Bröring et al. [2015] proposed a solution to acquire vehicular data and made it available to the community, showing applications to fuel consumption and emissions. However, with these data in a large covered area, the possibilities exceed that initial purpose. An adaptive and comprehensive scheme for data acquisition in VSNs was proposed by Yuan et al. [2016], opening a variety of applications based on these data.

A smartphone app DriveSafe is available on the Internet [Bergasa et al., 2014] to detect the level of safety while driving. Furthermore, these data can be used to understand the safety of the driver and the safety of the road or area as well. There are initiatives [OpenXC, 2012; MirrorLink, 2017; Ford, 2010; Magister54, 2015] that made available vehicular sensor data, which allows the industry and research groups to develop their solutions. A prototype to provide vehicular data access through a website was developed by Reininger et al. [2015], which allows access to the vehicle speed, RPM, fuel consumption, GPS and altitude, making possible to design a variety of applications based on these data. Another data source that can be used as a general purpose is an international collaboration between Japan, Italy, Singapore, Turkey, and the USA, UTDive [Angkititrakul et al., 2009]. The aim was to develop a framework for building models of driver safety behavior. Moreover, they made the data collected available to the community, allowing the wide developing of applications.

Table 2.7: Vehicular data space focus on general purpose applications.

Application Group	Goals	Authors	Vehicular Data Space																					
			Location	Speed	RPM	Acceleration*	Brake Pedal	Engine Load	Throttle Position	Gear	Fuel	Emissions	Engine Temp*	Turning*	ATM*	Radar	Video/Audio	Light	Infrastructure	Questionnaire*	Media*	Car Features		
General Purpose	Data Acquisition		✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓							
	Data	[Bröring et al., 2015]	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓							
	Available	[Bergasa et al., 2014]	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓							
	Data	[OpenXC, 2012; MirrorLink, 2017; Ford, 2010; Magister54, 2015]	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓						
		[Angkititirakul et al., 2009]	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓							
		[Reininger et al., 2015]	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓							
		[Yuan et al., 2016]	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	
Count			9	7	7	6	5	6	6	6	6	6	6	1	1	1	6	0	1	0	0	1		
Result																								

Acceleration = longitudinal/3-axis, Engine Temp = Engine Coolant Temp, Turning = Rotation Angle;
 ATM = Altitude/Atmospheric Pressure, Infrastructure = wind speed/humidity/temperature, and traffic light/inductive loop;
 Questionnaire = socio-cultural factors, gender/age/driving experience, personality, and cognitive style;
 Media = Social Media, News, and Government;

Table 2.8: Availability of Vehicular data space.

Availability	Vehicular Data Space																			
	Location	Speed	RPM	Acceleration*	Brake Pedal	Engine Load	Throttle Position	Gear	Fuel	Emissions	Engine Temp*	Turning*	ATM*	Radar	Video/Audio	Light	Infrastructure	Questionnaire*	Media*	Car Features
Partially Public									✓	✓							✓		✓	✓
Public	✓	✓	✓	✓	✓	✓	✓				✓	✓	✓							
Private					✓	✓	✓	✓				✓	✓	✓	✓	✓	✓	✓	✓	✓

Acceleration = longitudinal/3-axis, Engine Temp = Engine Coolant Temp, Turning = Rotation Angle;
ATM = Altitude/Atmospheric Pressure, Infrastructure = wind speed/humidity/temperature, and traffic light/inductive loop;
Questionnaire = sociocultural factors, gender/age, driving experience, personality, and cognitive style;
Media = Social Media, News, and Government;

2.4.5 Infotainment

Infotainment is a term used in the vehicular context to provide services to the driver and passengers, based on a combination of information and entertainment. A variety of applications can be developed to achieve this goal. For instance, it is common that drivers bring their data in smartphones through apps, either local or on the cloud. However, when they are driving, the use of smart devices becomes a risk to themselves and other drivers. Furthermore, a traditional hands-free approach has limitations in several applications. In this way, it is convenient to think that the apps in a driver's smartphones can become useful through the dashboard display and multimedia kit inside the cars. Many companies and research groups are investing in solutions to better involve drivers and the environment around them. In the following, we describe some initiatives and studies in that way.

GM developed OnStar [GM, 2011] a solution to maintain its customers connected with their own cars. OnStar uses an integrated cellular service to connect the car to the Internet, allowing drivers and passengers to use the car audio interface to contact OnStar representatives for emergency services, vehicle diagnostics, and directions or personalized trip information. Moreover, GM customers can use a smartphone app to take control of their vehicles, for instance, lock doors, send an alarm to locate it, find it on a map, send a trip to navigate through the GPS embedded in a car, and also monitor it along the time. Similarly, Audi offers the Audi Connect [Audi, 2014] to give drivers more control over their vehicles, main-

taining them connected all the time to the Internet through the 4G-lite cellular network.

Some automakers have invested to provide customers with highly integrated connected experiences through connected in-vehicle infotainment systems to smartphone applications. To achieve this goal, automakers in partnership with other companies like Apple, Google, Pioneer, and Sony, for instance, have developed a way to create that connectivity environment. A recent initiative created by Ford, named Smart Device Link (SDL) [SmartDeviceLink Consortium, 2017], aims to enable existing smartphone applications to interface with vehicles. Through an open source community, using a standard set of protocols and messages that connect applications of a smartphone to a vehicle head unit. There are initiatives [OpenXC, 2012; MirrorLink, 2017; Ford, 2010; Magister54, 2015] that allow industries and research groups to develop their solutions using an in-vehicle data and connectivity. Cheng et al. [2011] analyzed communication protocols and their suitability for infotainment and safety services in VANET.

Generally, these approaches aim to safely permit the user to interact with apps installed in their smartphones while driving, exhibit the results on the dashboard display and hear the audio via the car's speakers. Another important issue is related to the variety of car models, not being restricted to one brand or model. The applicability can be diverse, for instance, get directions, make calls, send and receive messages, navigate on the Internet using voice recognition, and listen to music. In that direction, Apple developed the CarPlay [Apple, 2014] solution for their customers. The Car Connectivity Consortium (CCC) developed the MirrorLink [MirrorLink, 2017], which enables to establish a connection with a list of compatible cars, smartphones and apps. Toyota and BMW have also an infrastructure for the users of Toyota Touch 2 [Toyota, 2015] and BMW ConnectedDrive [BMW, 2014], respectively.

2.4.6 Data Availability

An important issue in the initiatives and studies discussed above is the data availability. This can allow new investigations based on to use of such data. Table 2.8 summarizes the availability of a given data as follows: (i) *Partially Public*: not all

data is available to the general public. It can be delivered with a reduced sampling rate or a low-frequency rate, with specific features blocked, and also with some sort of noise; (ii) *Public*: data is available to the general public, with no restrictions; (iii) *Private*: data is only available for closed groups or people ready to pay to have full access. Most available VDS data are free for the public or partially accessible by them. On the other hand, there are datasets provided by private companies, governments or even research groups with restrict access to the general public.

It is possible to see the partial availability of *fuel* and *emissions* data due to restrictions of vehicle sensors' data access applied by some automakers. The access to the *infrastructure* data is also restricted to a set of sensors such as camera and road speed of reduced areas. The availability of *Social Media* data can be classified into three groups: full access; short sample of the dataset; and only paid access. Thus, initiatives and research groups that plan to use social media should be aware of these possibilities.

Based on the relationship between *Cost* and *Granularity* depicted in Figure 2.7b, and the *Data Availability* analysis, we evaluated each application group in terms of these three metrics. Figure 2.9 presents the cost and granularity, considering the data sources of a given data used by an application of VDS (IVD and EVD). Moreover, the evaluation of the data availability used by applications provides an access scale between *Public* and *Private* for a given application group.

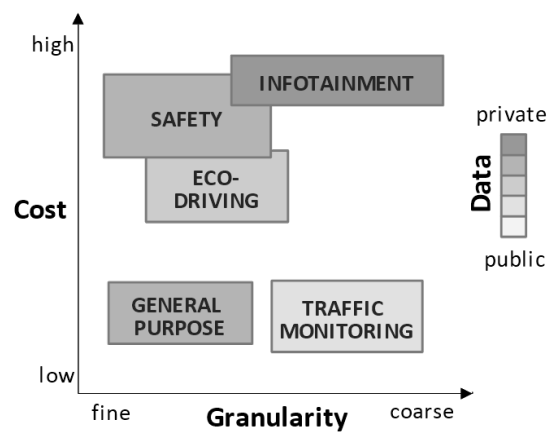


Figure 2.9: Overview of application groups based on their granularity, financial costs and data availability.

We noticed that safety applications used fine-grained data to obtain high-quality results, but introduced a high cost due to the quality of the used sensors and the fact that datasets are non-public. Traffic monitoring applications typically have a reduced cost, given the use of low-cost sources and public data. However, these applications have to deal with coarse-grained data which may reduce their accuracy. Another important group of applications is the infotainment. The data availability related to that class becomes essential to provide personalized infotainment solutions to drivers and passengers. This will probably demand a thorough study to understand the drivers' behavior, traffic, consumption trends of information and products, among other issues. The associated costs will depend on the data granularity, quality and availability.

2.4.7 Overview

The safety application group described in Table 2.4 reports 38 studies that list the most used data to detect the driver's behavior and events on the roads, but disregard the location and acceleration (longitudinal/3-axis). Driver's behavior applications also use the turning angle, which differs from the 3-axis acceleration due to its reduced noise. However this type of data comes from the ECU, and its access is not promptly available. On the other hand, IMU devices or smartphones can provide 3-axis acceleration, which provides a low-cost solution to detect the driver's behavior.

In event detection applications, locations play an essential role to identify the event on the map, and the speed and acceleration (longitudinal/3-axis) can provide semantic data from those locations. In insurance and fleet monitoring applications, there is a need for different sensor data, and possibly from smart devices as well, which will somehow identify the kind of behavior or status expected for the respective application.

It is important to notice that different data sources will have different roles in these applications (event detection, and insurance and fleet monitoring), and others as well. Sensor data such as fuel, emissions and light will possibly have no or little contribution to these previous applications. Social media data might be used to help identify the user's behavior and feelings, and, thus has the potential

to be very useful in this case.

The eco-driving application group described in Table 2.5 reports 14 studies that list the the most used data to detect the driver's style, events on the roads and evaluate an efficient fuel use, but disregard the location, speed and fuel consumption. This fact shows the intuitive relationship between vehicle speed and fuel consumption. Besides these data, RPM also contributes to these applications. The combined use of fuel consumption and brake pedal can offer a different solution for eco-driving applications. Social media and infrastructure can provide support for applications such as the shortest route, and near and cheapest gas station, which reduce emissions and fuel consumption.

We also observed that media data becomes an important data source in the traffic monitoring application group, where 9 of 14 studies use it to achieve their goals (see Table 2.6). This fact shows its capacity to describe events on the road from a user's perspective, which was not possible before. This is an opportunity to better manage the whole traffic and people's mobility.

Some studies showed the capacity of smartphones to measure movements and detect the driver's behavior. The comparison with the vehicle sensors from ECU is natural, making these smart devices an inexpensive way of instrumenting a vehicle. Moreover, smartphones have advanced sensors, allowing them to recognize the driving style, road and traffic conditions, and vehicle condition. On the other hand, there are substantial challenges involved in detecting movements using smartphones. The first one is the noise that comes from the vehicle movement and the uneven road. Besides, the position of the device can affect the results. Failures can occur considering that these devices are for general purpose. For instance, notifications of some applications can have a higher priority to the operating systems, and, then, the real-time measurement can be interrupted. Last but not least, real-time data is an essential feature for a driving analysis. However, continuous sensing and processing can drain the battery, making it impracticable for the users.

2.5 Chapter Remarks

The development of new applications and services for the ITS environment depends on the availability and study of large amounts of data, which leads to the Vehicular Data Space (VDS).

In this chapter, we survey recent studies describing services and applications for ITSs, but focused on the data used by them. We introduced the concept of VDS, which is used to describe the vehicular scenario from the data perspective. We proposed a taxonomy, according to the Vehicular Data Source (VDSource), discussed the different data sources currently used in ITSs. Furthermore, we discussed the relationship between *Costs* to develop and use each VDSource and its respective *Granularity* and *Scalability*. We also categorized the applications (Security, Eco-driving, Traffic Monitoring and Management, General Purpose, and Infotainment), noticing that 64% and 16% of them only used Intra-Vehicle Data (IVD) and Extra-Vehicle Data (EVD) to develop their applications, respectively, whereas 20% dealt with both groups. This clearly shows some interesting opportunities to explore the EVD and the fusion between IVD and EVD.

We also discussed the use of heterogeneous datasets to provide accurate methods for ITS applications. Thus, data fusion techniques have the potential to improve the accuracy of those applications, when there are several related descriptors. Some typical sensors used to model and identify the driver's behavior are acceleration longitudinal/3-axis, GPS, turning, and vehicle speed. Also constitute an opportunity, the generation of CO₂ emissions and fuel consumption reports, based on the investigations that use Intra-Vehicular Sensor (IVS). These reports can be sent to authorities who will be better informed when taking their decisions.

Our comprehensive literature review also showed that most of the data available in the VDS are freely available for the public or partially accessible by them. It is also clear that novel ITS applications will benefit from multiple heterogeneous datasets. Of course, this does not mean that a single variable represents a less descriptive scenario. On the contrary, in some cases the longitudinal acceleration, for instance, can identify dangerous driving maneuvers in real time, being a good solution for insurance companies.

Considering the Vehicular Data Space (VDS), the main contributions of this

work are: (i) the need of more investigations to recognize driving styles, relating them to individual and sociocultural factors; (ii) real driving observations need more spatiotemporal coverage; (iii) the need to expand and test applications in real-time environments; (iv) acceleration longitudinal/3-axis, GPS, turning, and vehicle speed are the most used sensor data to model driving behavior; (v) there is a complexity inherent in the processing of heterogeneous data since there is no standardization; (vi) heterogeneous data fusion is a fundamental challenge to leverage the ITS field.

Chapter 3

Heterogeneous Data Fusion

This chapter discusses the data fusion aspects of the Vehicular Data Space (VDS). We identified several issues in the data, which means that they must be treated before the data fusion process. Hereafter, we highlight some fundamental knowledge concerning Intelligent Transportation System (ITS), heterogeneous data fusion, challenges and opportunities in the field.

3.1 Contextualization

ITS integrates information and communication technologies to develop newer applications and services to boost the efficiency of transportation systems and mitigate their issues. Any ITS instance conducts one or more of the following intuitive steps: collection, processing, integration and providing information. ITS include at least four subsystems [Bazzan and Klügl, 2013; Faouzi and Klein, 2016]: i) Advanced Transportation/Traffic Management Systems (ATMS) to control and manage traffic devices (signals, monitoring, and safety devices), manage emergency situations, and other apparatus that support the system. ii) Advanced Traveler Information Systems (ATIS) to collect data and process it to improve understanding of traffic conditions and derive indicators which guide the traveler. iii) Automatic Incident Detection (AID) to apply algorithms for automatic incident detection as soon as possible to increase safety and reduce users perception of traffic disruption. iv) Advanced Driver Assistance Systems (ADAS) to apply technologies in

transportation system components (e.g., vehicles and roads) to reduce accidents and improve safety of the users. For instance, ADAS cover collision avoidance and driver assistance. Also, ITS involves others systems, such as Network Control, Traffic Demand Estimation and Forecast.

In this context, the demand of precise traffic information is an increasing challenge for public administrators and private businesses. ITSs subsystems are powered by data as much as possible. Traditional traffic sensors, usually, are installed to measure traffic flows at a given point, however they are ineffective when used alone. Nevertheless, there are other data sources on road infrastructures, such as cameras, GPS, smartphones and probe vehicles. All these multiple sources may provide complementary data and can be used to extract more comprehensive and detailed information about the traffic conditions. Thus, timely and precise traffic information allows ITS to provide traffic status and manage processes and services built to optimize the efficiency and safety of the transportation system.

Data information is at the heart of ITS. Indeed, there is no way to build ITS subsystems without data analysis. Usually, the data is heterogeneous (such as cameras, GPS, smartphones tracking, and probe vehicles). Thus, heterogeneous data fusion techniques are suitable in such situation [Nakamura et al., 2007]. There are many frameworks and models available in the literature to perform data fusion [Nakamura et al., 2007; Ayed et al., 2015; Khaleghi et al., 2013b]. There are three main approaches to perform data fusion: statistical, probabilistic and artificial intelligence [Faouzi and Klein, 2016].

Several issues make data fusion a challenging task, especially those regarding heterogeneous data. Most of the issues arise in the *Data Preparation* and *Data Processing* stages. In particular, data fusion aspects are extensively discussed by Khaleghi et al. [2013b]. For the authors, the data are naturally imperfect due to conversions (analogical/digital) or associations with some degree of uncertainty. They conducted a comprehensive study of methodologies that aim to solve problems related to heterogeneous data fusion. They elaborated a taxonomy of data fusion aspects describing problems such as outliers, conflict, incompleteness, ambiguity, correlation and disparateness. In this context, in this thesis, we focused on two main stages of the data life-cycle. The *Data Preparation* stage, which represents the most critical stage in studies related to ITS. Also, the *Data Processing*,

which deal with transforming the treated data into valuable or more informative data that can be used by applications and services.

The rest of this section is organized as follows. Section 3.2 represents the most critical stage of any study in ITS, dealing with data treatments. Section 3.3 highlights the process to transform the treated data into valuable or more informative data. In Section 3.4, we conducted a case study over vehicular data to show data issues and treatments that may be conducted before the fusion process. Finally, in Section 3.5, we conduct a discussion about heterogeneous data fusion in ITS, specially using vehicular sensor data.

3.2 Data Preparation

The data preparation is a critical stage of any study in ITS, since it is in this step that datasets are prepared to be used in different applications. It is at this stage that designers could consider to have “reliable datasets” that will have a strong impact on the final results.

Despite the relevance of this stage, just over half of the analyzed studies in this thesis explicitly mention the data preparation, whereas the others do not clarify the steps to prepare the data for the processing stage. One typical data preparation procedure is the reduction of variables, which aims to keep the most relevant features of the dataset [Hallac et al., 2016; Martinez et al., 2016; Castignani et al., 2015]. After that, most of the data from the VDS include spatial or temporal aspects, and the necessity to filter them depends on the application goals, making the resulting dataset adequate to its use.

The second non-trivial procedure of data preparation is to perform its corrections based on the data aspects, which almost all studies mention. These problems are more related to the data itself than to the methodologies used to combine them, mostly because the data collected from sensors are inherently imperfect. Based on these facts, the efforts to develop applications to an ITS usually depend on the role of each heterogeneous data to the application goal. Moreover, there is an inherent complexity in processing these data, which typically does not have any standard. This may become a barrier to do research in ITS.

In the following, we describe some of the data problems commonly found in the VDS, and propose some solutions. A fine data granularity usually allows a more valuable information about the entities of interest. The data granularity is a concerning aspect of data fusion, especially when dealing with applications that use rough sets and neither fine-grained nor coarse-grained information is beneficial for the final process.

Vagueness occurs in datasets where attributes are not well defined. The loose definition of attributes allows subjective measures, i.e., “fast” or “slow”. This issue commonly occurs in data sources like Questionnaire and Media from Vehicular Data Source (VDSource). The subjectivity of data present in social media, for instance, calls for strategies that allow its understanding. Using a Natural Language Processing (NLP) approach [Gu et al., 2016] and its algorithms, such as Term Frequency-Inverse Document Frequency (TF-IDF) [Kurkcu et al., 2017], Spell correction and Stop-word filter [Sinha et al., 2017], Latent Dirichlet Allocation (LDA) [Lau, 2017], and regular [Shekhar et al., 2016] expression, it is possible to reduce the noise and subjectivity of texts written by users. Fuzzy logic may also be used to remove the subjective aspect of these datasets.

Another issue in data preparation is the identification of outliers, i.e., extreme values that may do not belong to the solution. This process is completely data dependent and different techniques can be used to perform this filtering process. If outliers are left in the dataset, they may undermine the final solution, leading to imprecise results. Some of the filtering techniques to address this problem are Kalman [Bergasa et al., 2014; Ma et al., 2017] and Particle Filtering.

Incomplete data is, intuitively, data with missing parts. These missing parts may lead to incorrect conclusions and, thus, must be addressed. A possible strategy is to use probabilistic solutions whenever a data is missing. Ambiguity in datasets is a manifestation of its imprecision, and happens when two occurrences in the dataset are assumed to be precise and exact. However, they differ from each other.

There are other common methods to filter and correct the raw data. A Simple Moving Average (SMA) can be used to smooth out the effect of unwanted noise from the sensor data [Rettore et al., 2018a; Engelbrecht et al., 2014; Saiprasert et al., 2017], for instance. Besides, a band-pass and low-pass filter may remove

sensor noise [Chu et al., 2014; Engelbrecht et al., 2014]. The GPS incomplete data may be treated using a simple linear interpolation [Hallac et al., 2016; Saiprasert et al., 2017]. As a general way to prepare the raw data, we noticed the use of equations and thresholds (e.g., Max, Min, Mean, Median, Standard Deviation, Derivative, and Variance) to obtain particular results [Corcoba Magaña and Muñoz Organero, 2016; Ma et al., 2017; Gu et al., 2016].

All data sources, especially sensors, have a confidence degree. Whenever this confidence is lower than 100%, data is considered uncertain. Solutions to this problem include statistical inference and belief functions. The VDS is inherently disparate since there are sensors that assess different aspects in different units and scales. Using large quantities of diverse data allow the extraction of contextual information unable to be captured by physical sensors.

In summary, an important challenge in this stage is to find the best algorithm/method to apply to the raw data, aiming to treat and prepare the dataset for the next step. The key points we highlight at this stage are: (i) find the best way to fit and fix the data to be used in the proposed solutions; (ii) perform a variable reduction to keep the most relevant and descriptive features of the dataset; (iii) correct the dataset, by identifying outliers, conflict, incompleteness, ambiguity, correlation, and disparateness; (iv) apply heterogeneous data fusion techniques to also fit and fix the raw data; (v) use whenever possible standards to overcome the complexity of this problem domain and facilitate the research in ITS.

3.3 Data Processing

The data processing of VDS leads to various new descriptive data, giving vast possibilities of ITS applications, as mentioned in Section 2.4. In the data processing stage, the operation forms new aspects from raw or treated data. Depending on the investigation aims, a set of methods (e.g., mathematical operations, algorithms, models) can be applied to the data to produce a high-level data, allowing the development of new applications and services. Even considering the relevance of this stage to the whole data process, not all studies mentioned in this thesis made clear the description of the data processing stage.

The research in the ITS field involves interdisciplinary expertise once the dataset come from a variety of sources and each one is frequently used and maintained by specific groups. For instance, the weather data are supervised by meteorology institutes, although it can be used to alert risks on the road. Another data source that influences the traffic flow is provided by the department of transportation as a semaphore and speed limit. These data can be used to measure or identify the traffic flow. Furthermore, we can consider the weather data as a data layer to the whole transportation system. This means that each data point of other datasets present in the VDS might be associated with a weather data point (weather condition at that point). This can help to understand the traffic behavior from the point of view of weather conditions. Thus, a challenge here is to extract useful information from Intra-Vehicular Sensor (IVS) to perform some correlation with Extra-Vehicular Sensor (EVS), leading to personalized services for drivers in ITS.

In this scenario, data fusion becomes a tremendous challenge given the heterogeneity among the Vehicular Data Source (VDSource), asynchronous sensor operation, sensor errors and sensor noise. Furthermore, the computational infrastructure and the spatiotemporal aspects contribute to the efforts to fuse heterogeneous data. Rettore et al. [2017] developed a methodology to recommend the best gears by fusing the speed data, engine Revolutions Per Minute (RPM) data and throttle position data, based on a mathematical function to achieve low fuel consumption and CO₂ emissions. Almost all reviewed studies, which developed applications such as driving behavior and road event detection, deal with, somehow, a heterogeneous data fusion technique [Hallac et al., 2016; Martinez et al., 2016; Fox et al., 2015] that integrates multiple data sources to produce a more useful information than the individual data. Some of them applied Intra-Vehicle Data (IVD) fusion and others Extra-Vehicle Data (EVD) fusion to achieve their goals. However, the joint treatment of both fusion strategies is scarcely explored, being an important research topic for future of ITSs.

Another common aspect related to this stage is the use of Machine Learning (ML) techniques in data processing. Almost half of the studies aim to detect the driving behavior or road event using a machine learning technique. Leveraging the ideas discussed by [Ferdowsi et al., 2017; Chen et al., 2017], Table 3.1 shows the

Table 3.1: Most used classes of machine learning algorithms by the ITS applications.

Authors	Machine Learning Algorithms					
	Classification	Regression	Clustering	Dimensionality Reduction	Neural Network	Time Series
[Aoude et al., 2011; Chu et al., 2014; Elhenawy et al., 2015; Fox et al., 2015] [Hallac et al., 2016; Kumtepe et al., 2016; Zhang et al., 2016; Sinha et al., 2017] [Johnson and Trivedi, 2011; Lau, 2017; Hong et al., 2014; Aloul et al., 2015] [Martinez et al., 2016; Kurkcu et al., 2017; Corcoba Magaña and Muñoz Organero, 2016; Rettore et al., 2018a] [D’Agostino et al., 2015]		✓				
[Andrieu and Pierre, 2012; Castignani et al., 2015; Hallac et al., 2016; Rettore et al., 2018a]				✓		
[Andrieu and Pierre, 2012; Guo and Fang, 2013; D’Agostino et al., 2015; Hallac et al., 2016]		✓				
[Johnson and Trivedi, 2011; Engelbrecht et al., 2014; Aloul et al., 2015; Saiprasert et al., 2017]						✓
[Guo and Fang, 2013; Ly et al., 2013; Aloul et al., 2015]			✓			
[Meseguer et al., 2013; Elhenawy et al., 2015]					✓	

classes of ML algorithms used by the literature review we conducted in this thesis.

Next, we highlight the methods/algorithms applied by them: Extreme Learning Machine (ELM) [Martinez et al., 2016], Random Forest/Decision Trees [D’Agostino et al., 2015; Hallac et al., 2016; Rettore et al., 2018a], Support Vector Machines (SVMs) [Kumtepe et al., 2016; Zhang et al., 2016; Hallac et al., 2016; Elhenawy et al., 2015; Fox et al., 2015; Chu et al., 2014; Aoude et al., 2011; Sinha et al., 2017; Lau, 2017] to classify pothole, turn, driver, and driving behaviour. Logistic Regression [D’Agostino et al., 2015; Hallac et al., 2016; Andrieu and Pierre, 2012; Guo and Fang, 2013] to predict the driver, drivers’ risk, recognition of driving events. K-mean clustering [Ly et al., 2013; Guo and Fang, 2013; Aloul et al., 2015], Dimensionality Reduction Algorithms like Principal Component Analysis (PCA) [Hallac et al., 2016; Rettore et al., 2018a; Andrieu and Pierre, 2012; Castignani et al., 2015], Viterbi and Baum–Welch algorithms [Aloul et al., 2015], Artificial Neural Network (ANN) [Meseguer et al., 2013; Elhenawy et al., 2015], Adaboost [Elhenawy et al., 2015], K-Nearest Neighbors (KNN) classifier [Johnson and Trivedi, 2011; Lau, 2017], Naïve Bayes (NB) method [Corcoba Magaña and Muñoz Organero, 2016; Hong et al., 2014; Kurkcu et al., 2017; Lau, 2017], and, finally, Hidden Markov Models (HMM) to define different driver’s behavior based on observations [Aoude et al., 2011]. We also observed the use of algorithms to treat the temporal data aspects of VDS. The Dynamic Time Warping (DTW) algorithm aims to find an optimal alignment among signal vectors, allowing to detect and distinguish driving events, driver styles [Johnson and Trivedi, 2011; Aloul

et al., 2015; Engelbrecht et al., 2014; Saiprasert et al., 2017].

The key points we highlight at this stage are: (i) find the best algorithms/methodologies for data processing is an important and hard-task to the proposed solutions. (ii) extract useful information from Intra-Vehicle Data (IVD) to correlate them with Extra-Vehicle Data (EVD) to allow personalized services. This will become one of the top trends for future ITSs; (iii) data fusion plays an essential task in data processing given the data heterogeneity among the Vehicular Data Source (VDSource), and other aspects that need to be considered such as asynchronous sensor operation, sensor errors and sensor noise; (iv) machine learning (ML) techniques have a special role in data processing, mainly in classification and prediction tasks.

3.4 Vehicular Sensor Data Fusion

In this section, we conducted an exploratory analysis over the real vehicle data to show for each listed data issues (i.e. imperfection, correlation, inconsistencies, among others) which of them have been found in our experiment. Indeed, we found out several issues in the data implying that they must be treated before fusion process. We point out some fundamental knowledge concerning ITS, heterogeneous data fusion, challenges and opportunities in the field.

We examined the vehicular sensor data aspects in ITS context. We show challenges, useful data, as well as some methods to handle issues related to the data. In particular, our focus is on heterogeneous data fusion using intra-vehicle sensor data by collecting it from the Engine Control Unit (ECU) of a car. Although several papers presents reviews of heterogeneous data fusion [Nakamura et al., 2007; Ayed et al., 2015; Khaleghi et al., 2013b] or data fusion in ITS [Faouzi and Klein, 2016], our work provides the reader an illustration of the listed data fusion aspects with examples based on the conducted case study.

3.4.1 Vehicular Data

Modern vehicles rely heavily on data acquired through embedded sensors to improve the quality of their control systems. In order to better control the vehicle's

behavior, manufacturers invest both in quantity and quality of the sensors they use [Fleming, 2001]. Some of the sensors embedded in a modern vehicle include throttle pedal position, fuel pressure, and oil pressure. The sensors on a car communicate with the ECU through an internal wired network [Qu et al., 2010], and the data they output is accessible using the On-Board Diagnostic (OBD) interface.

Table 3.2: OBD Signaling Protocols

Protocol	Transfer Rates
SAE J1850 PWM	41.6 kbit/s
SAE J1850 VPW	10.4 kbits/s
ISO 9141-2	10.4 kbits/s
ISO 14230 KWP 2000	10.4 kbits/s
ISO 15765 CAN	250 or 500 kbits/s

There are five signaling protocols allowed on OBD interface, as shown in Table 3.2. All these protocols use the same OBD port. However, the pins are different except for those that provide power supply. The data collected from the sensors in the car are available through OBD Parameter IDs (PIDs). In Table 3.3, we show some of the sensors whose readings are available using the combination of OBD and smartphone. There are also other hundreds of sensors that can be accessed using OBD's parameter ID's - some of which are defined by the OBD standard, and the manufacturers define others.

3.4.2 Heterogeneous Data

Even though data collected from sensors embedded in a vehicle come from the same entity - the vehicle itself - it should not be considered homogeneous. The information is collected from different sensors spread across different parts of the vehicle's body in different measuring units. The heterogeneity of vehicular sensor data does not mean that there aren't relationships between the readings of different sensors since all of them monitor the same entity.

It is also possible to extract contextual information from data acquired by vehicular sensors. For instance, observing a car's speed over time, the traffic condition on its location can be inferred based on aspects like average speed and

Table 3.3: Sensors Collected from OBD and Smartphone

Sensors					
Engine load	Vehicle speed	Torque sensor	Fuel pressure	Oxygen sensors	Fuel Tank Level
Kilometers per liter	Intake air temperature	Ambient air temperature	Catalyst temperature	Relative throttle position	Accelerator pedal position
Fuel flow rate	CO2	Ethanol fuel %	Engine oil temperature	Fuel injection timing	O2 sensor monitor
Voltage	Distance traveled	Fuel remaining	Fuel rail pressure	Hybrid battery pack remaining life	Evap. system vapor pressure
Engine RPM	Engine coolant temperature	Fuel type	Malfunction indicator lamp	Exhaust gas recirculation error	Mass Air Flow Sensor
Altitude	GPS location	Collision sensor	Automatic brake actuator	Steering angle sensor	Rear camera
GPS speed	Gravity XYZ	luminosity sensor for headlights	Active park assist	Water in fuel sensor	Airbag sensor
Barometric Pressure	Time	Cost per mile/km	Front object laser radar	Night pedestrian warning IR sensor	Tire pressure sensor
Microphone sensor	Pressure sensor	Drowsiness sensor	Shock sensor	Rain-Sensing Windshield Wipers	Motion sensor

time stopped. These aspects represent peculiarities of traffic jams, where the average speed is low, and most vehicles are stopped for long periods.

3.4.3 Problems of Heterogeneous Data Fusion: Case Study

We considered as a case study the sensors data collected from vehicles and its relationship. We used an OBD Bluetooth adapter to collect data from a car. The logs of this vehicle consist of 55 trips of 40 km with an average time of 50 minutes each. Hereafter, we return to discuss the categories of data fusion problems but highlighting a practical view. Thereunto, we choose examples observed during the data collected from the vehicles, as our initial work.

3.4.3.1 Granularity

Granularity is related to the ability to derive valuable information about entities of interest on a dataset. It is a concerning aspect on data fusion, especially when dealing with rough sets, when neither fine and coarse-grained information is beneficial for the final process. Fine-grained information will not take advantage of the rough set techniques, on the other hand, a coarse-grained data may not be enough to derive useful information.

To characterize the granularity problem in vehicular sensor data, we investigate traces of taxis, buses, cars, and their respective time interval of data collection. In the literature, it is usual to find traces with measure between every 10 and 60 seconds. Thus, we measure the speed of a vehicle from its ECU each second and GPS speed each minute. Figure 3.1 shows an example of a car trace along almost 40 minutes, Figure 3.1(A) and (B) present the speed vehicle and GPS speed, respectively. Figure 3.1(C) shows GPS speed measured every minute. It is noted that in Figure 3.1(A) the vehicle speed is represented as fine-grained. Hence more detailed vehicle behavior is perceived. For instance, looking at the begin and end of the trace, it is clear to observe the stops-and-goes. This information reveals a particular behavior in a specific environment, urban area. On the other hand, Figure 3.1(C) represent the GPS speed in coarse-grained. Hence it can not address the same behavior mentioned before.

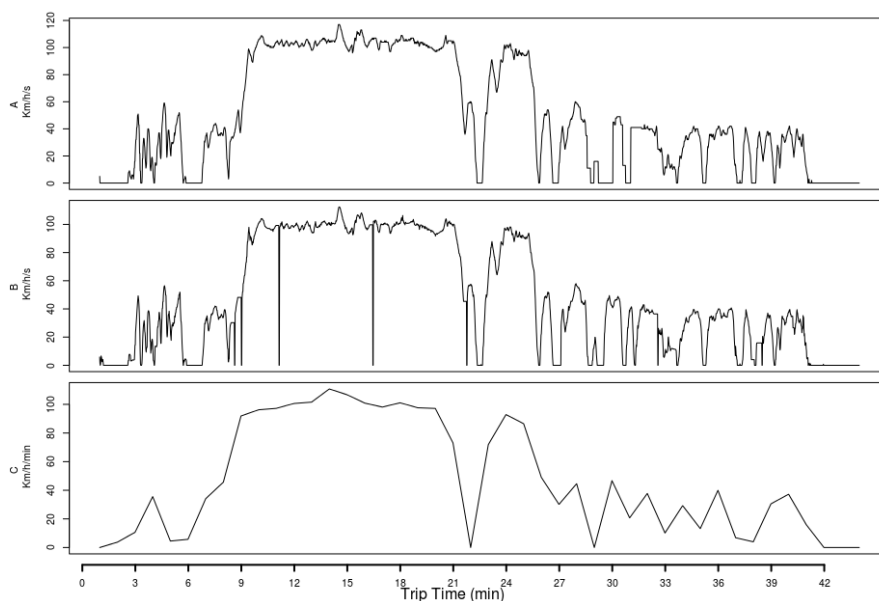


Figure 3.1: Comparison Between Vehicle Speed and GPS Speed Collected Every Second and Every Minute.

3.4.3.2 Vagueness

Vagueness occurs in datasets where attributes are not well defined. The loose definition of attributes allows subjective measures and the Fuzzy Logic may be a

way to remove the subjective aspect.

The vagueness in a vehicular data context may be intended as the speed of vehicle. In other words, it is not well defined by the speed, "fast" and "slow", of the vehicle. For instance, in Figure 3.1(A), the highway environment is characterized by the vehicle's speed behavior, which does rise above 80 km/h and below 120 km/h. Thereby, 80 km/h speed can be slow in a highway environment, but fast in the urban environment, where the vehicle's speed behavior does not rise above 60 km/h, due to legislation and traffic density.

3.4.3.3 Outlier

Outliers are extreme values that do not belong to the solution. These situations are often caused by errors in the sensors that generate it, or even unexpected values measured. When those data are considered false, it makes dangerous to data fusion systems, mainly because it leads statistical inferences to imprecise results. However, outliers may also describe particular events, becoming relevant data aspects and needs due attention.

The environment perception from sensors may come with incorrect data. These data represent points that distorter among the major data collected. Figure 3.1(B) shows the GPS speed along the trace. However, it is noted some distorter points with 0 (zero) values between high values collected. For instance, approximately in 10 minutes, the values are around 100 km/h and instantly changes to 0 km/h, returning to 100 km/h after that. Similar occurrences are shown along the trace and are called outliers.

3.4.3.4 Conflict

The same phenomenon, when observed by two or more sensors or specialists should be perceived in the same way by all of them. However, divergent specialists' opinions or punctual errors in sensor readings happen and cause conflicts in data observations. A simple, yet questionable, conflict solution is the Dempster combination rule [Yager, 1987].

In Figure 3.1, the conflicts appear when two sensors are related to describing the speed of the vehicle. Figure 3.1(A) shows, approximately, in 10 minutes

the values are around 100 km/h speed. However, in that same time interval, Figure 3.1(B) shows 0 km/h speed. The challenge of this topic is: which one may be considered for the data fusion?

3.4.3.5 Incompleteness

Incomplete data is, intuitively, data with missing parts. These missing parts may lead to incorrect conclusions based on the data and, thus, must be addressed. A solution to deal with this type of data is to treat the data in a probabilistic way.

The log used in our case study was obtained using an OBD Bluetooth adapter and a smartphone. However, interferences among electronic devices inside the vehicle, or barriers in the environment as tunnels, sometimes, cause the loss of communication. Consequently, gaps are introduced in a trace and made the dataset incompleteness as showed in Figure 3.2. Figure 3.2(A) shows the vehicle speed collected from ECU, and Figure 3.2(B) shows in three different moments, gaps caused by interruption of communication, ignoring important information and making the results inconsistent.

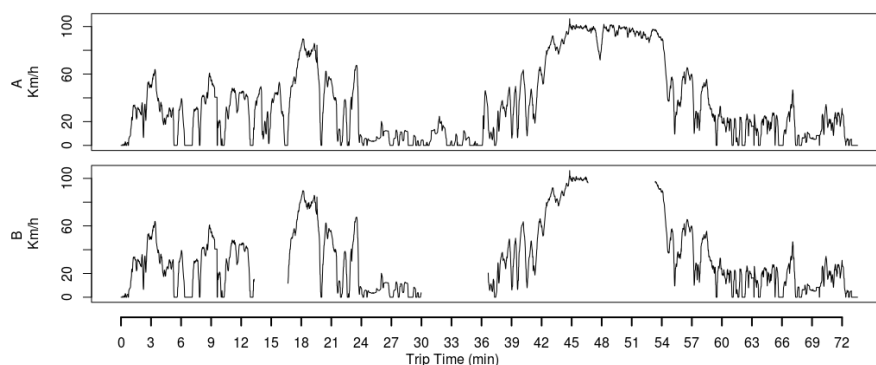


Figure 3.2: Comparison of GPS Speed and Incomplete GPS Speed Data.

3.4.3.6 Ambiguity

Different sensors can be considered as vehicle speed by ECU and GPS. In this case, the ambiguity manifests when both sensors present the same data to the same observation of environment. In Figure 3.3a, we show a histogram of the absolute difference between vehicle speed and GPS speed. The major frequency

of this difference is concentrated in 0 (zero), implying that both sensors collected the same speed. Furthermore, the values different to 0 implies that vehicle speed shows the current speed and GPS speed a different or conflicting value.

3.4.3.7 Uncertainty

Data collected from sensors or external sources are associated with a confidence degree. Whenever this confidence is lower than 100%, the data is considered uncertain. Solutions to this problem include statistical inference and belief functions.

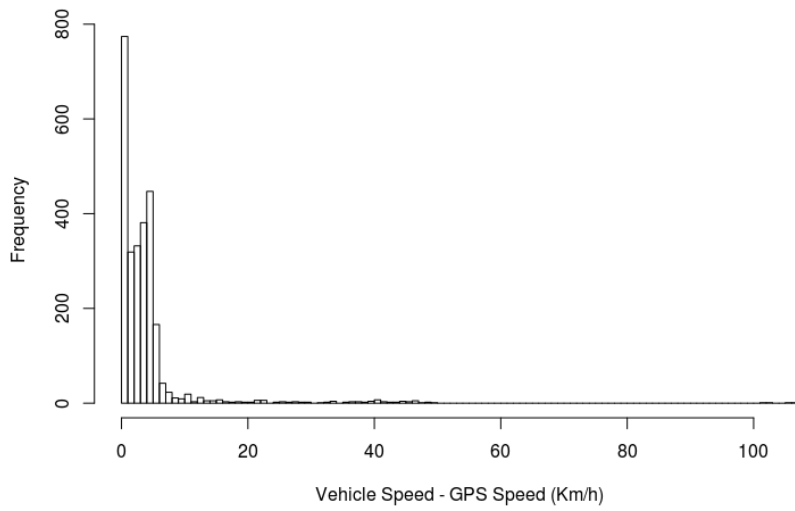
In the case of sensors, the uncertainty is always present, in other words, it is an inherent property of any sensor. Even though sensor data are collected directly from the vehicle by OBD, these data are not considered an absolute truth to provide a low uncertainty degree.

