# Fusion on Vehicular Data Space: An Approach to Smart Mobility

**Paulo H. L. Rettore**[1,3], **Guilherme Maia**[1],
**Leandro A. Villas**[2], **Antonio A. F. Loureiro**[1]

[1]Depto. de Ciência da Computação – Universidade Federal de Minas Gerais (UFMG)
Av. Antônio Carlos, 6627 – Belo Horizonte, MG – Brasil

[2]Instituto de Computação, Universidade de Campinas (UNICAMP)
Av. Albert Einstein, 1251 – Campinas, SP – Brasil

[3]Communication Systems (KOM), Fraunhofer FKIE
Zanderstraße, 5 – Bonn, North Rhine-Westphalia – Germany

{rettore,jgmm,loureiro}@dcc.ufmg.br,

leandro@ic.unicamp.br

***Abstract.*** *Urban mobility deals with the movement of people and cargo in urban environments and has 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 Intelligent Transportation System (ITS). However, the development of new applications and services for the ITS environment to improve the mobility depends on the availability of vast amounts of data. This thesis aims to explore data from a vast number of sources from the ITS context to provide directions to improve mobility in urban scenarios. 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 for SM, aiming to fuse those data considering their aspects, highlighting the most relevant methods and techniques to achieve the application goals.*

## 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 became important challenges to advance transportation systems, paving the way for Smart Cities. The development of transportation systems for smart solutions 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

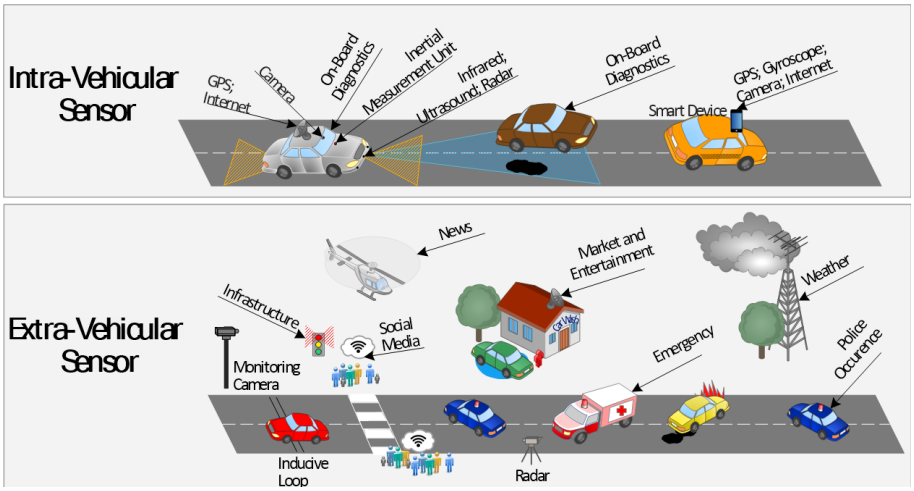



Figure 1: Data provided in urban area [**Rettore, P. H. L.** et al. 2019b].

vast range of data sources, as depicted in Figure 1, to deal with those aspects. In this sense, this thesis aims to provide concepts, applications, and services to improve the current state-of-the-art transportation systems, through the use of methods and techniques that apply heterogeneous data fusion.

### 1.1. Problem Statement

Based on these various data related to the transportation systems, a relevant research challenge emerges aiming to answer the following *"How those data can be used to improve people's life quality in large cities, especially regarding mobility and traffic?"*. If we go further and analyze these data, we face the problem of poor data quality and coverage. In this sense, noticing a lack of both the availability of data and on the data quality. Then, we aim to answer the question *"How to handle 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 data fusion we can improve the data quality, providing methods and applications to achieve Smart Mobility (SM)"*. The integration of multiple data sources becomes an essential process to provide consistent, accurate, and useful information to applications in Intelligent Transportation System (ITS). Such a process constitutes a challenging task especially when considering heterogeneous data and their spatiotemporal aspects.

