UNIVERSITÉ DU LUXEMBOURG

PhD-FSTC-2019-15
The Faculty of Sciences, Technology and Communication

DISSERTATION

Presented on 11/03/2019 in Esch-sur-Alzette

to obtain the degree of

DOCTEUR DE L'UNIVERSITÉ DU LUXEMBOURG

EN SCIENCES DE L'INGÉNIEUR

by

Bogdan TOADER
Born on 27 October 1987 in Ploiești (Romania)

MOBILITY ANALYSIS AND PROFILING FOR SMART MOBILITY SERVICES: A BIG DATA DRIVEN APPROACH

An Integration of Data Science and Travel Behaviour Analytics

Dissertation defense committee

A-Prof. Dr. Vincent Koenig, chairman
Professor, University of Luxembourg, Luxembourg, Luxembourg

Dr. Roderick Mc Call, vice-chairman
Research Scientist, Luxembourg Institute of Science and Technology, Luxembourg

A-Prof. Dr. Francesco Viti, dissertation supervisor
Professor, University of Luxembourg, Luxembourg, Luxembourg

Prof. Dr. Yves Le Traon, member
Professor, University of Luxembourg, Luxembourg, Luxembourg

Prof. Dr. Francisco Camara Pereira, member
Professor, Technical University of Denmark (DTU), Denmark

Abstract

Smart mobility proved to be an important but challenging component of the smart cities paradigm. The increased urbanization and the advent of sharing economy require a complete digitalisation of the way travellers interact with the mobility services. New sharing mobility services and smart transportation models are emerging as partial solutions for solving some traffic problems, improve the resource efficiency and reduce the environmental impact. The high connectivity between travellers and the sharing services generates enormous quantity of data which can reveal valuable knowledge and help understanding complex travel behaviour. Advances in data science, embedded computing, sensing systems, and artificial intelligence technologies make the development of a new generation of intelligent recommendation systems possible. These systems have the potential to act as intelligent transportation advisors that can offer recommendations for an efficient usage of the sharing services and influence the travel behaviour towards a more sustainable mobility. However, their methodological and technological requirements will far exceed the capabilities of today's smart mobility systems.

This dissertation presents a new data-driven approach for mobility analysis and travel behaviour profiling for smart mobility services. The main objective of this thesis is to investigate how the latest technologies from data science can contribute to the development of the next generation of mobility recommendation systems.

Therefore, the main contribution of this thesis is the development of new methodologies and tools for mobility analysis that aim at combining the domain of transportation engineering with the domain of data science. The addressed challenges are derived from specific open issues and problems in the current state of the art from the smart mobility domain. First, an intelligent recommendation system for sharing services needs a general metric which can assess if a group of users are compatible for specific sharing solutions. For this problem, this thesis presents a data driven indicator for collaborative mobility that can give an indication whether it is economically beneficial for a group of users to share the ride, a vehicle or a parking space. Secondly, the complex sharing mobility scenarios involve a high number of users and big data that must be handled by capable modelling frameworks and data analytic platforms. To tackle this problem, a suitable meta model for the transportation domain is created, using the state of the art multi-dimensional graph data models, technologies and analytic frameworks. Thirdly, the sharing mobility paradigm needs an user-centric approach for dynamic extraction of travel habits and mobility patterns. To address this challenge, this dissertation proposes a method capable of dynamically profiling users and the visited locations in order to extract knowledge (mobility patterns and habits) from raw data that can be used for the implementation of shared mobility solutions. Fourthly, the entire process of data collection and extraction of the knowledge should be done with near no interaction from user side. To tackle this issue, this thesis presents practical applications such as classification of visited locations and learning of users' travel habits and mobility patterns using historical and external contextual data.

Acknowledgments

First of all, I want to thank God for giving me the opportunity to be surrounded by wonderful people from which I benefit of their support during this challenging life experience that goes beyond research, experimentations, and scientific writing.

I would like to express my sincere gratitude to my supervisor, Prof. Dr. Francesco Viti. Without his help, this PhD would have never been possible. I appreciate a lot the liberty which was offered to me to do research by exploring different directions and scientific domains. He challenged me to conduct research that contributes to the state of the art and he always encouraged me, and supported me throughout these years. I have learned a lot from his rigorous scientific guidance as a researcher and from his positive, motivating, and creative attitude. He was more than just an academic supervisor for the MobiLab group in which he invested so much work and passion, giving us many good advices beyond the research domain. He was like a godfather for us and the productivity results can be clearly seen on many directions. For example, my two wonderful children were born during the PhD period.

My special thanks goes to the members of my dissertation committee, Prof. Dr. Vincent Koenig, Dr. Roderick Mc Call, Dr. Thomas Hartmann, Prof. Dr. Yves Le Traon and Prof. Dr. Francisco Camara Pereira, for their time to attend all the CET meetings, review my work and for providing interesting and valuable feedback.

I would also like to express my warm thanks to all the co-founders of the DataThings company which provided one of the best open source technology for both scientific and industry domains. They inspired me to think out of the box, helped me to improve my programming skills, and thought me how to work and collaborate in an efficient way. I am very happy as well about the friendship we have built up during these years.

Finally, and more personally, I would like to express my warmest thanks to my family and my friends for their continuous and unconditional support during these last years. I would like to thanks to the friends that made possible this PhD experience, encouraged me and I developed special friendship relationships: Assaad, Natalia, Remus, Martin. My heartfelt thanks goes to all the brothers and sisters in Christ from the Christian Community Church who continuously supported me in payers and especially for Rev. Timothy Heijermans who helped me with the proofreading of this thesis.

Last but not least, I would like to thank God for my wife Mioara and our greatest blessings, David and Victoria, who always encouraged, loved and cared for me. They are my endless resources of motivation and energy, who helped me always to be joyful even in the hardest moments.

Bogdan Toader
Luxembourg, February 2019

Contents

List of abbreviations and acronyms

4IR Fourth Industrial Revolution. 4, 5

AGI Artificial General Intelligence. 42, 46

AI Artificial Intelligence. 4, 5, 13, 35, 36, 41–44, 46, 47, 149

ANI Artificial Narrow Intelligence. 41

ASI Artificial Super Intelligence. 42

CM Collaborative Mobility. 36, 52, 78, 144

CMS Collaborative Mobility Solution. 6–10, 21, 23–25, 36–41, 47, 143, 144, 146

EMF Eclipse Modeling Framework. 37, 38

GIS Geographic Information System. 8, 11, 14, 31, 32, 34, 114, 117, 118, 128, 129, 133, 137, 144–147

GPS Global Positioning System. 9, 11, 14, 21, 25–31, 34, 35, 38, 44, 145, 147

GSM Global System for Mobile communications. 29

ICT Information and Communications Technology. 4, 36, 41

IoT Internet of Things. 4, 7, 9, 20, 39–41, 43, 45, 147

ITS Intelligent Transportation System. 5–10, 13, 20, 23, 25, 32, 33, 35, 37–40, 43, 51, 72–74, 77, 78, 142, 144, 148

KMF Kevoree Modeling Framework. 38, 80

LBRS location based recommendation system. 43, 44, 46

MaaS Mobility as a Service. 6

MAD magnetic, agile, and deep. 40

MDE Model-driven engineering. 37

MDE extensible markup language. 37

ML Machine Learning. 5, 10, 12, 13, 26, 32, 36, 39–43, 45, 47, 78, 143, 147, 149

MoD Mobility-on-demand. 24

MoDu “Mobilité Durable,” Sustainable Mobility. 6

MWG Many World Graph. 82, 85, 145

NN Neural Network. 32

OLAP Online Analytical Processing. 39, 40

OSM OpenStreetMap. 117

P2P Peer to Peer. 39

POI Point of Interest. 44, 46

RQ Research Question. 7, 9, 10, 12, 16

RS Recommendation System. 7, 12, 16, 21, 25, 34, 36, 42–47, 144–148

SVM Suport Vector Machine. 32

TIR Third Industrial Revolution. 4

TRS Travel recommendation system. 44, 46

List of figures

List of tables

Part I

Introduction and State of the Art

1

Introduction

This chapter begins by explaining the context and motivation of this dissertation, followed by the objectives of the thesis and the challenges addressed. Finally an overview is presented of the contributions of the thesis and its structure.

Contents

1.1 Context and motivation

1.1.1 A digitally empowered next generation of transport systems

The transportation industry is on the edge of unprecedented change which will impact the way people work, plan activities, organise their schedule and travel for leisure or work. These changes will contribute to a revolution in the way society evolves in synergy with new communication technologies and new sources of energy. The European Union calls for a digitally connected smart Europe and is laying the groundwork for the Third Industrial Revolution (TIR) [138] (called also the Digital Revolution) which considers transportation as an important element that can enable a new all-encompassing economic paradigm all over Europe. Starting from the 1980s, TIR refers to the advance of digital technology available today, including the personal computer, the internet, and Information and Communications Technology (ICT). It is foreseen that the mobility as we know it today will be completely digitalised and changed.

It is estimated that there will be more than 100 trillion sensors connecting people, infrastructure and environment by 2030, allowing the entire human population to collaborate directly with one another but also with a wide range of devices and autonomous services, decentralising economic life [138]. The transportation industry is expected to spend $85 billion on the Internet of Things (IoT) solutions by 2020. Moreover, regarding which technologies will impact the transportation domain the most, a survey shows that 81% of respondents believe IoT will revolutionise the transport sector [115]. The advent of IoT during the last decade has enabled the complete and continuous connectivity of people, goods, means of transportation and the entire transportation infrastructure. The result is an unparalleled amount of data (three Zettabytes of data in the digital universe collected each year) delivered at revolutionary speed (by 2020 1.7 megabytes of data created every second for every person on earth) and in continuous expansion (data production will be 50 times greater in 2020 than it was in 2010) [144].