3.4.3.8 Correlation

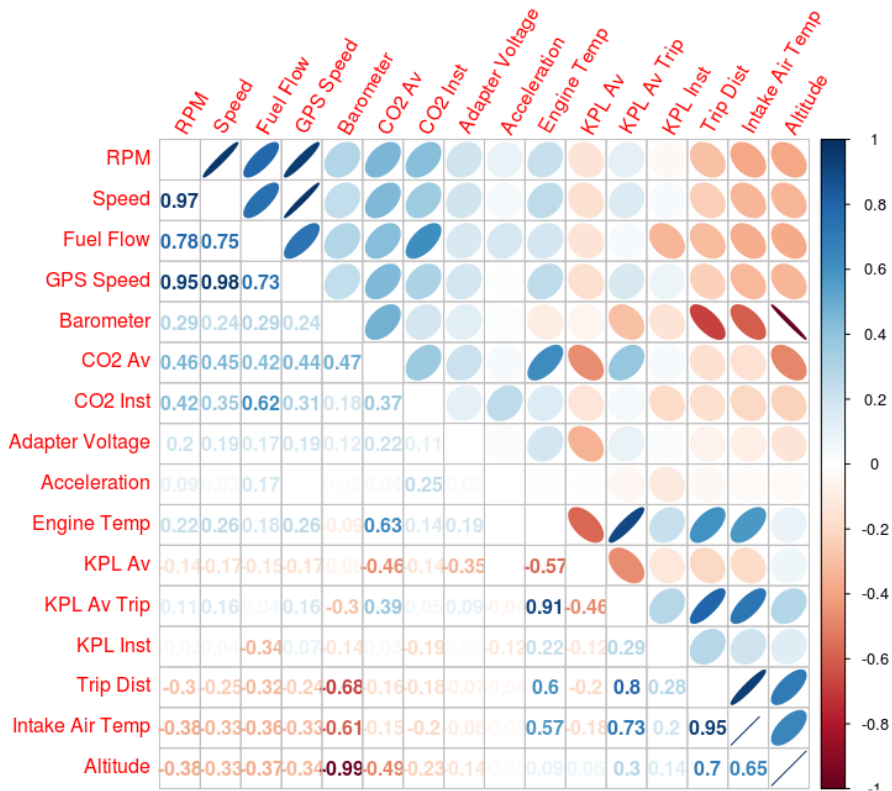
Data correlation is problematic in data fusion since it can either enhance or attenuate some aspects due to data incest. Data incest is a situation when correlated data is fed multiple times to the data fusion system, multiplying its importance on the final result.

We perform the Pearson Product Moment Correlation (PPMC), between all sensors readings in the data collected during a trip of one vehicle, as shown in Figure 3.3b. Since the correlation matrix is symmetric, on one side, it shows the explicit values of the correlation, and on the other side, the same value is visually shown as the ellipse that is expected from a bivariate distribution with the same correlation value. Thus, visually, ellipses close to straight lines represent two tightly linked sensors, which can be directly or inversely correlated, depending on the line direction. On the other hand, sensors with a small relationship will be represented by an almost invisible circle, due to the color scale. We considered a high correlation value between 0.5 to 1.0 or -0.5 to -1.0 , the medium correlation between 0.3 to 0.5 or -0.3 to -0.5 , low correlation between 0.1 to 0.3 or -0.1 to -0.3 and no correlation when 0.

In the high correlations, it is possible to see that revolutions per minute (RPM), Speed and GPS Speed represent the vehicle motion. So that, these data



(a)



(b)

Figure 3.3: Difference Between Vehicle Speed and GPS Speed (a) and Correlation Between Sensors Data in a Vehicle (b).

can be reduced to only one variable as Speed, for instance. However, there is a less explicit yet important relationship, like RPM and speed, which is governed by the transmission system of the car. Other possible reduction can be made in the relation between altitude and the atmospheric pressure, labeled as "Barometer". It is physically proven that the atmospheric pressure is inversely proportional to the altitude. Thus, this two variable can be explained using only one.

3.4.3.9 Disorder

When processing continuous data sources, sometimes measurements arrive out of their order and raise a natural question: what to do with this piece? A simplistic way of treating disordered data is to discard it simply. However, this tactic would ignore the contributions of the discarded piece. A more costly solution is to store all received data and reorder the entire set once an out of order observation arises.

This problem is not common in our scenarios, because the process of data collect is synchronous and the smartphone starts it. The other point is that the communication protocol deals with this problem.

3.4.3.10 Disparateness

Vehicular sensor data is inherently disparate since there are sensors that assess different aspects in different units and scales. Using large quantities of diverse data allows the extraction of contextual information unable to be captured by physical sensors.

As mention before, the vehicular sensor data is inherently disparate. In the vehicle, there are since sensors to measure the engine temperature until sensor to measure the fuel level. For instance, in Figure 3.4, it shows a dissimilarity between two sensors as revolution per minute (RPM) and carbon dioxide emission (CO_2). It may be possible to study the behavior of these two variables, but they remain disparate.

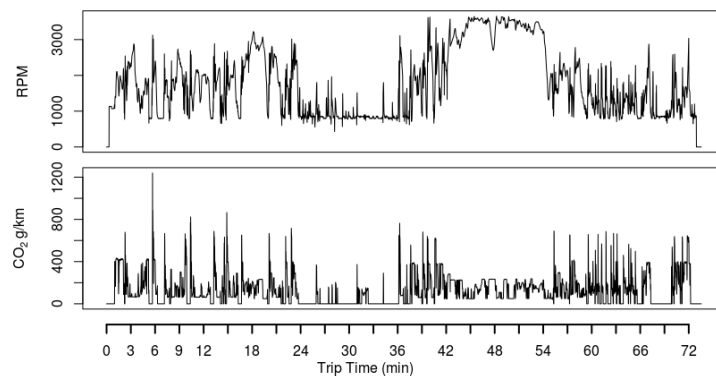


Figure 3.4: Disparateness Between Revolution per Minute and Carbon Dioxide.

3.5 Chapter Remarks

With the constant growth of the global population, urban mobility aspects and problems have become more challenging. Given the need of people to make their commutes quicker and safer in big cities, their current traffic infrastructures, and the elevated costs of restructuring it, a new approach to handle these issues is needed. Current information technologies and systems are capable of acquiring and processing massive volumes of data and outputting results with minimal delays, which makes them suitable for managing and planning new intelligent transportation systems for major cities.

Smart Mobility (SM) in ITS can be boosted by taking in account heterogeneous data collected from several sources as much as possible. However, in general, the data comes with some issues (*i.e.*, imperfection, correlation, inconsistencies, among others) making difficult heterogeneous data fusion process. In this thesis, we conducted an exploratory analysis of real vehicle data to show, for each listed data issues, which of them were found in our dataset. Indeed, we found out several issues in the data implying that they must be treated before the fusion process. Besides, understanding the vehicular sensors correlations allows providing solutions to optimize the vehicle use, reducing fuel consumption, emissions and vehicle maintenance. Which in terms directly influence the efforts to provide a Smart Mobility (SM) in a city.

Chapter 4

Intra-Vehicular Data Fusion

As defined in Section 2.3.1 the intra-vehicular data corresponds to the subset of sensors data that describe the main interactions between a vehicle and its driver, passengers or its surrounding environment, from the perspective of the vehicle itself. This section shows the Intra-Vehicular Sensor (IVS) allowing the exploration of heterogeneous data collected from several sensors to development of services and applications which may boost the Smart Mobility (SM) based on fuel efficiency, emissions, and safe driving.

4.1 Vehicular Sensor Data: Characterization and Relationships

Many technologies have been developed to provide effective opportunities to enhance the safety of roads and improve the transportation system. In the face of that, the concept of Vehicular *Ad-hoc* Network (VANET) was introduced to provide Intelligent Transportation System (ITS). In this chapter, we propose the use of an On-Board Diagnostic (OBD) Bluetooth adapter and a smartphone to gather data from two cars. Then we analyze the relationships between RPM and speed data to identify if this reflects the vehicle's current gear. As a result, we found a coefficient that indicates the behavior of each gear along the time in a trace. We conclude that this analysis, although in the beginning, suggests a way to determine

the gear state. Therefore, many services can be developed using this information as, the recommendation of gear shift time, eco-driving support, security patterns and entertainment applications.

We noticed that the data from a single sensor is not able to provide highly detailed contextual information about the vehicle's surroundings. However, some sensors are highly correlated with each other. As an example, fuel flow and revolutions per minute (RPM) are two highly related sensors and this relation is explained naively by the nature of combustion engine: each revolution involves a series of combustion on the cylinders. Thus, more revolutions mean more fuel consumption. In this section, we show that there are other relationships between individual sensors that can lead to a better understanding of a vehicle's state on a specific moment of a trip. The sensors relationships is an important aspect since it can provide useful information and insights for the vehicle's driver and occupants, and nearby vehicles as well.

We proposed the use of the OBD to identify the vehicle's current gear. The main contributions of this proposition are threefold: (I) characterize the dataset collected from vehicles' sensors, (II) show possible relationships between pairs of sensors, and (III) present specific relationships between linear speed and RPM, which is translated into the vehicle's current gear.

The remainder of this work is organized as follows. Section 4.1.1 presents the related works. Section 4.1.2 describes the characteristics of vehicular data. Section 4.1.3 we present our case study and illustrate the issues regarding the fusion of the data collected. We present the results in Section 4.1.4, and finally, in Section 4.1.5 we conduct a discussion about heterogeneous data fusion using vehicular sensor data and present our conclusions.

4.1.1 Related Work

There are several aspects involving a vehicle's operation that are not explicitly sensed, yet acquiring knowledge about these aspects would improve the reliability of vehicles' control systems. Faezipour et al. [2012] say a vehicle's Controller Area Network efficiency benefits from the number of sensors available to it. However, the solution is not as simple as adding as many sensors as possible to the vehicles:

the first obstacle is connecting all sensors to the controlling units. Wireless sensors would solve this connection problem, and Lu et al. [2014b] describe solutions to make it possible. Another feasible solution to replace physical sensors and expand sensing ability on an environment is virtual sensing [Kuo and Zhou, 2009]).

Virtual sensors calculate their output by taking readings from physical sensors and feeding them into mathematical models. Since the basis of virtual sensing is physical sensing, it is worth investigating the available sensors on a regular vehicle, as did Fleming [2001]. The author divided the vehicle into three main sensing areas: powertrain, chassis, and body and described the characteristics of the sensors used in specific components of these areas. Rodelgo-Lacruz et al. [2007] did not present sensor technology specifically. However the authors expanded the division of the vehicle's areas by adding the human-machine interface and multimedia, and telematics. This new division stresses the importance given to the drivers and their interaction with the cars' systems.

Since there are variables for which there are no physical sensors, some virtual sensors were developed to monitor the vehicular environment better. Ahmed et al. [2011] proposed a virtual sensing schema to monitor the health of physical sensors using virtual sensing of engine fault codes. Stephant et al. [2004] compared four virtual sensors that measure the sideslip angle of vehicles on a curve. The authors state that on normal conditions, where lateral acceleration is low, the sensors estimate the angle satisfactorily, on the other hand, for unusual conditions, where lateral acceleration is high, better models are needed.

The models used to develop virtual sensors may vary from neural networks to statistical methods. Atkinson et al. [1998] proposed a neural network model to predict aspects of a vehicle's behavior that cannot be directly assessed using values of other sensors with high accuracy. Another technique to implement virtual sensors, as shown by Wenzel et al. [2007], is Kalman filter. In work, the authors described the use of extended Kalman filters to determine variables such as yaw rate and lateral acceleration of a vehicle. With a similar approach, Brundell-Freij and Ericsson [2005] examined the effect on driving behavior of different driver categories and local environmental characteristics using a dataset of over 14,000 driving patterns.

Considering the cost of the sensors to measure the sideslip angle directly, the

authors Boada et al. [2016a] proposed a novel observer based on ANFIS, combined with Kalman Filters to estimate the sideslip angle, which in turn is used to control the vehicle dynamics and improve its behavior. The authors [Jin and Yin, 2015]) developed an estimation method to accurately estimate the vehicle sideslip angle and the lateral tire–road forces using in-vehicle sensors. Another interesting issue is utilizing smartphone sensors to estimate the vehicle speed, especially when GPS is unavailable or inaccurate in urban environments. This topic is discussed by the author [Yu et al., 2016]) that proposed an accurate vehicle speed estimation system, SenSpeed, which senses natural driving conditions in urban environments including making turns, stopping, and passing through uneven road surfaces.

In the same direction of most work-related, but considering a different approach, we analyze the relationships between RPM and speed data to identify if this reflects the vehicle’s current gear. Thereunto, we characterize the data collected from vehicles’ sensors, and we show that the specific relationship between linear speed and RPM is translated into the vehicle as a current gear.

4.1.2 Characteristics of Vehicular Data

Contextual information from vehicles is fundamental to better understand traffic patterns, drivers behavior and mobility patterns in a city. An example of contextual information generated by data collected from cars’ sensors [Ganti et al., 2010]), where the fuel consumption in the entire city scale was inferred from the readings of a few cars. To determine which sensors – individually or combined – better represent the vehicle’s context, we first need to characterize their readings in previously known contexts. In order to do this, annotated datasets are fundamental.

To the best of our knowledge, there are no publicly available datasets containing a significant number of car sensors’ readings, so we installed an OBD Bluetooth adapter in two vehicles to collect sensor readings. To characterize the sensor data, we selected a sample commute in our dataset that comprises a trip between two cities – namely Belo Horizonte, Brazil and Pedro Leopoldo, Brazil 40 km away from each other – with no abnormal traffic conditions.

In the collection process, an important step is to identify the data that pro-

vides valuable information about the vehicle. In our case, 25 variables were monitored, but only 16 out of these were analyzed. Some are direct readings from the vehicle's sensors; others are calculations based on data collected from the car and others are measured using the smartphone's sensors. These variables represent both lines and columns of the matrix in Figure 3.3b.

The variables that are directly read from the vehicle's sensors (through OBD) are:

1. *Intake Air Temp*: temperature of the air used in the air and fuel mixture.
2. *Engine Temp*: current temperature of the engine coolant liquid.
3. *Adapter Voltage*: voltage in the control module.
4. *CO2 Inst*: instant CO₂ emission of the engine.
5. *Fuel Flow*: fuel used by the engine on an instant.
6. *Speed*: speed shown by the odometer.
7. *RPM*: number of engine revolutions per minute.

The variables obtained by calculations are:

1. *Trip Dist*: distance traveled on the current log.
2. *KPL Av Trip*: average fuel consumption per kilometer on the current log.
3. *KPL Av*: average fuel consumption per kilometer of every logs.
4. *Acceleration*: speed variation between two observations.
5. *KPL Inst*: instantaneous fuel consumption per kilometer.
6. *CO2 Av*: average CO₂ emission of the engine.

Finally, variables obtained by sensors embedded in the smartphone are:

1. *Altitude*: instantaneous altitude of the vehicle.
2. *Barometer*: instantaneous atmospheric pressure.

3. *GPS Speed*: current speed measured by GPS sensors.

In a more detailed observation of the correlation matrix, we pointed in Figure 4.1 the four different types of correlation in its specific degrees. For instance, Figure 4.1a represents a high correlation between GPS speed and Vehicle speed, evidencing that relation is linear. Some points are not aligned with the relationship, this happened because of errors and differences in the readings of sensors. Another high correlation example is in Figure 4.1b, which shows the relation between atmospheric pressure (labeled as "Barometer") and altitude. It is physically proven that the atmospheric pressure is inversely proportional to the altitude. Thus the relationship is almost linear $-0,99$.

Figure 4.1c shows the correlation between -0.1 to -0.3 , that is a low correlation. However, the curiosity is that the scatter plot presents something similar to an exponential distribution. This Figure shows that the fewer liters are consumed per kilometer, the more gases are emitted. The other point is that the lowest carbon dioxide emissions happen with the lowest fuel consumption (more kilometers per liter) and it may characterize moments where the driver stops accelerating.

Finally, in the extreme of the correlation matrix, we show in Figure 4.1d a pair with no correlation, represented by -0.08 Pearson correlation coefficient. The relation between the battery voltage and intake air temperature does not represent relevant information. Since the battery voltage behaves considering the vehicle acceleration. In other words, the alternator works with the vehicle movement, and it is used to charge the battery and to power the vehicle electrical system. At the same time, the intake air temperature sensor it is not affected by the battery voltage.

During the time of collection, we were able to capture a variety of traffic situations: urban environments with various traffic levels, highways, strikes and roadblocks. An example of a Vehicle 1 observation, some of the sensor readings of its trip is illustrated in Figure 4.2 and represent its current Vehicle state. We consider the vehicle state the perception of the context, in which it is located, through its sensors readings. In the graphic, the colors of the columns divide the timeline in the three scenarios: urban traffic in the origin city, highway traffic, access routes to the destination city – called "Transition" and urban traffic in the

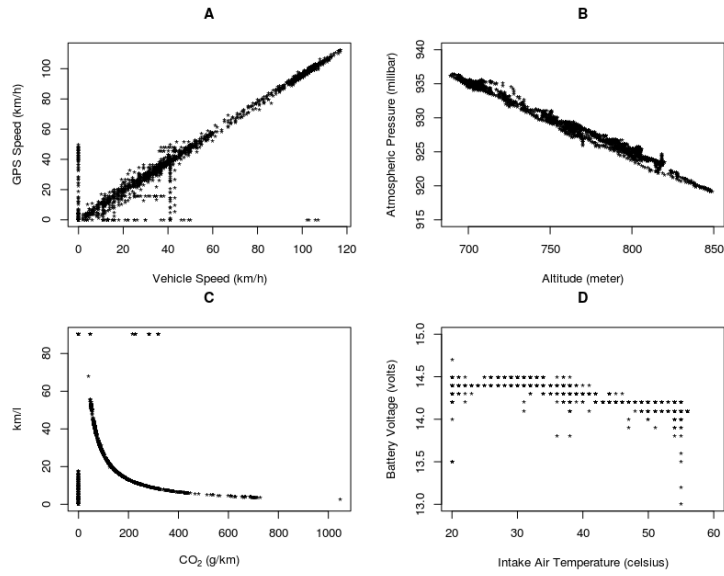


Figure 4.1: Correlation between sensors.

destination city.

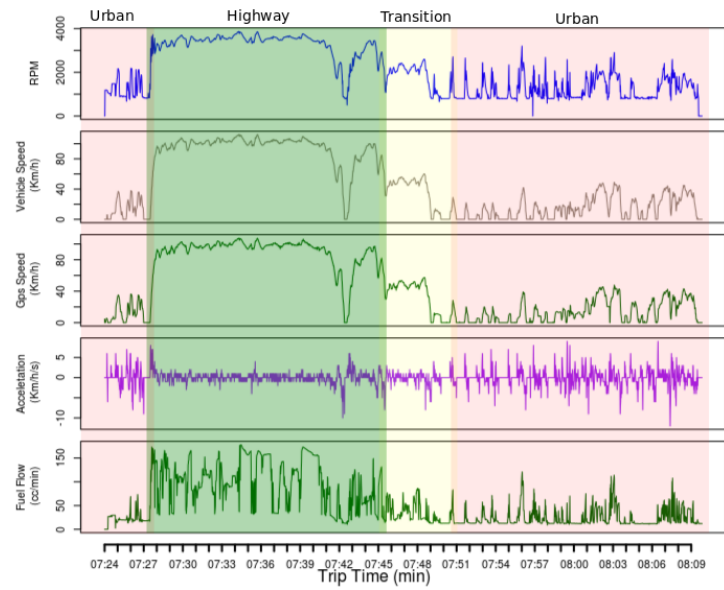


Figure 4.2: Vehicle sensor data behavior along the trace.

The urban environment is characterized by the vehicle’s speed behavior, which does not rise above 60 km/h, due to legislation and traffic density. Traffic density is also noticeable at the end of the timeline when the destination city’s

traffic is heavier, and, thus, the cars move in a stop-and-go fashion, stopped by traffic lights or road crossings and moving at every opportunity, until they are stopped again. This kind of behavior reflects into the horizontal lines at 0 km/h in the urban environments, followed by small peaks in speed. Acceleration, which is the variation of speed over time is also different in these situations. Due to constant acceleration and breaking of the car, the speed variation is higher in such situations.

On the other hand, the highway part of the trip shows a different behavior. The speed is constantly high, and there are no big acceleration or deceleration intervals, and the speed rarely drops below 60 km/h. To keep the vehicle moving at such high speeds, the engine must also work harder, translated into higher RPMs, which also present different values from the urban scenarios. Even though, there are some points in urban traffic where the engine revolves at more than 3000 times per minute, these occurrences are rare and do not last as long as the highway, where for approximately 15 minutes the revolutions did not go much lower than this value. A unique aspect of the highway part in this data is the fuel flow, which is significantly higher, but not as constant as the RPM or the speed. This behavior may reflect the road condition, altitude variations and atmospheric pressure, but it requires further investigation.

So, a more detailed characterization can be done with a more detailed study of these data. For instance, identification of traffic jam, strikes, roadblocks and accidents in an urban area and highway is an important issue to solve and require more investigation. Therefore, in this work, we first focus on the characterization of both urban and highway environment's providing high-level vision as shown in Figure 4.2.

4.1.3 Case Study

The characterization of the data acquired from the two vehicles revealed pairs of sensors that have a strong relationship. More specifically, the relationship between the readings of the RPM and speed sensors is close to linear, and to investigate it, we collected data from two cars to analyze their RPM and speed throughout the time. The two vehicles are in the same category, yet their manufacturers

and engine power are different. Their main characteristics of data acquisition are presented in Table 4.1. The logs of Vehicle 1 consist of 40 trips, each one of 40 km with an average time of 50 minutes. The logs of Vehicle 2 consist of 15 trips, each one of 10 km and with an average time of 30 minutes.

	Vehicle 1	Vehicle 2
Engine	1.0 16v	1.6 16v
Max RPM	7000	7000
Transmission	5	5
Power	76	122
Weight	1025 kgf	1000 kgf
Manufacturer	Renault	Hyundai
Model	Sandero	HB20
Trips	40	15
TripTime	50 min	30 min

Table 4.1: Data acquisition characteristics

The collected variables are presented on the scatter plots present in Figure 4.3. A visual inspection of the points reveals a relationship between the two sensors that is, indeed, close to linear as stated previously. However, there are clear groupings of points that share a stronger relationship that is equivalent to the gear ratios of the vehicle. Figure 4.3a presents the plot for the vehicle 1 that travels 40 km on urban environments and highways, where the fifth gear is used more often and the five different lines show the gears. Vehicle 2 travels on urban environments only and, because of this, rarely uses its fifth gear, which justifies the absence of the fifth line on the vehicle’s readings, shown in Figure 4.3b.

To isolate the groups that represent the vehicle’s gears, we reduced the two analyzed variables to obtain a unique view of them as their coefficient. Since they are distributed in well-defined lines, it is expected that the reduction through the division reveals the gears’ speed to RPM relation.

4.1.4 Results

As result of this case study, we calculated a coefficient that indicates the behavior of each gear and plotted it along the time as we show in Figure 4.4. We emphasize the constant occurrences of the coefficient with horizontal colored lines. These

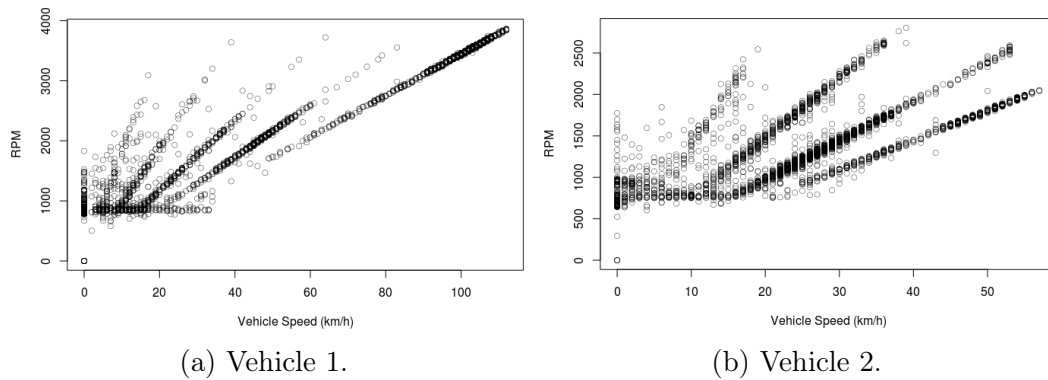


Figure 4.3: Correlation between vehicle speed and RPM.

lines represent the groups of points in the scatter plot, indicating the active gears. The lines represent the gears' RPM regarding speed in a crescent order, thus the 1st gear is in red, 2nd gear in purple, 3rd gear in yellow, 4th gear in gray and 5th gear in green. As Vehicle 2 does not use the 5th gear very often, because it moves only in the urban environment, there is a difference in the number of gear lines between Figures 4.4a and 4.4b.

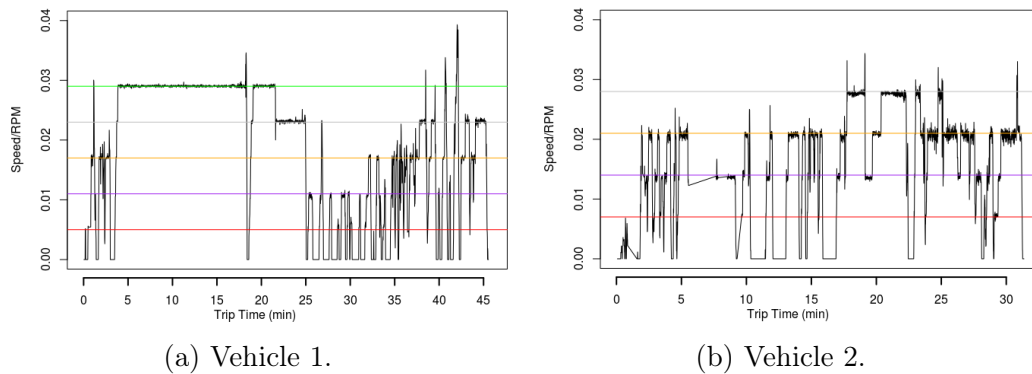


Figure 4.4: Vehicle's speed and RPM relation in a time series.

Another point that must be noted in Figure 4.4 is the difference in the horizontal values of the gear relations. Since we are comparing two different vehicles, from different manufacturers and different engines, their gearbox is also assumed to be different. Thus the gear ratios are also different. We have not had the opportunity to compare these results with readings from other vehicles from the same model, manufacturer or engine power. However, we believe that the same

relationship will hold for cars in the same model and, probably, from the same manufacturer due to the same parts being used in different models to reduce costs and supply chain complexity. Also, it is difficult to determine the first gear (in red), for a simple reason that its use is for a short time, to leave of inertia.

We evaluated all collecting trips, and we show that the same coefficient represents exactly the gear state. In other words, the coefficient evidence which gear is used and along the time it is possible to understand the vehicle context such as driver behavior better.

4.1.5 Section Remarks

In this work, we analyzed data collected from two cars using the OBD port. In the first analysis, we characterize the collected data by showing the correlation between the reading of the sensors and later we showed how the sensor readings behave in different scenarios. We present that the sensed values in urban environments are different from those captured when the same vehicle is on a highway. We trust that with the deeper investigation, it is possible to determine on which kind of environment - highway or urban traffic - a vehicle is based its sensor readings. Moreover, we also trust that the current traffic condition of a given vehicle reflects on its sensed data and is possible to determine the intensity of the traffic based on sensors from the vehicles.

The second analysis focuses on finding the effect of the vehicle's current gear in their speed and RPM. To do this, we collected data from these vehicles and found multiple close to linear relationships in the RPM and speed scatter plot. These relationships are effects of the gears which have specific ratios, that varied between our two test vehicles. The coefficient of each gear is directly linked to the slope of the lines that represent each gear. We believe that it is possible to determine the specific values that represent the lines' equations for any given dataset containing RPM and speed values. By discovering the current gears of a vehicle over time, we add a new variable for which there are no sensors available in the OBD data.

In summary, Figure 4.5 shows how our design of fusion on Vehicular Data Space (VDS) worked in this study. Where, the OBD vehicular sensors feed the

fusion process, the data preparation deal with data aspects showed in Chapter 3, data processing covers the related methods, and finally resulting in a gear virtual sensor as the data use. Moreover, this new virtual sensor, as well as many others, may boost the SM due to its contribution to understanding better the vehicles' state and the development of new systems and services, such as recommendations of gears based on fuel efficiency, emissions, safety driving and entertainment applications.

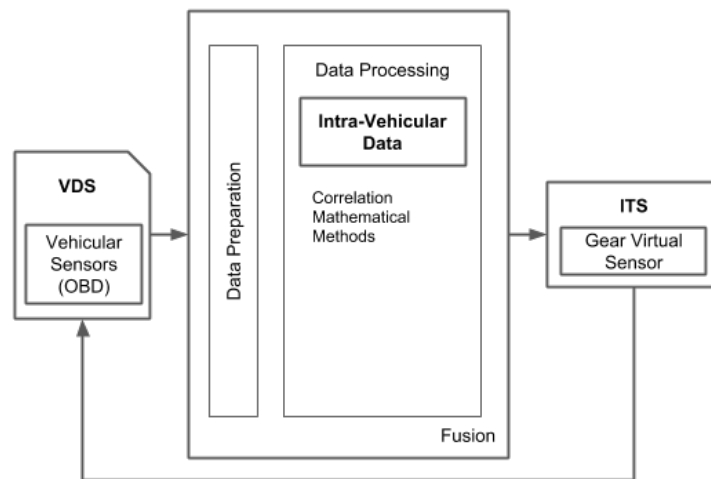


Figure 4.5: Design of fusion on VDS for gear virtual sensor.

4.2 Vehicular Virtual Sensor

Physical sensors are important parts of control systems, especially vehicular control systems. Sensor readings help drivers control their vehicles as well as their internal systems while keeping a vehicle stable and running. Currently, a modern luxury car carries hundreds of diverse and precise sensors and not all of them are visible to the conductor. However, there are phenomena and aspects for which there are no physical sensors available. Virtual sensors combine readings from multiple sensors in order to develop their output values based on conditions and models, and, eventually, substitute and monitor failing physical sensors, as well as sense complex variables. Designing a virtual sensor is usually a difficult process due to the complexity of the different processing stages it comprises. This sec-

tion presents a study on the process of creating and prototyping vehicular virtual sensors, describing development stages and presenting examples of virtual sensors created with a framework developed to facilitate the design process.

A problem that rises when using sensor data to monitor and control entities, especially vehicles, is its reliability regarding both availability and quality of information. A sensor must output correct readings constantly, and control systems depend on these characteristics to operate properly, however, every sensor has an inherent probability of presenting a malfunction on each one of these aspects. A solution to monitor physical sensors or temporarily replace them is a virtual sensor, which collects data from other sensors and outputs data according to models or formulas.

Virtual sensors are useful alternatives to monitor aspects, variables, and phenomena for which there are no physical sensors. There are cases where physical sensors are unavailable, and a virtual sensor can replace them, given that the variable they monitor is mathematically described or highly correlated to other monitored variables. In fact, a virtual sensor may substitute several physical sensors, used to monitor a single complex aspect for which there is no physical sensor, by combining their information using models and outputting the desired information. Additionally, a virtual sensor can produce new and higher level sensor information.

The process of designing a virtual sensor may be summarized to three steps, as illustrated in Figure 4.6: (1) collect and treat input sensor data, (2) define and apply methods and models to combine multiple input data sources to (3) output new calculated data. Collecting and treating input data is a particularly challenging step since there are several sources of problems related to sensor data, such as incompleteness and inconsistency. The second step, which consists of defining the way the virtual sensor will treat input data to generate new information is especially important and requires technical knowledge from the designer to determine and implement models and formulas. Finally, outputting calculated data requires the designer to format it to fit standards.

We discuss the design process of vehicular virtual sensors. First, we present related works that leverage vehicular sensor data to produce new data on Section 4.2.1. Section 4.2.2 presents the collection process performed by the authors using

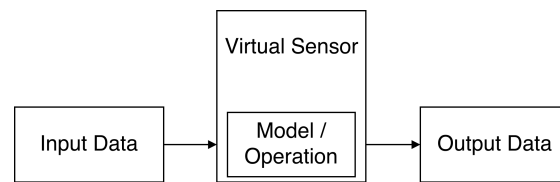


Figure 4.6: Virtual sensor design scheme.

two cars and problems related to sensor data. In Section 4.2.3, we will use a virtual sensor prototyping framework developed to facilitate the virtual sensor design process and the possible operations on sensor data. In Section 4.2.4.3, we will present virtual sensor examples as a way to demonstrate their design process and the new information that they provide. Finally, we present our conclusions and future works in the Section 4.2.5.

4.2.1 Related Work

The basis for diagnostic systems are physical sensors – even though virtual sensors are an alternative [Li et al., 2011; Stephant et al., 2004]), they still depend on physical sensors – thus, it is worth investigating the available vehicular sensors. AbuAli [2015] collected data from vehicular sensors using the OBD interface and used it to detect hazardous driving situations, like hard braking, speeding and traffic weaving. Jeong et al. [2013] proposed a methodology that identifies this kind of driving, as well as lane changes using a gyroscope embedded in a test vehicle. Imkamon et al. [2008] used video image processing to identify the density of vehicles nearby and turning directions to detect potentially hazardous situations.

Collecting vehicular fuel consumption and emission data can lead to applications that help drivers optimize these aspects in their driving styles. Ganti et al. [2010] used participatory sensing to induce fuel consumption from roads of a city using data collected from few vehicles and trace the most fuel-efficient route between two points. Ahn and Rakha [2008] stated that highways and high-speed routes not always are as fuel-efficient as less crowded arterial streets and to endorse this, Ericsson et al. [2006] developed a driver support tool that recommends the route that consumes the least fuel and points towards the importance of taking real-time traffic information in these recommendations.

Chen et al. [2014] and Eriksson et al. [2008] used taxis to collect data from the road conditions using accelerometers and GPS receivers. Using the data acquired, they were able to determine the condition of road surfaces as well as the location of potholes with high accuracy. Zan et al. [2010] assumed that the road condition could be delayed to reduce communications overhead and proposed a system that benefits from geocaching to forward sensed data when convenient.

Driving analysis is a topic of interest due to the increase of a safety issue in vehicles. To address it, several works focused on driving style recognition [Johnson and Trivedi, 2011; Bergasa et al., 2014; Carmona et al., 2015; Martinez et al., 2016; Hallac et al., 2016]. Some of these works identify who is the driver and others classify the driver behavior, as aggressive and normal, for instance. In both cases, we can apply the concept of virtual sensor design, proposed in this work. In other words, the input data considered are vehicular sensors, virtual vehicular sensors, and sensors embedded on smartphones. The model focuses on identifying drivers and their behavior based on a set of procedures encapsulated as a virtual sensor. Finally, the new sensors will output a driver's identity or behavior.

4.2.2 Data Acquisition

Mobile *Ad-hoc* Networks (MANETs) are powerful environment sensing tools, due to their capacity of deploying nodes on wide areas. Such benefits come at a cost, though: energy is a limited resource and should be used carefully, and the connection is not always available and costly activity, thus transmitting sensed data is a delicate process.

More recently, a new derivation of MANETs had emerged when vehicles were given communication capabilities in a Vehicular *Ad-hoc* Network (VANET). VANETs differ from MANETs as their characteristics are more specific to the vehicular environment. Thus the nodes are expected to move in well-defined patterns and concentrate in higher density urban regions. In fact, the vehicle is the most powerful sensing platform in any MANET, since it contains various types of highly reliable sensors while almost eliminating energy constraints, due to its rechargeable battery while driving, and having communication capabilities through cellular and wireless networks on urban areas.

The OBD system was first introduced to regulate emissions, but nowadays its applications have grown from helping aftermarket maintenance services to Eco-driving applications. To access sensor information using the OBD system, there are Parameter IDs (PIDs) that identify individual sensors. Some PIDs are defined by regulatory entities and are publicly accessible. However manufacturers may include other sensors' data under specific and undisclosed PIDs. In our case study, all vehicular sensors were collected using public PIDs.

Table 4.2: Data acquisition characteristics

	Vehicle 1	Vehicle 2
Engine	1.0 16v	1.6 16v
Max RPM	7000	7000
Gears	5	5
Power(hp)	76	122
Weight	1025 kg	1000 kg
Manufacturer	Renault	Hyundai
Trips	26	8
Drivers	5	4

To illustrate the general process of collection and preprocessing of raw sensor data, we will describe a case study conducted by the authors which involved sensor data from two different cars. Both vehicles were used in daily commutes conducted by multiple drivers in urban environments, and the trips were logged using a Bluetooth OBD adapter and a smartphone. Table 4.2 introduces characteristics of the collection process, such as the vehicles' and trips' characteristics.

Vehicular sensor data is subject to errors from two sources: the sensors themselves which are naturally uncertain and the collection process that suffers from communication and storage problems. In the face of this, it is important to preprocess data before submitting it to operations and models. When dealing with vehicular sensor data, the main problems [Rettore et al., 2016a]) that one should look for are missing data caused by communication absence or interruption when logging readings and outlying values from erroneous sensor readings because of communication problems or even sensor malfunction.

Incomplete data is a challenge when fusing sensor data since it may lead to incorrect assumptions and, consequently, conclusions. Virtual sensors may not

output correct values or even not work at all because their input data or part of it is missing or incomplete, hence the importance of identifying and treating this problem. Treating the data probabilistically may resolve incompleteness issues with the data. However it is not guaranteed that all vehicular sensor data will follow a known probability distribution. For our testing purposes incomplete sensor data is invalidated, so if a virtual sensor requires multiple sensors as input and one of these has experienced any problem that caused a missing value in a time interval, the virtual sensor will not be able to output values in this interval.

Identifying outlying values is a difficult problem and consists of a separate field of study, which advocates for its complexity and importance. A virtual sensor that takes outlying values might produce equally incorrect values, stressing the importance of detecting and treating these values before they are fed to virtual sensors. In our test data, outlying values were identified and treated manually, given the difficulty of these processes.

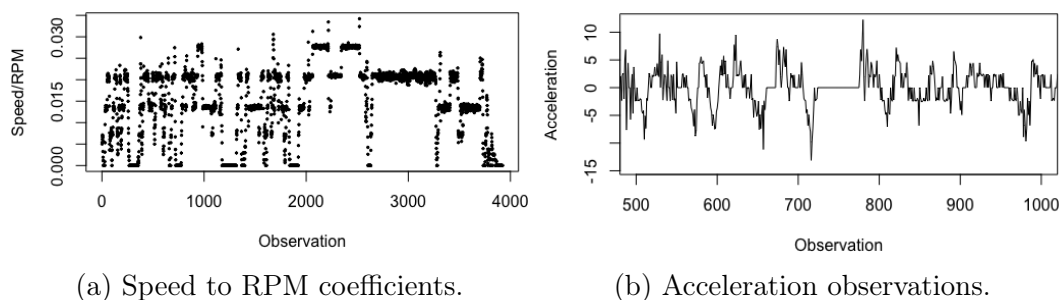
4.2.3 Operating Vehicular Data

In this section, we discuss the basic operations used to combine data from multiple sensors and create new information. Part of these operations was implemented in a framework we developed to allow rapid prototyping of the virtual sensors, we will further present, while other operations were used to treat and combine information in other related works. The operations are divided in three categories: (1) mathematical, (2) logical and (3) models, they are further discussed in the sequence.

4.2.4 Mathematical Operators

Arithmetic operators may seem simple and not used in a virtual sensing context, but they allow the creation of virtual sensors based on simple operations like sum, division, and derivation. These sensors measure aspects like a variation of a variable's values individually and related to other variables' and also produce transformed values, allowing them to be normalized and fitted to specific scales.

Figure 4.7 illustrates the results of calculations performed on raw sensor data to obtain information about the vehicle's gear and acceleration. Figure 4.7a



(a) Speed to RPM coefficients.

(b) Acceleration observations.

Figure 4.7: Example of calculated virtual sensors.

presents the result of the division $Speed/RPM$, used to investigate the relationship between these sensors' readings, controlled by the vehicle's transmission system. These results give us the signs of gear use when we observe different horizontal groupings. To further investigate driving behavior, an important measure is an acceleration, which is the variation of speed(S) on time(t), mathematically defined by $\Delta S/\Delta t$. The results of this division are shown in Figure 4.7b and will be further investigated in the upcoming sections.

4.2.4.1 Logical Operators

Logical operations are key to monitor values and combine conditions, that is, one might be interested in monitoring different variables for abnormal values to generate an alert, which is achieved by monitoring individual conditions and combining them to generate the higher level alert. In fact, in the vehicular context, monitoring as many aspects about a car's operation as possible is the way to identify and diagnose mechanical issues that may appear. The problem with monitoring values resides in determining limits and values to distinguish common situations from abnormal conditions.