### 1.2. Challenges

Based on our hypothesis, we faced some of the following challenges as part of the data cycle (Figure 3b): I) the high costs to embed sensors in a vehicle or on the roads reduces the spatiotemporal data coverage; II) different devices can create the same data with different levels of quality; III) the scientific/commercial value of data reduce its availability to the community; IV) security and privacy reduces the sample available; V) the need for a computational infrastructure to store and process large amounts of data; VI) social media platforms impose restrictions on their collecting process; VII) fix the dataset, by identifying outliers, conflict, incompleteness, ambiguity, correlation, and disparateness;

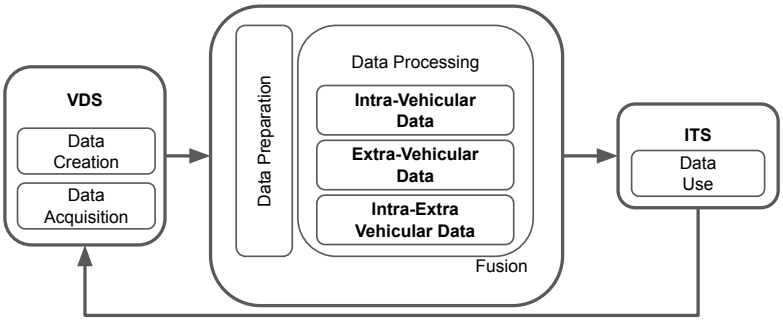



Figure 2: Design of fusion on VDS.

VIII) performs feature reduction to keep the most relevant and descriptive features on the dataset; IX) find the best algorithms/methodologies to the proposed solutions; X) extract useful information from Intra-Vehicle Data (IVD) to correlate them with Extra-Vehicle Data (EVD). This will become one of the top trends for future ITSs; XI) deal with the data heterogeneity issues that come with the asynchronous sensor operation, sensor errors, and sensor noise.

### 1.3. 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 2 depicts the refinement of this goal by showing the design of our fusion process in the ITS. We create the concept of a Vehicular Data Space (VDS) as the input data to this design (refer to Section 2). 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 combinations. The IVD only uses the data provided by vehicles. The EVD focuses on fusing data surrounding vehicles, while the Intra and Extra-Vehicle Data (IEVD) aims to combine both data types. The output of these three types of data fusion approaches are applications and services that can improve current mobility, or can be used as input data for a new data fusion cycle, as depicted in Figure 3b.

The fusion process depends on the data availability and the data preparation, which aim to deal with data issues (refer to Section 3). 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.

### 1.4. 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 temporal and spatial data fusion techniques using the same or other data sources available for VDS.

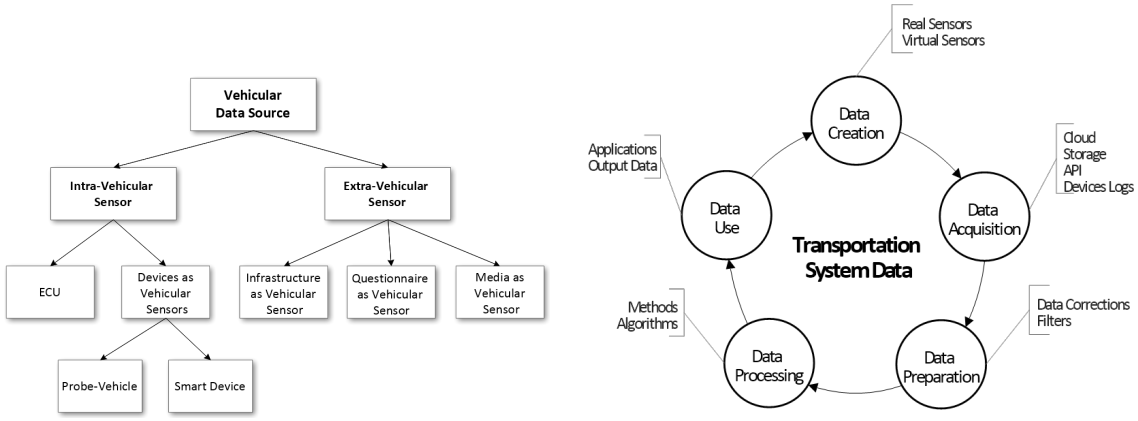
In that direction, we use basically mathematical methods, geostatistics and machine learning techniques in the following contributions: (i) a methodology to develop applications and services for SM based on the ITS data cycle stages; (ii) an IVD fusion technique and the corresponding methodology to detect a legitimate/illegitimate driver; 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; (iii) based on the vehicle's surrounding data, we designed the EVD fusion technique that combines the user's viewpoint and road data. We proposed the Road Data Enrichment (RoDE) framework with two main services: route service and event service; (iv) an IEVD fusion technique called Traffic Data Enrichment Sensor (TraDES) that fills the road spatiotemporal data gaps, using vehicular trace and road data, which allows to improve the data quality; Besides, we conducted a vast literature review where we discuss the concept of VDS and analyze the state-of-the-art applications and services developed for ITS.