All the afore-mentioned advances have brought us to the threshold of the Fourth Industrial Revolution (4IR) [72] or the Industry 4.0 built on the Digital Revolution. The present is marked by emerging technology breakthroughs in a number of fields, including robotics, Artificial Intelligence (AI), IoT and autonomous vehicles, representing new ways in which technology becomes embedded within societies and even the human body [72]. Using the IoT, cyber-physical systems communicate and cooperate with each other and with humans in real-time, with advances in communication and connectivity coupled with new technologies [146]. One of the design principles is that the systems must be able to support decentralized decisions, which enable the ability of cyber physical systems to make decisions on their own and to perform their tasks as autonomously as possible. This provides an enormous wealth of information that can be exploited to improve organisational decision making but at the same time raises big challenges. Some of the challenges in the transportation domain are explored throughout this thesis and practical solutions are provided. One of them is the autonomous capability that the transportation systems must have to manage efficiently an enor-

mous number of users, goods and transportation resources, with very short and stable latency times and with a minimum of information as possible (in order to not stress out the users or to be blocked because of the lack of information). In this sense, the latest information and communication technologies like big data analytics, Machine Learning (ML), AI and cloud computing are the means by which the 4IR will be implemented and explored throughout this thesis.

It is well-known that people desire to be better informed so as to be able to take better decisions independently and in near real-time. For example, people that are better informed about all the mobility alternatives and traffic conditions, can choose shared mobility options and also alternative routes in order to save money, reduce their environmental impact and their travel time. In the same time, a multitude of options can make it difficult for the users to find the best solution(s). This calls for intelligent systems that can choose the best options based on the users' preferences. Consequently, Intelligent Transportation Systems (ITSs) must evolve from classic systems that receive, store, analyse, process and send outcomes, to complex intelligent agents which must continuously analyse their context to autonomously take actions in near real time for a meaningful synchronisation of all the above entities. This requires the sensing systems that we daily interact with (*e.g.,* the sensors from smartphones) to be coordinated with all the users and the transportation resources. In this scenario, the benefits would be enormous: saved time and money, lower environmental impact and even saved human lives.

Therefore, the main motivation of this thesis is to investigate how the latest methodologies and technologies from data science and computer science domains can contribute to the compulsory transition of ITSs in order to meet the standards of the 4IR. The main contribution of this thesis is the development of multi-dimensional data-driven methodologies and tools which can realise the fusion of two distinct domains: on the one hand the transportation academia and industry; and on the other hand the latest data science concepts, methods and technologies (big data management, knowledge discovery, ML, AI). Throughout this dissertation, all the above entities are combined in a single framework, which in turn can drive the required dynamic and near real-time analytic processes.

1.1.2 Towards a clean and efficient transport: the era of sharing economy

At the same time, the transportation industry remains the sector with the fastest-growing concerns in terms of emissions. Passenger and freight transport volumes will continue to expand and due to the combination of population growth, urbanisation, and globalisation, carbon dioxide emissions from transport are expected to increase 60% by 2050. As new technologies and changed behaviour lead to significantly less CO2 being emitted, an improved technology can provide about 70% of the possible CO2 reductions until 2050. The final recommendations are that we need to both accelerate innovation and make radical policy choices to decarbonise transport [87]. Therefore, there are imperatives for the transportation industry to undergo technological changes in order to counteract the increased global movement especially by developing and

fostering/promoting shared mobility solutions, changes in supply chains and even new transport modes.

Working in this direction, the Luxembourgish government published in 2015 a plan which contains the objectives for the future of sustainable mobility called "Mobilité Durable," Sustainable Mobility (MoDu) [80]. For Luxembourg this is an urgent priority, as such small country has the highest number of cars per capita in Europe (0.672, while the average is 0.486) [69] and according to MoDu the mobility is composed of 72.5% trips done by motorized individual vehicles. In order to accelerate this process, MoDu 2.0 was released in 2018 which reveals the objectives for 2025 [81]. One of the four main objectives is the fostering of sustainable mobility through shared transport in a multimodal environment. Solutions like car sharing, carpooling and parking sharing have a potential to reduce traffic, reduce the cost of mobility and use the available space more efficiently. MoDu implementation strategy foresees connecting and involving public and private players (*i.e.,* employers, the government, citizens and local administrations) in a common effort to realise the transition to a shared economy where the main focus is to use the existing transportation resources in a more efficient manner.

Worldwide, sharing mobility is experiencing an exponential growth and the demand is projected to increase five-fold by 2020 [138]. This creates a high scale of disruption caused by the rapid shift towards a sharing economy. For example, Uber was able to operate in more than 250 cities in only five years and was valued at $41.2 billion, which is more than the market capitalization of the largest airlines companies [170]. In the same time, the implementation of massive decentralised platforms for collaborative mobility challenges analytics platforms to discover knowledge from data in motion, extract travel habits and provide reliable and faster sharing mobility services in dynamic contexts. The general direction is tailored to a Mobility as a Service (MaaS) paradigm which should completely change mobility as we know it today [19]. This enables disruptive services and technologies to emerge in an accelerated trend. An example can be seen in some prototypes of Autonomous Travel Suite concepts [5] or Toyota e-Palette concept [14]. This integrates transportation and hospitality through a driver-less, mobile suite offering door-to-door transportation service. Using autonomous driving technology, the travel suite takes passengers to multiple destinations, serving as a personal vehicle and simultaneous mobile hotel room but also mobile home as they can work, sleep, wash, eat and relax while they are travelling. Accordingly, the ITSs must be able to dynamically adapt in short time to new concepts of *home*, *work* and totally new travel behaviour and decision-making mechanisms. This will inevitably change the ITSs from conventional technology-driven systems into a more powerful multifunctional data-driven ITSs [229] which can enable an efficient shared mobility paradigm.

The above-mentioned trends constitute another motivation for this thesis. Throughout this dissertation different Collaborative Mobility Solutions (CMSs) are used as use cases (*e.g.,* car sharing, carpooling, parking sharing services and solutions). In particular, we demonstrate that using a combination of state-of-the art technologies from the data science domain coupled with methodologies from the transportation domain, it is possible to implement with parsimonious resources (in order to be possible the implementation at the level of low-resources mobile devices), the next generation of

autonomous sharing mobility services (*i.e.,* long term and on demand parking sharing, combinations of car sharing and ride sharing). Powered by intelligent Recommendation Systems (RSs) which process the data in motion, the objective is to autonomously match people and transportation resources in an efficient way in order to make sharing system more attractive and in turn reduce traffic congestion, save time and money, and efficiently use the available space by reducing the parking spaces. This requires a suite of complex operations which are examined throughout this thesis like the extraction from raw data, without any user input and in near real time valuable knowledge (*i.e.,* to learn travel habits and location visit patterns, perform location labelling, activity classification) and match people with transportation resources while considering personal preferences, the flexibility and limitations of each individual.

1.2 Objectives of the thesis

As we discussed in Section 1.1, the future of mobility points towards a hyperconnected paradigm through the sensing systems and IoT which will generate and rely on a tremendous amount of data. Consequently, ITSs must evolve and adapt to this new scenario. Data collected must be analysed in a dynamic way and for some services (*e.g.,* dynamic ride sharing) in near real-time. Automatic knowledge discovery processes must extract meaningful information from raw data, thus enabling autonomous reasoning and decision-taking tasks. In order to explore the latest methodologies and technologies from other domains which can contribute to the evolution and optimisation of the CMS, a series of objectives translated in Research Questions (RQs) can be then formulated, presented in the remaining of this section.

Thus, the main RQ can be formulated as follows:

Main RQ.
How data science-driven methods can be leveraged to dynamically analyse, profile and match people in order to synchronize and optimize collaborative mobility services and exploit shared mobility solutions?

Throughout this dissertation, we explore different facets of the Main RQ and the presented contributions in Part II, which provides new ways for using data science-driven methods to perform the dynamic analysis of data generated by the sensing systems in the transportation domain. First, Chapter 3 presents a data-driven indicator for assessing the compatibility of users with different shared mobility solutions, which will be used and complemented in the next chapters by data-driven methods and technologies. Because the indicator was tested only with a small dataset and a low number of users, a modelling framework for big data analytics is developed in Chapter 4. This makes use of specific data science-driven methods and technologies that enable the introduction of a complete framework capable of offering the required features of the ITSs (*e.g.,* fast data storage and access through temporal graphs, what-if analysis, deep search and query capabilities). Then, as the CMSs require a more user-centric approach (in order to address the problems of each user) and less input data from the user, Chapter 5 describes a new method for dynamic profiling in time and space which

can extract users' travel habits, using the latest methods and technologies for indexing, processing and location/activity classification. Finally, Chapter 6 presents new methods for improving the location/activity profiling presented in the previous chapter, using heuristic and Bayesian rules, coupled with Geographic Information System (GIS) external contextual data.