Determining a limit for acceptable values from a given variable may be as simple as using arbitrary values, which is a valid approach for certain use cases. However, more refined applications of logical operators require also a deeper knowledge about the monitored variables and their expected values, which can be achieved using probability distributions and statistical tools that determine how likely is a value and how distant from common readings is it. Figure 4.8a presents the distribution of acceleration values calculated from speed sensor data, which roughly fits

a normal distribution represented by the red curve. According to the density function of this distribution, in a sufficiently large collection, only 8% of the values will be greater than 7 km/h/s and smaller than -7 km/h/s, thus, a reasonable value to distinguish abnormal accelerations and decelerations would be these limits, given their low probabilities.

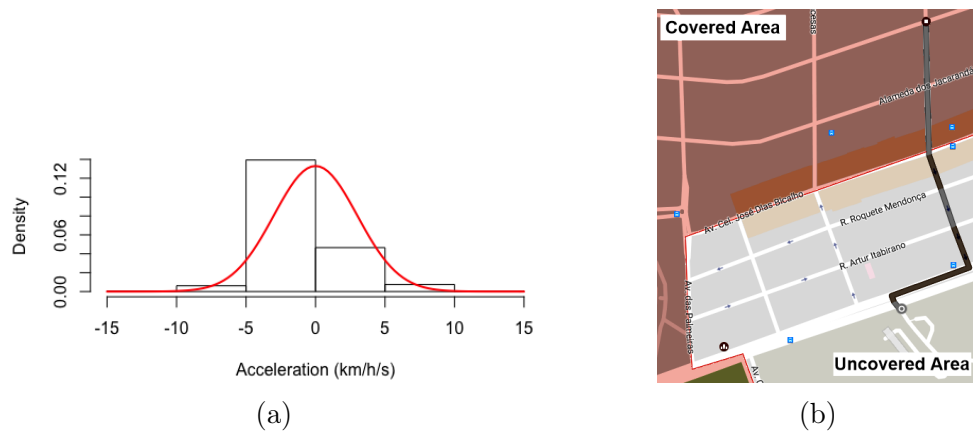


Figure 4.8: Distribution of acceleration values (a) and Route between areas delimited using the geofencing technique (b).

An example of a logical operator the application of conditions to location data is a technique called geofencing. The geofencing technique establishes a geographical region – real or imaginary – and determines which points or observations occur inside or outside this region. Among the applications of geofencing is monitoring an entity – in our case, a car – through time and determining when it was driving through a determined region. Figure 4.8b illustrates this example, determining a red region, which could account for the operation area of a transportation service where its vehicles are only supposed to traffic.

4.2.4.2 Models

The model category of operation comprises elaborate statistical and machine learning methods that transform raw sensor data to produce refined information about the vehicle, its driver, environment and context. These methods normally benefit from large collections of data, whether to distinguish different groups and categories or to train models that predict values from new inputs. Given this charac-

teristic, virtual sensors based on these operations will produce as better results as higher quality training data is provided.

Since collections of vehicular sensor data play an even more important role for these operations, it is necessary to discuss desirable attributes of data collections to produce better quality results. Naturally, larger datasets will improve virtual sensors' results since they are more likely to contain a larger diversity of situations, and it is expected to contribute to predictions and clustering accuracy. However, large collections of data may contain many observations of few events, instead of few – yet sufficient – observations of many different events, which will result in biased predictions. It may seem that comprehensive dataset are always desirable for predictive and clustering models. In fact, they are fundamental to detect as many variations of a given feature as possible. However highly focused collections are beneficial when looking for very specific and subtle variations, thus the decision between diversity and specificity is up to the designer, who should understand how their virtual sensor would benefit from each attribute.

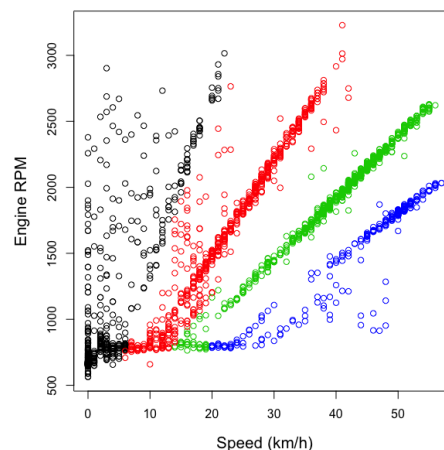


Figure 4.9: Speed and RPM relationship defined by gears.

Clustering algorithms gather individual elements according to one or multiple characteristics. Grouping sensor readings may highlight similar situations, tendencies, and profiles, which may count as simple insights on raw data or new and valuable information about vehicles and their operating context. An application of clustering techniques is illustrated in Figure 4.9, which shows sensor data from speed and engine's revolutions per minute sensors. The variables these sen-

sors monitor are related mechanically by the transmission system and its gears, that transmit engine revolutions to speed in different ratios. In Figure, there are four groupings of points, each of these represent an active gear, which is identifiable clustering these points by their revolutions to speed ratio.

Other types of algorithms and models also produce valuable results from raw sensor data. For instance, supervised learning algorithms are capable of identifying drivers based on a labeled set, a mixture of Gaussian probabilities is capable of identifying events based on collaborative sensor data [Chen et al., 2016]) and Kalman filters measure important variables to stability control systems [Wenzel et al., 2007; Boada et al., 2016b]).

4.2.4.3 Using Processed Data

In this section, we explore the uses of vehicular sensor data by virtual sensors. The examples we present will explore a vehicle’s operation state, drivers and context as these aspects influence sensor readings and, thus, are indirectly sensed. To develop the virtual sensors, we used vehicular sensor data captured using the OBD port as well as other sensors from smartphones that are absent in a vehicle (e.g., accelerometer).

4.2.4.4 Road Artifacts

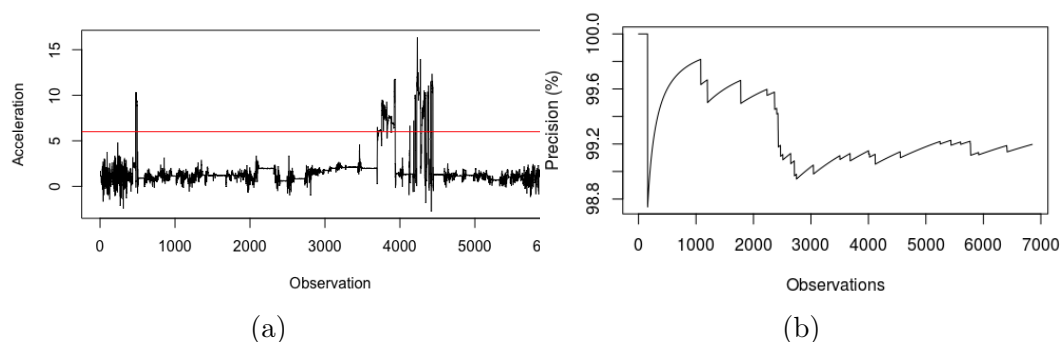


Figure 4.10: Accelerometer readings on trips (a) and Cumulative precision of driver identification (b).

Given modern vehicles’ ubiquity and their sensors’ variety and quality, they represent an important sensing tool for environments where they traffic. Direct use

of such pervasiveness is the ability to sense road artifacts in a larger scale to create a road state vision that will allow routes to be traced using better quality roads and city administrators to plan maintenance services where and when they are needed. Our sensor data collection system had access to accelerometers embedded in the smartphone, which lead to the basic identification of potholes.

Figure 4.10a shows accelerometer readings during a trip on which some potholes and rough roads were experienced. Higher acceleration values represent more intense vibrations sensed by the accelerometer and, thus, are more likely to represent an actual pothole or other disturbance on the road. It is important to notice that even though the readings captured are precise, a single trip is not enough to ensure the presence of an artifact. Numerous factors can produce similar vibrations to those of road artifacts and can produce false positive results. An alternative to reduce false positive results and ensure more accurate locations of road artifacts is described in [Chen et al., 2016]), which uses collaborative sensing to determine pothole locations on roads using multiple sensing vehicles.

4.2.4.5 Driver Identification

The set of steps developed to identify who is the driver follows three steps. The first one is to eliminate features, on the dataset, that contain missing values or that are not influenced by driver behavior (e.g., engine temperature, altitude). Secondly, we reduce the number of features, based on its variability, to eliminate correlated sensors data. Finally, we identify the drivers using supervised classification algorithms.

The processing steps were applied to vehicular sensor data to develop a new virtual sensor that identifies the current vehicle driver among a set of known drivers. We performed the Principal Component Analysis (PCA) to reduce the features to the most variable features. In the next step, we applied the moving average on the dataset and classified the drivers using the Extremely Randomized Tree algorithm. Finally, we output the current driver identity with an accuracy above 98%.

Figure 4.10b shows the output of driver ID sensor. As we can see, in the begin, the methods achieve 100% of precision and drops to over 98% while the

driver behavior resembles to the others along the observations, resulting in false positives.

4.2.4.6 Driver Behavior

The U.S. Department of Transportation's recently showed the number of deaths in motor vehicle crashes in 2015, which is above 35 thousand people [Administration, 2016]). They also argue that alcohol, speeding, lack of safety belt use and other problematic driver behaviors are contributions to the death in motor vehicle crashes. The driver behaviors vary considerably depending on age and gender, drugs consumption, the type of road used, distracted driving attitudes [Schroeder et al., 2013]), and other factors. For these reasons, the study of driver style has emerged to increase driving safety and, as a consequence, reduce deaths in traffic.

Considering as input data accelerations, breakings and turnings collected from accelerometer sensor of smartphone, it is possible to list its angular and lateral acceleration with the vehicle angular and lateral acceleration, once the smartphone is inside the vehicle. Then, different maneuvers can be detected by thresholds on these measurements. In that way, the rules to define a driver style can be defined by applying thresholds on the z -axis (representing acceleration and breaks), aiming to identify abrupt peaks that indicate aggressive increases of speed or harsh braking. Additionally, excessive speed in left or right turns is detected by thresholds on the x -axis acceleration, which outputs higher values in these occasions.

Figure 4.11 presents an application of these rules, as an example to identify aggressive and non-aggressive driver behavior. The virtual sensor outputs which kind of behavior the driver is having on each observation. As an example, we isolated the abnormal observations considering 90% under normal bi-variate density. The observations outside the ellipse can be classified as aggressive driver behavior, once the z -axis shows the acceleration and breaks with peaks between more than 3.5 and less than -0.5. Besides that, the x -axis shows different accelerations in the right and left turns, that can be associated with the vehicle acceleration to find evidence of vehicle losing traction, for instance.

On the other hand, the understanding of drivers' emotional state can pro-

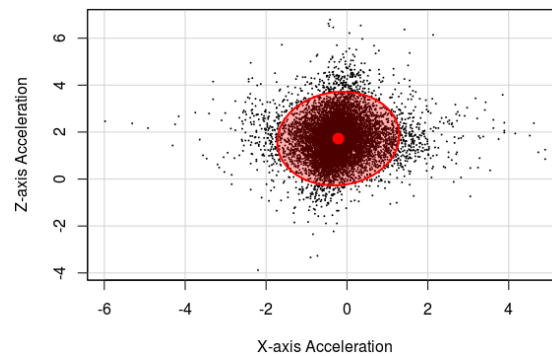


Figure 4.11: Smartphone accelerometer sensor with thresholds to determine driver behavior.

vide extra information about their driving style. Also, their feelings can be used to supply a vast set of recommendations. As an example, we can consider the environment temperature, the noise inside the vehicle and the driver sweating, and develop a simple rule that list these aspects to provide a virtual sensor that output, if the driver is getting dehydrated and needs to drink some water or turn on the air conditioner. In that case, the input data can be collected from the sensor as a microphone to detect the level of noise inside the car and indicate if the windows are open, wearable sensors on wristbands or headbands that measure and detect skin temperature, can be used to show if the driver is sweating, and finally the temperature sensor.

4.2.4.7 Good and Bad Driver

Telling apart good and bad drivers is a subjective task and quantifying this difference requires elaborate methods. In this example, we define a set of rules that may indicate the driving quality of a driver. The rules describe what is expected from a good driver in an urban environment. Thus, a driver will be judged as badly as many rules are broken at any given moment of a trip. The rules that will be used to define good driver are:

1. Speed values are below 100km/h
2. Acceleration between 7km/h/s and -7km/h/s
3. No driving after 23:00

4. No aggressive driving style
5. Engine revolutions below 60% of vehicle's capacity

Rules (1), (2) and (4) measure the driver's tendency to exceed speed limits, accelerate abruptly and generally behave aggressively in traffic, rule (3) is related to general safety, since crimes and accidents are more expected to be more frequent late in the night and rule (5) indicates conscious use of the vehicle's engine, which accounts for fewer maintenance costs and engine related problems.

To verify a driver's compliance to these rules, data from both physical and virtual sensor must be analysed. Rules (1), (3) and (5) are direct verifications of speed, time and RPM sensors using rules as described in section 4.2.4.1, rule (2) is verified by defining the limits also discussed in section 4.2.4.1 to the acceleration data calculated in section 4.2.4 and rule (4) is the interpretation of the virtual sensor presented in section 4.2.4.6.

Figure 4.12 presents an application of these rules to realistic data generated from our real sensor dataset. In this example, we enhanced sensor readings to force rule breaks and produce a scoring system that measures how many rules were broken. Since rules define the behavior of what was defined as a good driver, according to this scoring system, drivers will be as bad as many rules they break.

With this example, we showed that applying an elaborate set of rules to a dataset formed by both physical and virtual sensors can produce high quality and complex information. The definition of good and bad drivers, even though is based on threshold values for sensors, shows that is possible to measure abstract aspects of drivers, which may contribute to services like insurance models that charge customers based on how much and how good they drive.

4.2.5 Section Remarks

We presented the design process of vehicular virtual sensors, which are sensors that output new data based on input from other sensors and models or operations defined by designers. Modern vehicles have hundreds of accurate sensors distributed on their bodies for internal controlling purposes and the data these sensors output is available through a diagnostic port – OBD – that permits this data to be logged

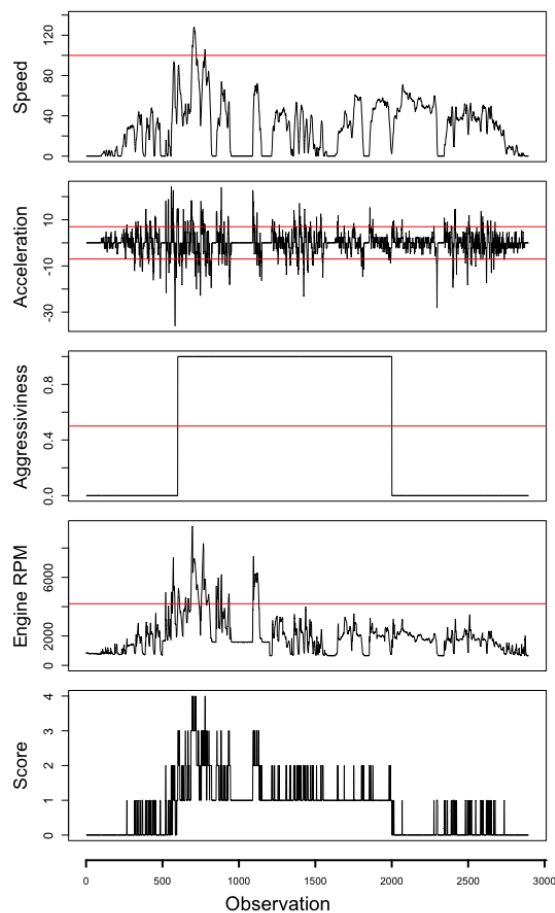


Figure 4.12: Instant precision of driver identification.

and processed. The variety and level of detail of vehicular sensor data allow virtual sensors to produce new, accurate, and complex information about the vehicle, its driver and context.

The design process of vehicular sensors was summarized into three stages: collection, operation, and presentation of new data. The collection process involves gaining access to sensor data using the OBD port and the problems related to sensor reading one might face when conducting a collection of vehicular data. To depict a collection process and the data related problems, we presented a collection we conducted using two cars and multiple drivers, the issues encountered and the solutions we used to minimize them.

Operating vehicular sensor data leads to various new information: from ac-

celeration rates to gear states. In this stage, we presented some operation forms that leveraged new aspects from our collection of sensor readings. For mathematical operators, we presented derivations from multiple data sources that produce insightful data about a vehicle's operation, for logical operators we showed the usefulness of determining a range of values – for sensor and location data – and also the importance of choosing adequate limits to isolate abnormal values. Finally, to exemplify the usage of models and algorithms on vehicular sensor data we developed a method to identify the current gear of a vehicle's transmission system by clustering sensor data.

The final step in the design process is outputting calculated data to users and other systems. In this stage, we presented examples of virtual sensors created using operations we defined. The sensors presented take advantage of the volume of data available using the OBD interface to extract information about a car's context and drivers.

In summary, Figure 4.13 shows how our design of fusion on VDS worked in this study. Where, the OBD vehicular sensors feed the fusion process, the data preparation deal with data aspects showed in Chapter 3, data processing covers the related methods, and finally resulting in a set of vehicular virtual sensors as the data use. As a result, the SM is benefited with the development of an approach to providing virtual sensors, which allows reducing the costs to embedded new physical sensors on the vehicle, decreasing its weight, hence reducing fuel consumption and emissions in a city.

4.3 A Method of Eco-driving

The development of actions to reduce fuel consumption and emissions and increase transportation systems' efficiency has become a huge challenge. Thus, a low-cost solution to improve fuel efficiency and reduce environmental damages is eco-driving, a group of behaviors focused on improving these aspects. Fuel consumption varies according to different factors: two different vehicles are expected to consume more or less fuel according to their engines' sizes or depending on the person who is driving them. In this section we present a gear virtual sensor for manual transmission cars, which adds information to understand drivers' habits,

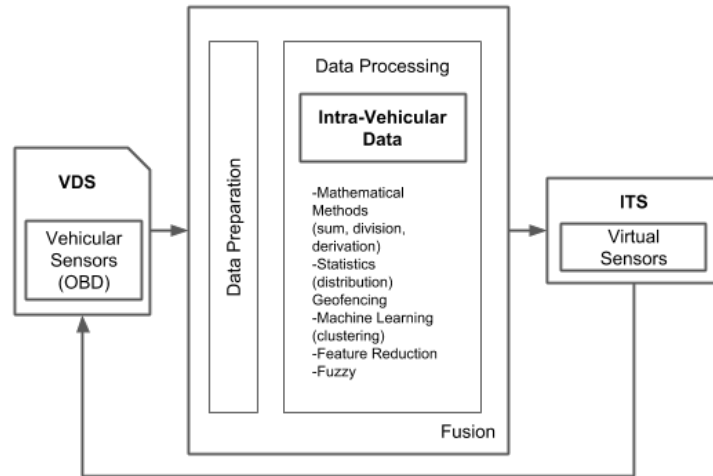


Figure 4.13: Design of fusion on VDS for vehicular virtual sensors.

allowing to analyze each gear individually about consumption. Our methodology developed gives the driver recommendations of the best gear considering speed and torque, reaching up to 29% averaged of efficiency in the fuel consumption and 21% averaged in CO₂ emissions reduction.

Fuel consumption is a factor that varies according to drivers' habits. Two different vehicles are expected to consume more or less fuel according to their engines' size. However, the same vehicle may behave differently depending on the person who is driving it. As an example, someone who drives a car aggressively and accelerates it more than another person who uses it more consciously is expected to consume more fuel. From an environmental - and even economic - point of view, it is desirable that drivers interact with their vehicles in a way that is as fuel efficient as possible, which reduces costs with refueling and greenhouse gases emissions.

Eco-driving is a set of types of behavior and techniques designed to reduce fuel consumption, which include recommendations on a person's driving style, the way and frequency a vehicle is used, its configuration, accessories, and maintenance. Eco-driving is part of a comprehensive approach to reduce the transport sector's contribution to a greenhouse effect. In order to increase a driver's fuel efficiency on a car, and thus, reduce gas emissions, we developed a method that analyses historical vehicular sensor data to suggest a gear shift that will result in

less fuel consumption.

Modern vehicles' control systems rely heavily on sensor data to control their stability and contribute to a safer driving experience. These sensor data are available through the OBD port. For the experimental setup of this work, we used Bluetooth adapters to record OBD data using smartphones.

Vehicular sensor data by itself does not present valuable information to the drivers since most of this data is used by the Engine Control Unit (ECU) to tune it and does not have a clear meaning to an inexperienced driver (*e.g.*, oxygen and fuel pressure sensors). Moreover, the portion of sensors that indicate meaningful information to the regular driver is naturally presented by the vehicles' gauges (*e.g.*, engine revolutions per minute and current speed). A challenge that arises is to present useful and valuable information as well as to provide services to drivers based on the readings of their vehicles' sensors.

This section presents a virtual sensor to provide a new service to drivers who share a common vehicle. The sensor identifies the current gear in a manual transmission vehicle. This information is useful to identify situations in a trip that increase fuel consumption. Having the gear information in a dataset of multiple drivers, we propose a method to give recommendations as to the best gear to drive at a given speed to improve the vehicle's fuel efficiency.

The remainder of this work is organized as follows. Section 4.3.1 presents the related work. Section 4.3.2 describes the collection process and characteristics of the data we acquired from our test vehicles. Section 4.3.3 discusses the steps and processes applied to our sensor data before using it. Section 4.3.4 explains the gear virtual sensor. Using the new sensor and other vehicle's sensors, we propose a method to recommend gear shift in Section 4.3.5, aiming for fuel economy. Section 4.3.6 shows the results of gear shift service simulation. Section 4.3.7 explore the applicability of recommendations system in a distributed scenarios. Finally, Section 4.3.8 presents our conclusions and future work.

4.3.1 Related Work

There are several studies in the literature related to driver behavior and efficient fuel consumption. Driving analysis is a topic of interest due to the increase of the

safety and efficiency issues in vehicles. Many companies are investing in specialized services of eco-driving to teach their employees, to reduce fuel consumption. For instance, Pañeda et al. [2016] characterized an efficient driving process for companies of the road transport sector. Their method allows ranking accurately each driver, allowing an individualized learning process, to reduce fuel consumption with a low investment.

The CGI Group Inc [CGI, 2014]), conducted a study based on more than 3 million Scania Truck trips, across seven European Union countries. They compare the impact of eco-driving coaching for different fleets and countries. Moreover, they proposed an estimated effect of coaching (EEOC), which provides a realistic estimate of the fuel savings to be gained from eco-driving coaching.

Corcoba Magaña and Muñoz Organero [2016] proposed a solution to reduce the impact of such events on fuel consumption. They developed a system to detect traffic incidents and provided an optimal deceleration that improved the fuel consumption up to 13.47%. Jeffreys et al. [2016] compared drivers in Australia who learned to apply fuel efficiency techniques to drivers who did not. They monitored 1056 private drivers over seven months, among them 853 drivers received education in eco-driving techniques, and 203 were monitored as a control group. The results showed that drivers who received eco-driving instructions presented a reduction of 4.6% in fuel consumption. Ruddy et al. [2014b] conducted a similar study in Canada, resulting in a decrease of fuel consumption and CO₂ emissions by up to 8%.

Differently, of the previous studies, we combined the efficient fuel consumption approach and the driver identification to achieve a personalized eco-driving recommendation service better. This allows introducing game strategies as ranking users of the same car based on their efficiency, for instance.

4.3.2 Data Acquisition

Nowadays, modern vehicles have high technology embedded systems to improve their driving safety, performance and fuel consumption, the latter is measured in Kilometer per Litre (KPL). To achieve these improvements, manufacturers have invested both in quantity, and quality of sensors vehicles possess. Currently, a

vehicle collects information from hundreds of sensors that are connected to the ECU through an internally wired sensor network [Qu et al., 2010]) and the data they output are accessible using the OBD interface.

The data collected from the sensors in the car are available through OBD Parameter ID (PID). Table 4.3 shows some of the data collected from sensors whose readings are available using the combination of smartphone, vehicle, and virtual sensors. There are also other hundreds of sensors that can be accessed using PIDs – some of which are defined by the OBD standards and others defined by the manufacturers. In this work, we are interested in data collected from the vehicle and also data provided by virtual sensors.

Table 4.3: ECU data, smartphone and virtual sensors

Collected Data				
Smartphone		Vehicle		Virtual Sensor
Device	Trip	Torque *	Engine RPM *	Acceleration *
Time	Distance			
GPS	Fuel Remaining	Fuel Flow *	Speed *	Reaction Time
Location				
Speed (GPS)	Ambient Air Temp	Engine Coolant Temp *	CO ₂ Average *	Air Drag Force
GPS HDOP	Cost km Inst	Adapter Voltage *	CO ₂ Instant *	Speed/RPM Relation *
GPS Bearing	Cost km trip	KPL Instant *	Pedal *	Gear *
Gyroscope	Barometer	Intake Air Temp *	KPL Average	
Altitude (GPS)		Trip KPL Average	Fuel Level	

(*) selection to the data processing stage

Table 4.4: Characteristics of data collected

	Vehicle 1	Vehicle 2
Engine	1.0 16v	1.6 16v
Max RPM	7000	7000
Transmission	5	5
Power	76 cv	122 cv
Weight	1025 kg	1000 kg
Manufacturer	Renault	Hyundai
Model	Sandero	HB20
Trips	36	8
Trip Time	28 hours	3 hours
Type of Trip	Naturalistic	Controlled
Drivers	10	4
Gender	6 M, 4 F	2 M, 2 F
Age	25–61	20–53

For the experimental setup of this work, we collected sensor data from two vehicles shared between multiple drivers using Bluetooth OBD adapters and smartphone. Table 4.4 summarizes information about the vehicles and the collection process. An important aspect of the process regards the type of trips logged for both vehicles: all four drivers sharing Vehicle 2 were asked to drive through two different routes, whereas Vehicle 1’s drivers used it for various purposes in their daily routines.

4.3.3 Data Preparation

We conducted our analysis considering the premise of only using vehicular sensor data or variables calculated from them. The goal is to answer the following question: *"Are vehicular sensor data capable of providing information about drivers, their behavior, and even further, ways they could improve the vehicle's fuel consumption?"*

After that, to address our premise, we avoided data collected from a smartphone as shown in Table 4.3, which lists 14 features collected from vehicle sensor data. In this step, we also created extra data based on vehicle to provide more explanation about the vehicle and the driver's behavior. The work [Rettore et al., 2016a]) guided us to better understand vehicular data after processing it. This work leads us to eliminate and treat data problems such as outliers, conflict, incompleteness, ambiguity, correlation, and disparateness.

4.3.4 Gear Sensor

In combustion engine vehicles, torque is transmitted to wheels by a transmission system composed of multiple gears with different ratios. Figure 4.14 illustrates the different relationships between the engine's number of revolutions per minute (RPM) and the vehicle's speed, as measured in our test vehicles using the OBD system. In both graphs, points concentrate in multiple lines, which represent different gear ratios.

Even though the current gear engaged is valuable information to describe the driver's habits, it is not available in any signaling protocol of the OBD interface in cars with manual transmission. In order to identify the current gear of a vehicle through OBD data, we evolved our previous work [Rettore et al., 2016b]) to develop a solution based on clustering algorithms that explore the different gears linear relationship between speed and RPM, using a previous virtual sensor created from an instantaneous relation $Speed/RPM$. This method allows us to separate each group of points and label them to extract gear information.

Since the points belonging to the same gear are grouped in a strongly correlated group, their speed to RPM quotient is also close in value. Our method to label a vehicle trip requires a driver to supply a single dataset that comprises

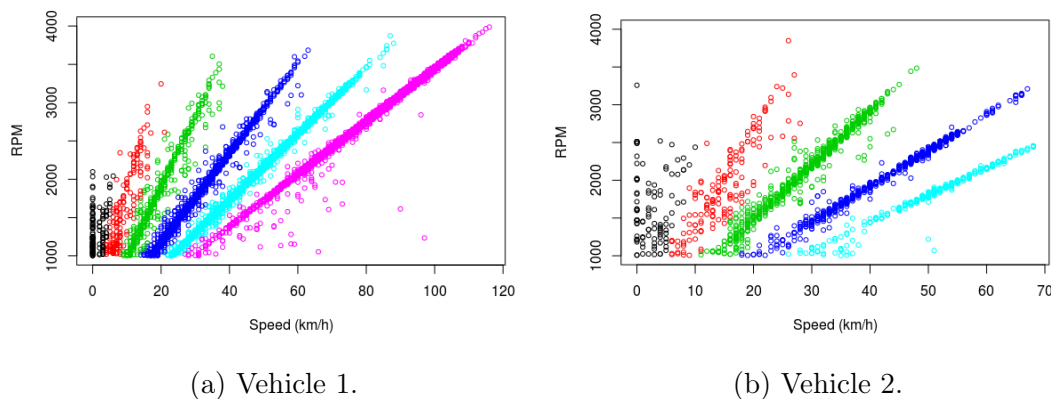


Figure 4.14: Correlation between vehicle's speed and RPM after clustering.

all gears of the same vehicle as training data. Having this dataset, a *k-means* algorithm clusters it in $n + 1$ groups, where n is the number of gears previously informed, and the extra gear state represents a situation where no gear is engaged. The outcome of this process is a new column in the dataset that indicates the current active gear of each observation, shown by different colors in Figure 4.14. Figure 4.14b presents another peculiarity, which is the absence of the fifth gear. Even though the vehicle has five gears, the last of them was not used in the trip that generated the plot.

4.3.5 Efficient Gear Change Service

Once there is data labeled by drivers about fuel consumption, it is possible to provide motorists with valuable insights regarding ideal gears aiming a better fuel consumption. The recommendation is based and targeted solely on fuel consumption data. Thus other aspects of vehicle operation are not taken into consideration. In fact, by accepting a gear change recommendation, the car is expected to have only a better consumption performance, which may have the opposite impact on torque availability and overall performance.

The process of recommending a gear shift is based on historical vehicular data, particularly the fuel consumption on specific gears and speeds. Given a current speed, gear, and consumption, the recommendation assesses whether there

is a gear for which average consumption is better than the current gear state. As illustrated in Figure 4.15, the recommendation map establishes a xy plane on z -axis based on the current vehicle speed and checks if there is a gear in y -axis for which the average x values of fuel consumption are better than the current setup. Besides that, the information of torque is also applied in a recommendation function. Another important point of that recommendation map, it concerns to isolate observations less than 1000 RPM. In general, these values are related to the synchronizations time between gears and add noise to the service.

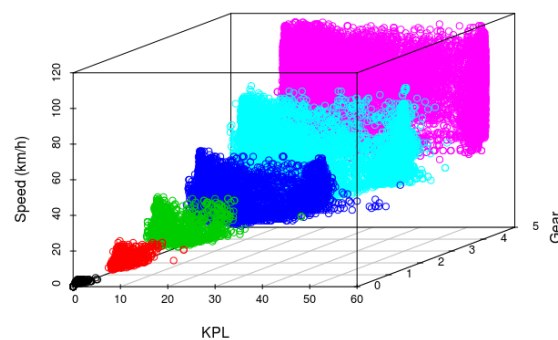


Figure 4.15: Speed and fuel consumption relationship for different gears of Vehicle 1.

It is important to notice that, since the recommendation process is based on historical data, for instant speeds higher than previous higher historical speeds, there is no recommendation available. However, as the process immediately includes the analyzed data in the historical dataset, new observations on the same speed will be eligible for recommendations. Figure 4.16 shows the historical scenarios of each vehicle regarding gear frequency at every speed. As mentioned in Section 4.3.4, the context of data collection is different in both vehicles, resulting in a different number of gears used between them. Another observation is the initial speed of the first gear, which has different values of 0 km/h. This situation is highlighted in the Vehicle 2 data, starting the first gear in 5 km/h, which reflects the recommendation restriction of 1000 RPM, or an inconsistent clustering performed in the previous step, which it can be explained by an insufficient dataset to label the data properly.

Relating the use frequency of each gear at a given speed, it is possible to

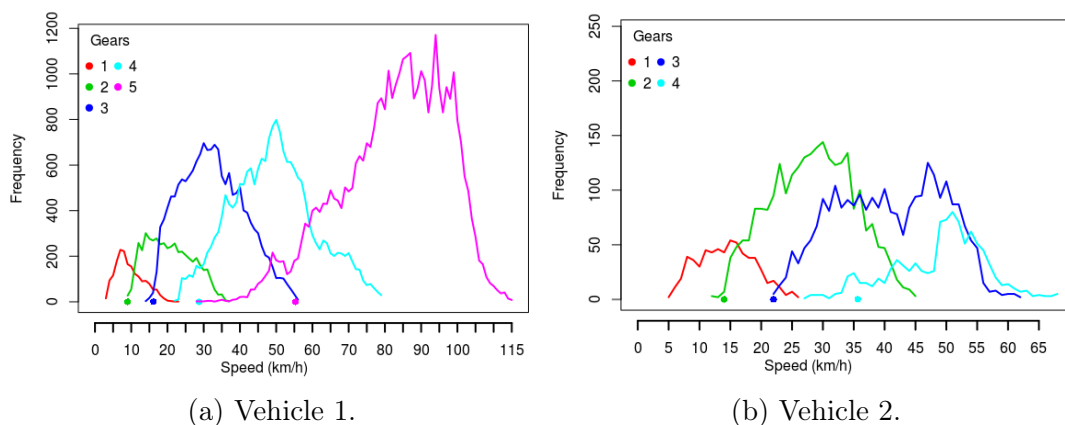


Figure 4.16: Gear frequency at a given speed.

observe the overlaps between them. This situation represents the use of different gears in the same minimum and maximum speed range. This information provides opportunities for recommendations, is based on economical driving or even driving to maximize the vehicle power. This work focuses on offering the driver the efficient fuel consumption. Thus, to obtain the speed limits of each gear, the equation (4.1) is applied, where the minimum speed of each gear ratio X is calculated, such that the *torque* is the smaller of each relation and the provided by the method. Moreover, also that minimum speed needs to respect the condition express in (4.1). The representation of these minimum thresholds between each gear is highlighted at the colored points on the x axis of Figure 4.16, where the torque is not less than 50%.

$$\begin{aligned} \minSpeed(x_{gear}, torque) &= \min(speed_x | \min(torque, \max(torque_x))) \\ \minSpeed(x_{gear}, torque) &\geq -2 * sd(speed_x) + mean(speed_x) \end{aligned} \quad (4.1)$$

This equation ensures that the minimum speed of each gear considers a specific torque, allowing that the recommendations relate to a medium power threshold of the vehicle. This threshold is dependent on the vehicle engine and, therefore, its generalization may not absorb the maximum efficiency of the recommendation method. For instance, vehicles with different engines react differently with the application of 50% of torque, i.e., the time to reach given speed and the final speed

in both vehicles are different.

After applying the recommendation method to the entire historical vehicular data, where the minimum torque is 50%, it is possible to note in Figure 4.17, a new frequency distribution of gears at a given speed. The average of instantaneous consumption of each gear indicates which gear best represents the fuel consumption ratio. It is expected that the higher gears represent this relationship. In other words, the recommendation seeks to advance the gear whenever the lower threshold of subsequent gears are reached.

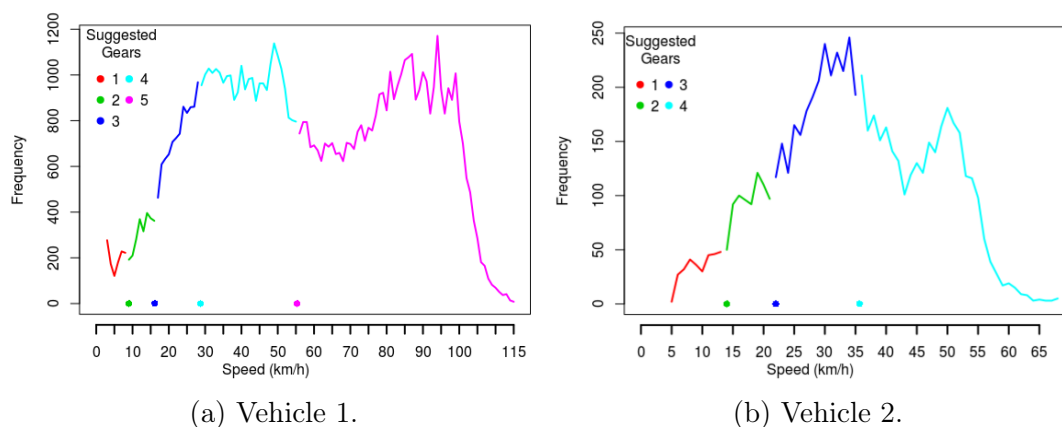


Figure 4.17: Gear frequency at a given speed.

It is important to note that torque was added to the recommendation system, to increase the effectiveness of the suggestion. However, the correct application of this torque depends on the individual characteristics of each vehicle, i.e., the torque required for the vehicle to move on rough terrain may vary due to these characteristics. In this way, this generalist approach may not recommend an effective gear, considering the characteristics of the terrain (strong descent and rise), for example.

In addition, to recommend gear shifts to improve fuel consumption performance, another strategy can take advantage of drivers identification, ranking users of the same car based on different parameters such as fuel consumption, aggressiveness, and vehicle care. This rank is usually present in games and strategies that use gaming elements to encourage multiple users to improve some desired aspect of their behavior.

4.3.6 Results

Table 4.5: Evaluation of gear recommendation system

Vehicle 1											
	Drivers										
	1	2	3	4	5	6	7	8	9	10	Total
KPL Average	14.81	14.63	14.15	15.40	15.38	10.84	11.24	11.66	11.86	12.68	13.27
KPL Average After Recommendation	15.34	15.95	14.70	19.57	16.32	13.82	13.50	14.88	13.86	13.80	15.17
Fuel Economy (%)	3.56	8.98	3.92	27.04	6.10	27.55	20.05	27.55	16.89	8.87	15.05
CO ₂ Reduction (%)	3.83	7.65	5.04	18.65	7.43	19.62	14.34	16.47	12.79	6.22	11.20
Vehicle 2											
	Drivers										
	1	2	3	4	Total						
KPL Average	6.67	6.81	6.38	6.60	6.61						
KPL Average After Recommendation	9.49	7.59	8.31	8.73	8.53						
Fuel Economy (%)	42.27	11.54	30.32	32.23	29.09						
CO ₂ Reduction (%)	26.36	12.20	23.89	24.52	21.74						

The evaluation of the recommendation system was made on each vehicle and drivers separately. The first step was to aggregate all trips performed by the different drivers, forming a unique set of data that characterizes the historical vehicle behavior. Then, the process of identifying the lower speed (based on a specific torque) threshold of each gear was performed, and also the average fuel consumption per gear. The next step was to apply the recommendation using the average fuel consumption per gear. Given the difference in final consumption between the original approach and the recommended approach, the final fuel consumption is estimated based on the overall fuel consumption average per trip. Table 4.5 presents the results for the gear shift recommendations.

The average fuel consumption and CO₂ reduction after the recommendation reached more than 15% and 11%, and 29% and 21% in the Vehicle 1 and 2, respectively, considering historical data. It is noted the situations where the recommendation resulted in significant improvements and situations where the improvements were not very significant. The lowest contribution of the recommendation (Fuel: 3.56% and CO₂: 3.83%) occurred with the Driver 1 of the Vehicle 1, and it is explained by the trips recorded, that is, the stored trips of this driver present mostly highways and, thus, the gear with higher frequency of use is the one that best presents the relation between fuel consumption and emissions, consequently.

On the other hand, the highest contribution of the recommendation system (Fuel: 42.27% and CO₂: 26.36%), with the Driver 1 of the Vehicle 2 is explained by the excessive use of a given gear exceeding the lower threshold of subsequent gear. The result of this behavior exploits as much as possible the recommendation of the greatest relation between gear and fuel consumption. The recommendation for Driver 4 of Vehicle 2 also achieved a high economy (Fuel: 32.23% and CO₂: 24.52%), explained by the trips in the urban environment, reducing the gear shifts and keeping high gear until the lower speed thresholds are reached for the gear reduction occur.

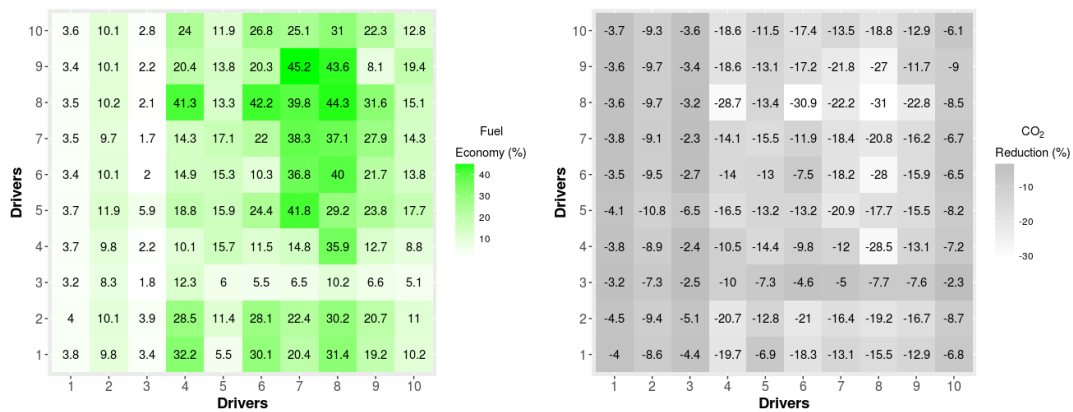
4.3.7 Collaborative Recommendation Service

Vehicular *Ad-hoc* Network (VANET) are an application of mobile networks concepts to urban environments, more specifically, vehicles. An important aspect regarding nodes in a VANET is their moving speed and communication radius, which results in short contact periods. However, despite short communication times, VANET can disseminate large volumes of data leveraging communication between vehicles and between vehicles and roadside infrastructure.