## 2. Background

Given the importance of data to ITS, we reviewed and analyzed recent studies describing services and applications for ITSs, but focused on the data used by them (refer to the list of publications, #1). 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), as shown in Figure 3a, discussing the different data sources currently used in ITSs. Furthermore, we discussed for each Vehicular Data Source (VDSource) the relationship between development *Costs* and its respective *Granularity* and *Scalability*. We also explored and 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 the current state-of-the-art and interesting opportunities that guide this thesis and possible further investigations to explore the EVD, IEVD and the respective data fusion.

Another contribution to the state-of-the-art is the discussion of challenges and open issues related to the main topics we observed based on the data cycle of the VDS (*Data Creation*, *Data Acquisition*, *Data Preparation*, *Data Processing*, and *Data Use*) as depicted in Figure 3b. Considering the Vehicular Data Space (VDS), the main contributions of this work are: (i) the need for 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 data standardization; and (vi) heterogeneous data fusion is a fundamental challenge to leverage the ITS field.

1. **Rettore, P. H. L.**, Maia, G., Villas, L. A., and Loureiro, A. A. F. (2019b). Vehicular data space: The data point of view. *IEEE Communications Surveys Tutorials*, 21(3):2392–2418 [Qualis A1, Impact factor: 22.973]




(a) Taxonomy of VDSource on the VDS.

(b) The data cycle on the VDS.

Figure 3: Most used data source in VDS and the data cycle [**Rettore, P. H. L.** et al. 2019b].

## 3. Heterogeneous Data Fusion

Before discussing how to apply data fusion techniques to increase the VDS data quality, we investigated and identified several issues in the data, which must be treated before the data fusion process (refer to the list of publications, #2). We also argued that the ITS can be boosted by taking into account heterogeneous data collected from several sources. In a real environment, in general, data comes with some issues (i.e., imperfection, correlation, outlier, conflicts, vagueness, granularity, inconsistencies, incompleteness, ambiguity, uncertainty, disorder, disparateness among others) making difficult the process of fusing heterogeneous data. In this sense, our exploratory analysis of real vehicular data showed several issues in the data implying that they must be treated before the fusion process. Besides, the understanding of correlations of vehicular sensors allows to provide solutions to optimize the vehicle use; reduce fuel consumption, emissions and vehicle maintenance; route description and others, which directly influence the efforts to provide a Smart Mobility (SM) solution for a city.

2. **Rettore, P. H. L.**, Santos, B. P., Campolina, A. B., Villas, L. A., and Loureiro, A. A. F. (2016b). Towards intra-vehicular sensor data fusion. In *2016 IEEE 19th International Conference on Intelligent Transportation Systems (ITSC)*, pages 126–131 [Qualis B2, Conference]

### 3.1. Intra-Vehicular Data Fusion

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. At this point, the Intra-Vehicular Sensor (IVS) allows the exploration of heterogeneous data collected from several sensors to the design of services and applications that may boost the SM based on fuel efficiency, emissions and safe driving.

In our first exploratory study (refer to the list of publications, #3), we proposed the use of an On-Board Diagnostic (OBD) Bluetooth adapter and a smartphone to gather data from two cars. Afterwards, we created a real testbed to analyze the correlation among the Intra-Vehicular Sensor (IVS) and identified a special correlation between RPM and speed data that reflects the vehicle's current gear. As a result, based on a simple mathematical model, we found a coefficient that indicates the behavior of each gear along the

time in a trace. This study supported our eco-driving approach (refer to list of publications, #4). Initially, we extended our dataset to cover more than 40 trips, 14 drivers and 30 hours of the experiment. We also proposed a low-cost methodology (applicable for any car with an OBD interface) that gives the driver recommendations for the best gear considering speed and torque, reaching up to 29% average of efficiency in the fuel consumption and 21% average in $CO_2$ emissions reduction. Following these experiments, we started an investigation to authenticate the drivers based on their behavior (refer to the list of publications, #5). We improved this approach creating a virtual sensor to differentiate a legitimate driver from a suspected one as an extra factor to authenticate, with over 98% accuracy (refer to the list of publications, #6). We also demonstrated that the presence of illegitimate vehicles might compromise the quality of essential services provided by Vehicular *Ad-hoc* Networks (VANETs), once they are capable of modifying the data which is being disseminated to the entire network. In addition, the dataset created to support these studies was motivated by the lack of OBD trace available on the internet.