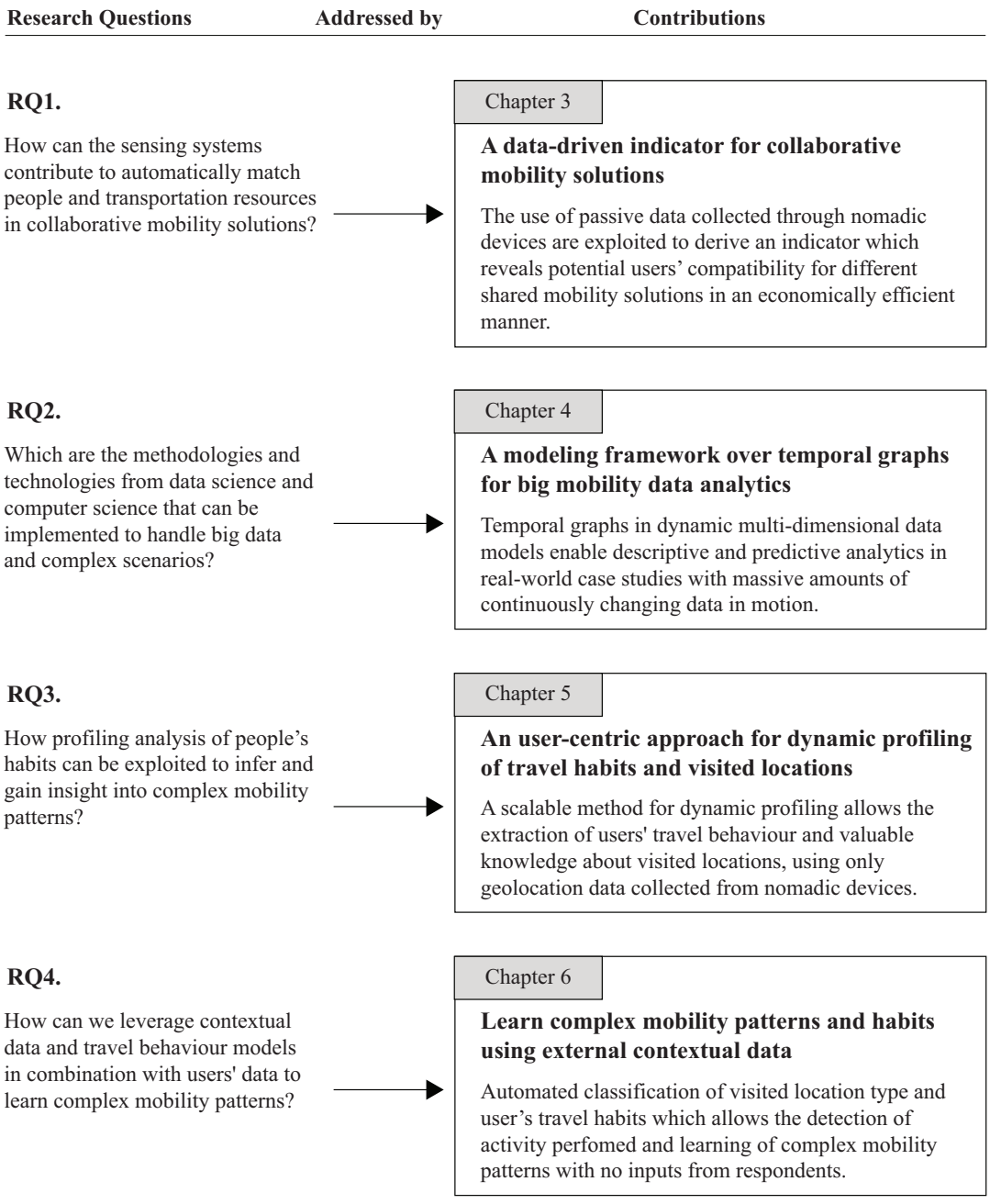


Figure 1.1: Research questions and contributions

As the Main RQ is an extended and generalised objective which applies to all the contribution from Part II of the thesis, a series of secondary RQs can be formulated and individually answered with respect to each contribution. In the rest of this section we discuss the RQ in relation to the chapters where each of them are addressed. A flowchart with the RQs and the corresponding chapters where each of them is addressed

can be seen in Figure 1.1.

RQ1.
How can the sensing systems contribute to match people and transportation resources in CMS?

The IoT is adopted at scale in the transportation domain as drivers and passengers are connected through their sensing systems by nomadic devices, vehicles are permanently connected to the internet and goods that need to be transported have identification tags or even chips that allow near real-time tracking. This environment generates important data regarding the position and status of each entity in the smart mobility environment.

The RQ1 is addressed in Chapter 3. Using data (*i.e.,* Global Positioning System (GPS) position, WiFi and Bluetooth connections) collected through the nomadic devices (smartphones, smartwatches), the objective of Chapter 3 is to propose a methodology that can be used to extract individual travel habits and derive an indicator for revealing potential collaborative mobility options between individuals. This indicator can be used by a recommendation system to find all combinations of collaborative mobility sharing systems (*e.g.,* carpooling, parking sharing, car sharing) which can be used by groups of people. As the proposed indicator is able to take into consideration individual preferences, schedule an entire chain of activities and remain sensitive to dynamic changes in different scenarios, compatible people can be matched for sharing services and transportation resources (*e.g.,* cars, parking spaces).

Massive amounts of data must be collected, analysed, and queried, and knowledge must be extracted from raw data to use it in complex problems and possible combinations of mobility solutions between humans, vehicles and goods. Similarly, in the contributions from this thesis, data collected from nomadic and wearable devices is used to extract insights and knowledge that allows the implementation and optimisation of CMS.

This leads to the next RQ, which can be formulated as follows:

RQ2.
Which are the methodologies from data science (*i.e.,* efficient data indexing, simulations) and technologies from computer science (*i.e.,* database management, fast and dynamic frameworks) that can be implemented to handle data at scale and explore complex scenarios?

In order to meet future requirements, ITSs need to become increasingly intelligent and to learn from historical data as the CMS and use case scenarios which are sometimes repeatable over time. Even if some of them will evolve/change over time, some situations can be predictable at design time [174]. For example, a travel assistant for shared mobility solutions can advise a carpooling user to change the regular schedule in order to find matching users to perform *e.g.,* a ride sharing. In the hypothetical case that the user cannot change the schedule for a specific activity, the recommendation system can offer other options for using CMS (*e.g.,* carsharing, bike sharing, public transport). This is possible if the system automatically learns that this is a good solution, from previous experiences with other users. In order to be able to implement this type

of intelligent and autonomous systems, the solution is to combine domain knowledge from the transportation industry with data science methods and techniques that allow extraction of data knowledge, as well as reason and autonomously react or adapt to new, unpredictable situations.

In order to provide a solution to the RQ2, Chapter 4 proposes an adapted data-driven methodology from data science, implemented through a data modelling framework that can process massive amounts of data in motion, which can be used in the smart mobility complex scenarios. The features of the proposed framework meet the requirements for massive data collections and complex scenarios. First, the framework allows data modelling through temporal graphs which speed up the deep search and query capabilities in large data collection but also with data that is collected in a continuous manner through the nomadic devices. Second, it allows the *what-if* analysis by exploring different alternatives, performing multiple parallel simulations in order to find the best solutions for complex smart mobility scenarios. Finally, the scalability and possibility of implementing any ML algorithm make this framework a perfect solution for the smart mobility domain. The practical usage and benefits are explained in a case study of a complex collaborative mobility scenario. The framework's performance is tested with a large-scale dataset, performing complex tasks and providing interactive real-time data visualization.

Even if ML methods and algorithms can help to extract commonalities over big datasets, there are cases where different entities behave very differently and common behavioural model can be unsatisfactory. In this case, each entity (*e.g.,* user) is transformed in a so-called "system of systems" [174]. The solution is to design the system in a more user-centric approach by individual profiling of users' travel habits, activities and visited location. Then appropriate solutions can be found by searching for personalised solutions that match users' travel habits and preferences. Consequently, this approach calls for the next RQ which can be formulated as follows:

RQ3.
How can profiling analysis of people's habits give new insights and offer a new perspective on the study of people's behaviour

The transportation industry advocate the necessity of shifting from classical ITS to Smart Social Mobility Services [181]. This implies an integrated and cooperative approach to sense users' individual patterns, interactions and offers user-centred mobility services. The main idea is to dynamically sense the specific mobility needs of users and recommend the best possible solutions through CMS. Such processes consist of a series of tasks, starting with detecting users' mobility needs, identifying a set of solutions to address the needs, and sending personalised recommendations for transportation sharing services to similar users with similar profiles. Profiling users and locations reveals valuable insights, travel patterns and changes of habits which are extremely useful for CMS in order to match people and transportation resources, taking into consideration personal preferences and the overall impact at the system level for any given recommendation.

Chapter 5 answers RQ3 by proposing a scalable method for dynamic profiling. The proposed method allows extraction of individual users' travel behaviour and valuable

knowledge about visited locations, travel habits and patterns in an automatic way, using only geolocation data collected from mobile devices, without any user input. Through this chapter, we explore and provide the foundation for the next generation of smart mobility recommendation systems, which can extract knowledge from raw data and then profile each individual, register each activity performed and each location visited accordingly. The profiling is done without any user input (except the GPS data from Google Map) and can be used simultaneously by a multitude of shared mobility applications (*e.g.,* , ride sharing, parking sharing, car sharing applications). Each of them can search for solutions to match people with similar travel patterns and habits, in order efficiently to share transportation resources (*e.g.,* cars, parking spaces).

Logically, the next research question can then be formulated as follows:

RQ4.
What is the impact of the proposed contributions in practical applications?

In Part II of the dissertation, different case studies are presented which demonstrate the usage of the profiling method, through practical applications in the shared mobility domain. Chapter 6 delves into practical application by demonstrating how the profiling is successfully used to extract users' travel patterns and habits, based on the historical visit patterns (*i.e.,* time of the day and duration of each location visit). The main contribution is that the profiling is done automatically, without any user input or intervention, using only the GPS data collected passively (without any user intervention) through the nomadic devices (*i.e.,* , via Google Map data). This means that the presented methodology leaves the door open to the next generation of recommendation systems, which will be able to make use of the automatic knowledge discovery, travel habit and patterns from raw data. The methodology combines the modelling framework proposed in Chapter 4 with the dynamic profiling method from Chapter 5 to automatically perform the classification of location type and activities performed on each location. The end result is represented by the probability of performing a certain activity at a certain location. We demonstrate also that additional rules (*i.e.,* a heuristic rule and a Bayesian update rule) can improve the estimations by considering the value of the information over time. Moreover, better results have been obtained when the final result is coupled with GIS data about the number of facilities located in a certain area. This information can be downloaded in near real-time from existing GIS platforms (*e.g.,* OpenStreetMaps) to further improve the overall estimation by looking at the existing facilities that match the probabilities obtained. Then, since the location profiling reveals the users' visit pattern of any location, different reasoning actions can be automatically performed *i.e.,* location type identification, labelling, classification of each location (*e.g.,* home, work, shopping, restaurant). The knowledge extracted automatically from raw data can enable the autonomous recommendation systems for sharing mobility, which will use this information to match users and transportation resources in an efficient way.