In a network through which vehicles can exchange information between themselves, and with infrastructure, vehicles will be able to share data to analyze fuel consumption and apply gear recommendation. Having access to information of different vehicles and drivers, it is expected that new driving profiles improve overall recommendation impact in the network. With driving habits information of other drivers, new gear utilization limits are expected to be discovered and, consequently, suggested to drivers that do not utilize them.

When recommending gear shifts to reduce fuel consumption, data from drivers of similar vehicles are desired given its positive impact and mechanical similarity. Our simulations used data from drivers in common vehicles as a base to calculate limits from which gear should be selected. By adding new data, these limits were different from those determined with local vehicular sensor data because other drivers utilize different gears in lower speeds. As a consequence, simulated fuel consumption from collaborative recommendation was lower than that from local data only.

Our service evaluation in a distributed context (*e.g.*, VANET) was conducted considering drivers of the same vehicle as different drivers in separate vehicles, thus, in our first case, there will be ten drivers in ten separate identical vehicles and our second case four drivers in four individual vehicles (see Table 4.4). Furthermore, we assume that in some moment all vehicles will have exchanged sensor information with each other and from this moment on, the recommendation system recalculates gear limits to suggest more fuel efficient gears to drivers.



(a) Fuel economy after driver pairs share data. (b) CO₂ emissions after driver pairs share data.

Figure 4.18: Pairs of drivers sharing data.

Figure 4.18a and 4.18b present a fuel economy and CO₂ emissions as seen by each driver, meaning that it shows the effect of recommendations as contacts between individual drivers happen. The contacts measure the fuel economy and CO₂ reduction from the source driver perspective, in other words, the contact between Driver i and Driver j enable the exchange sensor information, and a new gear recommendation is performed. Later, the recommendation is applied in both vehicles and evaluated in the point of view of the Driver i . An aspect related to local recommendation worth noting is that it applies to a single driver's historical sensor data and can also benefit from new collections of sensor readings. This occurs because the service looks for the lowest speed and torque to recommend a gear change and new lowest speed situations may appear due to behavioral changes or even new roads. We can see these situations in the diagonal, where $i = j$, and we have Driver i contacting themselves.

Our evaluation considers fuel economy and CO₂ emissions from network contacts which allow historical sensor data to be exchanged. However, in a wider perspective, individual fuel consumption improvements may lead to an overall greenhouse gases emissions and fuel consumption reduction. Moreover, in a Smart Cities, especially, ITS, such improvements may scale to cost reductions for drivers, suppliers, environment, and administrators of these systems.

4.3.8 Section Remarks

We propose a gear shift recommendation service aiming to improve the fuel consumption. To do so, we developed a virtual gear sensor for a manual transmission. Our method analyses the vehicle's historical sensor data to suggest a gear shift that results in more efficient fuel consumption. Our gear shift recommendation service reached up to 29% averaged of efficiency in the fuel consumption and 21% averaged in CO₂ emissions reduction.

In summary, Figure 4.19 shows how our design of fusion on VDS worked in this study. Where, the OBD vehicular sensors feed the fusion process, the data preparation deal with data aspects showed in Chapter 3, data processing covers the related methods, and finally resulting in a eco-driving suggestion as the data use. The recommendation system benefits from a distributed scenario, such as in SM in ITS layer, for instance. As the vehicle historical data is aggregated with the driver's behavior, our suggestion can identify the non-existent speed limits and, modify the previous recommendation. The benefit of these distributed scenarios is in the variation of this historic by vehicle. Being able to achieve the definition of eco-driving profiles for each vehicle in a network (VANET).

4.4 Driver Authentication in VANET

Given the number of vehicles traveling on the streets and highways around the world, new challenges and opportunities arise in the face of the progress of cities and society. Understanding vehicles' mobility can lead to better information about their efficiency, maintenance, and, in a broader context, traffic situations, events, and pollution. Moreover, modern control systems embedded on vehicles rely on

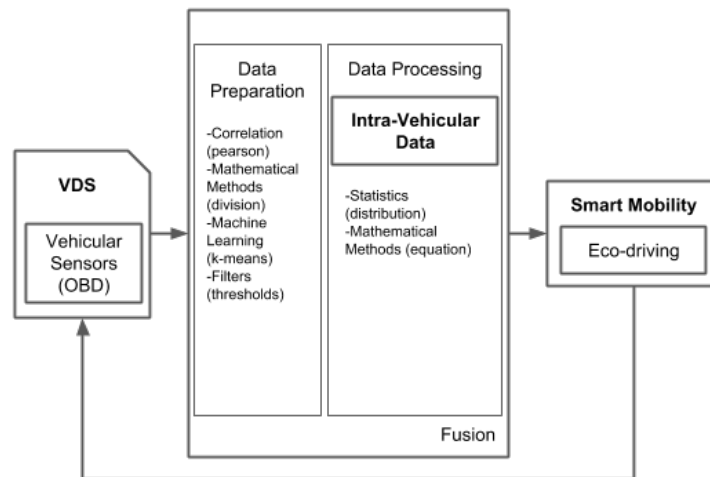


Figure 4.19: Design of fusion on VDS for eco-driving.

sensors to make the driving experience safer and more comfortable to the driver. Data from these sensors are available through the OBD port. Among the challenges associated with accessing such data is to present useful information as well as providing drivers with services and a vehicular network based on the sensors readings.

In this case, VANETs use vehicles' communication and sensing capabilities to provide applications and services with data from the surrounding environment. Moreover, a VANET contributes to the improvement of Advanced Driver Assistant Systems (ADAS) and ITSs, which offer a variety of services, including traffic safety, and comfort to drivers and passengers, such as access to social networks, video streams, and route suggestion. Many of these systems need to authenticate their users before providing them with content. However they do so in a way that an illegitimate driver can use the vehicle.

With this issue in mind, this work presents a Virtual Sensor (VS) to authenticate drivers based on their behaviors. This sensor is then used to differentiate a legitimate driver from a suspected one. The identification is treated as an extra factor to authenticate a driver and has two goals: to provide local services and network services. The VS uses data collected from embedded sensors to identify the person who is driving the vehicle, given a previously labeled dataset. Based on the driver's legitimacy, the VS can enable local and network services. To achieve these

goals, we employed a methodology to identify drivers, with over 98% accuracy. We also demonstrated that the presence of illegitimate vehicles might compromise the quality of essential services provided by VANETs, once they are capable of modifying the data which is being disseminated to the entire network.

The remaining of this section is organized as follows. Section 4.4.1 presents the related work. Section 4.4.2 discusses driver authentication and concerns about data privacy and security. Section 4.4.3 describes the collection process and the characteristics of acquired data. Section 4.4.4 describes the process of data corrections and the steps to reduce the number of features used to identify the driver. Section 4.4.5 presents the Virtual Sensor (VS) to identify legitimate and suspected drivers, as well as its evaluation. Section 4.4.6 analyzes the results when a suspected driver disseminates data in a vehicular network. Finally, Section 4.4.7 presents the conclusions and future work.

4.4.1 Related Work

There are studies in the literature related to both the driver behavior and the driver identification. Driving analysis is a topic of interest due to its importance in providing safety in vehicles. In order to address it, several studies have focused on driving style recognition [Johnson and Trivedi, 2011; Fazeen et al., 2012; Meseguer et al., 2013; Bergasa et al., 2014; Engelbrecht et al., 2014; Vaiana et al., 2014; Riener and Reder, 2014; Castignani et al., 2015; Martinez et al., 2016; Hallac et al., 2016; Kumtepe et al., 2016; Saiprasert et al., 2017]). Some of them identify who the driver is, whereas others classify the driver behavior as aggressive and normal, for instance. Zhang et al. [2016] developed a driver identification model using sensors available on a smartphone and the vehicle, through the OBD. They evaluated three vehicles in two different environments, controlled and ordinary. Considering only the vehicular sensors, the classification model obtained an accuracy of 30.36% in the controlled environment with 14 drivers and 85.83% in the ordinary environment with two drivers per vehicle. In contrast, we evaluated two vehicles in both environments with five and four drivers, respectively, and we obtained an accuracy above 98%.

Carmona et al. [2015] proposed a novel tool to analyze the driver behavior, providing detection of aggressive maneuver in real time. Aoude et al. [2011] developed algorithms for estimating the driver behavior at road intersections. They introduced two classes of algorithms that classify drivers as compliant or violating. They also validated their approach using ordinary intersection data, collected through the US Department of Transportation Cooperative Intersection Collision Avoidance System for Violations (CICAS-V). Ly et al. [2013] showed that there is a potential in using inertial sensors to distinguish drivers. The feature acceleration did not play a significant role in this, but the features associated with braking and turning events showed the opposite, the use of these sensors can potentially identify a driver.

Other studies aim to strengthen the authentication between drivers and vehicle. Most notably, some studies propose mechanisms to authenticate drivers based on biometric features. For instance, Yuan and Tang [2011] proposed an authentication mechanism based on the driver palm prints and palm vein distribution. Similarly, Silva et al. [2012] proposed an authentication mechanism based on electrocardiogram (ECG) readings, using sensors placed on the vehicle steering wheel.

Similar to our work, Burton et al. [2016] used a simulator to monitor driving patterns. They monitor features like pedal pressure, average trip distance and the steering wheel pattern. They used Support Vector Machines (SVM) to identify and authenticate drivers based on the extracted data. Similarly, Salemi [2015] proposed an authentication mechanism based on data obtained through the OBD port. That work extracted seven features from the data and applied SVM to identify and authenticate drivers, obtaining an accuracy of up to 94%.

Our work differs from the previous identification and authentication proposals in the following aspects: it only considers data extracted from the vehicle itself (e.g., unlike Burton et al. [2016]) and considers the driver behavior instead of stationary biometric data (e.g., unlike Yuan and Tang [2011] and Silva et al. [2012]). Besides, our work differs from the Salemi [2015] in the adopted methodology to identify drivers, which obtained higher accuracy (over 98%). We also combine the authentication of drivers to provide customized assistance systems to legitimate drivers and the network itself.

4.4.2 Extra Factor for Driver Authentication

In this section, we discuss and propose an approach to authenticating drivers based on their driving habits. It is necessary to identify the drivers from a set of data of a particular shared vehicle. Once the dataset of individual drivers is labeled, identifying drivers is a classification problem. A driver authentication methodology enables new vehicular services, both internally and externally. Intra-vehicle services regard the customization of ADAS, such as entertainment, ergonomics, and fuel efficiency services. In the extra-vehicle services, a vehicular network may allow message exchange, entertainment and personalized route suggestions based on the vehicle's driver authentication.

The process of identifying drivers is divided into six stages. Given the collected data, the first stage prepares the data by correcting and eliminating variables that contain missing values or that are not influenced by the driver behavior. In the second stage, we use the Principal Component Analysis (PCA) to reduce the analysis space, keeping data with greater variability. In the third stage, we partition the data into a training base and a test base, considering both a random partitioning and a trip partitioning, which considers the start and end characteristics of each trip. The fourth stage classifies drivers using the Extremely Randomized Tree (Extra-Trees) algorithm. At the end of this step, it is possible to identify the driver and provide data for the next stage that verifies if the driver a legitimate one.

The fifth stage disregards the real driver identity and classifies the driver as authentic or suspect. Finally, in the sixth stage, we perform an exploratory analysis to try to improve the classifier accuracy. To do so, we treat the input data in different ways to analyze the classification response. We use the raw data (without data treatment), data normalized and data with windows between 30 and 180 seconds with a moving average. In addition, the importance of each variable is checked using the random forest algorithm package, which maintains the variables that most contribute to the prediction accuracy. Figure 4.20 shows the identification flow to identify a legitimate/suspected driver.

It is worth mentioning that these steps describe the methodology that supports this proposal. This work uses the vehicular sensors themselves to determine

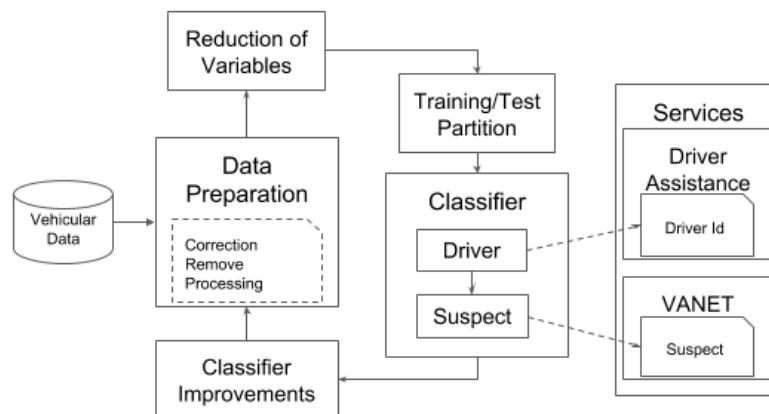


Figure 4.20: Identification of a legitimate/illegitimate driver.

the driver identification, and, consequently, allows to enable (or not) local and network services, with the authentication and identification of suspects, differently of Salemi's work [Salemi, 2015]) which does not focus on ADAS or VANETs services, for instance.

4.4.2.1 Privacy and Security of Vehicular Data

Currently, the main authentication mechanism between a driver and the vehicle is its key. In this mechanism, the key acts as an authentication token: any user with the token is considered legitimate. This mechanism is highly insecure since the token can be stolen together with the vehicle, granting illegitimate full control over the vehicle. For instance, an intruder with the ignition key can access sensitive private data from the drivers like their route preferences and exchanged messages. The illegitimate can also use the stolen vehicle to attack the network, impairing routing systems (by spreading fake messages) or driver safety systems (dropping or ignoring safety messages).

One of the goals of this work is to strengthen the security of the authentication system, using the driver behavior as a second authentication factor. The advantage of this approach is that authentication becomes based on features inherent to the driver, something an illegitimate cannot steal or replicate. However, because it relies on the driver behavior, the solution becomes reactive, identifying an illegitimate only after he/she bypasses the primary authentication mechanism.

At this point, blocking the illegitimate driver access to the vehicle becomes unfeasible as it may cause an accident or harm the transport system as a whole. Still, the identification of an illegitimate driver through the mechanism proposed in this work enables a set of security measures. These features may be both intra-vehicular (e.g., limiting the maximum driving speed) or inter-vehicular (e.g., notifying an insurance company, the vehicle owner or the police of the theft and the current location).

At any rate, the best course of action is to allow the vehicle to block the illegitimate access to ADAS partially. In this approach, all applications that are not vital for the vehicle or the network are blocked. That is, all entertainment and comfort applications, as well as applications that contain sensitive information, are affected. Again, messages related to the driver safety and the vehicle location cannot be blocked due to the risks to other drivers. To complement this approach, we also proposed that the vehicle periodically warn others whenever the current driver is illegitimate. Upon receiving this warning, neighboring vehicles forward the alert to others until it reaches a proper authority, who can take the appropriate measures.

4.4.3 Data Acquisition

The collection process uses the OBD-II interface as the means of accessing the vehicle data, transferring them via Bluetooth connection to a smartphone with the Android, where the data is processed and stored through an app. Table 4.3 shows some of the data collected from sensors whose readings are available using the combination of mobile phone, vehicle, and VSs. We are interested in data from the vehicle and also data provided by VSs.

Moreover, we aim to answer the following question: *Is the vehicular sensor data capable of identifying the driver, based on its behavior?* Thus, we focus on the data collected from both vehicle and VSs, which are designed using existing physical sensor data. A VS receives as input data from different physical sensors and eventually other data sources, to generate more sophisticated data using an algorithm. For example, the OBD interface may not provide a current gear of the vehicle to its driver. Thus, we can design a VS that receives data from physical

sensors, such as speed and motor revolution per minute, to infer the gear at a given instant.

In this work, we also performed a case study to answer the question above, using sensor data collected from two vehicles shared by fourteen drivers. Table 4.4 presents the setup of the data collection process¹. An important aspect of this process concerns the types of trips recorded by both vehicles: all four drivers sharing Vehicle 2 were asked to drive through two different routes (controlled experiment), while the ten drivers of Vehicle 1 used it for several ends in their daily routines (natural experiment). The whole dataset size contains above to 90 thousand observations.

4.4.4 Data Preparation

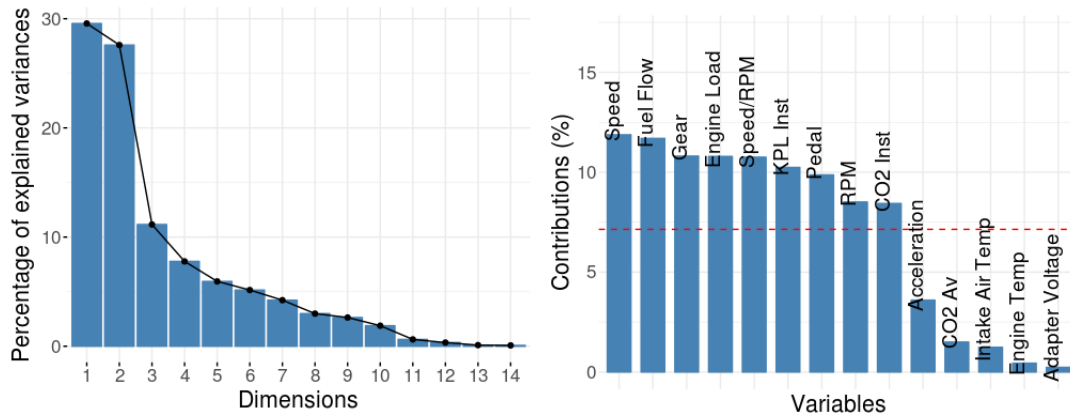
We conducted our analysis considering the premise of only using vehicular sensor data or variables calculated from them (VSs) in order to provide valuable information about the driver identity and behavior. Based on this premise, we discard the collected data that presents invalid values or does not reflect the driver behavior such as the air friction force and fuel level. Thus, fourteen variables out of 40 were preserved. Table 4.3 highlights the selected variables (*) for the next stage of data preparation. In that step, we developed the gear sensor [Rettore et al., 2017]) and performed the data treatment process [Rettore et al., 2016a]), which eliminated and treated data problems such as outliers, conflict, incompleteness, ambiguity, correlation, and disparateness.

The preparation stage treats and reduces the number of features. The latter is an important task, given that processing time tends to increase significantly with the number of dimensions of data. Thus, we first eliminated the features that contain missing values, which interfere in the next steps. Afterward, we used the Principal Component Analysis (PCA) to extract a set of relevant features. This process identifies the most variable information from a multivariate dataset and expresses it as a set of new features – Principal Components (PCs). These PCs represent the directions along which the variation in the data is maximal.

¹We encourage the community to explore the data acquired in this work, which is available at <http://www.rettore.com.br/prof/vehicular-trace/>, such as its description and further information.

The choice of PCA instead of Factor Analysis (FA) is due to the components are actual orthogonal linear combinations that maximize the total variance.

Figure 4.21a shows the percentage of variance in 14 PCs (number of evaluated features). The first principal component has the largest possible variance. In other words, the first PC contains as much of the variability in the data as possible. Each following component contains the largest possible variance smaller than its predecessor. The resulting vectors are an uncorrelated sorted set.



(a) PCs sorted by percentage of explained variance. (b) Data relevance considering the first two PCs.

Figure 4.21: The most representative variables of the dataset.

Considering the first two PCs, we can explain over 90% of the dataset variance as depicted in Figure 4.21b, which illustrates the features variance explained between this first two principal components (also called dimensions). The red dashed line would indicate the expected average value if the contributions were uniform. As we can see, each feature variance is explained by its contribution, and nine of fourteen features represent the most data variability. Therefore, these features can help to determine the driver behavior and his/her identity once these features vary between among the drivers.

4.4.5 Identification of Drivers and Suspects

A challenge in solving a machine learning problem is to find the right algorithm for it. That is because the best suitable algorithm depends on the set of data and the

problem. Therefore, the choice of an algorithm depends on the expected results, time constraints, data size, its quality, and nature. Based on these issues, we should solve them using tools that guide us to select a machine learning algorithm and its hyperparameters automatically. Thus, among the most known AutoML – Auto Machine Learning – tools, there is the TPOT [Olson et al., 2016]), a tool to explore thousands of possible machine learning algorithms and hyperparameter settings.

Before using the TPOT, we analyzed the data to split it. Two partitioning approaches were created: (i) Trips: all available trips were considered, dividing them into training (70%) and test subsets (30%). This partitioning considers the start of all trips as the training data, and the end of trips as the test data. It also allows capturing a more comprehensive set of behaviors for each driver between their trips. Due to its temporal observation. This dataset can identify the driver in different environments; (ii) Random: partitioning conducted randomly aims to eliminate the bias that may be introduced to the partitioning by the trips. Subsequently, the driver training and test data were grouped, resulting in one training base and one test base, respectively.

After performing TPOT, considering the type of partition, the best-chosen algorithm was the Extremely Randomized Tree or Extra-Tree (ET) [Geurts et al., 2006]). This algorithm is used to perform classification or regression and requires that all predictors to be numeric, and does not allow missing values. The Extra-Tree algorithm builds a set of unpruned decision trees, using a top-down strategy. Moreover, ET chooses randomly the cut-points and uses the whole learning sample to grow the trees.

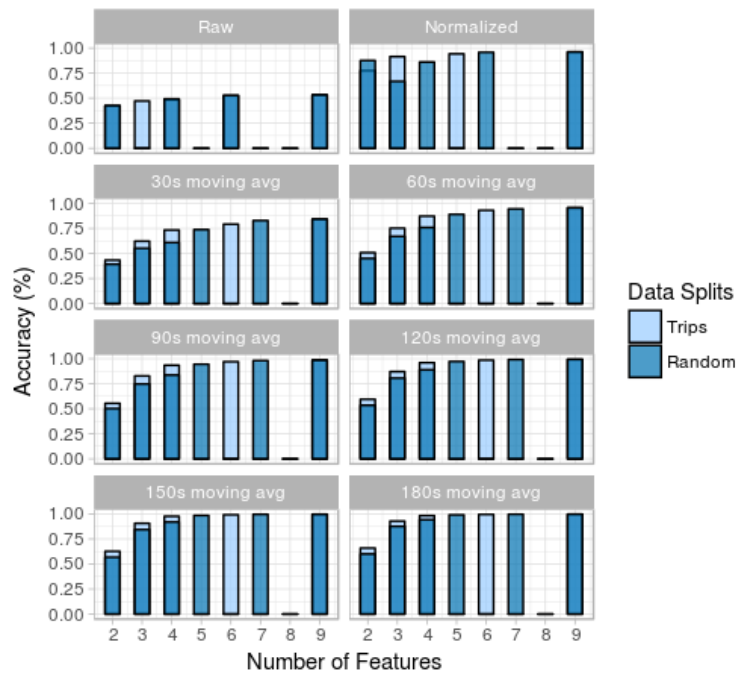
We evaluated the Extra-Tree algorithm regarding accuracy and number of features to determine a trade-off between them. Thus, we first performed the classification using raw data, but the results were not satisfactory to achieve our goal towards a personalized ADAS and network services. Consequently, we evaluated nine features to reduce them, based on their importance to the classifier. To do that, we used the feature importance metric included in the standard random forest packages. One way to calculate it is by counting the number of times a data pass through a node whose decision is based on a given feature. Using its frequency, we calculated the feature contribution to the prediction function.

We also applied a temporal window observation, similar to [Zhang et al., 2016; Carmona et al., 2015; Aoude et al., 2011]), to process the dataset and create a new subset, which is averaged by moving the average window. In that way, we explored the sizes of the moving average and its importance to the classifier. We evaluated raw data, normalized data and the moving average considering 30, 60, 90, 120, 150 and 180 seconds of observation, as well as two to nine features. Besides, the two data partitioning metrics were used (trip and random) to assess the validity of the approaches. These settings were chosen considering the importance of each feature, above 85%, to the prediction function. For that reason, the type of features are highlighted differently among vehicles, drivers, type of data treatment, and data split, absorbing the maximum description of drivers in a specific vehicle, making that process a customized approach to identify drivers and suspects.

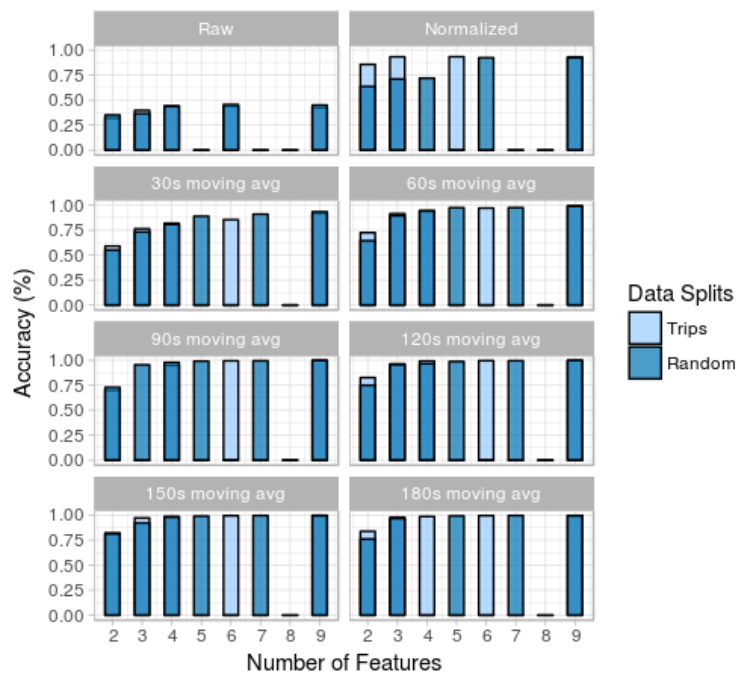
4.4.5.1 Evaluation of Driver Identification

Considering the trip data partition, we evaluated the classification method using the raw data (untreated) and observed that the accuracy reached 54% with nine variables, dropping to 43% when only two of them were considered, for Vehicle 1, as depicted in Figure 4.22a. Otherwise, Vehicle 2 showed 42% accuracy with nine variables dropping to 39% with two variables, as depicted in Figure 4.22b. After that, we analyzed the results for the normalized data, which aims to evaluate the classifier behavior and determine the ideal cut point for each vehicle. In that evaluation, Vehicle 1, with nine features, showed 96% of accuracy dropping to 77% with two features, whereas Vehicle 2 showed an accuracy of 93% dropping to 85% with nine and two features, respectively.

We also evaluated the dataset using the moving average between 30 and 180 seconds. This process allowed to increase the classifier accuracy and reduced the number of evaluated features. We noticed that the instantaneous sensor data makes the decision a difficult and confusing task. Thus, by applying a 30-second moving average to Vehicle 1, the accuracy was higher than 83% with nine features, 79% with six, 73% with four, 62% with three and 43% with two features, showing the same result of the raw data with two variables. By increasing the window size to 60 seconds of observation, there was an improvement in the accuracy that



(a) Vehicle 1.



(b) Vehicle 2.

Figure 4.22: Accuracy vs. number of features using different data treatment techniques.

reached 95% with nine features and 50% with two. This improvement continued as the window size increased, making it possible to identify the scenario where the classification accuracy reached over 98%. We considered the window size of 120 seconds for the moving average, resulting in 99% accuracy with nine features, above 98% with six and reaching 60% with two variables. This scenario repeats for Vehicle 2. However, it is possible to maintain a precision above 99% with only four variables.

That investigation showed the trade-off between accuracy and number of features. Because of that exploratory analysis, we chose the best relation for each vehicle, as being six features and moving average of 120 seconds of observation, for Vehicle 1, and four features and moving average of 120 seconds of observation, for Vehicle 2. This configuration led to an accuracy of 98% and 99% for Vehicles 1 and 2, respectively. When we considered both vehicles, the classifier accuracy achieved over 98%. We assigned this difference, and also the performance aspects (resources used – not discussed in this work), between these two vehicles to the use of different routes and the amounts of collected data. Vehicle 2 was used in a controlled route with eight trips and four drivers, and Vehicle 1 was conducted in ordinary routes with twenty-six trips and ten drivers. Besides, Vehicle 2 allowed a more significant variation of its driving, based on its superior motorization compared to Vehicle 1, resulting in a better distinction between the drivers.

The results allowed to define the best configuration of a classification method for each vehicle, leading to the development of personalized driver assistance services such as entertainment, ergonomics, route services and fuel efficiency services. Also, this result serves as an input to the suspect identification (illegitimate drivers) module, which aims to support the services in VANET, such as exchange messages between vehicles, entertainment, and personalized route suggestion. The data partitioning according to trips was considered part of the configuration step, contributing to improving the results, where we had a moving average of 120 sec, six and four variables for Vehicle 1 and Vehicle 2, respectively. These analyses and results depend on the setup step to record an initial driver data from the shared car.

4.4.5.2 Evaluation of Suspect Identification

We included suspects among known drivers considering that there is no knowledge about their driving habits. This condition results in suspects driving similarly to various legitimate drivers from a driver identification point of view. To simulate an illegitimate driver, each known motorist was treated as unknown at a time, and their data were removed from the training phase.

Using data produced by the classifier described and evaluated in Section 4.4.5.1 with ten drivers, trained with the full dataset, as well as with datasets missing individual drivers, it was possible to simulate and identify suspects driving vehicles. Inspecting the individual's behavior, it is possible to notice differences in its precision and results in distribution when mixed with legitimate and suspect data. Figure 4.23 shows the probability distributions in two cases: when driver 10 is identified in a trip and when the same driver is treated as an intruder in the dataset.

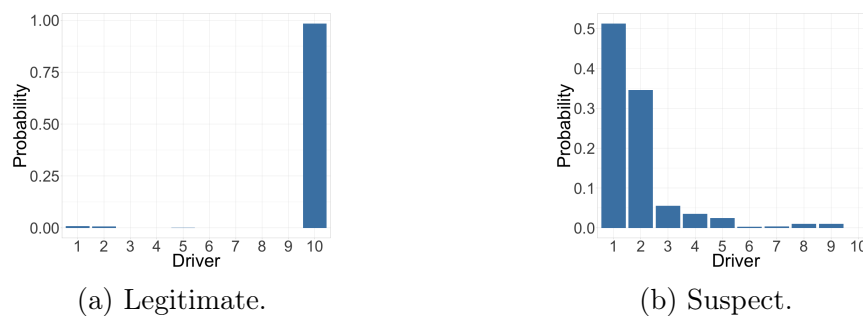


Figure 4.23: Classifier results when treating driver 10 as a legitimate and suspect driver.

Although there are visible differences between the distributions, they cannot always differentiate. Thus, we designed a new classifier to differ the probability distributions generated by the driver identifier when fed with an authentic or suspect data. This classifier takes as input probability distributions of all values obtained from the previous identification step. Training the second classifier with distributions generated by known drivers, as well as suspects, allowed us to identify suspects with over 99% precision correctly. An important aspect in this identification step is that telling apart known drivers and suspects is a task that

does not depend on data shared on a network, thus, allowing it to be performed offline.

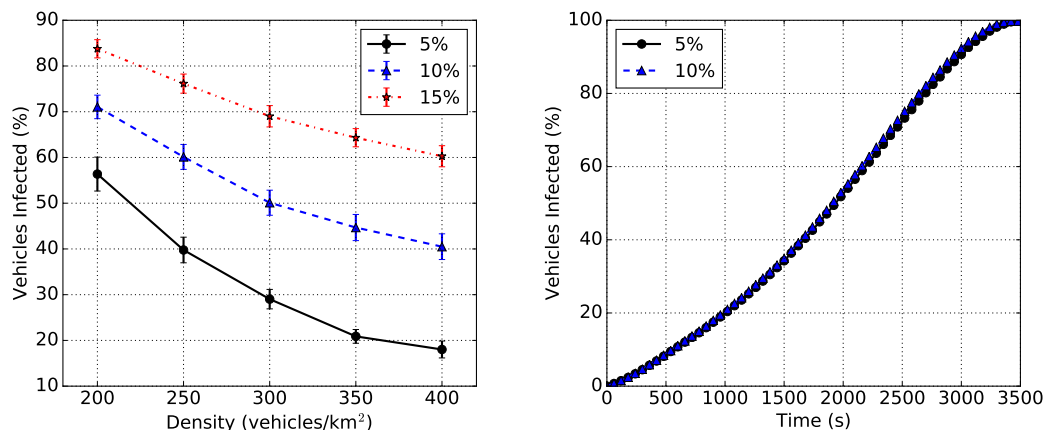
4.4.6 Suspicious Vehicles in VANETs

Aiming to assess the impact that suspicious vehicles might have on inter-vehicle communication services, we present a study considering two different scenarios. In the first one, a source vehicle, which is not a suspicious vehicle, disseminates 100 data packets to all vehicles in a Manhattan Grid with ten evenly-spaced vertical and horizontal double-lane streets in an area of 1 km^2 . Traditional flooding is used as the dissemination protocol. We varied the density of vehicles (200, 250, 300, 350 and 400 vehicles/ km^2) and the percentage of initially suspicious vehicles in the network (5, 10 and 15%).

In the second scenario, we considered a one-hour mobility dataset (6:00 am to 7:00 am) that covers an area of about 400 km^2 in the city of Cologne, Germany [Uppoor and Fiore, 2011]). Such dataset is realistic considering both macroscopic and microscopic viewpoints. We varied the percentage of initially suspicious in the network (5 and 10%). In this scenario, no data packets are being disseminated. Instead, vehicles exchange beacon messages with their neighbors at a rate of one beacon per second. It is worth noticing that in both scenarios, once a non-suspicious vehicle receives a non-duplicated data packet or beacon message from a suspicious vehicle, it also becomes a suspicious/infected vehicle. Our goal is to assess the spread of suspicious data on a VANET through multi-hop communication.

We implemented both scenarios using the simulation framework OMNeT++ 4.2.2, the IVC simulator Veins 2.1 and the mobility simulator SUMO 0.17.0. As main parameters, we set the bit rate at the MAC layer to 18 Mbit/s and the transmission power to 0.98 mW, resulting in a transmission range of about 200 m. We performed replications to reach a confidence interval of 95%.

Figure 4.24a shows the spread of infected vehicles during the data dissemination in the Manhattan Grid scenario. A vehicle becomes infected if it receives non-duplicated data directly from a suspicious vehicle or if it receives non-duplicated data that has been relayed by a suspicious vehicle during the dissemination pro-



(a) Data dissemination in the Manhattan grid. (b) Beacon exchanges in the Cologne scenario.

Figure 4.24: The spread of infected vehicles in VANET scenarios.

cess. As we can see, under lower densities of vehicles, the presence of a small number of suspicious vehicles (5%) results in more than 50% of vehicles becoming infected. As the density increases, the amount of infected vehicles decreases. This is because under higher densities the probability of having non-suspicious vehicles participating in the dissemination process increases. However, depending on the number of suspicious vehicles in the network, the number of infected vehicles can be over 40%.

Figure 4.24b shows the spread of infected vehicles during beacon exchanges in the Cologne scenario. Here, a vehicle becomes infected once it receives a beacon message from a suspicious vehicle or from a vehicle that has been infected. As we can see, even small amounts of initially suspicious vehicles in the network leads to almost 100% of vehicles becoming infected. This is due to the fact that as suspicious vehicles move around the city, they start to infect other vehicles, which will then infect other vehicles, thus reaching almost the entire network.

These results show that the presence of suspicious vehicles may compromise the quality of essential services provided by VANET. For instance, suspicious vehicles can modify sensitive data that is being disseminated to the entire network. Therefore, we can argue that being able to identify suspicious vehicles is paramount to the proper operation of VANET.

4.4.7 Section Remarks

Modern vehicles can communicate and sense their environment, which allows us to design a variety of applications and services to manage and provide greater security to people in transit, as well as comfort services for drivers and passengers. Many of these systems should authenticate their users to offer a directed content, but currently, they do not do so allowing a suspect driver to access and use those services.

This work proposed a VS to determine, locally, the driver of a vehicle at a given moment. We explored the driver identification as an extra factor of authentication to benefit driver assistance systems and vehicular networks services. The proposed methodology proved to be efficient and straightforward, maintaining its accuracy above 98% for a case study considering six features of Vehicle 1 and four features of Vehicle 2 with a 120-second moving average. The classifier was used to recognize legitimate and illegitimate drivers. We observed the different behaviors of the driver classifier when we submitted the legitimate driver data and the illegitimate one. This behavior reflects different probability distributions. The result of the trained classifier to distinguish between the two types of distributions reached precision above 99%. In addition, we discussed the importance of this approach in the VANETs context, simulating a scenario where the suspect driver is identified in the network and its potential impact on the data dissemination, since this suspect can modify the information, compromising the network.

Identifying who is the driver allows offering a personalized content and car adjustments to this driver. Considering the projections of SM, car-sharing will become a new mode for people move. Besides that, based on the driver preferences a more natural, fast, relax or low-cost route may be suggested. On the other hand, identifying a driver suspect may add a new and smart security layer to the ITS. Moreover, protecting services and applications which uses the VANET to broadcast its self.

In summary, Figure 4.25 shows how our design of fusion on VDS worked in this study. Where, the OBD vehicular sensors feed the fusion process, the data preparation deal with data aspects showed in Chapter 3, data processing covers the related methods, and finally resulting in a driver authentication as the data

use.

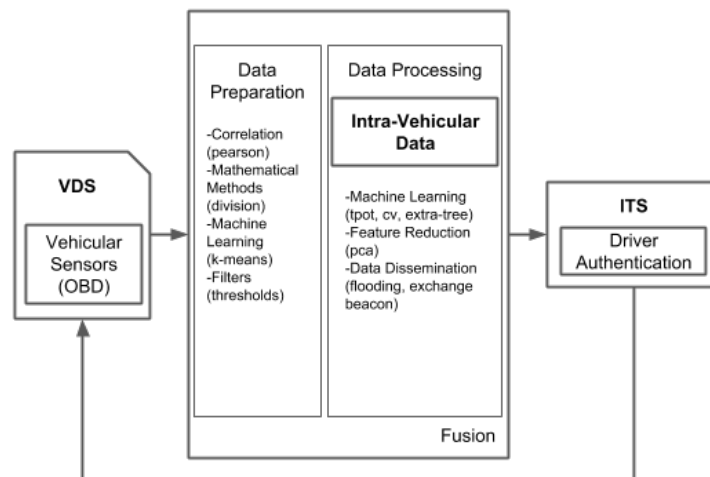


Figure 4.25: Design of fusion on VDS for driver authentication.

4.5 Chapter Remarks

In this chapter we proposed Intra-Vehicle Data (IVD) fusion approaches which consolidate the idea that there is a vast range of possibilities to develop applications and services, aiming comfort to drivers and passengers, infotainment, and safe driving. We noticed that, those applications need a correct and specific methodology to achieve their goals. However, we identify that the data preparation requires a combination of methods to first characterize the data such as statistical methods (data distribution, features reduction, mathematical methods, correlations), visualization methods, and filter to delimits the space of observation. After that, depending on the application goal a set of methods and techniques may be used. Although, we also noticed a trend to use machine learning techniques to deal with problems related to the ADAS, security, eco-driving and infotainment.

A topic that needs more attention is related to IVD privacy. Once the data comes from private vehicles the lack of data privacy reduce its availability and as a consequence more applications are developed to achieve a specific target, reducing its generalization capability and reach.

Chapter 5

Extra-Vehicular Data Fusion

As defined in Section 2.3.2 the extra-vehicular data corresponds to the subset of real and virtual sensors data that seek to describe the driver behavior or the environment around the vehicle by a variety of sources individually or fused. This section shows the Media as Vehicular Sensor (MVS), specifically the use of Location-Based Social Media (LBSM) to enrich the road data, allowing to explore the Smart Mobility (SM) opening new ways to build routes based on people preferences such as sentiment, event detection, and event description.

5.1 Enriching Road Data Based on Social Media

Nowadays, to plan and manage transportation systems are crucial tasks to promote the growth of a given city. Governments, researchers, and industries make efforts to understand mobility patterns in a city in order to develop solutions to reduce traffic issues and incident events [Bazzan and Klügl, 2013]. In this sense, an Intelligent Transportation System (ITS) emerges as a feasible way to improve real-time decision-making by leveraging the availability of information and communication technologies, thus providing applications and services to boost transportation systems. ITS depends on the availability of huge amounts of data and communication technologies. However, timely access to such data may present a limitation on the real-time traffic analysis performed by those systems, since only a set of companies have access to such data (e.g., data from inductive loops, traf-

fic cameras, semaphores, and origin-destination matrix) or it is often out of date. This happens due to the commercial value that such data have for companies, and to the deprecated infrastructure employed to deliver such data to end users. These facts become a barrier to better understand urban mobility and the transportation scenario, thus requiring other solutions.

The information delivered to users, especially traffic and road events, arrives with a poor description or even out of date, thus decreasing the efficiency of route management, flow control and the spread of detailed and useful descriptions of a given event. Overcoming these issues and leveraging the use of transportation system data to improve traffic efficiency demands multidisciplinary expertise. For instance, in order to provide consistent, accurate and useful information, integrating multiple data sources becomes an essential process. Such process is called Data Fusion and constitute a challenging task specially when fusing heterogeneous data, the asynchronous nature of data, and the presence of noise and errors on data. Furthermore, spatiotemporal aspects increase the complexity of fusing these heterogeneous data.