These studies directly supported other studies, providing data and complementing their ideas (refer to the list of publications, #7). Being also part of a tutorial and a book chapter (refer to the list of publications, #8 & #9). In addition, the data created by these studies are available on the Internet for any research purpose[1]. In summary, as an overview of these investigations, we noticed a trend to use machine learning techniques to deal with problems related to the Advanced Driver Assistant Systems (ADAS), security, eco-driving and infotainment. In addition, a topic that needs further investigation is related to IVD privacy. Once the data comes from private vehicles the lack of data privacy reduces its availability, and, as a consequence, more applications are designed to achieve a specific target, reducing its generalization capability and reach.

3. **Rettore, P. H. L.**, Campolina, A. B., Villas, L. A., and Loureiro, A. A. (2016a). 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 [Qualis B2, Conference]
4. **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 [Qualis A2, Conference]
5. **Rettore, P. H. L.**, Campolina, A., Luis, A., de Menezes, J. G. M., Villas, L., and Loureiro, A. A. F. (2018a). Benefícios da autenticação de motoristas em redes veiculares. In *(SBRC 2018)*, Campos do Jordão, Brazil [Qualis B2, Conference]
6. **Rettore, P. H. L.**, Campolina, A. B., Souza, A., Maia, G., Villas, L. A., and Loureiro, A. A. F. (2018b). Driver authentication in vanets based on intra-vehicular sensor data. In *2018 IEEE Symposium on Computers and Communications (ISCC)*, pages 00078–00083 [Qualis A2, Conference]
7. 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 [Qualis A2, Conference]
8. Felipe, C., Maia, G., Celes, C., Pereira, B., **Rettore, Paulo H. L.**, Campolina, A., Guidoni, D., Souza, F. S., Ramos, H., Villas, L., Mini, R., and Loureiro, A. (2017). Sistemas de transporte inteligentes. In *(SBRC 2017 - Minicursos)* [Qualis B2, Conference]
9. Cunha, F., Maia, G., Ramos, H. S., Perreira, B., Celes, C., Campolina, A., **Rettore, Paulo H. L.**, Guidoni, D., Sumika, F., Villas, L., Mini, R., and Loureiro, A. (2018). *Vehicular Networks to Intelligent Transportation Systems*, pages 297–315. Springer Singapore, Singapore [Book Chapter]

## 3.2. Extra-Vehicular Data Fusion

The extra-vehicular data corresponds to the subset of real and virtual sensors data that seek to describe the driver's behavior or the environment around the vehicle by a variety of sources individually or fused. This section describes the Media as Vehicular Sensor (MVS), specifically the use of Location-Based Social Media (LBSM) to enrich the road

[1] http://www.prof.rettore.com.br/vehicular-trace/

data, allowing to explore smart mobility, and, thus, opening new ways to build routes based on people's preferences such as sentiment, event detection and event description.

We propose the RoDE, a framework that fuses data from heterogeneous data sources to enhance ITS services, such as vehicle routing and traffic event detection (refer to the list of publications, #10). We describe RoDE through two services: (i) Event service, and (ii) Route service. This study allowed us to understand the data quality of Location-Based Social Media (LBSM) and how to handle the data issues that emerged from that, proposing improvements to the current state-of-the-art. For the first service (refer to the list of publications, #11), we presented the Twitter Incident (T-Incident), a low-cost learning-based road incident detection and enrichment approach based on heterogeneous data fusion. Our approach used a learning-based model to identify patterns on LBSM data which is then used to describe a class of events, aiming to detect different types of events, achieving scores above 90%. As a result, the enriched event description allows ITS to better understand the LBSM user's viewpoint about traffic events (e.g., jams) and points of interest (e.g., restaurants, theaters, stadiums). In addition, the T-Incident approach supports the definition of the Participatory Social Sensor (PSS), a framework to acquire and analyze LBSM (refer to the list of publications, #12). PSS received honorable mention at the Demo track of the Brazilian Symposium on Computer Networks and Distributed Systems. For the second service (refer to the list of publications, #13 & #14), we present the Twitter MAPS (T-MAPS), a low-cost spatiotemporal model to improve the description of traffic conditions through LBSM data.