Throughout this dissertation, several challenges arose when seeking to achieve the objectives and answering the above RQ. In the following section an overview of the general challenges is presented, followed by the challenges addressed in each of the contributions from the Part II of this thesis.

1.3 Challenges

This section presents the challenges addressed in this dissertation. Each chapter discusses a challenge derived from the contribution section, pointing to the provided solutions. Each challenge corresponds to a concrete issue occurring in the transportation domain, particularly in the shared mobility topic. Very often solutions from data science domain are employed. Figures 1.2 presents a flowchart with the challenges and the addressed chapters.

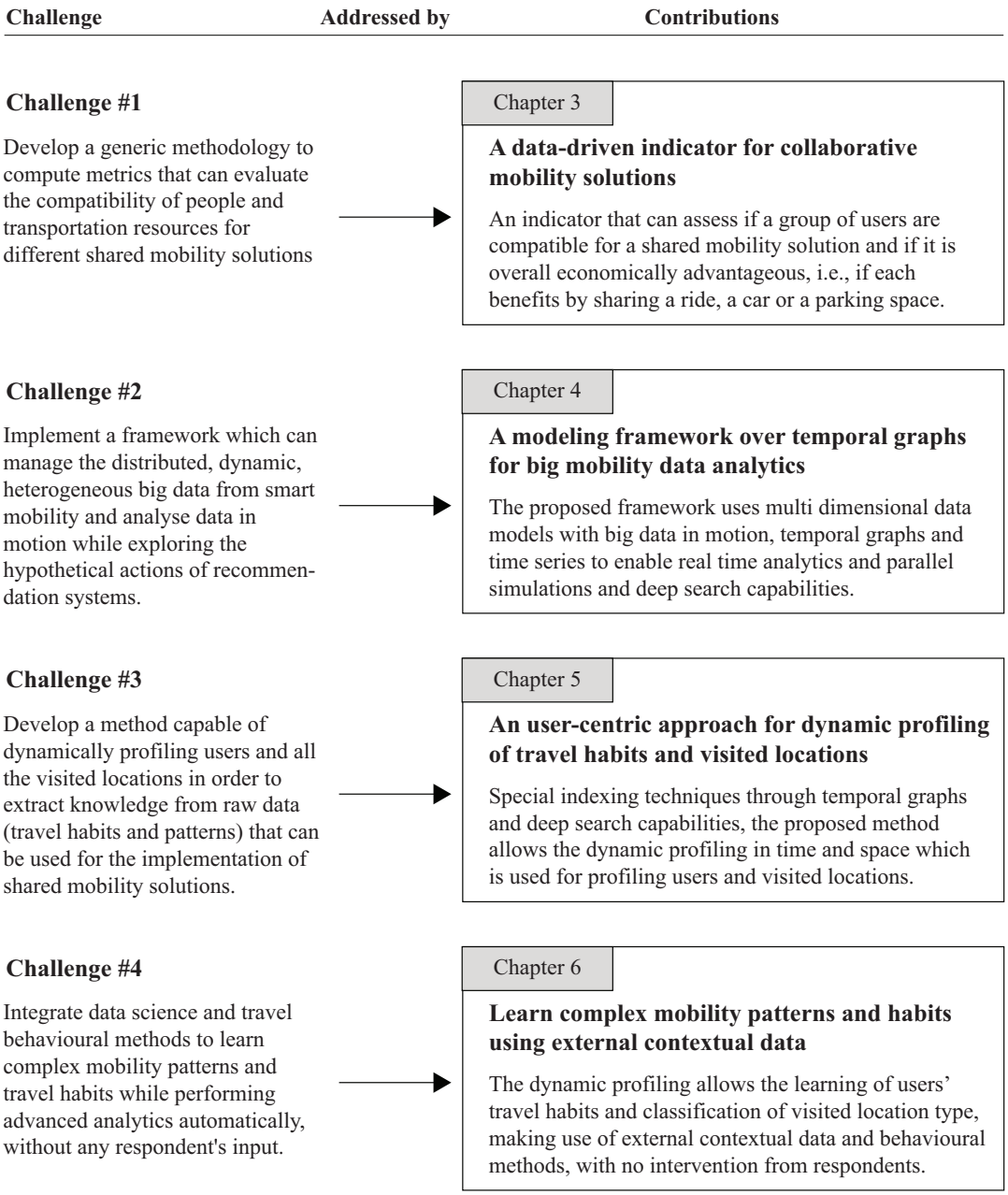


Figure 1.2: Challenges and contributions flowchart

Challenge #1:
Develop a generic methodology to compute metrics that can evaluate the compatibility of people and transportation resources for different shared mobility solutions.

Sharing travels, parking spaces and transportation resources (*i.e.,* cars, bikes) promise to be an effective way to increase the mobility resources usage rates and to reduce the number of cars and consequently road congestion. However, the literature shows many problems must yet be solved to achieve this objective, involving both operations issues (*i.e.,* how to best match users' in time and space) and behaviour challenges (*e.g.,* specific conditions that users consider when choosing to travel together, or arrangements for accepted detour and rescheduling) [211]. One of the biggest challenges is that different sharing services and solutions (*e.g.,* carpooling, parking sharing, car sharing) have different constraints, requirements and methods of finding compatible users, times and routes, all of which can satisfy all the travellers. At the same time, all the solutions found must reduce the cost and/or time of travel, at the level of both the individual and the exploitation system. Solutions must also offer a degree of flexibility in order to maintain the mobility service quality level and user satisfaction.

In order to respond to this challenge, the contribution from Chapter 3 is the introduction of an indicator that can be successfully used to assess if different users are compatible and whether is economically advantageous to share the ride or parking space for both the short or long term. The proposed method is designed in such a way that not only takes into consideration all the variables (*e.g.,* schedule, personal preferences, flexibility) and possible costs (*e.g.,* travel cost, parking fee), but also provides an implementation usable by future ML based RS.

Challenge #2:
Implement a framework which can manage the distributed, dynamic and heterogeneous big data from smart mobility and analyse data in motion, while exploring the hypothetical actions of recommendation systems.

One of the biggest challenges in the smart mobility domain is the use of data science as an enabler for implementation of large scale transportation sharing solutions. In particular, the next generation of ITS requires the combination of ML, AI and discrete simulations when exploring the effects of what-if decisions in complex scenarios with millions of users. This challenge is addressed in Chapter 4, which develops a multifunctional framework that can satisfy the requirements of descriptive and predictive analytics in real-world shared mobility scenarios (*i.e.,* carpooling, car sharing, parking sharing). We demonstrate that the proposed framework is able to handle massive amounts of continuously changing data coming from data in motion. Moreover, the proposed methodology is capable of merging the discrete simulations and statistical results in a single framework in a fast, efficient and complete architecture that can be easily deployed, tested and used.

Challenge #3:
Develop a method capable of dynamically profiling users and all the visited locations in order to extract knowledge from raw data (*i.e.,* travel patterns and habits) that can be used for the implementation of shared mobility solutions.

Statistical methods are widely used in the transportation domain to extract commonalities and similar behaviours for a large number of users [212], [191]. However, providing personalised shared mobility solutions for each individual using a common behavioural model cannot provide all the time satisfactory results. This challenge is addressed in Chapter 5, where, for the first time, a method is proposed for dynamic profiling of users and visited locations. The resultant profiling method is used to extract insights, travel patterns and habits, using only the GPS data collected from nomadic and wearable devices, without any travel survey or user input. Using this method, valuable knowledge regarding travel habits can be discovered, which can be used for recommendation systems to match people and shared mobility services in an autonomous, fast and dynamic way.

Challenge #4:
Integrate data science and behavioural methods to learn complex mobility patterns and travel habits while performing advanced analytics automatically, without any respondent's input.

Until now, extraction of daily and weekly activity/travel patterns and habits has been done using manually user reported information, coupled with GPS loggers that record the movement automatically [194], or the collection of data using smartphones [20]. However, all these methods have required a certain degree of user input in order to extract knowledge from raw data and to reason about the travel habits and patterns of each individual. The result: an expensive data collection process resulting in less data collected, due to a lower number of respondents willing to engage in this activity. Although a completely independent data collection and analysis method could be a solution, this represents a big challenge, as the data collected has no semantics interpretation or meaning in this case. This challenge is addressed in Chapter 6, where practical examples of learning complex mobility patterns and inferring urban mobility and habits are presented. Therefore, complex analytic and reasoning actions were performed automatically, without any user intervention *e.g.,* classification of the activity performed in each location and reconstruction of complex mobility patterns. We demonstrate also that the use of external contextual data from GIS information coupled with different rules can improve the overall accuracy of the proposed model.

1.4 Contributions and thesis structure

This thesis is composed of three parts. Part I introduces the thesis (with context, motivation, objectives of the thesis and the addressed challenges), background and state of the art. Part II proposes solutions to address the research questions and main challenges. Finally, Part III concludes the thesis. The overall structure of the thesis is presented in Figure 1.3.