Based on that, the Location-Based Social Media (LBSM) (e.g., Twitter, Instagram, and Foursquare) combined with navigation systems (e.g., Google Maps, Here WeGo, and Bing Maps) has become an alternative data source to study urban mobility. Social media platforms allow users to share their thoughts, viewpoints, and activities related to their feelings about almost everything, which include traffic conditions. Different research issues can take advantage of an LBSM as a low-cost data source [Bazzan and Klügl, 2013; Yin and Du, 2016; Ribeiro Jr et al., 2012; Kim et al., 2014].

In this work, we investigate the traffic scenario in the lens of LBSM and navigation platforms. In this sense, we propose a robust framework named Road Data Enrichment (RoDE) based on heterogeneous data fusion. Our framework, depicted in Figure 5.1, aims to deliver high-level information to navigation systems, road planners and general public, once a set of data sources pass through data fusion models, thus providing services as route and incident.

The RoDE framework provides two main services: (i) *Route Services*: We propose the Twitter MAPS (T-MAPS), a low-cost spatiotemporal grouping to improve the description of traffic conditions based on tweets. We compare Twitter

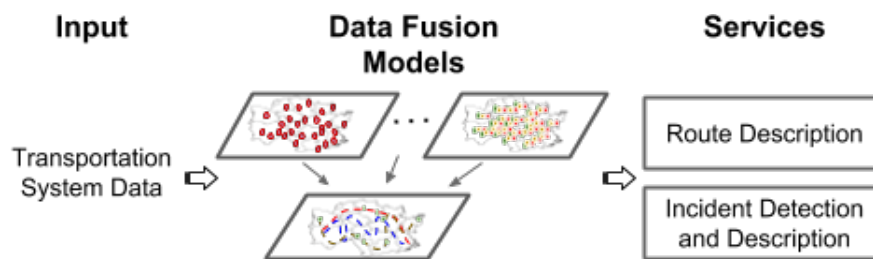


Figure 5.1: The design of RoDE.

MAPS (T-MAPS) routes with Google maps routes, and experiments show a high route similarity even though T-MAPS uses few and coarse-grained data. Moreover, we present three route description services over T-MAPS: Route Sentiment (RS), Route Information (RI), and Area Tags (AT) aiming to enhance the route information; (ii) *Incident Services*: We design the Twitter Incident (T-Incident), a low-cost learning-based road incident detection, and enrichment approach built using heterogeneous data fusion techniques. T-Incident enables incident detection and its description as RoDE services.

This chapter is organized as follows. Section 5.2 presents the related work. Section 5.3 details the first service of RoDE, Route Service, as well as the data collection process and its issues; the correlation between LBSM and traffic sensors data; the T-MAPS modeling process; a case study and the route description services. At the end of the route service, we present a short discussion. After that, Section 5.4 describes the RoDE: Incident Service and the data acquisition for such process; the incident data fusion approach that aims to enrich the incident data coverage; we explain the T-Incident design architecture, and the T-Incident evaluation. At the end of the incident service, we present a short discussion. Finally, Section 5.5 presents some concluding remarks and future work.

5.2 Related Work

The growth of the Internet and the proliferation of LBSM have enabled investigations on the huge amounts of data generated every single day. When considering the traffic and transit perspective, several studies have analyzed traffic conditions using LBSMs [Xu et al., 2018]. Many other studies focused on event detection

and diagnostics using Natural Language Processing (NLP) techniques [Ribeiro Jr et al., 2012; Crooks et al., 2013; Hasan et al., 2017].

Other studies performed sentiment analysis using LBSM data [Bertrand et al., 2013; Giachanou and Crestani, 2016]. Kim et al. [2014] proposed SocRoutes, a safe route recommending system, based on Twitter data. Unusual traffic events, based on social media, was investigated in [Giridhar et al., 2017]. Septiana et al. [2016] categorized road conditions with an accuracy up to 92%. Gu et al. [2016] explored tweets text aiming to extract traffic incident information providing a low-cost solution to existing data sources. They validated the Twitter-based incidents using data from RCRS (Road Condition Report System) incident, 911 Call For Service (CFS) incident, and Here WeGo travel time.

Yazici et al. [2017] showed that tweets collected from regular accounts are more likely to be irrelevant, though they can capture events that have just happened. On the other hand, tweets from specialist accounts are more valuable and structured, which are better when they are used to identify incident events. Also, they showed that the combination of both sources leads to better results when dealing with event detection. In the same way, Zhang et al. [2018] complemented the incident detection scenario by using social media data. They showed that social media data can be useful as an alternative way to improve traditional methods to detect traffic events in real-time.

Nguyen et al. [2016] developed the TrafficWatch, a real-time Twitter-based system aimed to leverage traffic-related information for incident analysis and visualization in Australia. They also developed a case study to detect road incidents before the Transport Management Centre (TMC) Log Time and those that are not reported by it. Pereira et al. [2013] made use of a reliable media available by traffic management centers, NLP techniques, featuring topic modeling, text analysis to improve the accuracy in measuring the duration times of an incident. They showed that the use of this source improves the prediction of an incident by 28% rather than its non-use.

This work extends and advances our previous study [Santos et al., 2018], which showed that LBSM feeds may offer a new traffic and transit layer to improve its current comprehension. Differently from most of the related work discussed above, we take a step forward by providing a model to clarify the traffic condition,

based on heterogeneous data fusion, aiming to add extra information to current navigation systems. Besides, RoDE provides a set of route and incident services such as Route Sentiment (RS), Route Information (RI), and Area Tags (AT). We also detail the spatiotemporal grouping, the features extraction process, as well as the ground truth of the incident and non-incident data to conduct our learning-based model with the LBSM data.

5.3 RoDE: Route Service

In order to provide a useful route service, we conducted a study to understand the relationship between the real traffic scenario and the data provided by Twitter, a very well-known and largely used LBSM platform. Initially, we focused on the data collection and its characterization. Then, we proposed the Twitter MAPS (T-MAPS), which intends to enhance the current navigation context by connecting LBSM data in different ways, for example, by evaluating tweets frequency or users' perspective of a region of interest

5.3.1 Data Acquisition

We collected tweets from New York City (NYC) demonstrating its coverage and the traffic factor correspondence. Then, we proposed and evaluated the T-MAPS applicability by showing its route similarity against Google Maps route recommendations. We also provided three route description services upon T-MAPS: Route Sentiment (RS), Route Information (RI) and Area Tags (AT). The motivation of RoDE: Route Services comes from the desire to expand the knowledge about the traffic conditions, in order to provide a more detailed scenario. Such issue has been little explored in the literature. Some applications may be proposed using social media to describe the traffic scenarios, such as the indication of the route's condition, the intensity of accidents and more detailed information about road event. This information may enrich the user's transportation experience, providing better assistance for decision makers when dealing with urban mobility.

An important question emerges from the inherent subjectivity of enriching the traffic description. To the best of our expertise, there is no ground truth for

the best route. For that reason, many tools aim to offer their traffic viewpoint like Google Maps, Here Wego, and TomTom maps. The main reason which motivated us to develop the T-MAPS was the desire to demonstrate the potential of using LBSM data, as a traffic data. Also, we aim to encourage the design of new applications, models, and analysis of urban mobility using LBSM.

Our dataset consists of 353,807 tweets from twenty-one manually selected users' accounts. Those accounts are maintained by departments of transport, specialists on traffic and transit reports such as news channels or dedicated companies. The number of tweets with geotagging is 307,020, most of them in NYC. Here, we explored Manhattan where has 38,112 tweets. The dataset was collected during the last three months of 2016. The dataset does not contain regular users due to the high user bias in their tweets regarding traffic feelings. Besides, some aspects which involve the use of LBSM data are highlighted in Section 5.3.2.

Figure 5.2a shows the spatial coverage of tweets in our dataset. Most tweets are over the road network, i.e., if we do zoom in, it is possible to see the I-95 highway with tweets along its extension. On the temporal point of view, Figure 5.2b shows the tweets' density along the hours for @NYC_DOT, @TotalTrafficNYC, and @511NYC users. Note that some peaks of tweets appear during rush times. For more details about the data acquisition process, please refer to [Santos et al., 2018].

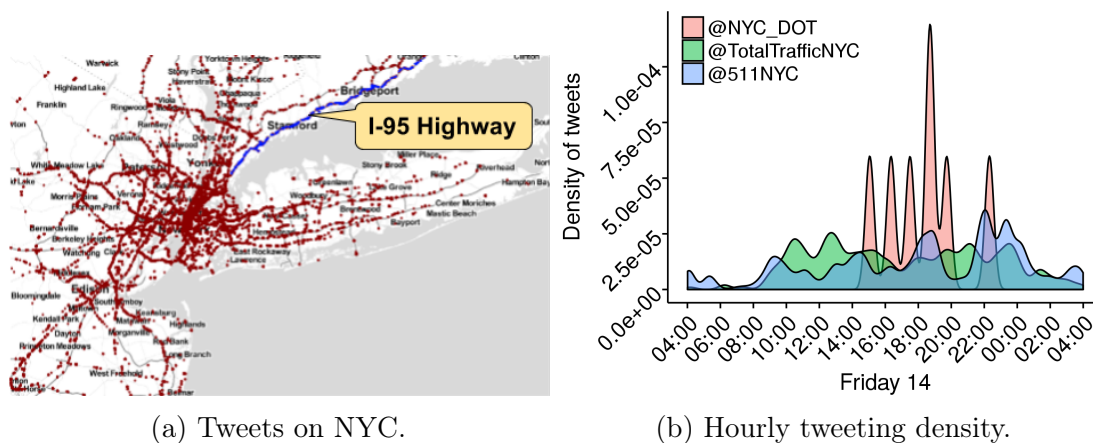


Figure 5.2: Route sentiment based on the tweets text analysis

5.3.2 What We Have Learned From The Data Aspects

Often, data from Twitter has aspects that lead to issues when using it on the traffic context. Here, we classify the data aspects into four classes: Data imprecision, User bias, Spatiotemporal assignment, and Inconsistencies. More extensive taxonomies can be found in [Rettore et al., 2016c; Khaleghi et al., 2013a].

Data Imprecision LBSM data comes with a certain degree of imprecision. Often, the data imprecision presents at least one of the characteristics: incomplete data, vagueness, granularity effects. The inherent heterogeneity of the data sources and “freedom” of data input on online platforms promote imprecision.

For instance, suppose the following tweet: “*Now 8:00 AM an accident at 100 W 33rd St #NYC #BadTraffic #creepedOut*”. One can obtain relevant knowledge about the event, e.g., the user’s sentiment, traffic condition, and the hour. However, the tweet lacks some information such as geotagging or event severity, being therefore *incomplete*. There are some techniques to mitigate data incompleteness. For instance, Pinto et al. [2017] proposed a record linkage approach to enrich incomplete data. Dubois and Prade [1994]; Yager [1982] used possibility theory and the probability of fuzzy events to handle imperfect data.

The *Vagueness* corresponds to an unclear description or data context. The above tweet shows vagueness due to the inability to precisely define the extension, position, cause or even those involved in the accident. Usually, a way to deal with vagueness is matching and fusing data from different sources.

The *Granularity* ranges from fine-grained to coarse-grained. In fine-grained data, it contains enough information to accurately describe the following items: event location, direction, the severity of accidents, and other information. Otherwise, coarse-grained provides a macro view of events with a broad description.

User Bias in the traffic and transit context, LBSM users can interpret the traffic congestion in different ways and use their freedom to post any information. For instance, suppose that Bob, a person from a small city, is in the traffic of a metropolis. Bob can interpret the regular traffic situation as a chaotic one, and then he posts on the online platforms his viewpoint. While Alice, a metropolis res-

ident, may understand as a typical situation. Consequently, the user’s perception may lead to bias introduction on data traffic from LBSMs.

The dedicated users (accounts which professionally report traffic condition) upon reporting traffic information can also introduce bias. Such users can, for instance, feed information for a specific audience or place. In this work, we picked manually dedicated users’ accounts to overcome regular users’ bias, but despite the diverse nature of users in the dataset (department of transport, news specialists, dedicated companies, so on), data may follow inherent bias of users interests and intentions.

Spatiotemporal Assignment the spatiotemporal assignment is a critical data aspect, particularly regarding traffic and transit context. The geolocation and temporal tagging allow traffic specialists to study and characterize a region at any instant or time interval. Below, we discuss some issues to extract the LBSM spatiotemporal information.

Spatial: it is fundamental to assign a location to the data, aiming to understand the context surrounding the information. However, deriving this information, even when present, is not always a trivial task. Suppose a tweet containing the spatial location in written form instead of a geotag, requiring a way to extract textual address location. Although such techniques already exist, the inherent unstructured form and freedom of writing (e.g., abbreviations, only 280 characters) on LBSMs turn a challenge the spatial textual extraction. Moreover, such particularities often result in information subjectivity or misinterpretation. There are research efforts to overcome these issues. Liu et al. [2011]; Finkel et al. [2005] used Natural Language Processing (NLP) techniques to obtain parts of speech and entity recognition to label sequences of words that are the name of the things. Li and Sun [2014] optimized NLP techniques to tweets text.

Information availability is another issue that affects the spatial data assignment. Some regions will have more spatial coverage than others due to several factors. For example, large cities tend to have higher spatial coverage than smaller towns. The cause of this may simply be due to the more substantial number of

users, companies and information traffic, or a complex social matter.

Temporal: associating a timestamp to the shared data is key to understand the past, present, and, possibly, the future scenario of the transport networks. LBSM platforms usually assign a timestamp when users input data to the system. However, this markup may not represent the same moment as when the event occurred. Thus, some open questions about temporal assignment are *What is the validity of data published by a user of LBSM? How can we characterize the delay between the event and the data input on LBSM platforms?*

Inconsistencies Here, we discuss two data inconsistencies: conflicts and out of order.

Conflict: the conflicting data from LBSMs appears when two or more data sources diverge about a specific event. For instance, suppose that Alice and Bob share their feelings about the same traffic event. Alice reports that nothing serious happened and the traffic flows well, while Bob reports that a severe accident happened which promotes a negative impact on the traffic. Based only on these two points of view, it is difficult to determine what happened. In the literature, the Dempster-Shafer evidence theory has gained notoriety in reducing data source divergences [Zadeh, 1984; Florea et al., 2009]. Also, it is possible to give a reputation weighting to users' accounts, and then apply rules to decide on the most credible information.

Out of order: the freedom offered by LBSM platforms allows users to enter traffic and transit information out of sequence into the system. These data appear as inconsistent to the systems that use them. Out of sequence data often is related to the temporal data dimension. For instance, a user may share information about a past traffic event. Therefore, we have to consider how to use such data properly. Usually, the trivial solution is to discard the out of sequence data. However, if the data was identified correctly and then sorted, it may be used as a feedback data at the cost of more processing and storage resources.

5.3.3 Twitter as a traffic sensor

To reveal the potential of LBSM data to enhance and complement the conventional ways to see traffic and transit, it is fundamental the understanding of how related the tweets are to the traditional traffic sensor. For example, if a conventional traffic sensor detects an anomalous event, can tweets explain such atypical event? In that way, to answer this question, we use the Jam Factor (JF) from HERE WeGo API ¹ as an aggregated traditional traffic sensor data. According to the Here documentation, the JF is a fused representation of traditional heterogeneous data. JF ranges from 0 to 1 (from free to congested). We chose Here JF since no other company provides such kind of data. We choose HERE WeGo JF due to the fact that other companies do not provide access to this kind of data.

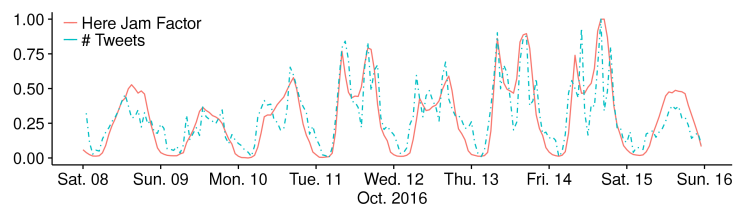


Figure 5.3: Tweets frequency and Here Jam Factor time series.

Figure 5.3 shows the correlation between Here JF and tweets in the dataset along a week in Oct. 2016. The time series in blue is the aggregated Here JF, and the orange one corresponds to the number of tweets. We re-scale the tweet time series to lie between 0 and 1, and aggregated each series hourly. Then, we observe that the curves are similar. We compute the Spearman's rank (ρ), a nonparametric correlation coefficient, to identify relationships between two variables. The ρ has a value between -1 and $+1$, where -1 means that the observations are entirely dissimilar and $+1$ the opposite. We apply Spearman's rank in the time series resulting in $\rho = +0.81$. It is possible to interpret that the #tweets tend to increase when the JF increases.

¹<https://wego.here.com>

5.3.4 T-MAPS Modeling Process

The T-MAPS is a low-cost spatiotemporal model which aims to clarify traffic events through tweets. This model allows the representation of the traffic scenario in different aspects by considering instantaneous or historical data, and its text mining. Below, we present the three steps of the modeling process as discussed in [Santos et al., 2018].

Data acquisition: this step consists of segmenting the area of interest and retrieving data from the LBSM platforms. We use a neighborhood segmentation to develop the T-MAPS approach.

Filtering and Data Fusion Process: this step aims to filter and bind LBSM data to the segmented region. We propose the use of a weighted time-varying digraph as a model to map these areas and data. The time-varying digraph is represented as a series of static networks, one for each time step. Formally, let R be the set of segments of the region, then a snapshot digraph is defined as $D_t = (V, E, m)$, where $V = \{r | r \in R\}$ denotes the segmented region, and $E = \{(u, v) \in V | u \text{ is adjacent to } v \text{ in } R \text{ segmentation}\}$ denotes the directed edges between physically connected regions, and m is the weights (discussed below). The T-MAPS time-varying digraph is a sequence of snapshot digraphs, thus T-MAPS(D) = $\{D_{t=t_{\min}}, D_{t+\Delta}, \dots, D_{t_{\max}}\}$, where t_{\min} and t_{\max} are the start and end time of the available dataset, and Δ can be adjusted conveniently.

Metrics: it consists of assigning cost weights to the directed edges. Formally, $m(u, w) : E \rightarrow value$, where $m(u, w)$ is a function mapping the directed edges to a metric cost. The metric function represents the analyzed traffic scenario using the LBSM data. Figure 5.4 illustrates a simple example of the T-MAPS modeling process. First, we segmented the NYC map into five regions of interest, then we collected LBSM available data. Next, we obtained the digraph $G = (V, E, m)$, where V is the set of regions, and E the directed edges between adjacent regions. Then, we bound Twitter's traffic data to the resulting regions graph. Finally, the weights are assigned to the edges using different metric functions. The resulting time-varying digraph allows us to analyze the traffic scenario condition and description. We present some metric functions below.

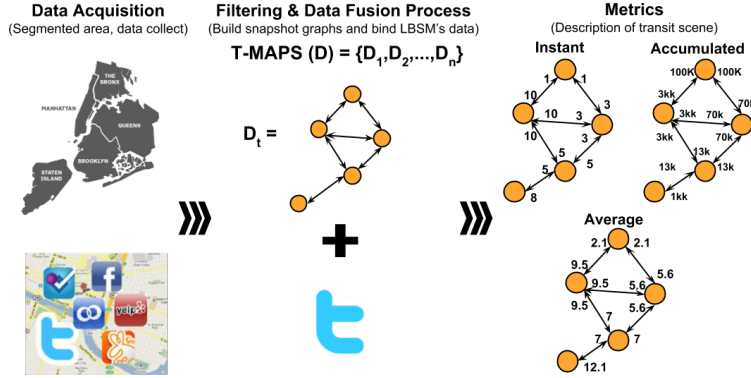


Figure 5.4: T-MAPS modeling process.

Instant: this metric function considers all tweets in each time t on a day by fusing and filtering them properly. This strategy corresponds to a snapshot view of the traffic at that moment. The smallest t must agree with the configured Δ of T-MAPS model. Usually, instantaneous data are sparse and cover poorly the region of interest. However, this data may highlight an event at a given time.

Accumulated: this metric considers all previously available data for a given time. It requires two parameters, t_{start} and $t_{\text{reference}}$, where $t_{\text{start}} < t_{\text{reference}}$ and must respect the temporal dataset availability. It accumulates all data between t_{start} and $t_{\text{reference}}$. One can interpret this metric as a historical metric looking to the past until the reference time point. In our experiments $t_{\text{start}} = t_{\text{min}}$.

Average: it uses the same approach of *Accumulated*. However, the values assigned to the edges are the average of tweets' occurrences over time, such as day, week and year. This information must be passed as a parameter to the metric function. One can interpret it as a typical traffic condition metric, putting into the account the historical information.

5.3.5 A Case Study

We conducted a case study to demonstrate the potential of T-MAPS. In that direction, we first compare the recommendation similarity of T-MAPS and Google Direction (GD) routes. Afterward, we present three route description services

demonstrating the T-MAPS potential as well as other opportunities to enhance and clarify the traffic scenario description. The Manhattan region was segmented into 29 official neighborhoods. Consequently, the T-MAPS digraph snapshot contains 29 vertices. Besides, the minimum time interval between two consecutive T-MAPS graphs corresponds to a $\Delta = 1$ hour. Although T-MAPS was designed to accommodate both data resolution (micro and macro), the case study used a macro viewpoint due to data coverage limitation.

5.3.5.1 T-MAPS Applicability

We evaluated the T-MAPS applicability by comparing its similarity, in recommended routes, with GD. Note that the T-MAPS route suggestion considers a macro resolution of the regions on the map, but our model is flexible enough to encompass fine-grained resolution if there is enough data for this. From a macro resolution, T-MAPS aims to recommend regions which have the best conditions regarding the applied metrics.

We query the T-MAPS and GD, 812 recommend routes in Manhattan neighborhoods. The routes were derived from the combination $2 \times C_k^n$, where $n = 29$ (Manhattan neighborhoods) and $k = 2$ (origin and destination). Note that we considered routes like $A \rightarrow B$ and $B \rightarrow A$. The routes start and end at the center of the region. Also, we rule out routes that start and end at the same region. We query the routes in three different moments (7:00 am, 3:00 pm and 7:00 pm of a day along one week, based on its rush hour representation).

The similarity technique measured the matched areas where the recommended routes by T-MAPS (using Dijkstra’s algorithm) and GD passed through. Figure 5.5 displays the similarity between routes along eight days in the dataset, considering three metric functions. The box-plots summarize 58,464 routes analyzed. T-MAPS with *Instant* metric showed a high variation of similarity rate, its median ranges from 50% up to 66.7%, while *Accumulated* metric shows 60% to 70% and *Average* metric 60% to 66.7%. It means that more than half of the evaluated routes overlapped the GD. We expected that *Instant* metric would pose the lowest similarity due to its intrinsic disparity with other metrics since it does not consider the historical data. As a global evaluation, the median of route similarity

reached 62% with Google Directions. Note that T-MAPS uses a macro view, while GD does not, which implies in fewer regions per route by T-MAPS than GD. The upper quartile (1/4 of the routes) until the maximum value exhibited a similarity between 75% and 100% between the T-MAPS and GD suggested routes.

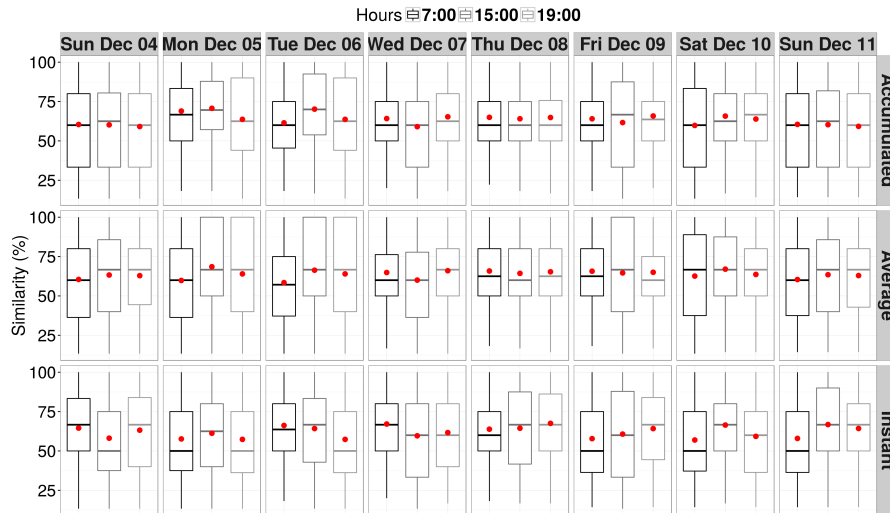


Figure 5.5: Route recommendation similarity between T-MAPS and Google Directions (dots represent the mean).

5.3.6 Route Description Services

Based on the applicability results, which demonstrated a possibility to aggregate extra information to a current route recommendation services, we move on to explore the tweet's texts. Initially, we performed the cleaning phase in the tweet (lowercase transformation, accents removal, tokens extraction, and filtering stops words, links, and special characters). Then, we applied three types of text mining to build the descriptions services over the T-MAPS model: Route Sentiment (RS), Route Information (RI), and Area Tags (AT). Figure 5.6 depicts a prototype to offer the T-MAPS services.

In Figure 5.6a, the RS service allows the user to observe the users' feelings (positive to negative) at a given area which they will pass through. The RI service explores each area providing a word cloud, Figure 5.6b, where the word size indicates its high-frequency over the route. The spread information enables the

users to see the big picture of highlight events in each area. Finally, we developed the AT service, Figure 5.7. For that service, we used the Term Frequency (TF) and Inverse Document Frequency (IDF) – (TF-IDF) – method to measure how important a word is to a set of tweets in given area of Manhattan. This technique allowed us to find words which are single for one explored area.

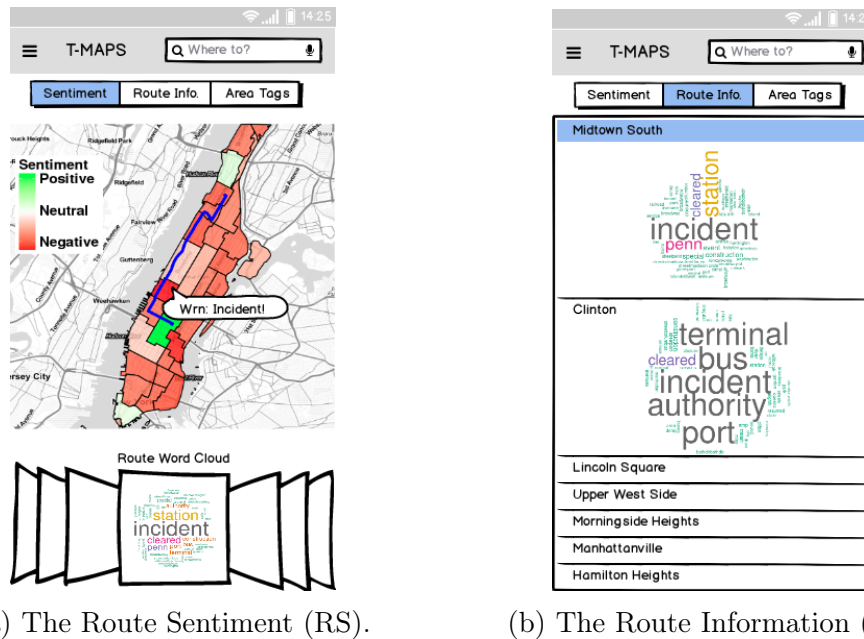


Figure 5.6: Route sentiment based on the tweets text analysis

The developed T-MAPS services used the *Accumulated* metric, aiming to characterize the Manhattan region, based on our observation window. Any other metric can be applied to provide a different description, achieving a different goal. With these services (sentiment, route information and area tags), the T-MAPS can enrich the current route recommendation systems, indicating to the users an extra path description or even providing routing based on these descriptions. For instance, the user may choose a route which expresses good feelings and beautiful environment. Alternatively, even routes with cultural activities.



Figure 5.7: The Area' Tags (AT) of each region of the path.

5.3.7 Discussion

In summary, the results of our RoDE are: *Route Services* showed the median of route similarity reached 62%, where T-MAPS uses region granularity while GD uses street granularity. For a quarter of the evaluated trajectories, the similarity achieved up to 100%. Also, we presented three route description services, based on natural language analyzes, Route Sentiment (RS), Route Information (RI), and Area Tags (AT), aiming to enhance the route information of current navigation tools.

5.4 RoDE: Incident Service

Once we have dealt with route services, we focus our efforts to improve current road incident event detection and description. We develop the T-Incident, a low-

cost learning-based road incident detection and enrichment approach built using heterogeneous data fusion techniques. For this purpose, we design a spatiotemporal grouping that fuses incident data from two different data sources (i.e., Here WeGo and Bing Maps), resulting in a new incident layer with more data coverage. Then, by using the same approach, we fuse (i) non-incident data (acquired from TripAdvisor), (ii) LBSM data (acquired from Twitter), and (iii) the new incident data layer obtained in the previous step. Moreover, we apply refined methods of NLP to extract patterns from social media data that may describe the incident event and its surrounding. Finally, we use a learning-based model to identify these patterns and detect the event types automatically. Thus, allowing the incident detection and its description as RoDE services. Notice that in our scenarios incident represents events which describes traffic issues such as accident, delays, weather, vehicle disable, and so on.

5.4.1 Data Acquisition

The lack of information in urban transport environments is one of the greatest challenges for those working in the transportation system area. Researchers are often restricted to theoretical studies or a short range of public data. Luckily, the current increase of online platforms, such as LBSM, make it possible for people to share their data, routines and opinions regarding a variety of aspects. T-Incident is an approach to accurately identify traffic events (incident and non-incident) and enrich their descriptions. The data acquisition process aims to combine different data sources, such as Here WeGo, Bing Maps², Tripadvisor³ and Twitter⁴ in both temporal and spatial dimensions to achieve those goals.

The dataset consists of 158,413 tweets acquired from 2018-09-14 to 2018-11-06. In that process, we crawled data from Twitter filtering tweets by set of words related to incident events, such as *congestion*, *accident*, *construction*, *planned event*, *road hazard*, *disabled vehicle*, *traffic*, *jam*, *car*, *weather*. All collected tweets are geolocated and most of them are in Manhattan-NYC. Moreover, we were interested in tweets from both regular (common accounts) and specialist

²<https://bing.com/maps>

³<https://tripadvisor.com/>

⁴<https://developer.twitter.com/en/docs>

(accounts controlled by corporations) users. We also discarded tweets posted as *retweet*. In other words, we collected the user’s impressions and not the spread of information.

LBSM data has several issues, as mentioned in Section 5.3.2, which we also deal with here. To collect as much incident events as possible, we acquired data from two different data sources: Here WeGo and Bing Maps. The incidents gathered from both platforms have temporal granularity of one hour. We have collected 9,784 distinct incidents acquired from Here WeGo and 1,924 distinct incidents acquired from Bing Maps. To use those incidents data, we fuse both data sources, filling the gaps that a data source has with the other one and vice-versa. Also, we combine common incidents from both data sources enriching them if possible, since each one can have different incident description (Section 5.4.2 details this process). All datasets overlap spatially and temporally.

Table 5.1: Data acquired from different data sources.

Source	Goal	Sample	Time Interval	Spatial Location
Twitter	Event Detection	158,413	2018-09-14 to 2018-11-06	Manhattan New York
Here WeGo	Incident	9,784		
Bing Maps	Incident	1,924		
Trip Advisor	Non-Incident	50		

In order to detect incidents, we also need to comprehend what is not an incident. First, we choose places with no incident evidence, collecting data from sources which deal with touristic places. For example, Tripadvisor, a travel website that shows places, hotels, restaurant reviews, and other travel-related content. Then, a set of the most popular places ranked by the tourists was chosen, such as museums, observatories, parks, pubs, theaters and so on. Table 5.1 summarizes the data collected and Figure 5.8 shows the spatial data coverage of each data source used to develop the T-Incident approach.

5.4.2 Incident Data Fusion

In this section, we present a method to increase the coverage of incident data and enrich its description by fusing data from different sources. We argue that the

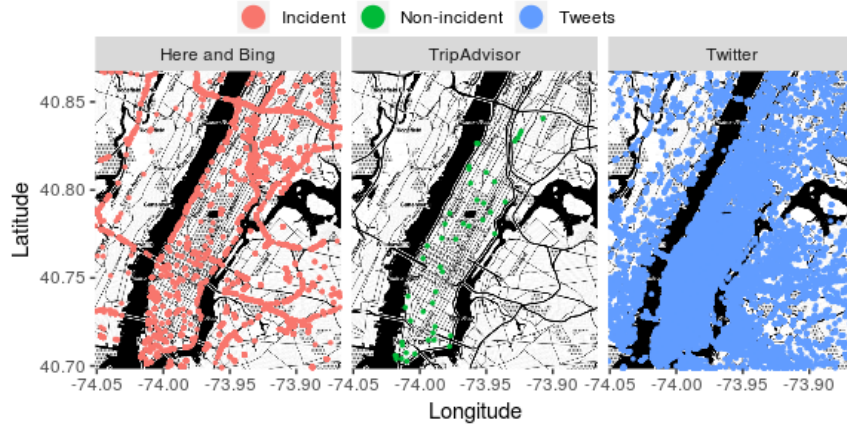


Figure 5.8: The spatial coverage by data sources used.

greater the number of incidents used, the more tweets can be grouped, benefiting our learning-based approach. After acquiring data from the Here WeGo and Bing Maps platforms, we pre-processed them to standardize their features.

Thereafter, we conducted a spatiotemporal grouping (see Section 5.4.3.1 and Algorithm 1 for more details). However, the goal here was to identify an incident event reported by both data sources, thus representing the same event. In this case, the temporal interval and the spatial location of them must be very close. We assume that two events are close, and, therefore, the same, if they start on the same day and hour but are also located at most 10 meters apart from one to another. We named these same events as *Intersection*. In other words *Intersection* is the data resulted of $(\text{Here} \cap \text{Bing})$. Figure 5.9 shows the frequency of each incident type by a given data source. Moreover, we can see the same events reported by both sources in the *Intersection* graphic.

We also evaluated the similarity of incident types from the *Intersection*. We found that the incident type similarity between Here and Bing reached 99.83%. In other words, both data sources labeled the incidents almost similarly. As a final step, we created a *New Incident Layer*, which combines the data coverage from both data sources and increases the information description about incidents, using the intersection of them. Since each data source has its individual way of reporting incident events, detailing the road name or a short description text, the fusion enriches the whole context.

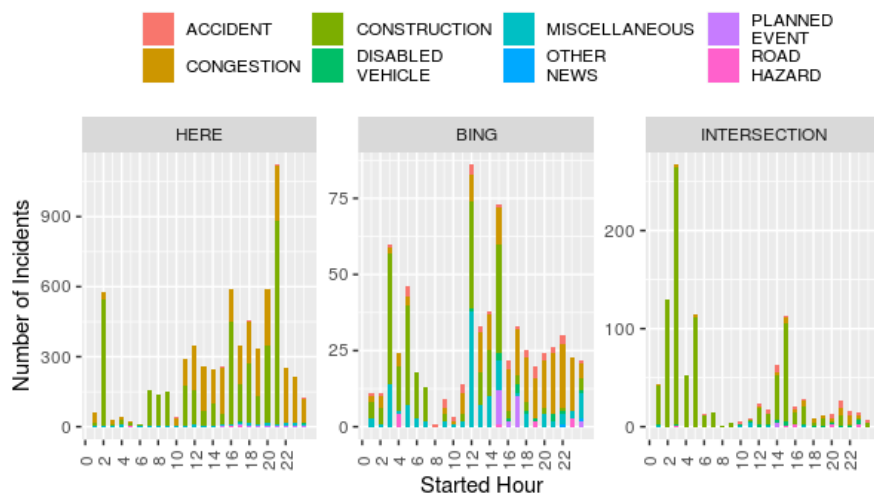


Figure 5.9: Hour of an incident by data source and the intersection of them.

Figure 5.10 shows the spatial data coverage of each data source and the intersection between them, during the process of data acquisition (2018-09-14 to 2018-11-06). It also shows the data representativeness for each source. For instance, Here WeGo corresponds to 80.53% of the whole data, while Bing Maps covers 8.31% and the *Intersection* corresponds to 11.16%. The *New Incident Layer* covers 100% of the entire data collected, thus enriching more than 11% of similar events with richer detailed information.

5.4.3 T-Incident Design Architecture

This section presents a learning-based incident detection approach based on heterogeneous data fusion. We conducted our analysis considering the premise that the LBSM can provide valuable information about the traffic and incident condition, as discussed in [Santos et al., 2018].

Based on the ITS data as an input to our design, we created a spatiotemporal grouping which aims to combine different data sources (see Section 5.4.1 in temporal and spatial dimensions). After that, we conducted a feature extraction process aiming to acquire the user’s viewpoint around the event which it was previously grouped. Then, we developed a learning-based model to identify potential incidents considering the user’s reports. Finally, we evaluated our approach us-

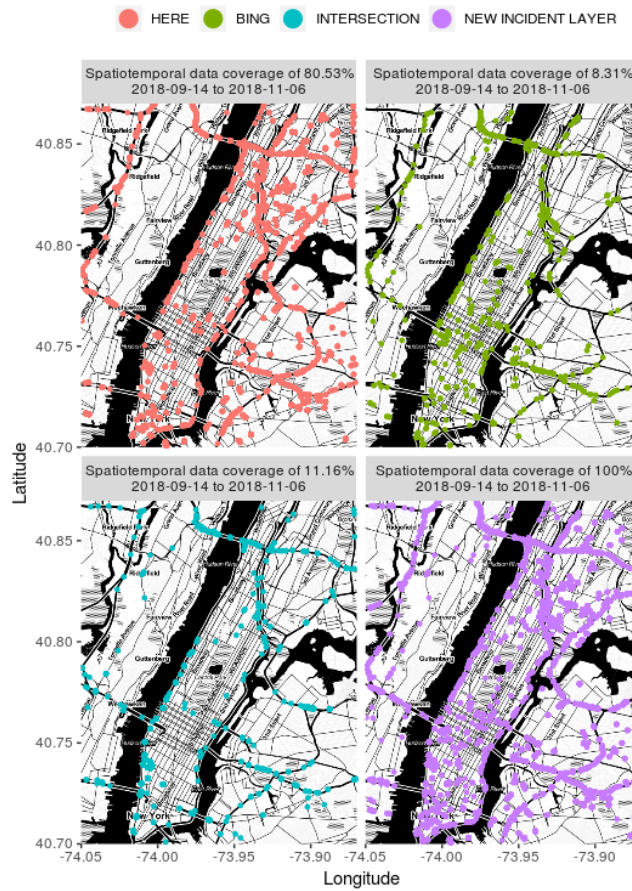


Figure 5.10: Spatial incident coverage per data layer.

ing different spatial grouping modes. In the following, we describe each stage of T-Incident as depicted in Figure 5.11.

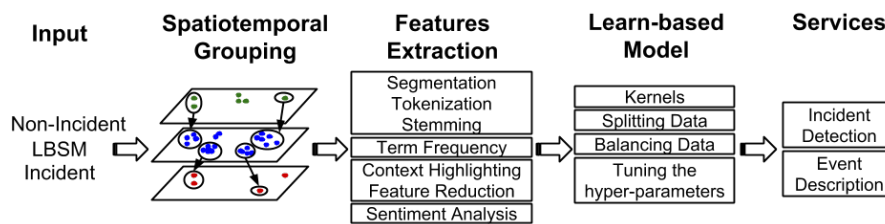


Figure 5.11: Design of T-Incident.

5.4.3.1 Spatiotemporal Grouping

The grouping mode considers the heterogeneity of the data sources used and its spatiotemporal coverage variation. Therefore, we proposed an approach which merges the incident/non-incident data layers with the tweets layer based on both dimensions. To do that, we considered the incident as an event and not each type of it, i.e., we grouped the incident types in only one event – *Incident*. Each incident has a start location, end location, and duration time. Our grouping considers only the incident start location as same to the events named – *Non-Incident*. Another characteristic of our data preparation consists in setting the non-incident time interval with the same interval of the Twitter data.

Based on the dataset of incident and non-incident, we are able to conduct a temporal filter which looks for the intersection between events and tweets. Once those data have merged, we perform a spatial filter based on the radius of each event location. We created a set of radii, aiming to identify the better grouping mode once we are dealing with user bias and the vast amounts of unrelated data. That methodology enabled to group a different number of tweets around the event (see Table 5.2, and, thus, the information surrounding the event can be more valuable to the context or more generalist to it.

Table 5.2: Number of tweets for each spatiotemporal grouping model.

Event	Radius (km)						
	0.01	0.05	0.1	0.2	0.3	0.4	0.5
Incident	121	959	3,098	9,467	30,085	63,853	68,877
Non-Incident	260	3,161	6,522	13,060	20,699	30,492	35,786

Even though the spatiotemporal grouping could be conducted in different ways (e.g., based on streets segment, neighborhoods and a grid dividing the geographical area), we chose the use of different radii around the incident, as our initial approach. Tweets, which were not grouped, were labeled as *Unknown* and removed. We noticed a trade-off to choose the radius size and the relevance of information floating around the event. In other words, a small radius implies in fewer data grouped, but relevant information about the event. A larger radius results in more data grouped, but less descriptive information of the event. That situation becomes a challenging task when there are reduced amounts of data acquired.