10. **Rettore, Paulo H. L.** , Pereira, B., Rigolin F. Lopes, R., Maia, G., Villas, L., and Loureiro, A. (2020). Road data enrichment framework based on heterogeneous data fusion for its. *IEEE Transactions on Intelligent Transportation Systems* [Qualis A1, Impact factor: 5.744]
11. **Rettore, P. H. L.**, Araujo, I., de Menezes, J. G. M., Villas, L., and Loureiro, A. A. F. (2019a). Serviço de detecção e enriquecimento de eventos rodoviários baseado em fusão de dados heterogêneos para vanets. In *Anais do XXXVII Simpósio Brasileiro de Redes de Computadores e Sistemas Distribuídos*, pages 363–376, Porto Alegre, RS, Brasil. SBC [Qualis B2, Conference]
12. Araujo, I., **Rettore, P. H. L.**, and de Menezes, J. G. M. (2019). Participatory social sensor: A framework to social media data acquisition and analysis. In *Anais Estendidos do XXXVII Simpósio Brasileiro de Redes de Computadores e Sistemas Distribuídos*, pages 17–24, Porto Alegre, RS, Brasil. SBC Honorable mention. [Qualis B2, Conference]
13. Santos, B. P., **Rettore, P. H. L.**, Ramos, H. S., Vieira, L. F. M., and A.F. Loureiro, A. (2017). T-maps: Modelo de descrição do cenário de trânsito baseado no twitter. In *(SBRC 2017)* [Qualis B2, Conference]
14. Santos, B. P., **Rettore, P. H. L.**, Ramos, H. S., Vieira, L. F. M., and Loureiro, A. A. F. (2018). Enriching traffic information with a spatiotemporal model based on social media. In *2018 IEEE Symposium on Computers and Communications (ISCC)*, pages 00464–00469 [Qualis A2, Conference]

### 3.3. Intra-Extra-Vehicular Data Fusion

At this point, we noticed a gap in the state-of-the-art when we described the fusion process of the VDS, considering the IVD and EVD aiming to improve the data quality, providing applications to promoting the SM.

Planning and managing transportation systems are crucial tasks to promote the growth of cities. Such a fact is pushing new initiatives from governments and private sectors to improve 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 to achieve the ITS goals. Based on that, we proposed TraDES, a low-cost traffic sensor for ITS (refer to list of publications, #15). TraDES aims at fusing data from vehicular traces with road traffic data to enrich

current spatiotemporal traffic data. In that direction, we proposed a robust methodology to group spatially and temporally these different data sources, producing a vehicular trace with its respective traffic conditions, which is given as input to a learning-based model. This study allowed us to explore vast possibilities once we are able to increase the spatiotemporal traffic data coverage. After the data enrichment, we can understand the driver's decisions, as our ongoing investigation (refer to the list of publications, #16), which aims to identify if the traffic influences the driver's decision to take different routes or if this decision has any correlation with vehicular sensors.

15. **Rettore, P. H. L.**, Rigolin F. Lopes, R., Maia, G., Aparecido Villas, L., and Ferreira Loureiro, A. A. (2019c). Towards a traffic data enrichment sensor based on heterogeneous data fusion for its. In *2019 15th International Conference on Distributed Computing in Sensor Systems (DCOSS)*, pages 570–577 [Qualis A2, Conference]
16. Findings on Driver' Decisions based on Heterogeneous Data Fusion [Under Submission to T-ITS]

## 4. Research Accomplishments

The main accomplishments of this research can be summarized as follows:

- In the list of journals, two are already published and one is under submission.
- Seven papers in SBC, IEEE and ACM national and international conferences.
- Part of this thesis has also contributed to other researches, where three papers in national and international conferences; one Demo which received honorable mention and one tutorial in the Brazilian Symposium on Computer Networks and Distributed Systems; and one book chapter in Elsevier.
- This thesis also supported one Bachelor project and one Master thesis.

## 5. Conclusion and Thesis Impact

This thesis tackled the data quality on the transportation system scenario, which is one of the fundamental problems that needs to be solved to achieve high-quality applications for Smart Cities, leading to the ITS. Herein, we proposed a general approach which shows methods and techniques to enable heterogeneous data fusion on the VDS, aiming to improve the data quality of ITS, achieving a set of SM goals. We highlighted methods and techniques to address those goals such as mathematical methods, threshold filters, statistics, 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.

The results of this thesis have impacted the literature in several ways. Our survey paper is the first literature review that provides a critical and comprehensive review from the data perspective. We categorized the data from the VDS into IVD and EVD perspectives, allowing to identify challenges and open issues to perform data fusion. This review supported the next steps of this thesis that 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 IEVD, which we also advanced the state-of-the-art. However, this is just the beginning, we ended this thesis opening a new and unexplored field of investigation, regarding the advance the IEVD fusion. 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.