Part I: Introduction and state of the art. Chapter 1, introduces the context of the thesis, the motivation behind the research, an overview of the general chal-

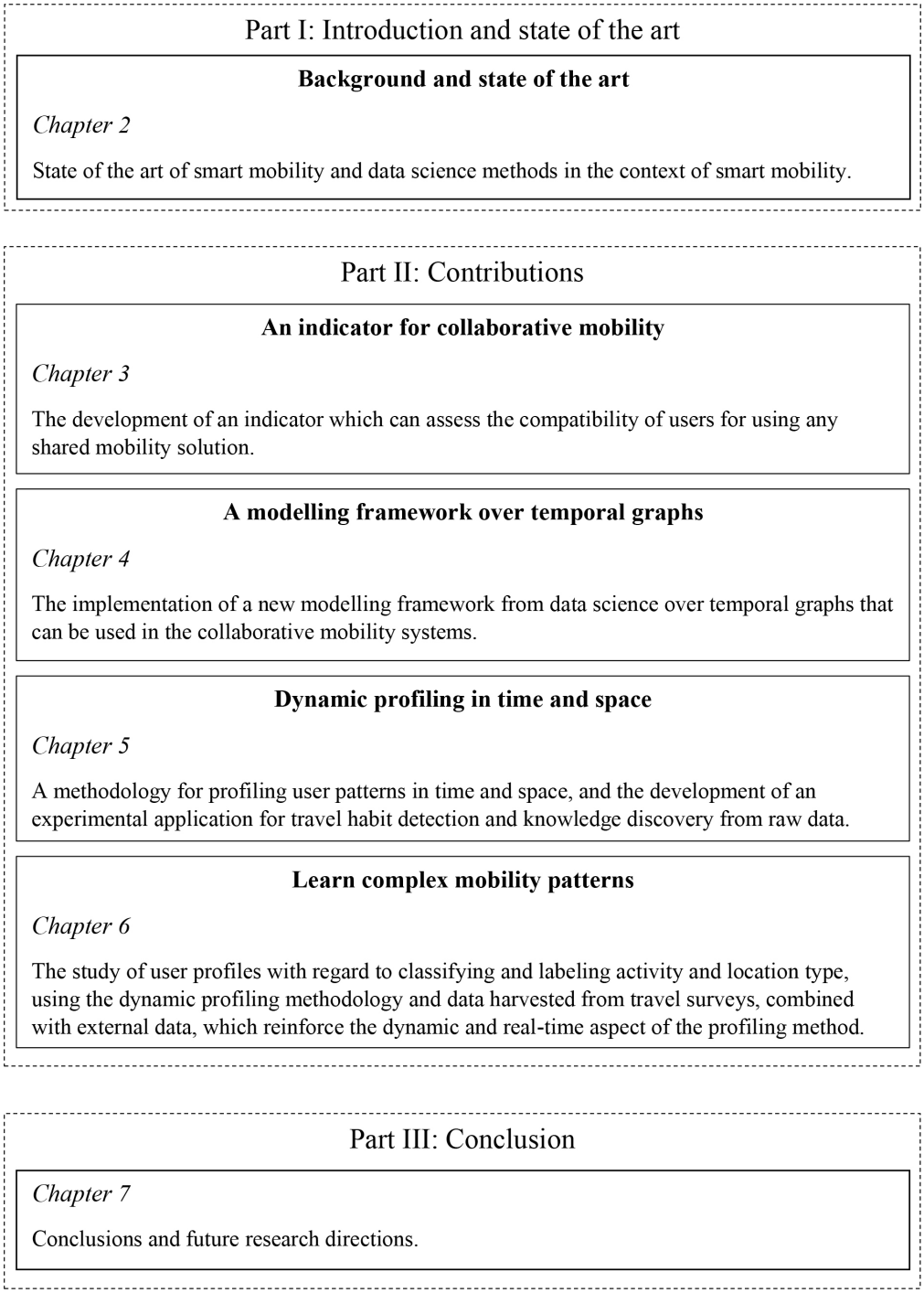


Figure 1.3: Thesis structure

lenges from the smart mobility and data science domains, and finally the challenges addressed in this thesis. Chapter 2 presents the background and the state of the art of transportation engineering and data science in the context of smart mobility.

Part II: Contributions. The second part of the thesis addresses all the RQs presented in Section 1.2 and challenges presented in Section 1.3. Chapter 3 focuses on the transportation domain methodology and presents the collaborative mobility issues and challenges that must be solved in the smart mobility and shared mobility domains. The main contribution is the theoretical foundation for an indicator that can match people with a combination of shared smart mobility services, which answers *RQ1* and *Challenge 1.* Experiments have been performed with a small dataset. The remaining challenges are the dynamic requirements and scalability of the proposed methodology which will be solved in the next chapters.

Chapter 4 solves the Challenge #2 and RQ2 by implementing a new modelling framework from data science over temporal graphs, adapted and modified to be used with geolocation data in the transportation domain and in particular in the collaborative mobility systems. The dynamic and scalable features of the framework are tested for large datasets and in a practical case study. The remaining challenges are represented by the need for a more centric approach (that deals with individual users' requirements and preferences), addressed in the following chapter.

The main contribution of Chapter 5 reside in the development of a methodology for profiling in time and space of mobility patterns, detection of habits and change of habits that can be used by future RSs. This chapter answers RQ3. Challenge #3 is addressed through the applications which are evaluated and results discussed for different shared mobility solutions and tasks *i.e.,* ride sharing and parking sharing user matching, activities and location classification, and profiling of non-recurrent trips (*e.g.,* holidays and business trips).

Chapter 6 presents the study of profiling users, activities and locations by using the methodology presented in the previous chapter and the users activity matrices from different travel surveys. This study answers RQ4, using only user data collected by smartphones. The main contribution of this chapter addresses Challenge #4 by proposing a new way of inferring urban mobility and user habits from passive smartphone data collection (*i.e.,* without any user input or intervention). In order to be more accurate, the results are empowered with extra data obtained from other external data sources, which reinforce the dynamic and real-time aspect of the proposed profiling method.

The contributions presented in this part are based on work that has been presented in the following papers:

- Usage of Smartphone Data to Derive an Indicator for Collaborative Mobility between Individuals
  B Toader, F Sprumont, S Faye, M Popescu, F Viti
  ISPRS International Journal of Geo-Information 6 (3), 62

- A new modelling framework over temporal graphs for collaborative mobility recom-

mendation systems
Bogdan Toader, Assaad Moawad, Francois Fouquet, Thomas Hartmann, Mioara Popescu, Francesco Viti,
2017 IEEE 20th International Conference on Intelligent Transportation Systems (ITSC)

- A Data-Driven Scalable Method for Profiling and Dynamic Analysis of Shared Mobility Solutions
  Bogdan Toader, Assaad Moawad, Thomas Hartmann, Francesco Viti
  IEEE Transactions on Intelligent Transportation Systems (submitted in 2018, unpublished to date)

- Using Passive Data Collection Methods to Learn Complex Mobility Patterns: An Exploratory Analysis
  B Toader, G Cantelmo, M Popescu, F Viti
  2018 21st International Conference on Intelligent Transportation Systems (ITSC)

- Inferring Urban Mobility and Habits from user location history
  Guido Cantelmo, Bogdan Toader, Constantinos Antoniou, Francesco Viti
  22nd EURO Working Group on Transportation Meeting, EWGT 2019, 18 - 20 September 2019, Barcelona, Spain (submitted in 2018, unpublished to date)

Finally, this dissertation is concluded in Chapter 7, where possible future research directions are discussed.

This research has been funded by the Luxemburgish FNR (Fonds National de la Recherche) through an AFR grant for the PLAYMOBeL project (9220491) and by the EU Marie-Curie-funded project InCoMMune(618234).

2

Background and state of the art

This chapter presents the general background for this dissertation, before the state of the art is discussed in the following chapter. It first introduces important terms and techniques for smart mobility and data science. The chapter then details relevant topics and a related literature review in order to correctly frame the contributions presented in the second part of the thesis.

Contents

2.1 Background

The rapid large-scale adoption of IoT has caused issues in different industries and domains. An example of a domain that uses a methodology similar to the one proposed in this dissertation is the electrical smart grid domain.

Big data collected continuously from smart meters must be handled efficiently. Using a fusion of methodologies and technologies from data science and domain specific knowledge, it was possible to build a proper knowledge representation of the context and to take adequate actions for continuously sensed data, analysing complex data in motion at scale with temporal graphs [108], [107]. The objective is to analyse data collected in a cyber-physical system in near real-time, and to ultimately support decision-making processes based on the results of this analysis [106].

Similarly, ITS drives the implementation of data science techniques for real-time data analytics in the transportation domain. This means that new methodologies must be able to handle not only data at rest applications (*i.e.,* data collected is analysed after the event occurs) but also data in motion (*i.e.,* the analytics occur in real-time as the event happens). Data in motion gathered from advanced sensing (such as built-in sensors from mobile devices) and other types of traffic information (such as traffic metering) can be combined to better analyse users' travel behaviour in near real time and derive specific mobility habits and travel patterns. Travel behaviour analysis is the core of modern smart mobility, from which sustainable solutions for collaborative mobility services can be derived. In order to study users' habits, mobility patterns must be extracted, analysed and solutions must be provided at different scales.

Over time, the literature [27] emphasizes that the contribution of ITSs can dramatically improve urban mobility. Following the recommendations from [87], additional research must be done as the ITSs must be prepared for analysing data in motion in near real-time, learning users' behaviour and performing fast searches in large datasets, which could instead contribute to a more integrated, fast and flexible method for implementing collaborative mobility services at different levels and for different needs.

2.2 State of the art

This section discusses the state of the art and literature review related to the work presented in this dissertation. The content from each section is relevant for a good understanding of the contributions from the second part of this thesis. In the first part, Section 2.2.1 presents the relevant state of the art from the smart mobility domain. In the second part, Section 2.2.2 presents a review of the data science methods, challenges and technologies used in the in context of smart mobility, presented in the contributions from Part II.

2.2.1 Smart Mobility

The smart mobility research domain is a wide topic, referring mainly to the study of methods and technologies related to shared and soft mobility, including ride sharing, car sharing, public transportation, walking, biking, and more. In order to correctly frame the research area of this thesis and the motivation behind the state of the art, we need to look at a complete application which is developed throughout this thesis. In Figure 2.1 we can observe a complete flow and main processes from a smart mobility recommendation system. In the rest if this section, we dedicated a subsection for each part of this process flow.