We describe the spatiotemporal grouping in Algorithm 1. The inputs to the grouping are the tweets, incidents and the radius. The expected result is an updated Tweet dataset containing the *event*, *incident id*, and *incident type*. We also developed an optimization process splinting the geographic area, latitudinally, in x sections, aiming to reduce the number of operations conducted in large areas with large amounts of data. After that, for each tweet and incident, we tested if they are in the same section or near with one hop up or down (Line 7). Satisfied that condition, the tweet must be between the incident start and end time (Line 8). For then, we measure the distance between the tweet and the incident, aiming to find the minimum distance to assign its new attributes (Lines 9-14).

5.4.3.2 Feature Extraction

We assume that the interest information floats around the observation location. Stressing the grouping based on a radius around the event, making it an intuitive and very powerful approach, as shown in Section 5.4.4. However, data from LBSM brings issues that can lead to other challenges such as data imprecision and users' bias. In that way, the feature extraction role aims to clean the tweet and provide a set of words which describe better the event's surrounding.

We first applied for each grouping and event class a set of NLP methods such as lowercase transformation, accents removal, tokens extraction, and filtering stop words, links, and special characters. After that, we reduced inflectional and derivational forms of a word to a common base form. Then, we analyzed the Term Frequency (TF) from the event, extracting a matrix of the most frequent words mentioned in that area. Moreover, we filtered that matrix based on the sparsity, i.e., we removed terms that were sparse than 0.98%.

We also introduced a context highlighting step for a specialist to reduce non-related words of a given event. This is because, even though we conducted the previous steps, the LBSM keeps noises which must be removed. We noticed, by experiments, that the Term Frequency-Inverse Document Frequency (TF-IDF) approach does not stress the words which describe each event' class accurately. Then, that analysis was not valuable in this work.

At the end of that process, we gathered the set of most important words

Algoritmo 1: Spatiotemporal LBSM Data Grouping

```

Input: tweets,incidents,radius
Result: tweets grouped by event, incident Id, and incident Type
1 /* The previous step split each dataset into x slices,
   reducing the computation */
2 initialization;
3 for each tweets do
4   currentIncidentId  $\leftarrow$  0;
5   currentIncidentTmp  $\leftarrow$  None;
6   currentDistance  $\leftarrow$   $\infty$ ; /* larger than radius */
7   for each incidents do
8     if equal(tweets.sec,incidents.sec) or diff(tweets.sec,incidents.sec)
9       is (+ 1 or - 1) then
10      /* Tweets between the incid. time */
11      if TemporalFilter(incidents.starttime, incidents.endtime,
12        tweets.timestamp) then
13        /* Distance from the radius */
14        distance  $\leftarrow$  SpatialFilter(tweets.coord, incidents.coord,
15          currentDistance, radius);
16        /* Record the less distance */
17        if distance < currentDistance then
18          currentIncidentId  $\leftarrow$  incidents.Id;
19          currentIncidentTmp  $\leftarrow$  incidents.Type;
20          currentDistance  $\leftarrow$  distance;
21        end
22      end
23    end
24  end
25  /* Assigning the event type(Incident, Non-Incident,
26    Unknown) for each tweet */
27 end

```

posted by common Twitter users. Figure 5.12 shows an example of a set of words grouped by radius between 0.01 km and 0.5 km. This indicated how specific or general could be the information around the event regarding its radius. Figs. 5.12a and 5.12b show more words, weighting them differently and reducing the intersection between incident and non-incident. However, upon increasing the radius we can see fewer words with high weights stressing common words between both

classes (see Figs. 5.12c and 5.12d). Our goal is to understand that behavior and train an algorithm to automatically identify those classes. Next, the set of words will feed a learning-based model, described below.

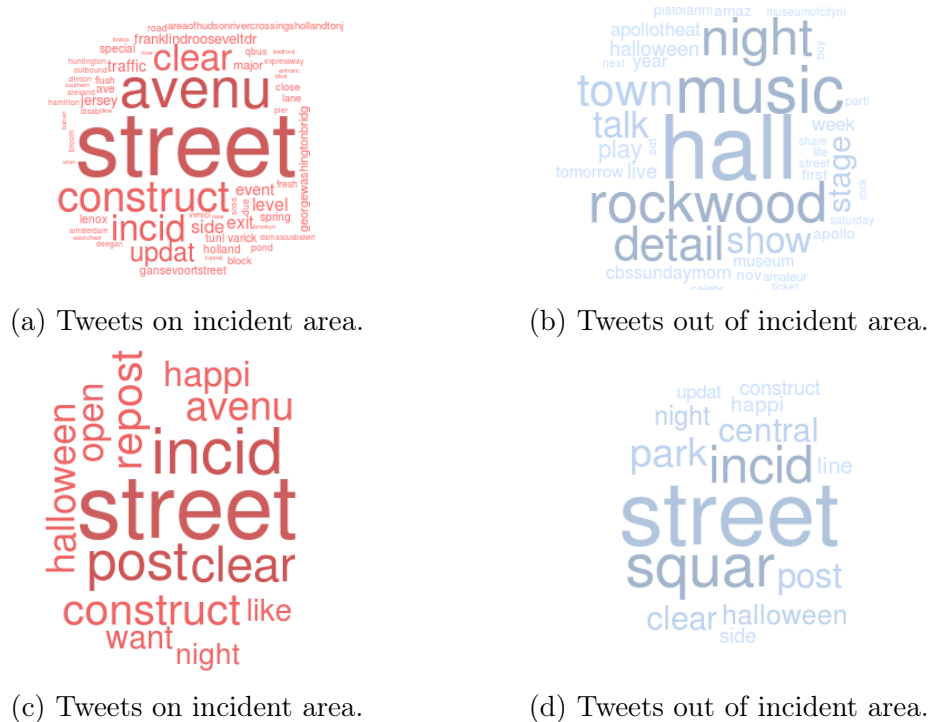


Figure 5.12: Spatiotemporal grouping based on a radius of 0.01 km ((a) and (b)) and 0.5 km ((c) and (d)).

a) Feature Reduction: The number of features obtained from the last stage may be large enough to introduce computational barriers as the processing time, memory and storage capacities. We conducted a method to reduce the number of features based on their importance and frequency. In other words, we initially developed two approaches to achieve that goal. The first one was the Principal Component Analysis (PCA) to extract a set of relevant features. This process identifies the most variable information from a multivariate dataset and expresses it as a set of new features – Principal Components (PCs). These PCs represent the directions along which the variation in the data is maximal. The second one was based on the ranking of the most frequent words.

Both methods output the results to the specialist who makes the decision.

Table 5.3: Relevant features based on radius of 0.01 km.

Event	Most Frequent Features					
Incident	traffic	side	exit	contract		incid accid
	avenu	street	updat	georgewashingtonbridg		clear
	jersey	event	major	franklindrooseveltdr		level
Non-Incident	town	night	year	apolloheat		show
	detail	hall	music	halloween		stage
	week	play	live	rockwood		talk

We noticed that the PCA did not catch a good set of words such as the use of most frequent words did. When the tweet dataset was acquired without a *track of words* (any tweet, without specific words), the PCA performs better than the use of the most frequent words. On the other hand, PCA is not suitable for tweets with a set of specific track words as mentioned in Section 5.4.1. As result, we performed the feature reduction for each grouping and event class, extracting only the most representative set of words from the previous stage. Table 5.3 shows an example of features obtained after ranking the most frequent words on spatiotemporal grouping based on a radius of 0.01 km.

b) *Sentiment Analysis:* The sentiment analysis was conducted for each tweet for each grouping and event class, allowing us to extract the feelings that Twitter users have about the event, in which they passed through. To derive the sentiment from the tweet’s text, we used a dictionary of words and its associated feelings [Jockers, 2017]. The sentiment depends on the number of words/feelings occurrences to calculate the score, and we can associate a sentiment (positive or negative) to the tweet. As result, for each tweet we extracted the set of feelings words and its frequencies, binding them with the set of words processed on the previous stage, for that same tweet.

5.4.3.3 Learning-Based Model

The last stage was responsible for extracting useful information which better describes a given class of event and feeds our learning-based model with a set of features labeled by the event. In this way, we started to deal with a classification problem. First, we chose the most common classification algorithms (*kernels*, used

in the same context of this work, based on the literature review [Xu et al., 2018]. To conduct this step, we used the following *kernels*: Support Vector Machine (SVM), k-Nearest Neighbors (KNN) and Random Forest Classifier (RF).

Next, we split the data into two sets, following the convention of the most machine learning approaches: Training Set, corresponding to 70% of the entire dataset; and Test Set, corresponding to 30% of the entire dataset. To validate the training process, we applied the cross-validation considering 10 folds split in 70% and 30% of the training and test, respectively. Our goal was to evaluate the training curve and the testing curve, avoiding possible over-fitting and under-fitting. That partition was conducted for each group.

Notice that the dataset exhibited an explained unbalancing, once the number of tweets around the non-incident areas is bigger than around the incident ones. In this case, we explored the re-sampling techniques which aim to balance classes either increasing the frequency of the minority class or decreasing the frequency of the majority class. Our goal was to obtain approximately the same number of observations for both classes.

We used a random under-sampling, aiming to balance the class distribution by randomly picking and eliminating the majority of class examples. That strategy helps to improve run-time and storage by reducing the number of training data samples once the training is huge enough, considering LBSM data. However, the classifier may suffer hard consequences since the potential useful information can be discarded. For that reason, this step is not limited to that approach, as it always depends on the quality and quantity of LBSM data acquired.

After that, tuning the hyper-parameter becomes a challenging task and an exploratory approach was adopted to deal with. We used a *GridSearchCV* class from Scikit-Learn API [Pedregosa et al., 2011], which takes a set of parameters and values to exhaustively combine them, aiming to find the best configuration. Knowing that the complexity of such search grows exponentially with the number of parameters, we defined a set of parameters for each *kernel* following some guidelines. For the SVM, we based on [Hsu et al., 2003], and for the other ones, we followed the user’s guide for Auto-WEKA [Kotthoff et al., 2017].

5.4.3.4 T-Incident Services

The results of the learning-based model allowed us to understand the best spatiotemporal grouping and the set of NLP methods to filter the LBSM texts, and, then, accurately outline the events. Based on that, we were able to output the incident and non-incident events detection service and the event description service.

Once we identified an event, we started to analyze its context. To do that, we conducted a text summarization process, aiming to create a short and coherent version of a longer document. We considered a document a set of tweets grouped by incident type, i.e., we applied the text summarization to a group of tweets labeled by incident type and hour, and by incident id. This process provides a short description for each group, allowing to give the users and traffic planners the viewpoint of the LBSM users regarding the transit events and points of interest.

In that area, there are two methods of text summarization: *Extractive* and *Abstractive*. The first one selects the tweets, ranking their relevant phrases and choosing only those which are meaningful to the event. The abstractive method aims to generate entirely new sentences to capture the meaning of the event. For this version of T-Incident, we developed the event description service, using the extractive text summarization method.

5.4.4 Evaluation

In this section, we describe T-Incident performance evaluation against the set of classifier algorithms and spatiotemporal grouping modes as outlined in Section 5.4.3. Then, we present T-Incident services to detect and enrich the event description.

5.4.4.1 Event Detection

Our incident detection approach was based on an exploratory analysis of classifiers algorithms, hyper-parameters and radius. Figure 5.14 shows the results regarding a Training and Test process. We validate our training process performing a *Cross-validation approach* which aims to split the training set in training and validation sets among 10 folds. Figure 5.13 shows the learning curve of each kernel performing

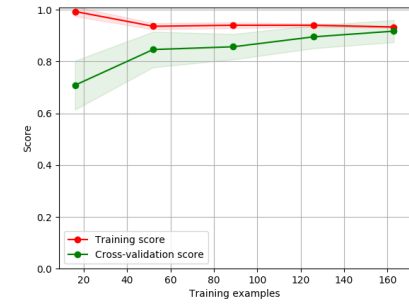
on a spatiotemporal grouping with radius of 0.01 km and 0.5 km, as an example. The main goal here is to study the generalization of a given model, avoiding over-fitting and under-fitting, and find out the best spatiotemporal grouping. We noticed that the radius of 0.01 km (Figs. 5.13a, 5.13c, and 5.13e) delivers the best score, around 90%, in most kernels after 140 training samples where we see the curves converging and the model stabilization. However, the reduced data limit the exploration of event description service.

Once we increase the radius, we were able to see the curves decreasing as depicted in Figs. 5.13b, 5.13d, and 5.13f. Using a 0.5 km radius, we observed a score between 58% and 65%. Decreasing the radius to 0.4 km, we noticed averaged scores above 61% and below 65%. A radius between 0.3 km and 0.2 km showed very close results as scores above 65% and below 70%, in average. Using 0.1 km, we obtained scores around 70%, and between 75% and 80% considering the radius of 0.05 km.

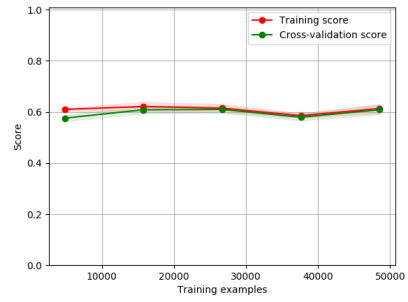
We deal with a trade-off between higher radius (more grouped data and smaller scores) and lower radius (fewer data and higher scores). The important lesson here is the application of a consistent methodology that was able to provide a generalization model to detect incidents.

Next, we evaluated three metrics from the Cross-validation and Test: i) *F1 Score*: is the weighted average of Precision and Recall. This score takes both false positives and false negatives into account ($2 \times Recall \times Precision / (Recall + Precision)$); ii) *Recall*: measures how good a test is at detecting the positives ($TP / (TP + FN)$); iii) *Precision*: is the ratio of correct predicted positive observations to the total predicted positive observations ($TP / (TP + FP)$).

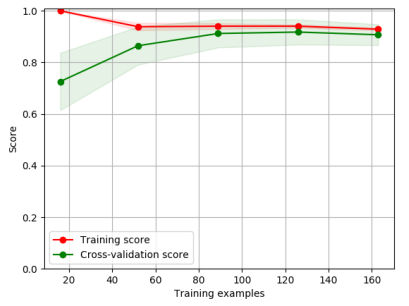
Figure 5.14 shows the best set of parameters that can feed the T-Incident. As noticed in the learning curves, the better spatiotemporal grouping could be the radius of 0.01 km which shows a Test score above 90% in all metrics evaluated. However, we considered a very good result scores above 70% due to the quality of LBSM data. Once assumed that, we can even use the radius of 0.1 km keeping the *F1 score*, *Recall* and *Precision* around 75% on average. After that spatiotemporal grouping, we observed a divergence among those metrics scores, and the decrease of scores, which can be explained by the increase of intersection between the incident and the non-incident set of features.



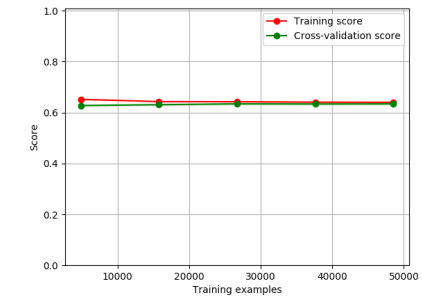
(a) KNN – radius 0.01 km.



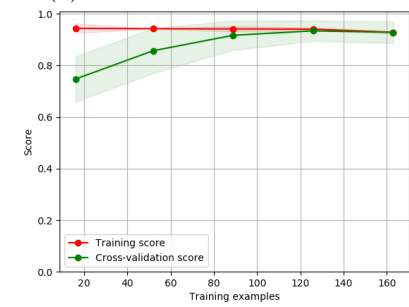
(b) KNN – radius 0.5 km.



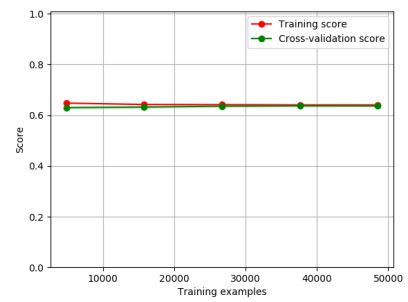
(c) RF – radius 0.01 km.



(d) RF – radius 0.5 km.



(e) SVM – radius 0.01 km.



(f) SVM – radius 0.5 km.

Figure 5.13: The learning curve of a given kernel and spatiotemporal grouping.

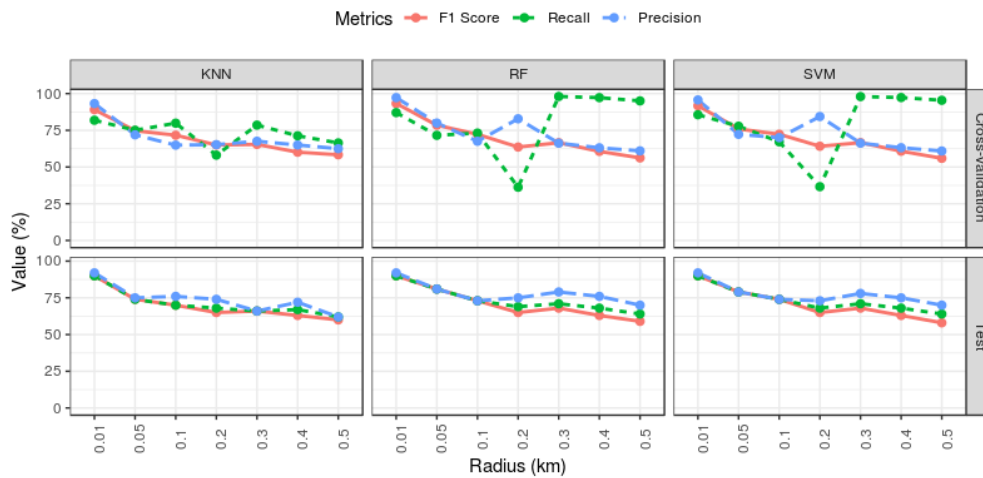


Figure 5.14: Classification results based on different kernels and evaluation metrics.

5.4.4.2 Event Description

The results observed in the detection stage, allowed us to identify the best spatiotemporal grouping which accurately outlines the event. In this sense, we conducted a text summarization process, based on the *Extractive* method, creating a short and coherent version of the event. Notice that we used for this analysis the spatiotemporal grouping with radii 0.01 km and 0.1 km, based on the trade-off between accuracy and size of the data sample.

As an example of the T-Incident description service with a radius of 0.01 km, the text below summarizes a specific incident event on *Franklin D Roosevelt Drive*. We highlighted the words to make this text clear for the reader to understand what happened there. With that analysis in hand, we aim to enable users and road managers to understand and decide what can be done about it.

Cleared: Construction on #FranklinDRooseveltDrive SB from Exit 9 - East 42nd Street to 34 street;
Updated: Incident on #FranklinDRooseveltDrive SB at Exit 9 - East 42nd Street; **Cleared:** Incident on #FranklinDRooseveltDrive SB at Exit 9 - East 42nd Street; **Incident** on #FranklinDRooseveltDrive SB at Exit 9 - East 42nd Street; **Closure** on #FranklinDRooseveltDrive NB at Exit 9 - East 42nd Street;
Cleared: Closure on #FranklinDRooseveltDrive NB at Exit 9 - East 42nd Street; **Construction** on #FranklinDRooseveltDrive Both directions at Exit 9 - East 42nd Street

At the same time, using the spatiotemporal grouping with a radius of 0.1 km, for instance, we analyzed a specific non-incident event, the *Town Hall* and its surroundings. The text below summarizes that area, highlighting the top trends of places which were extracted by users' impressions. In that way, this is possible

to find out cultural places to go, where to book a hotel' room and where to eat there.

Open House New York Sunday Stop 1! **Town Hall**. It was never taken over by the Broadway **theatre** giants because ther..; #30DaysForMyArt DAY 16: "Go see a **broadway show**." It's simple: There is NOTHING like a Broadway show. I've lived in; **Beastie Boys Book: Live**; Direct with Adam Horovitz; Michael Diamond: The **Town Hall**; Good morning **Times Square**. Bad I have to leave today! (**@Millennium Broadway Hotel** - @millenniumpr in New York, NY); Good night! (**@Millennium Broadway Hotel** - @millenniumpr in New York, NY); YEP! I Like Wrestling Podcast #45: WWE Super Show-Down Predictions, Raw; Smack; Show Time! (@Beautiful: The Carole King Musical in New York, NY); **Mooch's book party**. Really. (@Hunt; **Fish Club** in New York, NY); Head over heels wPeppermint!!! (@Hudson Theatre - @hudsonbway for Head Over Heels in New York, NY); I had the heirloom tomato lobster salad. **Kristine had the burger** (@Burger; Lobster in New York, NY);

Moreover, the T-Incident description service provides an overview of incident events in each area and day hour. The text below was summarized considering the spatiotemporal grouping with a radius of 0.05 km in Manhattan at 5 am, for instance. It delivers to the users and road manager a feasible and low-cost way to understand areas which may be avoided or even take better attention at that hour. Notice that, our analysis aims to focus on the top trends of incident events at a given day and hour, enriching the current context and delivering to the public a very short and summarized information.

Cleared: Construction on **#GeorgeWashingtonBridge** WB from New York SideLower Level to New Jersey SideLower Level; **Cleared:** Construction on **#WLine** Both directions from Whitehall Street-South Ferry Station to Ditmars Boulevard-Astoria Station; **Updated:** Construction on **#WLine** Both directions from Whitehall Street-South Ferry Station to Ditmars; **Cleared:** Construction on **#NY9A** SB from West 42nd Street to West 38th Street; **Cleared:** Closure on **#RiversideDrive** Both directions from West 145th Street to West 155th Street; **Cleared:** Construction on **#FranklinDRooseveltDrive** SB from Exit 9 - East 42nd Street to 34 street; **Cleared:** Construction on **#M42Bus** Both directions at 42 St at 12 Av and the 42 St Pier; **Closed** in **#NewYork** on **42nd St WB between Lexington Ave and Madison Ave, stop and go traffic** back to 3rd Ave **#traffic**; **Accident**, center lane blocked in **#HudsonRiverCrossingsGwb** on The G.W.B. Upper Level Outbound after The Harlem Riv; **Accident**, left lane blocked in **#HudsonRiverCrossingsGwb** on The G.W.B. Upper Level Outbound after The Harlem River

5.4.5 Discussion

The results of our RoDE can be summarized as follows: the *Incident Services* showed the best set of parameters that can feed our T-Incident approach, leading to the incident detection and event description services. The better spatiotemporal grouping mode considered the radius of 0.01 km, showing the incident detection scores above 90% in all evaluated metrics. However, we considered that a very good result presents scores above 70% due to the quality of LBSM data. Once assumed that, we can even use the radius of 0.1 km keeping the *F1 score*, *Recall*

and *Precision* around 75% in average. Based on that, the event description service allowed us to provide a summarized description for each group, providing users and traffic planners the viewpoint of the LBSM users regarding the transit events and points of interest.

5.5 Chapter Remarks

In this chapter, we presented the Road Data Enrichment (RoDE) framework, a low-cost approach to ITSs based on Heterogeneous Data Fusion. RoDE delivers a high-level information, allowing a navigation system, road planners and general public a more consistent, accurate and useful information, providing two main contributions: *Route Services* and *Incident Services*. RoDE is able to enhance the route information of current navigation tools, detect incidents on the road, and enrich the event description. It provides to users and traffic planners the viewpoint of the LBSM users and different traffic/transit data sources, regarding the transportation system.

In summary, Figure 5.15 shows how our design of fusion on Vehicular Data Space (VDS) worked in this study. Where, the LBSM, road map data, and point of interest feed the fusion process, the data preparation deal with data aspects showed in Chapter 3 and others which help to treat the data for the data processing which covers methods related to the application goals and finally resulting in the RoDE as the data use.

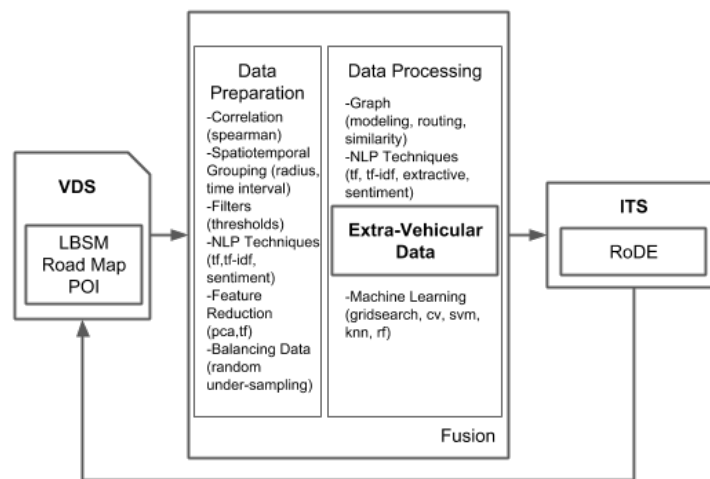


Figure 5.15: Design of fusion on VDS for RoDE.

Chapter 6

Intra-Extra-Vehicular Data Fusion

In this chapter, we describe the fusion process on the Vehicular Data Space (VDS), considering the Intra-Vehicle Data (IVD) and Extra-Vehicle Data (EVD) aiming to provide an application to promoting the Smart Mobility (SM).

6.1 Introduction

Planning and managing transportation systems are crucial tasks to promote the growth of cities. For instance, the number of fatalities and injuries on the road has achieved an alarming scenario. Such fact is pushing new initiatives from governments and private sectors to improve the road traffic efficiency and safety. However, the lack of traffic information provided by the transportation systems decreases the efficiency of route management, flow control and the spread of traffic descriptions. To provide accurate traffic information, the integration of data from multiple data sources are needed. Then, once again the heterogeneous data fusion becomes a feasible solution way to achieve the Intelligent Transportation System (ITS) goals.

Due to the lack of traffic data and considering vehicles as potential entities of participatory sensing, where communities can contribute with sensing traffic information, we propose Traffic Data Enrichment Sensor (TraDES), a low-cost traffic sensor for ITS based on heterogeneous data fusion. TraDES aims at fusing data from vehicular traces and its respectively embedded sensors with road traffic

data to enrich the current spatiotemporal traffic data. To do that, we propose a methodology to spatiotemporally group these different data sources and perform a learning-based approach to detect the traffic condition based on a set of vehicle sensors. As a result, the methodology outputs an enriched traffic sensor, with an accuracy of up to 90%, allowing the delivery of information about traffic conditions to navigation systems, road planners and the general public. Hence, the main contributions of TraDES are: i) *A scalable and low-cost approach*: we focus on free access data and a spatiotemporal grouping method, which enable to add more data layers to enrich available traffic data or even to produce another application. ii) *Increase the spatiotemporal traffic data coverage*: using vehicular participatory traces and road traffic data as input to a robust methodology allows us to infer the traffic condition for regions where there is no available information; iii) *Enrich the traffic data*: by taking advantage of vehicular sensors, we develop analyses that provide an overview of fuel consumption, emissions and so on for each traffic condition.

The rest of the chapter is organized as follows. In Section 6.2, we describe the related works to the traffic problems. Section 6.3 presents the data acquisition and characterization process. In Section 6.4, we present TraDES design. The evaluation is detailed in Section 6.5. Finally, in Section 6.6 we highlight the final remarks and conclusions.

6.2 Related Work

The issues related to transportation and traffic in huge cities are well known by governments and private sectors. These issues pushes new initiatives and investigations on ITSs to improve the road traffic efficiency and safety. Those investigations may be conducted by considering many different entities and its data from the ITS scenario. Using only GPS from smartphones, Goncalves et al. [2014] conducted a study and characterization of traffic and road conditions. They built the Iris Geographic Information System (GIS) platform using a Android smartphone on the client side and a server for collecting and storing data, pre/post processing, analyzing and managing the traffic condition.

Zuchao Wang et al. [2013] developed a system for visually analyzing urban

traffic congestion. They used GPS trajectories and speed data from taxis in Beijing to design a model to extract and derive traffic jam information in a realistic road network. The process consists of an efficient data filtering step based on spatiotemporal aspects, size and network topology to create a graph structure and its visualizations. Han et al. [2014] developed the SenSpeed, an accurate vehicle speed estimation system, to address an unavailable GPS signal or inaccurate data in urban environments. The authors relied on smartphone sensors, such as gyroscope and accelerometer to sense turns, stops and crossing irregular road surfaces. The results show that the real-time speed estimation error is 2.1 km/h, while the offline speed estimation error is 1.21 km/h, using the vehicle speed through the On-Board Diagnostic (OBD) as ground truth in their experiments. Ning et al. [2017] conducted a study to detect traffic anomalies based on the analysis of trajectory data in Vehicular Social Networks (VSN). The VSN is an integration of social networks and the concept of the Internet of Vehicle (IoV).

Using public data, Gu et al. [2016] explored the Twitter platform, aiming to extract traffic incident from users posts, thus providing a low-cost solution to increase the road information. Santos et al. [2018] also improved traffic and transit comprehension through the Twitter MAPS (T-MAPS), a low-cost spatiotemporal model to improve the description of traffic conditions using tweets. Differently from most of the related work discussed above, we take a step forward by providing a methodology to increase the spatiotemporal traffic data coverage. For that, we fuse free public access heterogeneous data, such as participatory vehicular traces and road traffic data, aiming to enrich the transportation scenario, thus feeding with data the current navigation systems, road planners and the general public.

6.3 Data Acquisition

Nowadays, there is a variety of entities on the urban transport environment that provides data to transportation systems. However, the spatiotemporal data coverage depends on huge infrastructures and policies for data access, such as security and privacy. In this sense, governments and academy initiatives to improve the transportation data coverage are essential for achieving the ITS view. TraDES is an approach to accurately identify traffic conditions (*Traffic* and *Non-Traffic*)

based on a set of features from vehicular traces with the aim of enriching the quality of road traffic data. The data acquisition process consists on fusing data from different data sources, such as Here WeGo ¹ (traffic map) and enviroCar ² (vehicular traces) in both temporal and spatial dimensions to provide a novel traffic sensor.

Bröring et al. [2015] presented the enviroCar platform, which aims to acquire vehicle sensors' data and provide free access to such data, thus enabling traffic monitoring and environment analysis through the Internet. Given the importance of sensors to a vehicle's operation, new vehicle models embed many high-quality sensors to get more reliable and diverse information about themselves. All data produced by sensors in a vehicle are delivered to its Engine Control Unit (ECU) through an internal network, named Controlled Area Network (CAN), which is accessible through the vehicle's OBD port. The OBD system was first introduced to regulate emissions. However, it is now used for a variety of applications. There are different signaling protocols to transmit internal sensor data to external devices through a universal port. Such a universal port is present in all cars produced since 1996 in the U.S. and Europe. There are Parameter IDs (PIDs) to access sensor information using the OBD, which identify individual sensors. Some PIDs are defined by regulatory entities and are publicly accessible. However, manufacturers may include other sensors' data under specific and undisclosed PIDs.

Using Android smartphones and OBD adapters, the enviroCar collects a set of sensors data produced by vehicles and upload it to the web for free public access. The enviroCar dataset consists of 585,050 observations in almost 200 Germany cities acquired from 2017-01-01 to 2018-08-07. However, we were not able to acquire spatiotemporal traffic data with the same coverage. Hence, we reduced the vehicular traces to 255,743 observations and 1872 distinct trips, containing a set of cities for which there also is traffic data. All collected trips are geolocated and most of them are in Germany (subject of our study). In addition, the frequency of the sensor data acquisition is every 5 seconds. Table 6.1 shows some of the sensors data collected by enviroCar.

To collect as much traffic condition data as possible from a traffic map, we col-

¹<https://wego.here.com>

²<https://envirocar.org/>

Table 6.1: Features from vehicles and roads.

Data	Features		
Vehicle	Speed*	MAF*	RPM*
	Throttle*	Engine*	Intake*
	Position	Load	Air Temp
	CO ₂ *	Fuel*	O ₂ Lambda
	Intake Pressure		Voltage
Smartphone	Device Time	Altitude	GPS Location
	GPS	GPS	
	Speed	Features	
Road	JF*	FC	FF
Traffic	SP	SU	

MAF = Mass Airflow; RPM = Revolution per Minute;
GPS Features = HDOP, Bearing, VDOP, Accuracy,
and PDOP; JF = Jam Factor; FC = Current Flow;
FF = Free Flow speed; SP = Speed capped by
speed limit; SU = Speed not capped by speed limit;

lected data from 13 different cities in Germany with a temporal granularity of one hour. Since there is no historical traffic data available, we opened a data acquisition streaming from 2017-01-01 to 2018-08-07 to collect vehicular traces and traffic data from the same spatiotemporal interval. As a result, we collected 1,555,582 road traffic observations from Here WeGo from 5 cities in Germany, which also have reported vehicular traces. Table 6.2 summarizes the collected data, which will then be spatially and temporally fused. We also started a data acquisition process among different map sources, such as Bing Maps³ and MapQuest⁴, but there was not enough data reported in Germany.

³<https://bing.com/maps>

⁴<https://www.mapquest.com/>

Table 6.2: Data acquired from different data sources.

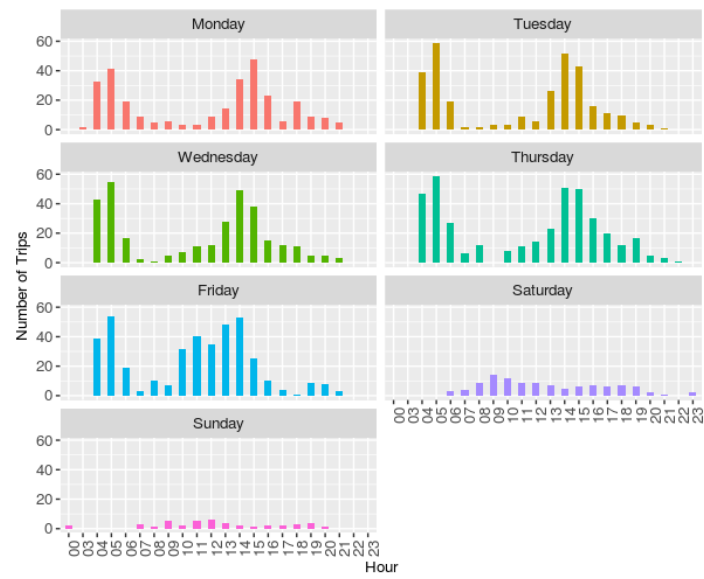
Source	Goal	Sample	Temporal Interval	Spatial Location
Enviro Car	Vehicular OBD Traces	255,743	2017-01-01 to 2018-08-07	Monchengladbach, Viersen, Willich, Dusseldorf, Korschenbroich, Wegberg, Munster, Neuss, Juchen
Here WeGo	Road Traffic	1,555,582		Monchengladbach, Viersen, Dusseldorf, Munster, Neuss

6.3.1 Data Characterization

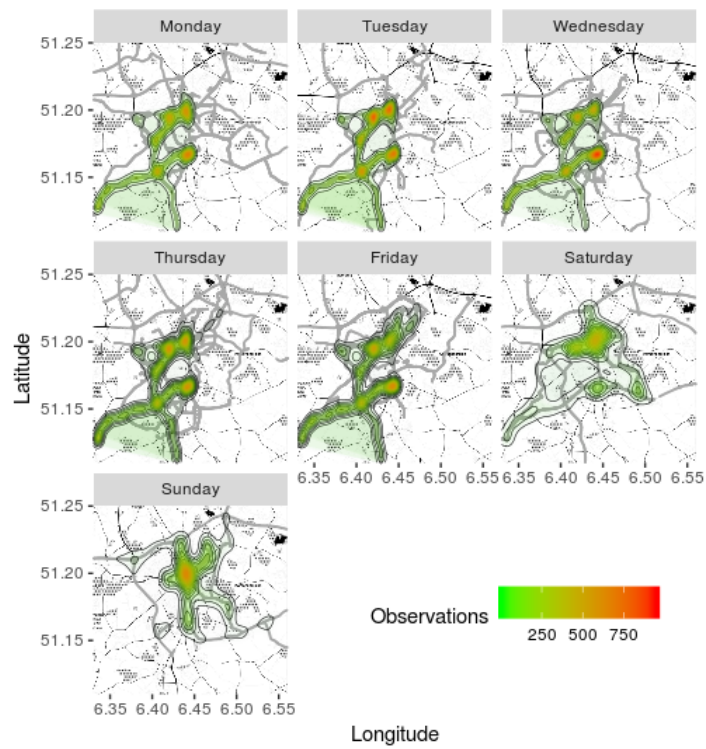
6.3.1.1 Trace

Contextual information from vehicles is fundamental to better understand traffic patterns, drivers behavior and mobility patterns in a city. In this sense, we explore the spatial and temporal aspects of the collected vehicular traces. As observed in our previous work [Santos et al., 2018], which take into account road traffic data and Location-Based Social Media (LBSM), the traffic and users have a similar behavior when considering the day of the week and hour of the day, as you can also see in Figure 6.1a. The number of trips increases in the beginning of the day, decreasing until the middle of the afternoon, when the curve returns to rise. That behavior reflects people in their workday during the week. Furthermore, in the weekend, people tend not to use their own vehicles and stay at home or use another vehicle to move.

Figure 6.1b shows the spatial coverage of the vehicular traces on the regions of Monchengladbach during the week. We can notice the areas during specific weekdays where there are more traces than others. Moreover, different areas of the city are explored during the weekend. Considering the sensors' data acquired from the vehicle and smartphone, as shown in Table 6.1, we can also analyze features, such as fuel consumption, emissions and level of noise in a given area of the map. Those observations may allow navigation systems, road planners and the general public a more descriptive overview of the transportation system.



(a) Frequency of trips per week and hour.



(b) Traces per week in Monchengladbach.

Figure 6.1: Spatiotemporal analysis of vehicular traces.

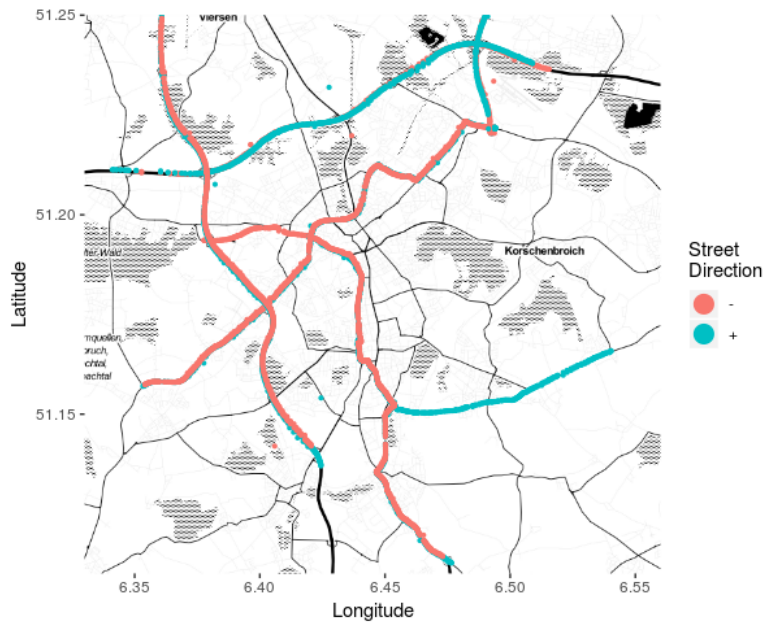
6.3.1.2 Traffic

After showing the potential of vehicular traces to provide better traffic comprehension in a given area, we conducted the same data characterization using road traffic data acquired from Here WeGo. Firstly, we can notice the limited data coverage on all cities observed. Figure 6.2a shows an example of a road map with reported Jam Factors (JF). That factor is a real number between 0 and 10 indicating the expected quality of travel. As the number approaches 10.0, the quality of travel tends to get worse, and when the JF reaches 10 it means that there is a road closure. That limited road data coverage implies that navigation systems may suggest routes based on insufficient traffic information, once only the main roads report traffic conditions, while adjacent ones do not.

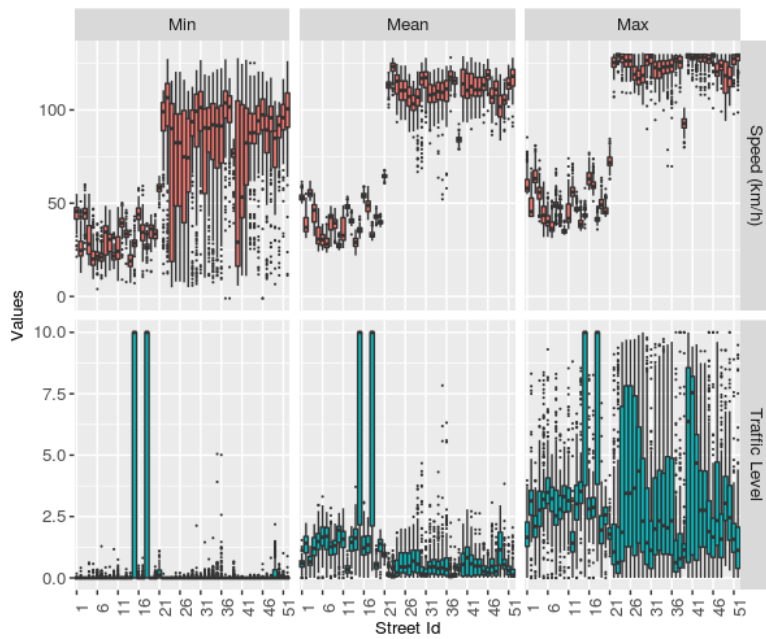
After observing the road traffic data coverage, we extract each street segment to analyze its average speed and JF. Figure 6.2b shows, considering each street segment, an overview of traffic condition in Monchengladbach. We can notice that there is a group of street segments with low speeds and high jam factors. That behavior may indicate areas close to downtown, and the opposite behavior indicates they are highways. In other words, these analyses may be used to classify the types of roads in a city according to their use. We also notice that there are two segments (15 and 18) that stay closed during our data acquisition process. However, we observe vehicular traces that use these segments, reinforcing the need to employ alternative approaches that consider different data sources, as the one proposed here, to better explain the current traffic condition.

6.4 TraDES' Design

This section presents an approach to enrich traffic data based on heterogeneous data fusion. First, we feed our proposed TraDES with ITS data. Next, we conduct a data preparation stage which consider a spatiotemporal grouping process, aiming to fuse data from different data sources (see Section 6.3 considering both temporal and spatial dimensions). Then, we filter data, fill missing values using imputation techniques, reduce the number of features and balance the data to feed the next stage. Thereafter, we develop a learning-based model based on Artificial Neural



(a) Road map data.



(b) Traffic level and speed per street Id.

Figure 6.2: Traffic data analysis in Monchengladbach.

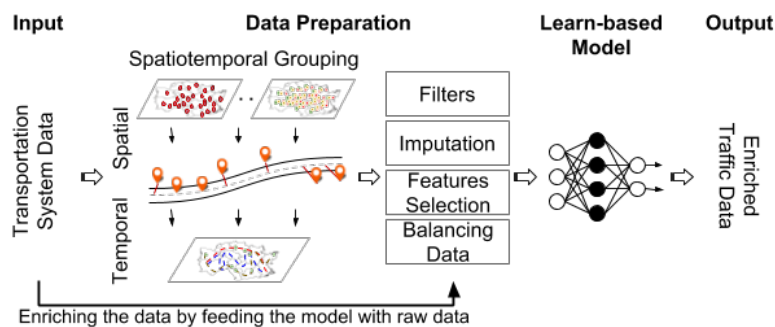


Figure 6.3: Design of TraDES.

Networks (ANN) to identify potential traffic conditions by considering the fusion of different data sources and a set of vehicular sensors data. Finally, we evaluate our approach by feeding the model with raw vehicular data and obtain as output enriched traffic data. Hereafter, we describe each stage of TraDES, as depicted in Fig. 6.3.

6.4.1 Input and Output Data

The TraDES methodology was developed to allow the entry of raw transportation system data and get as output enriched road traffic data. In a general way, our methodology do not pose any restriction on using different types of ITS data sources as input to the model. However, we conduct our case study with vehicular OBD traces and data from road traffic networks. The results of our data fusion process provide to end users and traffic planners a novel and enriched traffic sensor for the uncovered road traffic networks.

6.4.2 Data Preparation

6.4.2.1 Spatiotemporal Grouping

The proposed spatiotemporal grouping takes into account heterogeneous data sources and their spatiotemporal coverage. Therefore, we develop an approach that merges the vehicular traces layer with the road traffic layer based according to both dimensions (i.e., spatially and temporally). We describe the spatiotemporal grouping in Algorithm 2, where the inputs are the vehicular trace, traffic

data, and the traffic street coordinates. The result of such process is an updated vehicular trace dataset containing the traffic condition.

Algoritmo 2: Spatiotemporal Traffic Data Grouping

Input: trace data, traffic data, traffic street coordinates
Result: traces grouped by road traffic condition

```

1 initialization;
2 /* creating a time-stamp for each st. seg. id */
3 trafficStreetCoord.timesTamp ← trafficTime(trafficStreetCoord);
4 /* get the OSM street id by GPX data format */
5 traceData.streetId ← mapMatching(getGPX(traceData));
6 trafficStCd.streetId ← mapMatching(getGPX(trafficStCd));
7 /* merge the OSM street id to each traffic observation */
8 trafficData.streetId ← mergeTrafficStreetId(trafficStCd);
9 for each element in traceData do
10 | /* subset of traffic data with same streetId of traceData
11 | */
11 | traffic = subset(trafficData, streetId == traceData.streetId)
12 | for each element in traffic do
13 | | /* temporal filter by day or hour */
14 | | if TemporalFilter(traffic.timesTamp, traceData.deviceTime) then
15 | | | traceData.FF ← traffic.FF;
16 | | | traceData.JF ← traffic.JF;
17 | | | traceData.SP ← traffic.SP;
18 | | | traceData.SU ← traffic.SU;
19 | | end
20 | end
21 end

```

a) **Spatial:** The spatial grouping is performed by following the approach developed by Marchal et al. [2004], which aims to conduct a map-matching process to identify the route on transportation network that the GPS coordinate actually took. In Algorithm 2 (Line 2), we add a time-stamp to each street segment in the traffic street coordinates, modifying the data to a trace based format. After that, we convert the traffic street coordinates and trace data to a GPX format (Lines 3-4), where we have the following data structure [*'id'*, *'longitude'*, *'latitude'*, *'timestamp'*]. This allows it to be fed to the next step, the map-matching approach

(Lines 3-4) by using the TrackMatching API developed by Marchal et al. ⁵.

In the following, we briefly describe the map-matching approach (see [Marchal et al., 2004] for more details). We begin with a road network acquired from the OSM ⁶ and modelled as a directed graph $G(V, E)$, where V is the set of vertexes (coordinates of a given street segment) and E is the set of edges (link between those coordinates). Consider P_i as the set of coordinate points (x_i, y_i) and timestamp t_i ($i = 1 \dots n$), T_c as a stream of Trace, and T_f as a stream of Traffic Map, where $P_i \in T_c$ and $P_i \in T_f$. The distance is calculated using the euclidean distance between P and the oriented edge AB. Bellow, we define the distance:

$$d(P, AB) = \begin{cases} d_e(P, P') & \text{if } P' \in [AB] \\ \min\{d_e(P, A), d_e(P, B)\} & \text{elsewhere} \end{cases}$$

Where P' is the projection of P on the link AB and d_e denotes the euclidean distance. Based on that definition, the distance $d_{p,AB}$ is equally distant from the opposite segment direction $d_{p,BA}$. Then, we introduce a perpendicular shift λ to the road segment reflecting the distance between the middle of the road and the middle of the driving lanes. After calculating the distance between P and the street segments, the score of a path is measured in order to estimate the algorithm error.

Based on that approach, we conduct a map-matching of T_c and T_f , resulting in the accurate identification of each street segment ID for a given P with a precision of about 10 m/pt. Then, with both sets of data converged to the same street identification, we are able to spatially fuse the vehicular trace and the road traffic data.

b) *Temporal:* After the spacial grouping using a map-matching method, we conduct the temporal grouping to comprehend the vehicles' behavior and the traffic surrounding. In our Algorithm 2 (Lines 6-16), we select each element of the vehicular trace and submit it to the temporal validation together with the traffic data. The temporal data granularity can be coarse-grained (traffic summary per day) or fine-grained (traffic summary per hour) (Line 9). For this TraDES ver-

⁵<https://mapmatching.3scale.net/>

⁶<https://www.openstreetmap.org>

sion, due to the computational costs, we perform the temporal grouping based on coarse-grained traffic data. Therefore, the traffic information, such as FF (Free flow speed), JF (Jam Factor), SP (Speed capped by speed limit), and SU (Speed not capped by speed), is fused to the current road traffic data (Lines 10-13).

6.4.2.2 Filters

We conduct our analysis by considering the premise that even when using only vehicular sensor data it is possible to provide valuable information about the traffic behavior. Based on this premise, we eliminate from the collected data all variables that present issues such as outliers, conflict, incompleteness, ambiguity, correlation, and disparateness or does not reflect the traffic behavior [Rettore et al., 2016a]. Thus, nine variables out of 30 were preserved, where eight features corresponds to the vehicular data and one to the road traffic data. Table 6.1 highlights the selected variables (*) for the next stage of data preparation.

6.4.2.3 Imputation

When analyzing the vehicular sensors data we noticed that they had randomly spread gaps on the dataset. A problem that arises when using sensor data to monitor and control entities, especially vehicles, is its reliability regarding both availability and quality of information. A sensor must output correct readings constantly, and our approach depend on these characteristics to operate properly. However, every sensor has an inherent probability of presenting a malfunction on each one of these aspects.

In this sense, there are two possible solutions. First, it is to temporarily replace the real sensors by a virtual sensor, which collects data from other sensors and outputs data according to models or formulas. Second, it is to apply imputation techniques to fill the gaps on the data. In this work, we focus on imputation techniques, specifically interpolation methods. Once we have to deal with time-series and there is no seasonality on the vehicular traces, despite some trends, we use a simple linear interpolation. Then, for each car $C = (T, F)$, where T is the set of trips, F the set of features; f is a single feature, where $f \in F$ and i is its

index, next we present the interpolation equation:

$$f_{i+1} = f_i + (f_{i+2} - f_i)/2$$

As a result of the imputation stage, we are able to fill sensor data gaps, such as fuel, CO₂, and RPM, that presented reading errors, storage or sensor fails in the data acquisition process. This stage increases the amount of data that can be used in the data analytics process.

6.4.2.4 Features Selection

This step aims at identifying the best set of features to feed the TraDES and still obtain a high accuracy while maintaining lower computational costs, such as processing time, memory and storage capacities. Once the data has irrelevant features, they can decrease the accuracy of the models evaluated. In this way, performing the features selection process before modeling our data may reduce the over-fitting, improves the accuracy and reduce the training time.

We perform four techniques to reduce the number of vehicle's features. Table 6.3 show those techniques and the features selected by each one. The first one asks the *User* to choose the features the he/she guesses best describe the traffic condition. The second technique is the *Principal Component Analysis* (PCA) to extract a set of relevant features. This process identifies the most variable information from a multivariate dataset and expresses it as a set of new features – Principal Components (PCs). These PCs represent the directions along which the variation in the data is maximal.

We also apply the Recursive Feature Elimination (RFE) technique, which aims at selecting those features that fit a model resulting in high accuracy. The RFE rank those features by the model's coefficient or feature importances attributes, recursively eliminating the dependencies and collinearity that may exist in the model. Finally, the Feature Importance (FI) is calculated using the Extra Trees Classifier, which computes the relative importance of each feature. In other words, that technique calculates the probability of reaching a node as the number of samples that reach the node divided by the total number of samples. The higher the value the more important the feature.

Table 6.3: Set of features resulted by each selection technique.

Technique	Features			
User (ALL)	MAF	Speed	RPM	Engine Load
	Fuel	CO ₂	Throttle Position	Intake Air Temp
PCA	MAF	Intake Air Temp	RPM	CO ₂
	Fuel			
RFE	MAF	Speed	CO ₂	Intake Air Temp
FI	MAF	Speed	RPM	Intake Air Temp

6.4.2.5 Balancing Data

We noticed an imbalance on the dataset once we grouped the Jam Factors in two groups and the number of observations with *Traffic* is bigger than the *Non-Traffic* ones. In this case, we explored the re-sampling techniques, which aim at balancing classes either by increasing the frequency of the minority class (Over-sampling) or by decreasing the frequency of the majority class (Under-sampling). Our goal was to approximately obtain the same number of observations for both classes.

We combine two techniques to deal with imbalanced data. The Synthetic Minority Over-sampling Technique (SMOTE) uses the k-Nearest Neighbors (KNN) algorithm to find similar observations for minority class, and randomly choose one of the KNN to create the synthetic samples in the space. Next, we apply the Tomek links algorithm, which looks for pairs of opposite instances classes that are nearest neighbors and removes the majority instance of the pair. Tomek link aims at making clear the border between the minority and majority classes, making the minority regions more distinct.

These strategies help to improve the accuracy of our proposal, since a reduced amount of data may introduce bias to the learning-based model. For that reason, this step is not limited to these approaches, as it always depends on the quality and quantity of the acquired transportation system data.

6.4.3 Learning-based Model

The learning-based model is fed with the vehicular trace labeled with *Traffic* and *Non-Traffic* information. Even though the road traffic data provides levels between 0 to 10 (Jam Factors), we group these levels in two groups, where the *Non-Traffic* label corresponds to the traffic level 0 and the *Traffic* label corresponds to traffic levels between 2 to 10. Notice that, traffic level 1 is considered an intermediate traffic level (*Low-Traffic*, which introduces bias to our model since the vehicle behaves in the same way as in a traffic level 0 and in a traffic level between 2 to 10. In other words, that intersection makes it difficult to decide which traffic level better suits to the vehicular traces with level 1 of traffic. Then, the *Low-Traffic* was discarded in this approach due to the demand of more vehicular traces and traffic data spatiotemporally grouped. Table 6.4 summarizes the data that feed the learning-based traffic model.

Table 6.4: Data to feed the learning-based model.

Jam Factors	Traffic State	Sample	Goal
0	Non-Traffic	3,216	Training/Test
2 to 10	Traffic	9,291	
Not Covered		234,315	Traffic Detect

In this way, we start to deal with a data enrichment problem, which aims at training a model to identify the current traffic state (*Traffic* and *Non-Traffic* through vehicular features. First, we choose the most common classification algorithms (*kernels* to separate these two classes, such as Multi-Layer Perceptron (MLP), Support Vector Machine (SVM), KNN and Random Forest Classifier (RF).

Based on the previous stages, we conduct an exploratory approach to identify the hyper-parameters of each *kernels*, which results in better accuracy. We use a *GridSearchCV* class from Scikit-Learn API [Pedregosa et al., 2011], which takes a set of parameters and values to exhaustively combine them, aiming at finding the best configuration. Knowing that the complexity of such search grows exponentially with the number of parameters, we define a set of parameters for each *kernel* following some guidelines. For the SVM, we rely on [Hsu et al., 2003], and for the other ones, we follow the user’s guide for Auto-WEKA [Kotthoff et al., 2017].

After evaluating the hyper-parameters of each *kernels*, we notice that the MLP was able to separate the classes *Traffic* and *Non-Traffic* using the vehicular features while the other *kernels* showed limited results. The MLP is built on Neural Network (NN), which aims at performing information processing based on the brain neurons structures. Because the human brain is able to learn and make decisions based on learning, NN must do the same. Thus, a neural network can be interpreted as a processing scheme capable of storing knowledge based on learning (experience) and making this knowledge available to the application.

Therefore, we choose the MLP classifier that trains using Backpropagation [Ng et al., 2011] as TraDES's learning algorithm. MLP learns a function $f(\cdot) : R^v \rightarrow R^t$, where v is the vehicle features and t is the traffic state. One benefit of MLP is that it can learn a non-linear function for classifying more complex traffic contexts. Concerning the applicability on the ITS context, our experiments show that a MLP can be applied to accurately predict the traffic state using vehicular sensors data, thus increasing the traffic data quality, such as its spatiotemporal data coverage.

Thereafter, we split the data into two sets, following the convention of most machine learning approaches: Training Set, corresponding to 70% of the entire dataset; and Test Set, corresponding to 30% of the entire dataset. To validate the training process, we applied the cross-validation considering 10 folds split in 70% and 30% of the training and test, respectively. Our goal is to evaluate the training curve and the testing curve, avoiding possible over-fitting and under-fitting. That partition is conducted for each feature selection technique.

After training the NN, we can feed TraDES with vehicular trace data prepared according to Section 6.4.2 and input it to the learning model. At the end of the process, TraDES outputs the vehicular trace data with the current traffic state, thus enriching the road network data with traffic state, averaged fuel consumption, emissions, speed and so on.

6.5 Evaluation

In this section, we evaluate TraDES by considering the vehicle's features selection and spatiotemporal data coverage. After conducting an exploratory analysis of the

classification algorithms, hyper-parameters and the feature selection approach, we present the results regarding the Training and Test process in Figure 6.4. We validate our training process by performing a *Cross-validation approach*, which aims at splitting the training set in both training and validation sets among 10 folds. Figure 6.5 shows the learning curve of the MLP kernel, which is an essential procedure to prove the generalization of our model, avoiding over-fitting and under-fitting.

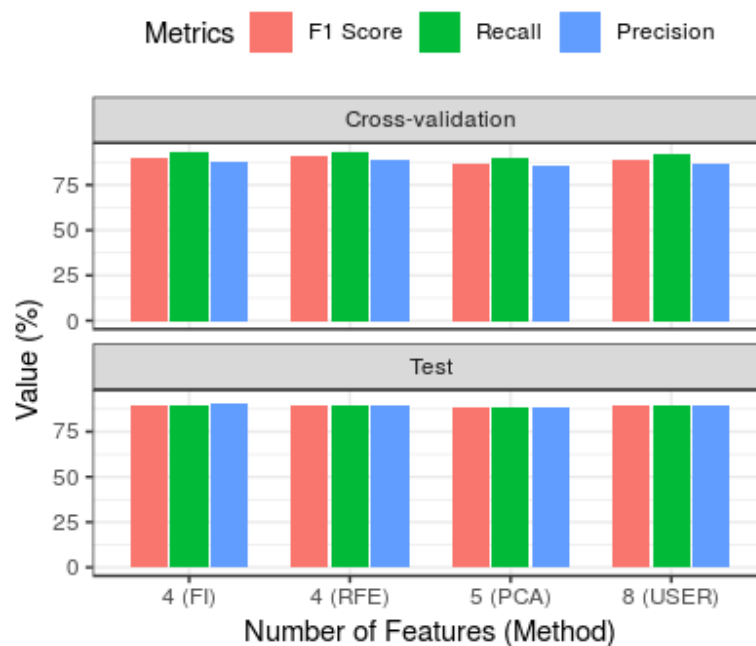


Figure 6.4: Metrics per set of features.

In our analysis, the RFE and FI algorithms selected the best set of vehicle features, with both achieving the same score, around 90%. An important lesson learned here is that the application of a consistent methodology is able to provide a generalization model to detect traffic condition.

Figure 6.4 shows the best set of features that can be used as input to TraDES. We evaluate our model using three metrics on the Cross-validation and Test, considering the confusion matrix created by each set of features. For instance, the *F1 score*, *Recall* and *Precision* report an accuracy around 90% on average for FI and RFE. All features report an accuracy around 89%, however by introducing higher computational costs when compared to the other ones. The accuracy decreases

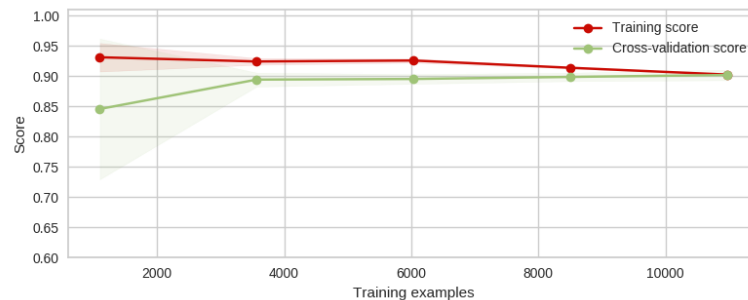
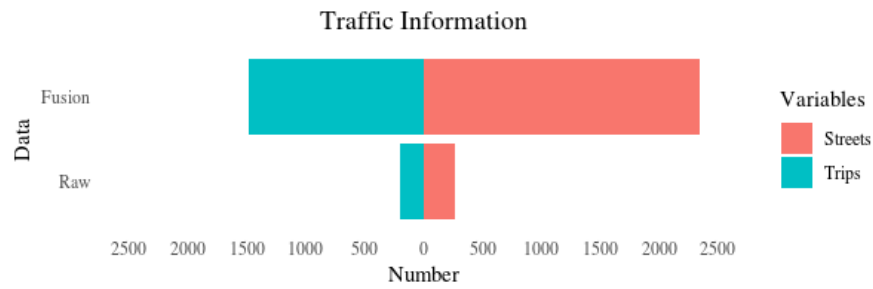


Figure 6.5: Learning curve of RFE algorithm.

to 87% when using PCA to remove non-representative features, which can be explained by the increase of conflicts among those features, i.e., the lack of the speed feature may turn the learning process difficult.

After validating the TraDES' methodology and proving the generalization of the learning-based model, we give it as input raw vehicular traces that do not show traffic conditions, as showed in Table 6.4. Thereafter, the traffic sensor outputs enriched traffic data, thus allowing the evaluation of the benefits of our heterogeneous data fusion approach for ITS. Figure 6.6 shows the coverage of street segments and vehicular trips, based on raw data and fusion data. The raw data consists of vehicular traces and traffic condition at the same time and space, while the fusion data consists of the whole traffic condition provided by the use of vehicular traces as input to a learning-model. These analyses enable us to see the macro and micro benefits of TraDES, that enabled to increase the number of trips from less than 300 to abounding 1,500, and the number of streets covered from almost 400 to around 2,400.

Figure 6.7 shows the spatial data coverage, highlighting the traffic condition when considering raw and fusion data. As you can see, there are specific streets that constantly have traffic jam (*Traffic* while others have free traffic (*Non-Traffic*). The benefits of TraDES' approach is clear when we look at the *Traffic* condition in raw and fusion data. For instance, consider that a navigation system makes use of one of those traffic sources (Raw and Fusion) to its routes suggestion services. Certainly, it performs differently when using each one of them. In other words, navigation systems with access to enriched traffic data is better equipped to suggest better routes by avoiding as much as possible bad traffic conditions, differently



(a) Number of trips and streets.



(b) Number of trips and streets per city.

Figure 6.6: Evaluation of trips and street segments between the raw data and the fused data.

from systems with access to only raw traffic data.

Notice that, TraDES increases the number of streets covered, but also the number of traces which pass through those streets. TraDES also allows exploring the whole sensors embedded on those vehicular traces. Then, we can find the amounts of kilometers, emissions, fuel consumption, hours spent on a given traffic condition, and so on. Figure 6.8a shows the frequency of street use by vehicles between the raw data and fusion one. Besides TraDES' evaluation, Figure 6.8b shows the sum of the total kilometers traveled, emissions (CO_2 , fuel consumption and hours spent on the roads, when considering raw and fusion data. With such

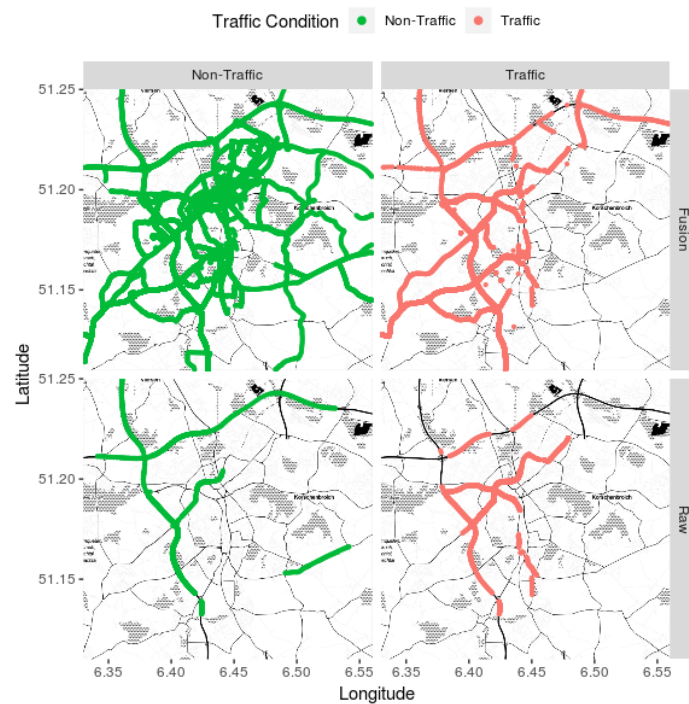
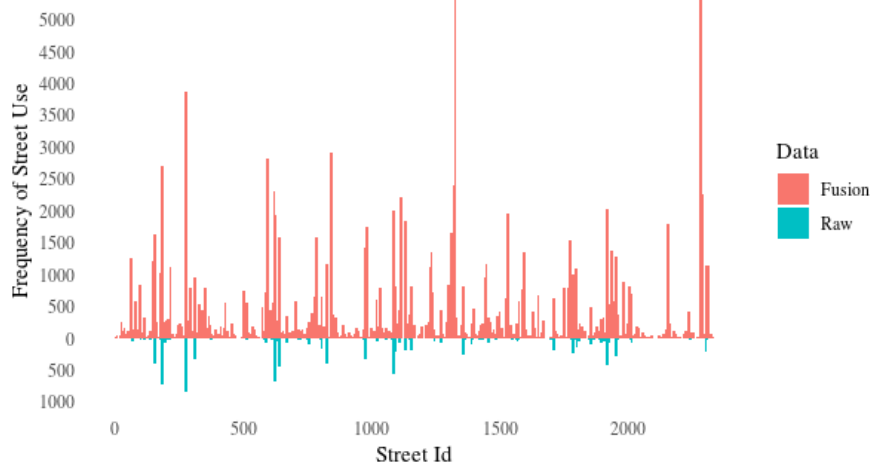


Figure 6.7: Traffic map coverage between the raw data and fused data in Monchengladbach.

analysis, we aim at enabling users and road managers to better understand the traffic behavior and help plan investments in a given area.

6.6 Chapter Remarks

In this chapter, we presented Traffic Data Enrichment Sensor (TraDES), a low-cost traffic sensor for ITSs based on Heterogeneous Data Fusion. TraDES is able to infer the traffic condition on regions that do not have any reported traffic data, thus providing navigation systems, road planners and the general public more consistent, accurate and useful information about the traffic in a given area. TraDES is also able to enhance the route information of current navigation tools, improving the road traffic data quality and enriching the current spatiotemporal data coverage. It provides to users and traffic planners an overview of the traffic condition, fuel consumption, emissions, streets' use frequency by fusing data from different



(a) Frequency of street segments use between the raw data and the fused data.



(b) Data coverage between the raw and the fused data.

Figure 6.8: Evaluation of trips and street segments between the raw data and the fused data.

data sources.

In summary, Figure 6.9 shows how our design of fusion on VDS worked in this study. Where, the OBD vehicular sensors and road map data feed the fusion process, the data preparation deal with data aspects showed in Chapter 3 and others which help to treat the data for the data processing which covers methods related to the application goals and finally resulting in the TraDES as the data use.

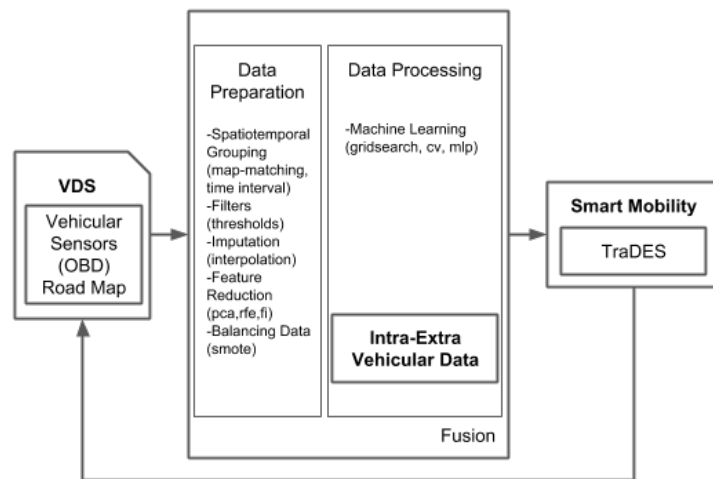


Figure 6.9: Design of fusion on VDS for TraDES.

Chapter 7

Final Remarks

7.1 Conclusions

In this thesis, we have proposed a general approach which organizes methods and techniques to enable heterogeneous data fusion on the Vehicular Data Space (VDS), aiming to achieve a set of Smart Mobility (SM) goals. We categorize the data from the VDS into Intra-Vehicle Data (IVD) and Extra-Vehicle Data (EVD) perspectives, allowing to identify challenges and open issues to perform data fusion. By showing a set of applications and services to improve the data quality of Intelligent Transportation System (ITS), we highlighted methods and techniques to address those goals such as mathematical methods (equations, operations), threshold filters, statistics (distributions), geofencing, fuzzy logic, feature reduction, machine learning (supervised and unsupervised classification), correlations, algorithms to deal with spatiotemporal data grouping, data balancing, graph modeling, natural language processing, and imputations methods. This thesis differentiates the data fusion into three main categories – Intra-Vehicle Data (IVD), Extra-Vehicle Data (EVD), and Intra and Extra-Vehicle Data (IEVD), which cover the whole applications and services in Intelligent Transportation System (ITS). We have also shown a lack of studies dealing with data fusion of EVD and Intra and Extra-Vehicle Data (IEVD), which we also advanced the state-of-the-art.

Our comprehensive study showed that the use of heterogeneous data fusion techniques have the potential to improve the accuracy of applications and services

of ITS when there are several related descriptors. It is also clear that novel ITS applications will benefit from multiple heterogeneous datasets. Through the use of different techniques, this thesis made the following contributions:

- A vast literature review to provide the concept of VDS and the state-of-the-art applications and services developed to ITS.
- We presented a methodology to develop applications and services to SM based on the ITS data cycle stages;
- We designed IVD fusion and proposed a methodology to detect a legitimate/illegitimate driver, resulting in an accuracy above 98%. We also developed a virtual gear sensor for manual transmission, and used it in an eco-driving methodology that analyzes the vehicle's historical sensor data to suggest a gear shift. The results showed more efficient fuel consumption, emissions, and reduced vehicle maintenance;
- Based on the vehicle's surrounding data, we designed Extra-Vehicle Data (EVD) fusion that combines the user's viewpoint and road data. We proposed the Road Data Enrichment (RoDE) with two main services: route service and incident service. The former service provides three route description services (Route Sentiment (RS), Route Information (RI) and Area' Tags (AT)) that aim to enhance the route information. The latter service proposes a methodology to detect road events achieving scores above 90%, allowing us to understand Location-Based Social Media (LBSM) user's viewpoint, regarding the transit events and points of interest; and (iv) Intra and Extra-Vehicle Data (IEVD) fusion, where we propose the Traffic Data Enrichment Sensor (TraDES) to fill the road spatiotemporal data gaps, using vehicular trace and road data, improving the data quality allowing a reliable route suggestion.

7.2 Future Work

There are different extensions that we can follow, based on Figure 1.2, given the richness of VDS and the combinations of data preparation and heterogeneous data

fusion techniques. For example, add contextual information to the datasets such as traffic conditions and driver's behavior to vehicular mobility traces. The deployment of virtual sensors may be used to validate the utility and operations of real sensors. In order to deploy these virtual sensors, a physical platform with access to an On-Board Diagnostic (OBD) port and a processing unit is needed.

We can further improve the gear virtual sensor to show the transitions between gears and add this feature to the analysis of the driver's behavior. This will eventually lead to an effective gear change service based on fuel consumption and torque. We can also evaluate the recommendation service simulation as a real-time service, through a smartphone application that provides the gear suggestion to the driver. We can then compare the results with the *Gear Shift Indicator (GSI)*, a solution developed by companies as *Ford*, *BMW*, *Renault* and *Fiat* to guide the best gear to use in order to reduce the fuel consumption. It is also possible to design gamification strategies to encourage multiple drivers to improve a desired aspect of their behavior and also evaluate how much the driver recognizes the suggested gear as a good option.

The driver behavior authentication may also be expanded by embedding the system to the vehicle and apply different machine learning algorithms as well as report evaluation metrics. In addition, we plan to investigate the authentication computational cost, taking into account the vehicle's features and evaluate solutions to circumvent the presence of suspects in Vehicular *Ad-hoc* Networks (VANETs).

The RoDE approach showed a great potential to explore new research ideas, such as the extension of Twitter MAPS (T-MAPS) route description by applying strategies to further increase the information quality. Besides that, we can employ regular users' accounts from LBSM and use reputation models to handle conflicting information. The incident service (Twitter Incident (T-Incident)) may be extended to web version. Moreover, adding more specialist accounts, and improving the current identification and description results. Another possibility is to develop strategies to eliminate the specialist intervention in the feature's extraction stage. Moreover, based on the T-Incident results, it is possible to design an incident prediction service and incident duration time. It is also possible to provide and evaluate different vehicular routes based on incident descriptions. Upgrade the

T-Incident to an online version can be a step forward to improve the current transportation systems.

Finally, the fusion of IVD and EVD allows to create the TraDES' route suggestion services based on the increased traffic data coverage. Also, we can expand the data analyses by considering other types of data and data sources, such as weather and social media, which may be beneficial to develop solutions to SM.

7.3 Comments on Publications

In the following, we list all publications obtained during the doctorate. Papers in Section 7.3.1 are direct results of this thesis. Results from collaborations in other research projects related to Internet of Things (IoT), which also considered data fusion concepts, are shown in Section 7.3.2

7.3.1 Contributions from the Thesis

Conference Publications:

- Rettore, P. H., André, B. P. S., Campolina, Villas, L. A., and A.F. Loureiro, A. (2016a). Towards intra-vehicular sensor data fusion. In *Advanced perception, Machine learning and Data sets (AMD'16) as part of the 2016 IEEE 19th International Conference on Intelligent Transportation Systems (ITSC 2016)*, , Rio de Janeiro
- Rettore, P. H., Campolina, A. B., Villas, L. A., and Loureiro, A. A. (2016b). Identifying relationships in vehicular sensor data: A case study and characterization. In *Proceedings of the 6th ACM Symposium on Development and Analysis of Intelligent Vehicular Networks and Applications, DIVANet '16*, pages 33–40, New York, NY, USA. ACM
- Campolina, A. B., Rettore, P. H. L., Machado, M. D. V., and Loureiro, A. A. F. (2017). On the design of vehicular virtual sensors. In *2017 13th International Conference on Distributed Computing in Sensor Systems (DCOSS)*, pages 134–141, Ottawa, Canada. ISSN

- Rettore, P. H. L., Campolina, A. B., Villas, L. A., and Loureiro, A. A. F. (2017). A method of eco-driving based on intra-vehicular sensor data. In *2017 IEEE Symposium on Computers and Communications (ISCC)*, pages 1122–1127, Heraklion, Greece. IEEE. ISSN
- Santos, B. P., Rettore, P. H. L., Ramos, H., Vieira, L. F., and Loureiro, A. A. F. (2017). T-maps: Modelo de descrição do cenário de trânsito baseado no twitter. In *(SBRC 2017)*
- Rettore, P. H. L., Campolina, A., Luis, A., de Menezes, J. G. M., Villas, L., and Loureiro, A. A. F. (2018b). Benefícios da autenticação de motoristas em redes veiculares. In *(SBRC 2018)*, Campos do Jordão, Brazil
- Rettore, P. H., Campolina, A., de Souza, A. L., Maia, G., Villas, L. A., and A.F. Loureiro, A. (2018a). Driver authentication in VANETs based on Intra-Vehicular sensor data. In *2018 IEEE Symposium on Computers and Communications (ISCC) (ISCC 2018)*, Natal, Brazil
- Santos, B. P., Rettore, P. H., Ramos, H. S., Vieira, L. F. M., and A.F. Loureiro, A. (2018). Enriching traffic information with a spatiotemporal model based on social media. In *2018 IEEE Symposium on Computers and Communications (ISCC) (ISCC 2018)*, Natal, Brazil
- Rettore, P. H. L., Araujo, I., de Menezes, J. G. M., Villas, L., and Loureiro, A. A. F. (2019). Serviço de detecção e enriquecimento de eventos rodoviários baseado em fusão de dados heterogêneos para vanets. In *SBRC 2019*, Gramado, Brazil

Journal Publications:

- Vehicular Data Space. IEEE Communications Surveys and Tutorials

Book chapters:

- Arya, K. V., Bhadoria, R. S., and Chaudhari, N. S. E. (2018). *Emerging Wireless Communication and Network Technologies*. Springer Nature. Chapter: Vehicular Networks to Intelligent Transportation Systems

Tutorials:

- Cunha, F. D., Maia, G., Celes, C., Guidoni, D., de Souza, F., Ramos, H., and Villas, L. (2017). Sistemas de Transporte Inteligentes: Conceitos, Aplicações e Desafios. In *(SBRC 2017 - Minicursos)*

Conference Publications Under Review:

- International Conference on Distributed Computing in Sensor Systems (DCOSS)- TraDES: Traffic Data Enrichment Sensor based on Heterogeneous Data Fusion for ITS

Journal Publications Under Review:

- IEEE Transactions on Intelligent Transportation Systems - RoDE: Road Data Enrichment Framework based on Heterogeneous Data Fusion for ITS

7.3.2 Other Publications**Conference Publications:**

- Santos, B. P., Rettore, P. H., Vieira, L. F. M., and A.F. Loureiro, A. (2019). Dribble: a learn-based timer scheme selector for mobility management in IoT. In *2019 IEEE Wireless Communications and Networking Conference (WCNC) (IEEE WCNC 2019)*, Marrakech, Morocco

Bibliography

- Abdelhamid, S., Hassanein, H. S., and Takahara, G. (2015). Vehicle as a resource (VaaR). *IEEE Network*, 29(1):12--17. ISSN 08908044.
- AbuAli, N. (2015). Advanced vehicular sensing of road artifacts and driver behavior. <http://ieeexplore.ieee.org/document/7405452/>.
- Abut, H., Erdoğan, H., Erçil, A., Çürüklü, A. B., Koman, H. C., Tas, F., Argunşah, A. Ö., Akan, B., Karabalkan, H., Çökelek, E., et al. (2007). Data collection with " uyanik": too much pain; but gains are coming.
- Administration, F. H. (2016). Highway statistics 2015.
- Agency, U. S. E. P. (2017). SmartWay - United States Environmental Protection Agency. <https://www.epa.gov/smartway>. Accessed: May 17, 2017.
- Ahmed, M., Saraydar, C. U., ElBatt, T., Yin, J., Talty, T., and Ames, M. (2007). Intra-vehicular Wireless Networks. In *2007 IEEE Globecom Workshops*, pages 1--9, Washington, DC, USA. IEEE.
- Ahmed, Q., Bhatti, A. I., and Iqbal, M. (2011). Virtual sensors for automotive engine sensors fault diagnosis in second-order sliding modes. *IEEE Sensors Journal*, 11(9):1832--1840. ISSN 1530437X.
- Ahn, K. and Rakha, H. (2008). The effects of route choice decisions on vehicle energy consumption and emissions. *Transportation Research Part D: Transport and Environment*, 13(3):151--167. ISSN 13619209.
- Aloul, F., Zualkernan, I., Abu-Salma, R., Al-Ali, H., and Al-Merri, M. (2015). iBump: Smartphone application to detect car accidents. *Computers & Electrical Engineering*, 43:66--75. ISSN 00457906.
- Andrieu, C. and Pierre, G. S. (2012). Using statistical models to characterize eco-driving style with an aggregated indicator. In *2012 IEEE Intelligent Vehicles Symposium*, pages 63--68, Alcalá de Henares, Spain. IEEE. ISSN 14746670.

- Angkititrakul, P., Hansen, J. H. L., Choi, S., Creek, T., Hayes, J., Kim, J., Kwak, D., Noecker, L. T., and Phan, A. (2009). *UTDrive: The Smart Vehicle Project*, pages 55--67. Springer US, Boston, MA.
- Aoude, G. S., Desaraju, V. R., Stephens, L. H., and How, J. P. (2011). Behavior classification algorithms at intersections and validation using naturalistic data.
- Apple (2014). Car Play - Apple. <https://www.apple.com/ios/carplay/>. Accessed: May 10, 2017.
- Aquino, A. L., Cavalcante, T. S., Almeida, E. S., Frery, A. C., and Rosso, O. A. (2015). Characterization of vehicle behavior with information theory. *The European Physical Journal B*, 88(10):257. ISSN 1434-6036.
- Araújo, R., Igreja, Â., De Castro, R., and Araújo, R. E. (2012). Driving coach: A smartphone application to evaluate driving efficient patterns. *IEEE Intelligent Vehicles Symposium, Proceedings*, 1(1):1005--1010. ISSN 1931-0587.
- Arya, K. V., Bhadoria, R. S., and Chaudhari, N. S. E. (2018). *Emerging Wireless Communication and Network Technologies*. Springer Nature.
- Atkinson, C. M., Long, T. W., and Hanzevack, E. L. (1998). Virtual sensing: a neural network-based intelligent performance and emissions prediction system for on-board diagnostics and engine control. *Progress in Technology*, 73(301-314):2--4.
- Audi (2014). Audi Connect. <https://www.audiusa.com/help/audi-connect>. Accessed: May 10, 2017.
- AXA (2013). AXA Drive. <https://www.axa.com>. Accessed: July 1, 2017.
- Ayed, S. B., Trichili, H., and Alimi, A. M. (2015). Data fusion architectures: A survey and comparison. In *2015 15th International Conference on Intelligent Systems Design and Applications (ISDA)*, pages 277--282. IEEE.
- Ayyildiz, K., Cavallaro, F., Nocera, S., and Willenbrock, R. (2017). Reducing fuel consumption and carbon emissions through eco-drive training. *Transportation Research Part F: Psychology and Behaviour*, 46:96--110. ISSN 13698478.
- Bank, W. (2017). The World Bank. <http://www.worldbank.org/>. Accessed: May 15, 2017.
- Bazzan, A. L. and Klügl, F. (2013). Introduction to intelligent systems in traffic and transportation. *Synthesis Lectures on Artificial Intelligence and Machine Learning*, 7(3):1--137.