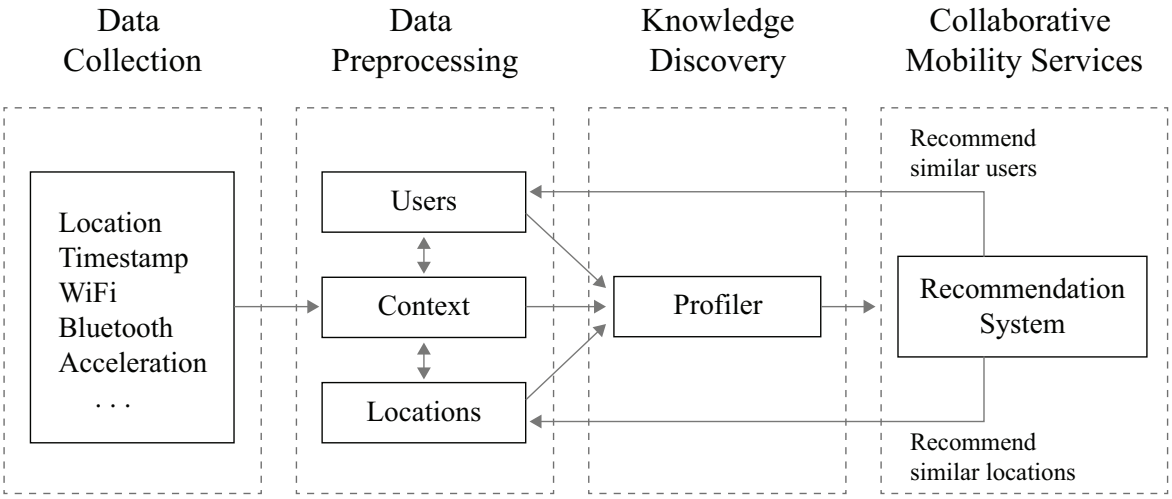


Figure 2.1: The phases and process flow of a smart mobility recommendation system

First of all, the main contribution of this thesis is related to research and development which can contribute to the advancement of CMS. Although this thesis does not aim to develop a functional recommendation system, the proposed study represents the foundation for future research in this direction. Therefore, Section 2.2.1.1 presents the state of the art of CMS, with a review of the issues and problems which need to be solved, concepts and methodologies which are used in the literature, and historical advances in this topic. This represents the theoretical foundation of the contributions from the second part of this thesis. Thus, a future RS can be defined as an intelligent system which can autonomously process a flow of data, reason about the input information and the context of all the involved entities, and provide personalised recommendations. For example, in the case of the application presented in Figure 2.1, the RS receives the input of users' profiles and travel habits. Therefore, the system can match users with similar mobility patterns to efficiently use shared mobility solutions or recommend similar locations to visit, based on the historical visited locations.

The above process flow has different phases and starts with the data collection phase. Data can be collected using different digital methods, devices and sensing systems. There is no restriction on the type of data which can be collected (*e.g.,* GPS location, Wifi and Blutooth connections, motion data). The type of data depends on the features of the devices used for data collection and the embedded sensing systems (*e.g.,* motion sensors, GPS modules, connectivity components). In order to have the basic

understanding and to motivate the choice of the methods and technologies used in the contributions from this thesis, Section 2.2.1.2 presents a historical summary and the evolution of data collection methods and technologies used, from the basic paper-based travel surveys right through to the latest passive sensing systems.

Once the data is collected, the following stage is the data preprocessing. In this phase, basic preprocessing tasks are performed *e.g.,* data cleaning and reconstruction of missing data, data structure (in the example from Figure 2.1, data structure defines the users, context and visited locations). Section 2.2.1.3 presents a review of the methodologies used in the literature for the automatic extraction of the activity location and duration. Based on this review, the contribution from Chapter 3 uses one of the presented methodologies to clean and reconstruct the missing data collected with the smartphones. This stage is important in order to provide the most accurate information for the profiler.

The following stage is the knowledge discovery from the data received as input from the preprocessing phase. Section 2.2.1.5 also presents a review of the methods used in the literature for profiling user travel behaviour and visited location. This review is useful for understanding the contribution from Chapter 5 a user-centric approach for dynamic profiling of travel habits and visited locations is proposed. Similarly, Section 2.2.1.4 presents a review of the methods used for inferring travel behaviour and trip information from raw data without any user intervention/input. Related to this, Chapter 6 presents a practical application to learn complex mobility patterns and habits using external contextual data. In the rest of this section we present a review of the main topics from Figure 2.1.

2.2.1.1 Collaborative mobility services

Nowadays, quick and easy transportation has been an essential part of modern society. Nevertheless, increased urbanisation makes traffic congestion worse unless there are big changes in citizens' travel behaviour to promote more efficient, sustainable and environmental travel alternatives [114]. In order to solve this issue, different solutions have been explored and it is important to know what actually is causing traffic congestion. Interesting to note, the average number of passengers per car (assuming standard vehicles with five passengers including the driver) for European countries is approximately 1,45 passengers. This means that vehicles are often running at low occupancy, with only 29% occupation rate [9].

Collaborative mobility solutions such as transportation sharing services (*e.g.,* carpooling, ride sharing, car sharing) caught the attention of a larger number of researchers. Numerous studies have already shown the impact of collaborative mobility on the environment and its effects on the behaviour of people (see *e.g.,* [152], [86], [121], [157], [58]). Other studies analyse factors influencing shared mobility activities. For example, [135] conducted a survey to analyse people's views regarding carpooling activities. The results showed that 55% of the respondents did not carpool because of difficulties in finding compatible users (with similar location and schedule) and 45% prefer the flexibility of solo driving. Another survey showed that the poor carpooling schedule

and trust level between strangers are two major obstructions for carpool activities [188].

At the same time, the study individual activity travel patterns has shown that the repetitiveness of individuals' activities is influenced by several factors *e.g.,* the types of activities, accessibility of different locations and different commitments. Also [197] showed that different types of activity have different pattern of repetition. Interestingly, [33] demonstrate that in a period of six week study, 70% of all the trips have a repetitive behaviour of visiting the same 2–4 locations. This means there are incentives to exploring travel behaviour in an attempt to extract travel patterns and habits which can be used by CMS to match people and transportation resources efficiently. The question that arise in this context is how we can integrate all the components in an intelligent system which can provide advisory/recommendation services in order to attract more people towards collaborative mobility for a more efficient use of transportation resources. To answer this question, it is important to understand which are the available shared mobility solutions and their issues, limitations and promising research direction to improve and integrate them in an unified and synchronised manner.

Different shared mobility solutions are available, in the form of private and public services, which involve different resources, conditions and operating procedures. We can categorize these services into *joint sharing* and *concurrent sharing services*. While in joint sharing (*e.g.,* carpooling) the objective is to group more people in fewer cars, in concurrent sharing (*e.g.,* carsharing, parking sharing) the target is to intensify the resource usage. The question that arise is how the collaborative mobility can be assessed in a combination of services in such a way that all those services can complement each other. Thus, the concept of CMS must be characterized by the following rules and features:

- Shareable resources among users can be both private resources (*e.g.,* private cars) and public or private third parties (*e.g.,* cars haring system, parking lots)
- The ITS collects on a continuous basis the travel patterns and preferences for all users in a closed environment system
- For joint sharing services users must be matched for simultaneous usage of a transportation resource (*e.g.,* a car). This requirement is applicable for both recurring trips (carpooling) and for instant ridesharing (dynamic ridesharing)
- For concurrent sharing services, compatible users must be matched so that they can use the resources without overlapping
- ITS acts as a recommendation/advisory system, measuring compatibility of different combinations of resources by a group of users
- The objective of CMSs at the system level is to match users in order to maximise the resources' usage, combining different services to be used by compatible group of users. The objective is to assure that the total cost per traveler is lower than if the passenger did not use the collaborative sharing system

Even if the existing literature has emphasized the importance of the above mentioned solutions to traffic congestion problems, much of the literature pays particular attention to only parts of those services and objectives and rarely in a combined methodology. In the rest of this section we present the challenges, additional work related to CMSs and we point to the solutions proposed in this dissertation.

Matching compatible users for shared mobility solutions has recently attracted more attention in the literature [218]. For example, some studies focused on the optimization problem of finding efficient matches between passengers and drivers [18]. Optimization algorithms for complex matching rides systems are also explored in [95]. A real challenge for high-dimension matching problem is the development of complex recommendation systems which can find compatible users and sharing solutions. Some studies explore the required functions that must be implemented in the future recommendation systems for scaling the ride sharing matching problem [101]. For example fast algorithms are proposed to generate the shortest path considering different requirements ([113], [79]), but they do not consider the involved cost evaluation of each provided solution. There are studies which evaluate the matching of individuals in shared mobility systems using only preferred departure times and different matching strategies in near real time ([23], [35]). Other studies quantified the reduction in travel costs when sharing rides ([94], [123]).

Another form of joint sharing services is the semi-organized ride sharing practice, defined as flexible carpooling in the literature ([132], [189], [48], [121]). This is gaining popularity especially where high occupancy vehicle lanes are implemented because individuals want to benefit from using those faster lanes and reduced tolls. In this type of shared mobility solutions, passengers and drivers meet spontaneously in specific locations without any notice or exchange of information. An advantage of the flexible carpooling scheme is the convenience provided without any specific commitment. The disadvantage is that requires a large number of users.

With the emergence of the internet, a number of private matching agencies emerged to provide diverse ride sharing services for travellers. An extended review of these agencies shows that still the ride sharing has continued to decline [92]. The main difficulties identified are: schedule incompatibility between users; cost-sharing difficulties; and the lack of methods for choosing a specific route which is advantageous for every passenger. These challenges demonstrate the need for innovative systems and services that can actually successfully redirect people's behaviour towards a more efficient, sustainable and friendly environment thanks to the sharing economy.

Recently, new user-centric services are transforming urban mobility by providing timely and convenient transportation through Mobility-on-demand (MoD)) systems, led by companies such as Uber, Lyft, Blablacar. Studies shows that ride-pooling services can provide substantial improvements in urban transportation systems [21]. These decentralised systems based on shared economy principles, provide a reliable mode of transportation which is focused, used and run by individuals who can both operate and access the services on demand. These services provide access to mobility by reducing the cost, the waiting times and the stress associated with travel.