- Beanland, V., Fitzharris, M., Young, K. L., and Lenn?, M. G. (2013). Driver inattention and driver distraction in serious casualty crashes: Data from the Australian National Crash In-depth Study. *Accident Analysis & Prevention*, 54:99–107. ISSN 00014575.
- Bengler, K., Dietmayer, K., Farber, B., Maurer, M., Stiller, C., and Winner, H. (2014). Three decades of driver assistance systems: Review and future perspectives. *IEEE Intelligent Transportation Systems Magazine*, 6(4):6–22. ISSN 15249050.
- Bergasa, L. M., Almería, D., Almazán, J., Yebes, J. J., and Arroyo, R. (2014). Drivesafe: An app for alerting inattentive drivers and scoring driving behaviors. In *2014 IEEE Intelligent Vehicles Symposium Proceedings*, pages 240–245, Dearborn, MI, USA. IEEE. ISSN 1931-0587.
- Bertrand, K. Z., Bialik, M., Virdee, K., Gros, A., and Bar-Yam, Y. (2013). Sentiment in new york city: A high resolution spatial and temporal view. *arXiv preprint arXiv:1308.5010*.
- BMW (2014). BMW ConnectedDrive. <http://www.bmwusa.com/standard/content/innovations/bmwconnecteddrive/connecteddrive.aspx>. Accessed: May 10, 2017.
- Boada, B., Boada, M., and Diaz, V. (2016a). Vehicle sideslip angle measurement based on sensor data fusion using an integrated anfis and an unscented kalman filter algorithm. *Mechanical Systems and Signal Processing*, 72:832–845.
- Boada, B., Boada, M., and Diaz, V. (2016b). Vehicle sideslip angle measurement based on sensor data fusion using an integrated ANFIS and an Unscented Kalman Filter algorithm. *Mechanical Systems and Signal Processing*, 72-73:832–845. ISSN 08883270.
- Board, J. T. S. (2017). Japan Transport Safety Board. <https://www.mlit.go.jp/jtsb/english.html>. Accessed: May 15, 2017.
- Brace, C., Hari, C. J., Akehurst, D., Poxon, S., and Ash, J. (2013). Development and Field Trial of a Driver Assistance System to Encourage Eco Driving in Light Commercial Vehicle Fleets. *Ieee-Inst Electrical Electronics Engineers Inc*, 14(2):796–805. ISSN 1524-9050.
- Bröring, A., Remke, A., Stasch, C., Autermann, C., Rieke, M., and Möllers, J. (2015). enviroCar: A Citizen Science Platform for Analyzing and Mapping Crowd-Sourced Car Sensor Data. *Transactions in GIS*, 19(3):362–376. ISSN 13611682.
- Brundell-Freij, K. and Ericsson, E. (2005). Influence of street characteristics, driver category and car performance on urban driving patterns. *Transportation Research Part D: Transport and Environment*, 10(3):213 – 229. ISSN 1361-9209.
- Burton, A., Parikh, T., Mascarenhas, S., Zhang, J., Voris, J., Artan, N. S., and Li, W. (2016). Driver identification and authentication with active behavior modeling. In *12th International Conference on Network and Service Management (CNSM)*.

- Campolina, A. B., Rettore, P. H. L., Machado, M. D. V., and Loureiro, A. A. F. (2017). On the design of vehicular virtual sensors. In *2017 13th International Conference on Distributed Computing in Sensor Systems (DCOSS)*, pages 134–141, Ottawa, Canada. ISSN .
- CarChip (2013). CarChip. <http://www.carchip.cc/>. Accessed: June 19, 2017.
- Carmona, J., García, F., Martín, D., Escalera, A., and Armingol, J. (2015). Data Fusion for Driver Behaviour Analysis. *Sensors*, 15(10):25968–25991. ISSN 1424-8220.
- Castignani, G., Derrmann, T., Frank, R., and Engel, T. (2015). Driver behavior profiling using smartphones: A low-cost platform for driver monitoring. *IEEE Intelligent Transportation Systems Magazine*, 7(1):91–102. ISSN 19391390.
- CGI (2014). Modeling the Relation Between Driving Behavior and Fuel Consumption.
- Chen, K., Lu, M., Tan, G., and Wu, J. (2014). CRSM: Crowdsourcing based road surface monitoring. *Proceedings - 2013 IEEE International Conference on High Performance Computing and Communications, HPCC 2013 and 2013 IEEE International Conference on Embedded and Ubiquitous Computing, EUC 2013*, pages 2151–2158.
- Chen, K., Tan, G., Lu, M., and Wu, J. (2016). CRSM: a practical crowdsourcing-based road surface monitoring system. *Wireless Networks*, 22(3):765–779. ISSN 15728196.
- Chen, M., Challita, U., Saad, W., Yin, C., and Debbah, M. (2017). Machine learning for wireless networks with artificial intelligence: A tutorial on neural networks. *arXiv preprint arXiv:1710.02913*.
- Cheng, H. T., Shan, H., and Zhuang, W. (2011). Infotainment and road safety service support in vehicular networking: From a communication perspective. *Mechanical Systems and Signal Processing*, 25(6):2020–2038. ISSN 08883270.
- Chu, H., Raman, V., Shen, J., Kansal, A., Bahl, V., and Choudhury, R. R. (2014). I am a smartphone and i know my user is driving. In *2014 Sixth International Conference on Communication Systems and Networks (COMSNETS)*, pages 1–8, Bangalore, India. IEEE. ISSN 2155-2487.
- ClickutilityTeam, Innovability, G. E. (2017). Smart mobility worlds. <http://www.smartmobilityworld.net/en/>.
- Commission, E. (2017). eSafety - European Commission. <https://ec.europa.eu/>. Accessed: May 17, 2017.
- Corcoba Magaña, V. and Muñoz Organero, M. (2016). WATI: Warning of Traffic Incidents for Fuel Saving. *Mobile Information Systems*, 2016. ISSN 1875905X.

- Corporation, M. S. (2010). CarSim. <https://www.carsim.com/>. Accessed: June 13, 2017.
- Council, E. T. S. (2017). European Transport Safety Council. <http://etsc.eu/>. Accessed: May 15, 2017.
- Crooks, A., Croitoru, A., Stefanidis, A., and R, J. (2013). # Earthquake: Twitter as a distributed sensor system. *Transactions in GIS*.
- Cunha, F. D., Maia, G., Celes, C., Guidoni, D., de Souza, F., Ramos, H., and Villas, L. (2017). Sistemas de Transporte Inteligentes: Conceitos, Aplicações e Desafios. In *(SBRC 2017 - Minicursos)*.
- D’Agostino, C., Saidi, A., Scouarnec, G., and Liming Chen (2015). Learning-Based Driving Events Recognition and Its Application to Digital Roads. *IEEE Transactions on Intelligent Transportation Systems*, 16(4):2155–2166. ISSN 1524-9050.
- de Francisco, R., Huang, L., and Dolmans, G. (2009). Coexistence of zigbee wireless sensor networks and bluetooth inside a vehicle. In *2009 IEEE 20th International Symposium on Personal, Indoor and Mobile Radio Communications*, pages 2700–2704, Tokyo, Japan. IEEE. ISSN 2166-9570.
- Deery, H. A. and Love, A. W. (1996). The Driving Expectancy Questionnaire: development, psychometric assessment and predictive utility among young drink-drivers. *Journal of Studies on Alcohol and Drugs*, 57(2):193–202.
- Detech, C. (2017). Crash Detech. <https://www.crashdetech.com>. Accessed: June 19, 2017.
- Dubois, D. and Prade, H. (1994). Possibility theory and data fusion in poorly informed environments. *Control Engineering Practice*.
- Elhenawy, M., Jahangiri, A., Rakha, H. A., and El-Shawarby, I. (2015). Modeling driver stop/run behavior at the onset of a yellow indication considering driver run tendency and roadway surface conditions. *Accident Analysis and Prevention*, 83(Ahfe):90–100. ISSN 00014575.
- Engelbrecht, J., Booysen, M. J., Bruwer, F. J., and van Rooyen, G.-J. (2015). Survey of smartphone-based sensing in vehicles for intelligent transportation system applications. *IET Intelligent Transport Systems*, 9(10):924–935. ISSN 1751-956X.
- Engelbrecht, J., Booysen, M. J., and Van Rooyen, G.-J. (2014). Recognition of driving manoeuvres using smartphone-based inertial and GPS measurement. In *The 1st International Conference on the Use of Mobile Information and Communications Technology in Africa (UMICTA 2014)*, pages 88–92, Stellenbosch, South Africa. SUNScholar.

- Ericsson, E., Larsson, H., and Brundell-Freij, K. (2006). Optimizing route choice for lowest fuel consumption - Potential effects of a new driver support tool. *Transportation Research Part C: Emerging Technologies*, 14(6):369--383. ISSN 0968090X.
- Eriksson, J., Girod, L., Hull, B., Newton, R., Madden, S., and Balakrishnan, H. (2008). The pothole patrol: using a mobile sensor network for road surface monitoring. In *Proceeding of the 6th international conference on Mobile systems, applications, and services - MobiSys '08*, page 29, New York, New York, USA. ACM Press.
- Faezipour, M., Nourani, M., Saeed, A., and Addepalli, S. (2012). Progress and challenges in intelligent vehicle area networks. *Communications of the ACM*, 55(2):90--100. ISSN 00010782.
- Faouzi, N.-E. E. and Klein, L. A. (2016). Data Fusion for ITS: Techniques and Research Needs. *Transportation Research Procedia*, 15:495--512. ISSN 23521465.
- Fazeen, M., Gozick, B., Dantu, R., Bhukhiya, M., and González, M. C. (2012). Safe Driving Using Mobile Phones. *IEEE Transactions on Intelligent Transportation Systems*, 13(3):1462--1468. ISSN 15249050.
- Ferdowsi, A., Challita, U., and Saad, W. (2017). Deep learning for reliable mobile edge analytics in intelligent transportation systems. *arXiv preprint arXiv:1712.04135*.
- Finkel, J. R., Grenager, T., and Manning, C. (2005). Incorporating non-local information into information extraction systems by gibbs sampling. In *43rd annual meeting on association for computational linguistics*.
- Fleming, W. J. (2001). Overview of Automotive Sensors. *IEEE Sensors Journal*, 1(4):296--308. ISSN 1530437X.
- Florea, M., Joussetme, A.-L., Bossé, É., and Grenier, D. (2009). Robust combination rules for evidence theory. *Elsevier Information Fusion*.
- Ford (2010). AppLink. <https://developer.ford.com/pages/applink>. Accessed: July 7, 2017.
- Fox, A., Kumar, B. V., Chen, J., and Bai, F. (2015). Crowdsourcing undersampled vehicular sensor data for pothole detection. In *2015 12th Annual IEEE International Conference on Sensing, Communication, and Networking (SECON)*, pages 515--523, Seattle, WA, USA. IEEE.
- French, D. J., West, R. J., Elander, J., and Wilding, J. M. (1993). Decision-making style, driving style, and self-reported involvement in road traffic accidents. *Ergonomics*, 36(6):627--644.
- Ganti, R. K., Pham, N., Ahmadi, H., Nangia, S., and Abdelzaher, T. F. (2010). GreenGPS. In *Proceedings of the 8th international conference on Mobile systems, applications, and services - MobiSys '10*, MobiSys '10, page 151, New York, NY, USA. ACM.

- Geurts, P., Ernst, D., and Wehenkel, L. (2006). Extremely randomized trees. *Machine learning*, 63(1):3--42. ISSN 1573-0565.
- Giachanou, A. and Crestani, F. (2016). Like it or not: A survey of twitter sentiment analysis methods. *ACM Computing Surveys (CSUR)*.
- Giridhar, P., Amin, M., Abdelzaher, T., Wang, D., K, L., George, J., and Ganti, R. (2017). ClariSense+: An enhanced traffic anomaly explanation service using social network feeds. *Pervasive and Mobile Computing*.
- Glendon, A., Dorn, L., Matthews, G., Gulian, E., Davies, D., and Debney, L. (1993). Reliability of the driving behaviour inventory. *Ergonomics*, 36(6):719--726.
- GM (2011). OnStar. <https://www.onstar.com/us/en/home.html>. Accessed: May 10, 2017.
- Goncalves, J., Goncalves, J. S. V., Rossetti, R. J. F., and Olaverri-Monreal, C. (2014). Smartphone sensor platform to study traffic conditions and assess driving performance. In *17th International IEEE Conference on Intelligent Transportation Systems (ITSC)*, pages 2596--2601, Qingdao, China. IEEE.
- Group, P. (1992). Vissim. <http://vision-traffic.ptvgroup.com/en-uk/home/>. Accessed: August 19, 2017.
- Gu, Y., Qian, Z. S., and Chen, F. (2016). From Twitter to detector: Real-time traffic incident detection using social media data. *Transportation Research Part C: Emerging Technologies*, 67:321--342. ISSN 0968090X.
- Guo, F. and Fang, Y. (2013). Individual driver risk assessment using naturalistic driving data. *Accident Analysis & Prevention*, 61:3--9. ISSN 00014575.
- Hallac, D., Sharang, A., Stahlmann, R., Lamprecht, A., Huber, M., Roehder, M., Leskovec, J., et al. (2016). Driver identification using automobile sensor data from a single turn. In *Intelligent Transportation Systems (ITSC), 2016 IEEE 19th International Conference on*, pages 953--958, Rio de Janeiro, RJ, Brazil. IEEE.
- Han, H., Yu, J., Zhu, H., Chen, Y., Yang, J., Zhu, Y., Xue, G., and Li, M. (2014). SenSpeed: Sensing driving conditions to estimate vehicle speed in urban environments. *Proceedings - IEEE INFOCOM*, 15(1):202--216. ISSN 0743166X.
- Hartenstein, H. and Laberteaux, K. (2009). *VANET vehicular applications and inter-networking technologies*, volume 1. John Wiley & Sons, Torquay, UK.
- Hasan, M., Orgun, M., and S, R. (2017). A survey on real-time event detection from the Twitter data stream. *Journal of Information Science*.

- Honda (2015). Honda Sensing.
- Hong, J.-H., Margines, B., and Dey, A. K. (2014). A smartphone-based sensing platform to model aggressive driving behaviors. In *Proceedings of the 32Nd Annual ACM Conference on Human Factors in Computing Systems, CHI '14*, pages 4047--4056, New York, NY, USA. ACM.
- Hsu, C.-W., Chang, C.-C., Lin, C.-J., et al. (2003). A practical guide to support vector classification.
- Imkamon, T., Saensom, P., Tangamchit, P., and Pongpaibool, P. (2008). Detection of hazardous driving behavior using fuzzy logic. *2008 5th International Conference on Electrical Engineering/Electronics, Computer, Telecommunications and Information Technology*, 2:657--660.
- Jeffreys, I., Graves, G., and Roth, M. (2016). Evaluation of eco-driving training for vehicle fuel use and emission reduction: A case study in australia. *Transportation Research Part D: Transport and Environment*, pages -. ISSN 1361-9209.
- Jeng, S. T. and Chu, L. (2015). Tracking heavy vehicles based on weigh-in-motion and inductive loop signature technologies. *IEEE Transactions on Intelligent Transportation Systems*, 16(2):632--641. ISSN 1524-9050.
- Jeong, E., Oh, C., and Kim, I. (2013). Detection of lateral hazardous driving events using in-vehicle gyro sensor data. *KSCE Journal of Civil Engineering*, 17(6):1471--1479. ISSN 12267988.
- Jeong, J. and Oh, T. (2016). Survey on protocols and applications for vehicular sensor networks. *International Journal of Distributed Sensor Networks*, 12(8):1--17. ISSN 15501477.
- Jin, X. and Yin, G. (2015). Estimation of lateral tire-road forces and sideslip angle for electric vehicles using interacting multiple model filter approach. *Journal of the Franklin Institute*, 352(2):686--707.
- Jockers, M. (2017). syuzhet: Extracts sentiment and sentiment-derived plot arcs from text. <https://cran.r-project.org/web/packages/syuzhet/>.
- Johnson, D. A. and Trivedi, M. M. (2011). Driving style recognition using a smartphone as a sensor platform. *IEEE Conference on Intelligent Transportation Systems, Proceedings, ITSC*, pages 1609--1615. ISSN 2153-0009.
- Kaplan, S., Guvensan, M. A., Yavuz, A. G., and Karalurt, Y. (2015). Driver Behavior Analysis for Safe Driving: A Survey. *IEEE Transactions on Intelligent Transportation Systems*, 16(6):3017--3032. ISSN 1524-9050.

- Karagiannis, G., Altintas, O., Ekici, E., Heijenk, G., Jarupan, B., Lin, K., and Weil, T. (2011). Vehicular Networking: A Survey and Tutorial on Requirements, Architectures, Challenges, Standards and Solutions. *IEEE Communications Surveys Tutorials*, 13(4):584–616.
- Khaleghi, B., Khamis, A., Karray, F., and Razavi, S. (2013a). Multisensor data fusion: A review of the state-of-the-art. *Information Fusion*.
- Khaleghi, B., Khamis, A., Karray, F. O., and Razavi, S. N. (2013b). Multisensor data fusion: A review of the state-of-the-art. *Information Fusion*, 14(1):28–44. ISSN 15662535.
- Kim, J., Cha, M., and Sandholm, T. (2014). SocRoutes: safe routes based on tweet sentiments. In *23rd International Conference on WWW*.
- Kotthoff, L., Thornton, C., and Hutter, F. (2017). User guide for auto-weka version 2.6. *Dept. Comput. Sci., Univ. British Columbia, BETA lab, Vancouver, BC, Canada, Tech. Rep.*, 2.
- Kumtepe, O., Akar, G. B., and Yuncu, E. (2016). Driver aggressiveness detection via multisensory data fusion. *EURASIP Journal on Image and Video Processing*, 2016(1):5. ISSN 1687-5281.
- Kuo, S. M. and Zhou, M. (2009). Virtual sensing techniques and their applications. *2009 International Conference on Networking, Sensing and Control*, pages 31–36.
- Kurkcu, A., Zuo, F., Gao, J., Morgul, E. F., and Ozbay, K. (2017). Crowdsourcing incident information for disaster response using twitter. *Transportation Research Board 96th Annual Meeting*.
- Laberteaux, H. H. and P., L. (2008). A Tutorial Survey on Vehicular Ad Hoc Networks. *IEEE Communications Magazine*, 46(June):164–171.
- Lau, R. Y. (2017). Toward a social sensor based framework for intelligent transportation. In *2017 IEEE 18th International Symposium on A World of Wireless, Mobile and Multimedia Networks (WoWMoM)*, pages 1–6. IEEE.
- Lee, U. and Gerla, M. (2010). A survey of urban vehicular sensing platforms. *Computer Networks*, 54(4):527–544. ISSN 13891286.
- Li, C. and Sun, A. (2014). Fine-grained location extraction from tweets with temporal awareness. In *37th ACM SIGIR*.
- Li, H., Yu, D., and Braun, J. E. (2011). A review of virtual sensing technology and application in building systems. *HVAC&R Research*, 17(November 2014):37–41. ISSN 10789669.
- Liu, X., Zhang, S., Wei, F., and Zhou, M. (2011). Recognizing named entities in tweets. In *49th AMACL*.

- Lu, N., Cheng, N., Zhang, N., Shen, X., and Mark, J. W. (2014a). Connected vehicles: Solutions and challenges. *IEEE Internet of Things Journal*, 1(4):289--299. ISSN 23274662.
- Lu, N., Cheng, N., Zhang, N., Shen, X., and Mark, J. W. (2014b). Connected vehicles: Solutions and challenges. *IEEE Internet of Things Journal*, 1(4):289--299. ISSN 23274662.
- Ly, M. V., Martin, S., and Trivedi, M. M. (2013). Driver classification and driving style recognition using inertial sensors. In *2013 IEEE Intelligent Vehicles Symposium (IV)*, pages 1040--1045. ISSN 1931-0587.
- Ma, C., Dai, X., Zhu, J., Liu, N., Sun, H., and Liu, M. (2017). Drivingsense: Dangerous driving behavior identification based on smartphone autocalibration. *Mobile Information Systems*, 2017.
- Machina Research and Telefonica (2013). Connected Car Industry Report - Part 2. Technical report, Telefonica.
- Magister54 (2015). OpenGauge. <https://github.com/Magister54/opengauge>. Accessed: July 7, 2017.
- Map, O. W. (2017). OpenWeatherMap. <https://openweathermap.org/api>. Accessed: May 27, 2017.
- Marchal, F., Hackney, J., and Axhausen, K. W. (2004). Efficient map-matching of large gps data sets - tests on a speed monitoring experiment in zurich, volume 244 of *arbeitsbericht verkehrs und raumplanung*. Technical report.
- Martinez, M., Echanobe, J., and del Campo, I. (2016). Driver identification and impostor detection based on driving behavior signals. In *2016 IEEE 19th International Conference on Intelligent Transportation Systems (ITSC)*, pages 372--378, Rio de Janeiro, Brazil. IEEE.
- Mednis, A., Elsts, A., and Selavo, L. (2012). Embedded solution for road condition monitoring using vehicular sensor networks. *2012 6th International Conference on Application of Information and Communication Technologies, AICT 2012 - Proceedings*, pages 0--4.
- Merritt, K. (2017). Socrata. <https://socrata.com>. Accessed: August 20, 2017.
- Meseguer, J. E., Calafate, C. T., Cano, J. C., and Manzoni, P. (2013). DrivingStyles: A smartphone application to assess driver behavior. In *2013 IEEE Symposium on Computers and Communications (ISCC)*, pages 000535--000540, Split, Croatia. IEEE. ISSN 15301346.
- Mike, P. (2013). Automotive sensors and electronics: trends and developments.
- MirrorLink (2017). MirrorLink - Car connectivity Consortium. Accessed: May 10, 2017.

- Nakamura, E. F., Loureiro, A. a. F., and Frery, A. C. (2007). Information fusion for wireless sensor networks. *ACM Computing Surveys*, 39(3):9–es. ISSN 03600300.
- Ng, A., Ngiam, J., Foo, C. Y., Mai, Y., and Suen, C. (2011). Backpropagation Algorithm. http://ufldl.stanford.edu/wiki/index.php/Backpropagation_Algorithm. (Accessed on 10/11/2018).
- Nguyen, H., Liu, W., Rivera, P., and Chen, F. (2016). Trafficwatch: Real-time traffic incident detection and monitoring using social media. In *Pacific-Asia Conference on Knowledge Discovery and Data Mining*, pages 540–551. Springer.
- Ning, Z., Xia, F., Ullah, N., Kong, X., and Hu, X. (2017). Vehicular Social Networks: Enabling Smart Mobility. *IEEE Communications Magazine*, 55(5):16–55. ISSN 0163-6804.
- of Transportation, U. D. (2017a). DSSS - UTMS Society of Japan. <http://www.utms.or.jp/english/system/dsss.html>. Accessed: May 17, 2017.
- of Transportation, U. D. (2017b). U.S. Department of Transportation. <https://www.transportation.gov/>. Accessed: May 15, 2017.
- Olson, R. S., Bartley, N., Urbanowicz, R. J., and Moore, J. H. (2016). Evaluation of a tree-based pipeline optimization tool for automating data science. In *Proceedings of the Genetic and Evolutionary Computation Conference 2016, GECCO '16*, pages 485–492, New York, NY, USA. ACM.
- OpenXC (2012). OpenXC. <http://openxcplatform.com/>. Accessed: July 7, 2017.
- Paefgen, J., Fleisch, E., Staake, T., Ackermann, L., Best, J., and Egli, L. (2013). Telematics strategy for automobile insurers : Whitepaper. Working paper, ITEM - Institute of Technology Management with Transfer Center for Technology Management (TECTEM).
- Paefgen, J., Kehr, F., Zhai, Y., and Michahelles, F. (2012). Driving behavior analysis with smartphones: Insights from a controlled field study. In *Proceedings of the 11th International Conference on Mobile and Ubiquitous Multimedia, MUM '12*, pages 36:1--36:8, New York, NY, USA. ACM.
- Paefgen, J. F. R. (2013). On the Determination of Accident Risk Exposure from Vehicular Sensor Data – Methodological Advancements and Business Implications for Automobile Insurance Providers. *University of St. Gallen, Business Dissertations*, (4170):1–147.
- Pan, B., Zheng, Y., Wilkie, D., and Shahabi, C. (2013). Crowd sensing of traffic anomalies based on human mobility and social media. In *Proceedings of the 21st ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems - SIGSPATIAL'13*, pages 334–343, New York, New York, USA. ACM Press.

- Pañeda, X. G., Garcia, R., Diaz, G., Tuero, A. G., Pozueco, L., Mitre, M., Melendi, D., and Pañeda, A. G. (2016). Formal characterization of an efficient driving evaluation process for companies of the transport sector. *Transportation Research Part A*, 94:431--445.
- Parker, D., Reason, J. T., Manstead, A. S., and Stradling, S. G. (1995). Driving errors, driving violations and accident involvement. *Ergonomics*, 38(5):1036--1048.
- Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., and Duchesnay, E. (2011). Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825--2830.
- Pereira, F. C., Rodrigues, F., and Ben-Akiva, M. (2013). Text analysis in incident duration prediction. *Transportation Research Part C: Emerging Technologies*, 37:177--192.
- Pinto, C., Pita, R., Barbosa, G., Bertoldo, J., Sena, S., Reis, S., Fiaccone, R., Amorim, L., Ichihara, M. Y., Barreto, M., Barreto, M., and Denaxas, S. (2017). Probabilistic integration of large Brazilian socioeconomic and clinical databases. *30th IEEE International Symposium CBMS*.
- Poó, F. M. and Ledesma, R. D. (2013). A Study on the Relationship Between Personality and Driving Styles. *Traffic Injury Prevention*, 14(4):346--352. ISSN 1538-9588.
- Qu, F., Wang, F. Y., and Yang, L. (2010). Intelligent transportation spaces: Vehicles, traffic, communications, and beyond. *IEEE Communications Magazine*, 48(11):136--142. ISSN 0163-6804.
- Reininger, M., Miller, S., Zhuang, Y., and Cappos, J. (2015). A first look at vehicle data collection via smartphone sensors. In *2015 IEEE Sensors Applications Symposium (SAS)*, pages 1--6, Zadar, Croatia. IEEE.
- Reis, S., Pesch, D., Wenning, B.-L., and Kuhn, M. (2017). *Intra-Vehicle Wireless Sensor Network Communication Quality Assessment via Packet Delivery Ratio Measurements*, pages 88--101. Springer International Publishing, Cham.
- Rettore, P. H., André, B. P. S., Campolina, Villas, L. A., and A.F. Loureiro, A. (2016a). Towards intra-vehicular sensor data fusion. In *Advanced perception, Machine learning and Data sets (AMD'16) as part of the 2016 IEEE 19th International Conference on Intelligent Transportation Systems (ITSC 2016)*, , Rio de Janeiro.
- Rettore, P. H., Campolina, A., de Souza, A. L., Maia, G., Villas, L. A., and A.F. Loureiro, A. (2018a). Driver authentication in VANETs based on Intra-Vehicular sensor data. In *2018 IEEE Symposium on Computers and Communications (ISCC) (ISCC 2018)*, Natal, Brazil.

- Rettore, P. H., Campolina, A. B., Villas, L. A., and Loureiro, A. A. (2016b). Identifying relationships in vehicular sensor data: A case study and characterization. In *Proceedings of the 6th ACM Symposium on Development and Analysis of Intelligent Vehicular Networks and Applications*, DIVANet '16, pages 33--40, New York, NY, USA. ACM.
- Rettore, P. H., Santos, B. P., Campolina, A. B., Villas, L. A., and Loureiro, A. A. (2016c). Towards Intra-Vehicular Sensor Data Fusion. *19th International Conference on ITS*.
- Rettore, P. H. L., Araujo, I., de Menezes, J. G. M., Villas, L., and Loureiro, A. A. F. (2019). Serviço de detecção e enriquecimento de eventos rodoviários baseado em fusão de dados heterogêneos para vanets. In *SBRC 2019*, Gramado, Brazil.
- Rettore, P. H. L., Campolina, A., Luis, A., de Menezes, J. G. M., Villas, L., and Loureiro, A. A. F. (2018b). Benefícios da autenticação de motoristas em redes veiculares. In *(SBRC 2018)*, Campos do Jordão, Brazil.
- Rettore, P. H. L., Campolina, A. B., Villas, L. A., and Loureiro, A. A. F. (2017). A method of eco-driving based on intra-vehicular sensor data. In *2017 IEEE Symposium on Computers and Communications (ISCC)*, pages 1122--1127, Heraklion, Greece. IEEE. ISSN .
- Ribeiro Jr, S. S., Davis Jr, C. A., Oliveira, D. R. R., and Meira Jr, W. (2012). Traffic observatory: a system to detect and locate traffic events and conditions using Twitter. In *5th ACM SIGSPATIAL*.
- Riener, A. and Reder, J. (2014). Collective Data Sharing to Improve on Driving Efficiency and Safety. *Proceedings of the 6th International Conference on Automotive User Interfaces and Interactive Vehicular Applications - AutomotiveUI '14*, pages 1--6.
- Rodelgo-Lacruz, M., Gil-Castineira, F. J., Gonzalez-Castano, F. J., Pousada-Carballo, J. M., Contreras, J., Gomez, a., Bueno-Delgado, M. V., Egea-Lopez, E., Vales-Alonso, J., and Garcia-Haro, J. (2007). Base technologies for vehicular networking applications: review and case studies. *2007 IEEE International Symposium on Industrial Electronics*, pages 2567--2572.
- Rutty, M., Matthews, L., Andrey, J., and Matto, T. D. (2013). Eco-driver training within the City of Calgary's municipal fleet: Monitoring the impact. *Transportation Research Part D: Transport and Environment*. ISSN 13619209.
- Rutty, M., Matthews, L., Scott, D., and Del Matto, T. (2014a). Using vehicle monitoring technology and eco-driver training to reduce fuel use and emissions in tourism: a ski resort case study. *Journal of Sustainable Tourism*, 22(5):787--800. ISSN 0966-9582.
- Rutty, M., Matthews, L., Scott, D., and Matto, T. D. (2014b). Using vehicle monitoring technology and eco-driver training to reduce fuel use and emissions in tourism: a ski resort case study. *Journal of Sustainable Tourism*, 22(5):787--800.

- Sagberg, F., Selpi, Bianchi Piccinini, G. F., and Engström, J. (2015). A Review of Research on Driving Styles and Road Safety. *Human Factors: The Journal of the Human Factors and Ergonomics Society*, 57(7). ISSN 0018-7208.
- Saiprasert, C., Pholprasit, T., and Thajchayapong, S. (2017). Detection of Driving Events using Sensory Data on Smartphone. *International Journal of Intelligent Transportation Systems Research*, 15(1):17--28. ISSN 18688659.
- Salemi, M. (2015). *Authenticating drivers based on driving behavior*. Rutgers The State University of New Jersey-New Brunswick.
- Santos, B. P., Rettore, P. H., Ramos, H. S., Vieira, L. F. M., and A.F. Loureiro, A. (2018). Enriching traffic information with a spatiotemporal model based on social media. In *2018 IEEE Symposium on Computers and Communications (ISCC) (ISCC 2018)*, Natal, Brazil.
- Santos, B. P., Rettore, P. H., Vieira, L. F. M., and A.F. Loureiro, A. (2019). Dribble: a learn-based timer scheme selector for mobility management in IoT. In *2019 IEEE Wireless Communications and Networking Conference (WCNC) (IEEE WCNC 2019)*, Marrakech, Morocco.
- Santos, B. P., Rettore, P. H. L., Ramos, H., Vieira, L. F., and Loureiro, A. A. F. (2017). T-maps: Modelo de descrição do cenário de trânsito baseado no twitter. In *(SBRC 2017)*.
- Satzoda, R. K. and Trivedi, M. M. (2015). Drive Analysis Using Vehicle Dynamics and Vision-Based Lane Semantics. *IEEE Transactions on Intelligent Transportation Systems*, 16(1):9--18. ISSN 1524-9050.
- Schrank, D., Eisele, B., and Lomax, T. (2012). Tti's 2012 urban mobility report. *Texas A&M Transportation Institute. The Texas A&M University System*, page 4.
- Schrank, D., Eisele, B., Lomax, T., and Bak, J. (2015). 2015 urban mobility scorecard. *Texas A&M Transportation Institute. The Texas A&M University System*, page 4.
- Schroeder, P., Meyers, M., and Kostyniuk, L. (2013). National survey on distracted driving attitudes and behaviors--2012. Technical report, National Highway Traffic Safety Administration.
- Septiana, I., Setiowati, Y., and Fariza, A. (2016). Road condition monitoring application based on social media with text mining system: Case Study: East Java. In *2016 International Electronics Symposium (IES)*, pages 148--153. IEEE.
- Shekhar, H., Setty, S., and Mudenagudi, U. (2016). Vehicular traffic analysis from social media data. In *2016 International Conference on Advances in Computing, Communications and Informatics (ICACCI)*, pages 1628--1634. IEEE.

- Silva, H., Lourenço, A., and Fred, A. (2012). In-vehicle driver recognition based on hand ecg signals. In *Proceedings of the 2012 ACM international conference on Intelligent User Interfaces*.
- Silva, M. J., Cavalcante, T. S., Rosso, O. A., Rodrigues, J. J., Oliveira, R. A., and Aquino, A. L. (2019). Study about vehicles velocities using time causal information theory quantifiers. *Ad Hoc Networks*.
- Sinha, M., Varma, P., and Mukherjee, T. (2017). Web and Social Media Analytics towards Enhancing Urban Transportations. In *Proceedings of the 2nd International Workshop on Network Data Analytics - NDA'17*, pages 1--7, New York, New York, USA. ACM Press.
- SmartDeviceLink Consortium, I. (2017). Smart Device Link - SDL. <https://www.smartdevicelink.com>. Accessed: May 10, 2017.
- StateFarm (2017). Drive Safe and Save. <https://www.statefarm.com/insurance/auto>. Accessed: July 19, 2017.
- Stephant, J., Charara, A., and Meizel, D. (2004). Virtual sensor: Application to vehicle sideslip angle and transversal forces. *IEEE Transactions on Industrial Electronics*, 51(2):278--289. ISSN 02780046.
- Taubman-Ben-Ari, O., Mikulincer, M., and Gillath, O. (2004). The multidimensional driving style inventory—scale construct and validation. *Accident Analysis & Prevention*, 36(3):323--332.
- Technology, S. (1999). Scope Technology. <http://www.scopetechnology.com/>. Accessed: July 10, 2017.
- Thyssenkrupp (2017). Urban-hub: People shaping cities. <http://www.urban-hub.com/smart-mobility/>.
- Tonguz, O. K., m. Tsai, H., Talty, T., Macdonald, A., and Saraydar, C. (2006). Rfid technology for intra-car communications: A new paradigm. In *IEEE Vehicular Technology Conference*, pages 1--6, Montreal, Que., Canada. IEEE. ISSN 1090-3038.
- Tonguz, O. K., Tsai, H.-m., Saraydar, C., Talty, T., and Macdonald, A. (2007). Intra-Car Wireless Sensor Networks Using RFID: Opportunities and Challenges. *2007 Mobile Networking for Vehicular Environments*, pages 43--48.
- Toyota (2015). Toyota Touch 2. <https://www.toyota-europe.com/world-of-toyota/articles-news-events/2016/toyota-touch-2>. Accessed: May 10, 2017.
- Truxillo, D. M., Macarthur, J., Hammer, L. B., and Bauer, T. N. (2016). Evaluation of a Supervisor Training Program for ODOT's EcoDrive Program. *Transportation Research and Education Center (TREC)*.

- Tsai, H.-M., Tonguz, O. K., Saraydar, C., Talty, T., Ames, M., and Macdonald, A. (2007). Zigbee-based intra-car wireless sensor networks: a case study. *IEEE Wireless Communications*, 14(6):67–77. ISSN 1536-1284.
- Tuohy, S., Glavin, M., Hughes, C., Jones, E., Trivedi, M., and Kilmartin, L. (2015). Intra-vehicle networks: A review. *IEEE Transactions on Intelligent Transportation Systems*, 16(2):534–545. ISSN 1524-9050.
- Twitter (2006). Twitter. <https://twitter.com>. Accessed: September 9, 2017.
- Uppoor, S. and Fiore, M. (2011). Large-scale urban vehicular mobility for networking research. In *IEEE Vehicular Networking Conference (VNC '11)*, pages 62–69.
- Vaiana, R., Iuele, T., Astarita, V., Caruso, M. V., Tassitani, A., Zaffino, C., and Giofrè, V. P. (2014). Driving Behavior and Traffic Safety: An Acceleration-Based Safety Evaluation Procedure for Smartphones. *Modern Applied Science*, 8(1):88. ISSN 1913-1852.
- van Huysduynen, H. H., Terken, J., Martens, J.-B., and Eggen, B. (2015). Measuring driving styles: A validation of the multidimensional driving style inventory. In *Proceedings of the 7th International Conference on Automotive User Interfaces and Interactive Vehicular Applications*, AutomotiveUI '15, pages 257–264, New York, NY, USA. ACM.
- Wang, S., Zhang, X., Cao, J., He, L., Stenneth, L., Yu, P. S., Li, Z., and Huang, Z. (2017). Computing Urban Traffic Congestions by Incorporating Sparse GPS Probe Data and Social Media Data. *ACM Transactions on Information Systems*, 35(4):1–30. ISSN 10468188.
- Wang, W., Xi, J., and Chen, H. (2014). Modeling and recognizing driver behavior based on driving data: A survey. *Mathematical Problems in Engineering*, 2014:20. ISSN 15635147.
- Waze (2006). Waze. <https://www.waze.com/>. Accessed: September 9, 2017.
- Weather (2017). Weather. <https://weather.com>. Accessed: May 29, 2017.
- Wenzel, T., Burnham, K., Blundell, M., and Williams, R. (2007). Kalman filter as a virtual sensor: applied to automotive stability systems. *Transactions of the Institute of Measurement and Control*, 29(2007):95–115. ISSN 0142-3312.
- Xu, S., Li, S., and Wen, R. (2018). Sensing and detecting traffic events using geosocial media data: A review. *Computers, Environment and Urban Systems*, (June). ISSN 01989715.
- Yager, R. R. (1982). Generalized probabilities of fuzzy events from fuzzy belief structures. *Information sciences*.
- Yager, R. R. (1987). On the dempster-shafer framework and new combination rules. *Information Sciences*, 41(2):93–137. ISSN 00200255.

- Yazici, M. A., Mudigonda, S., and Kamga, C. (2017). Incident detection through twitter: Organization versus personal accounts. *Transportation Research Record: Journal of the Transportation Research Board*, (2643):121--128.
- Yin, J. and Du, Z. (2016). Exploring Multi-Scale Spatiotemporal Twitter User Mobility Patterns with a Visual-Analytics Approach. *ISPRS International Journal of Geo-Information*, 5.
- Yu, J., Zhu, H., Han, H., Chen, Y. J., Yang, J., Zhu, Y., Chen, Z., Xue, G., and Li, M. (2016). Senspeed: Sensing driving conditions to estimate vehicle speed in urban environments. *IEEE Transactions on Mobile Computing*, 15(1):202--216.
- Yuan, Q., Liu, Z., Li, J., Yang, S., and Yang, F. (2016). An adaptive and compressive data gathering scheme in vehicular sensor networks. *Proceedings of the International Conference on Parallel and Distributed Systems - ICPADS, 2016-Janua*:207--215. ISSN 15219097.
- Yuan, W. and Tang, Y. (2011). The driver authentication device based on the characteristics of palmprint and palm vein. In *International Conference on Hand-Based Biometrics*, pages 1--5.
- Zadeh, L. A. (1984). Review of A Mathematical Theory of Evidence. *AI Magazine*.
- Zan, B., Sun, T., Gruteser, M., and Zhang, Y. (2010). ROME: Road monitoring and alert system through geocache. *Proceedings of the 2010 IEEE International Symposium on Parallel and Distributed Processing, Workshops and Phd Forum, IPDPSW 2010*, pages 1--8.
- Zhang, C., Patel, M., Buthpitiya, S., Lyons, K., Harrison, B., and Abowd, G. D. (2016). Driver classification based on driving behaviors. In *Proceedings of the 21st International Conference on Intelligent User Interfaces, IUI '16*, pages 80--84, New York, NY, USA. ACM.
- Zhang, Z., He, Q., Gao, J., and Ni, M. (2018). A deep learning approach for detecting traffic accidents from social media data. *Transportation research part C: emerging technologies*, 86:580--596.
- Zhao, J., Li, W., Wang, J., and Ban, X. (2016). Dynamic Traffic Signal Timing Optimization Strategy Incorporating Various Vehicle Fuel Consumption Characteristics. *IEEE Transactions on Vehicular Technology*, 65(6):3874--3887. ISSN 0018-9545.
- Zuchao Wang, Min Lu, Xiaoru Yuan, Junping Zhang, and Van De Wetering, H. (2013). Visual Traffic Jam Analysis Based on Trajectory Data. *IEEE Transactions on Visualization and Computer Graphics*, 19(12):2159--2168. ISSN 1077-2626.
- Zuckerman, M. (2002). Zuckerman-kuhlman personality questionnaire (zkpq): an alternative five-factorial model. *Big five assessment*, pages 377--396.