Privacy is another challenge in the sharing services scheme. One concern is the risk

of exchanging private information with strangers [60]. The loss of privacy due to systematic data collection of private information is another major concern [24]. We address this issue by proposing the implementation of sharing services schemes at the organisational level. In the second part of this thesis we explore different case studies in a closed environment where trust is much higher between individuals and organisations have motivation to securely keep sensitive anonymised information in-house. Moreover, the solutions provided also have the ability to hide sensitive information, using specific blurring techniques *i.e.,* for geolocation data, a blurring method is the limitation of accuracy at higher levels (*e.g.,* one kilometre).

Recent technological advances in nomadic and wearable devices (smartphones and smartwatches) combined with the market penetration of portable technologies and the latest developments in transportation are emerging as an attractive option for large-scale sensing of human behaviours which can contribute to the evolution of the ITSs [64]. Sensor technologies embedded in the mobile devices carried by travellers can generate unprecedented amount of data related to human mobility patterns [130]. One can argue that today we have the technologies and tools to solve those issues. The solution can be the combination of the collaborative mobility schemes with the existing technologies and the development of new data-driven ITS [228]. In this context, ITS should offer individual based, real-time information about all the users and sharing alternatives and advise them which sharing solution to choose in order to save money, time and have a decent degree of flexibility. In this context it is crucial to automatically collect and process the massive datasets generated, in order to detect travel patterns and possible interaction between the system participants without the intervention of the respondents [232].

Working in this direction, the contribution from Chapter 3 proposes a data-driven indicator for collaborative mobility which aims to extend the functionality of the RSs, not only for the first-come-first-served but also for long term services (*e.g.,* planned carpooling). In this case, it is possible to evaluate the recurrent but also the dynamic ride sharing compatibility between users in a single metric. This allows one to find the best option for a group of users, based on the compatibility score between individuals and available sharing solutions. Moreover, in the case that no matching has been found between a group of users, or that the compatibility score is low, a future RS can send personal advices on how travel behaviour can be changed (*e.g.,* rescheduling and reordering of activities) in order to increase the matching rate and the compatibility score between individuals.

In order to solve the above challenges and overcome related limitations, the entire process should be automated and a method must be devised to assess all service-specific variables and constraints in a single metric, which can consider individual preferences and maximize user flexibility. Automation of the entire process is attractive because it removes the users' burdens not only of manually inputting and constantly editing the journey plan, but also of searching for the best sharing solutions and compatible users for CMSs. Therefore, location, origin, destination, departure time and sequence of activities must be automatically extracted from the data collected. In Section 2.2.1.3 we inspect the methodologies for extraction of activity duration and location from raw GPS data in order to automatically identify the sequence of activities and individual travel patterns. The obtained patterns can be used by a future recommendation system

to automatise the extraction process, using the specific profiling techniques presented in the contribution from Chapter 5 and used in applications to learn complex mobility behaviours, as can be seen in the contribution from Chapter 6.

The very first step in this process is mobility date collection. The following section presents a summary of the evolution of data collection methods from the most primitive to the latest passive sensing systems.

2.2.1.2 Data collection methods and technologies

Thanks to the digital revolution in the transportation domain, classical data collection methods (such as traditional travel diary) are being replaced by an increasing number of digital surveys. As technology has evolved, different nomadic devices (*i.e.,* smartphones, smart watches) have been used in order to increase data collection accuracy and relevance, starting with the classic GPS-based logging surveys, to GPS data collection using smartphones and the latest smartphone prompted-recall based travel survey. These applications, including their usefulness and limitations, will be explored in the rest of this section.

Traditional travel data collection methods using paper-based surveys required significant contributions from survey respondents, which usually have to manually record all trip sequences and activities information. As a result, travel surveys usually run for a limited amount of time (not more than a few weeks) and, due to the self-reporting process, include a degree of under-reported or incomplete information (such as wrong departure times or missing activities) [236]. Such limitations make traditional surveying methods inadequate to capture variations in travel behaviour that do not occur within a short time period [15]. With the advent of technology, digital travel surveys have proven to be a valid alternative to overcome these limitations.

A number of studies [15], [207], [230] have shown how digital travel surveys or travel diaries (via *e.g.,* smartphones) reduce survey burdens while increasing data accuracy. In other words, we can collect data that is more accurate for a longer time period [15]. The reason is that automatic surveying leverages powerful ML techniques to infer activity and/or mode information, meaning that users have to validate and eventually correct their information instead of introducing it manually [160]. Moreover, after an initial phase of training, the system learns user preferences, thus further reducing respondent's efforts [230]. Hence, such survey methods can overcome the main problems of collecting travel diaries, mainly their high costs, their accuracy and the limited collection times.

The first digital data collections were represented by the GPS-based logging surveys. The usage of GPS devices has many advantages, such as reduction in respondents burden, higher data accuracy, detailed trip route and the ability to extract additional information such as vehicle speeds [61]. This type of survey has been widely implemented, most of the time as a complementary solution to household travel surveys [41], [59]. Even though there are obvious advantages to these passive travel surveys, equipping the respondents with those devices is not only expensive, but also generates

additional problems, such as the need to charge and carry on additional equipment, followed by a recollection phase which brings along other problems [194], discussed below.

Although device-based logging has been widely adopted worldwide, it still relies on some additional devices that users have to carry all the time [230]. In practice, there are two options. On one hand, the agency conducting the survey can provide a wearable device, such as a smartwatch, with a sufficient number of sensors and dedicated software [196]. However, good wearable loggers are expensive and become obsolete quickly. Moreover, users might forget to carry them, thus introducing errors and unreported trips.

A second option is using smartphone-based travel surveys, which rely on applications that can be downloaded and run in the background [230]. Recent technological advances in smartphones combined with the increased market penetration of portable technologies emerged as an attractive option for large-scale sensing of human behaviour [65]. Using the additional embedded sensors for proximity, motion and connectivity, generates unprecedented amount of data related to human mobility patterns [131]. Moreover, as most of the time respondents possess a personal smartphone, there is no need to carry additional devices or to recollect the data, this taking full benefit from the au- tomatic data transmission over internet [32], [47]. As respondents install the software on their personal smartphone, the probability of securing more comprehensive information rises. Additionally, smartphones have a variety of sensors, such as GPS, accelerometer and proximity sensor. These can infer both activity and mode information [15]. However, smartphones are powered from batteries with a limited capacity, mean- ing that power consumption is a limiting factor. Even though many applications claim to have low power consumption, battery duration, reliability, and lifetime depends on the manufacturers. Moreover, the higher the number of sensors, the higher the power consumption.

The latest generation of travel surveys combines the pervasive and advanced sensing smartphone data collection with a prompted-recall based travel survey [231], in order to extract additional information that was not possible to be extracted automatically, such as activity type performed in a specific location. Because the respondents still have the burden to manually complete or confirm part of the survey, there have been attempts to make the data collection process semi-automatically [161]. Different systems have been proposed such as SmartMo [37], MEILI [20] and rMove [100], but all of them require a certain degree of user reporting.

To exploit the big data when extracting the travel patterns, in the second part of this thesis we explore methods that make use of passive data collections (*i.e.,* contribution from Chapter 5 which presents a method for dynamic profiling and Chapter 6 which presents an application for learning complex mobility patterns). In these studies, the data is obtained through the sensing systems embedded in the nomadic devices. The novelty of our approach is that the extraction of travel patterns and habits is performed without any additional user-submitted information or intensive data processing. This not only reduces the respondent's burden but can reveal new complex insights that cannot be captured through traditional methods.

After the collection phase of mobility data, specific methodologies for extracting activity duration and location must be employed. The following section presents a review of the main methodologies found in the literature.

2.2.1.3 Methodologies for extracting activity location and duration

The study of travel behaviour and mobility patterns mandates the automatic extraction of activities locations and places of interest. As we discussed in the previous section, this must be accurately done by nomadic and wearable devices, which can collect a large amount of location points, represented by position coordinates as well as the date and time when the sensor captured the location information. In order to transform the raw position points data into knowledge that a machine can understand or humans can visualise and interpret, this data must be transformed into so-called places of interest and trip information/routes (defined as origin and destination points). In this section we present and compare the most relevant methodologies and algorithms used by researchers for extracting activity duration and location, using different data sources, devices, sensors and algorithms. Chapter 3 presents a contribution which makes use of a methodology presented in this section.

In the past, researchers used data obtained from different GPS devices and other traditional sources. For example, Ashbrook and Starner, 2003 [30] used a wearable GPS receiver and a GPS data logger to collect data from six users for seven months. Clusters of places using a variant of the k-means clustering algorithm has been used in order to detect users' locations and sub-locations. They integrate the results in a system that can incorporate these locations into a predictive model of the user's movements. Several potential applications of such models are presented, including single and multi-user scenarios. The precision of the method was not tested, but they argued that this methodology can be the basis of future prediction algorithms, as well as relative frequency and probability of locations in time.

Hariharan *et al.,* 2004 [105] collected data using hand-held GPS devices carried on by two persons for one year. Using the classical approach, the method employed takes the temporal sequence of recorded locations and uses a set of decision rules based on distance and time between points in order to identify clusters of GPS points which represent visited locations. This agglomerative algorithm iteratively tests GPS points to determine if they remain within a given threshold distance. If the time between the first and last observed point exceeds a predefined stay duration, a cluster is assigned. Probabilistic models were developed for modelling a location history and probabilities of being in a specific location, within a given recurring time interval. Even if the method precision is not evaluated, this may serve as a starting point for exploration of probabilistic models of location histories.

In a similar fashion, Agamennoni *et al.,* 2009 [17] used a speed threshold criterion in order to identify activity locations from GPS records of the trucks in an open-pit mining site. Even if the presented algorithm is very fast, this is mainly due to the fact that the calculations are very simple, with data obtained from low-speed areas having no high variations and without taking time into consideration. Because the method

accuracy was not tested, this would probably serve better as a complementary method for identifying GPS errors (so called "supersonic jumps"), rather than an efficient method for extracting activities location and duration.

Technological advances in mobile devices and sensing systems have made data collection and processing much easier and more accurate. The study done by Xiang *et al.,* 2016 [219] proposed methods to extract stops from single trajectories using the sequence-oriented clustering method. In this method, spatial and time information are adopted as input. The proposed algorithm is able to detect effective stops and discard the false positive stops. The reachability is represented in a graph which illustrates the clustering structure and different levels of a specific trajectory. Even if the algorithm has a very high precision of 91.3% in recognizing the effective stops and eliminating false positive stops, it was tested only on small datasets for short distances on a small scale. The proposed method requires high computation, loading all the GPS locations and then computing using different tools, being suitable for small and very accurate distances. This can be a good complementary method for detecting false positive stops.

Besides the GPS data, Ohashi *et al.,* 2015 [164] used also accelerometer data from smartphones, with five features derived on the basis of the sensors' characteristics and specific human-travel behaviour. They propose a novel method for automatically extracting trips on the basis of continuously collected data. While conventional methods based on detecting stay areas with a boundary suffer from errors for short-distance trips, the authors showed that the proposed method was able to correctly extract the trips and suppress outliers in classifying each GPS point either in a stay or trip point. Moreover, the method uses the GPS-positioning error as a positive feature in order to classify an indoor location as a stay point. Even if the proposed method showed a promising 89,4% precision and correctly classified short-distance trips, this is more suitable for extracting trips than locations. Nevertheless, a continuous and intensive use of the GPS and accelerometer sensors can have negative effects in the energy management and the resources used.

To the best of our knowledge, one of the most advanced methodologies to discover places-of-interest from multi-modal smartphone data is presented in a study by Montoliu *et al.,* 2013 [156] which consider two different levels of aggregation or clustering in order to obtain the points of interest. In the first level of clustering, the location points are grouped in places of interest using a time-based clustering method. In the second level, the stay points are grouped in stay regions, using a grid-based clustering algorithm. A client-server system has been installed on smartphones, which collects location information by integrating GPS, WiFi, Global System for Mobile communications (GSM)) - and accelerometer sensors, among others. The method employed an algorithm to learn places of interest not only from the GPS data but also from the WiFi, bluetooth and GSM cell phone towers. Data is stored in a local database of locations associated with each entity scanned on a continuous basis. This is an efficient method of obtaining location from multi-modal mobile phone data with good accuracy even from indoor locations. This strategy results in significant savings of battery life, switching to different power saving modes (*e.g.,* GPS being programmed to switch off automatically when the location is obtained with the WiFi map or the phone is static). Moreover, using this method it is possible to reconstruct the missing location data (*e.g.,* if the GPS is not turned on, the location is set by the WiFi network). This

Table 2.1: Comparison table between methodologies.

	Montoliu et al., 2013 [157]	Thierry et al., 2013 [199]	Ohashi et al., 2015 [165]	Hariharan et al., 2014 [106]	Agamennoni et al, 2009 [18]	Ashbrook and Starner, 2003 [31]	Xiang et al., 2016 [220]
Real data (not-artificial)	•		•	•	•	•	•
From smartphone	•		•				•
Time-based	•	•	•		•	•	•
Energy optimization	•						
Working with missing location points	•						•
No. of subjects/ experiment period	8/5 months	750 artificial GPS tracks	8/15 months	2/12 months	1 truck/1 day	1/4 months + 6/7 months	Only small distances
Precision	63%	92.3% with artificial points	89.4%				91.3%
Algorithm/ method used	Time-based clustering, grid-based clustering algorithm	Kernel-based algorithm, hotspot exploration	Classification based on GPS and accelerometer	Agglomerative clustering	Speed threshold criteria	k-means clustering algorithm	Sequence-oriented clustering method

was one of the most suitable methodologies reviewed, given that it has been tested in a case study similar to the contribution referred to in Chapter 3. The only drawback is that it is much harder to replicate the same results from this study because the framework and algorithms were implemented in a private mobile application. Replicating this would be an arduous, time-consuming task and it is not the core of this thesis.

In order to obtain similar results using fewer resources and less effort in implementation, we use the methodology developed by Thierry *et al.,* 2013 [198] in the contribution from Chapter 3. The proposed algorithm differs from the traditional approach because it does not analyse data points sequentially, but it uses GPS points to build a kernel density surface. The peaks are selected as possible location stops and the GPS points are categorised as belonging to a trip or a stop location. The proposed algorithm has a precision of 92.3%, tested with an artificial dataset. Moreover, the code is available as a tool that can be used together with ArcGIS 10.

A side by side comparison between the presented methodologies with their features and performances can be seen in Table 2.1.

In the following section we present a review of the methods for inferring travel behaviour and trip information from the information extracted from raw data.

2.2.1.4 Inferring travel behaviour and trip information

The increasing availability of data coming from different sources, processes and devices offers new perspectives for research in the transportation sector. As a matter of fact, in the last two decades extensive interest has emerged in data-based explo-

ration of all matters relevant to the smart mobility paradigm. For example, travel behaviour data analytics has recently received much attention, because of the complex and dynamic character of human mobility patterns. As oversimplified assumptions on user behaviour will always lead to oversimplified mobility patterns, a broader knowledge about mobility needs is required in order to properly evaluate effective mobility solutions e.g., collaborative mobility services (sharing mobility, mobility-on-demand solutions, etc.). This brings new challenges to both data collection and meaningful data extraction from big data. This dissertation aims to contribute in this direction, providing new methods for travel patterns and habits extraction from raw data.

Since travel behaviour analysis is the study of individuals' patterns, an increasing number of mobility issues were solved through a more user-centric approach. This trend has been highly associated with the future research directions for more powerful multifunctional data-driven intelligent transportation systems [229]. This approach highlights the study and understanding of the complexity of human mobility behaviour and activities by fusing transportation engineering, data science and computer science [200].

Even if complex models capable of considering these phenomena already exist, they need individual and accurate inputs in order to provide realistic outputs. Those inputs can be obtained by combining data collected through smartphone sensing capabilities with user-reported information, thus making possible the activity detection and modelling. Although the data obtained is more accurate and relevant, this increases the respondent's burden when prompted to recall, annotate and classify the activities performed in different locations.

In the last decade, inferring trip information through data fusion of GPS traces and external contextual sources like GIS data have been proposed by different authors [42], [136]. The main idea is that GIS information can be combined with some heuristic rule about activity scheduling and duration in order to infer activity location – in the case of services – and mode of transport – in the case of transport facilities. However as mentioned in the previous section, this GPS logging approach requires carrying an extra device and prior information about *home* and – usually – *work* locations [42].

Additionally, land-use (location of residences, work places and other activities) changes continuously over time, meaning that the GIS database needs also to be constantly updated. This is the main limitation when the main interest is to collect activity information over a long period. To avoid this limitation, other authors proposed smartphone-based applications that do not require an additional device [15], [230].

While almost any method successfully identifies *home* and *work* location, recent studies show that last generation surveying methods show better accuracy and higher resolution in representing leisure activities [160]. While different systems have been proposed [99], [230],[160], all of them require a certain degree of user reporting. By way of comparison, other authors suggested the use of ML techniques to extract activity location from trajectories, as suggested in [215]. Lastly, some authors tried to exploit mobile phone location history [50] to model individual human mobility. The study present a methodology to extract individual mobility patterns from mobile phone traces of millions of users. However, these data are provided by a phone operator, which limits

their usefulness in modelling behaviour at an individual or household level [50]. By contrast, in the case of contributions from Chapter 6 we use fewer traces that are available through automatic passive data collection applications (*e.g.,* Google Maps) but for a longer period, meaning that underlying behaviour can be detected.

Concerning the methods available for inferring trip information, existing works use different techniques to infer travel information, which may be classified into two main groups: *heuristic* and *learning-based approaches* [15]. The main difference is that heuristic models rely on simple rules related to recurrent user behaviour (such as activity scheduling or duration) to learn trip characteristics [205]. This group of models may be considered *model-driven*, as it combines traces with land-use (or GIS) data in order to exploit existing knowledge about the transportation system [205], [195].

The main limitation is that these approaches are usually not general as they depend on a specific region or transport system [15]. For this reason, learning-based models have also been developed to derive these rules from some data through ML or data mining techniques [49]. However, in this case, the main problem is that different machine learning techniques will classify data in a different way causing different errors, which can be difficult to identify and fix.

Some of the most common approaches involve Neural Network (NN) [49] or Support Vector Machine (SVM) [163]. The main difference is that the former provides probabilistic results whereas the latter yields a deterministic value. Even in this case, to detect the ideal algorithm is far from trivial. While SVM has more appealing mathematical properties in term of convergence, NN better represents hidden phenomena such as risks associated with biased information. Finally, unless a large number of sensors is adopted, these methodologies can identify only a limited number of features [15].

As we saw in this section, inferring travel behaviour and trip information offers new perspectives of research in travel behaviour and travel patterns. In the following section we present a review of the travel behaviour analytics methods presented in the literature.

2.2.1.5 Travel behaviour profiling and advanced analytics

In general, data-driven travel behavioural profiling applied to ITSs refers to the process of constructing and applying various learning techniques, using the mobility data generated by users and other entities (*e.g.,* sensors from different means of transportation, traffic counters). In the transportation domain, profiling is a method used in various topics, with different objectives. Driver behaviour has been profiled using advanced motion sensors from cars and smartphones to detect driving events and to classify drivers according to specific categories. Profiling methods are used in fleet management, insurance policies, fuel consumption optimization or gas emission reduction [56], [166], [119], [63], as well as in route choice in multimodal networks in order to consider the individual preferences in route recommendation systems [52] and in Internet oriented user centric ITSs [53].

More recently, attention has focused on understanding human mobility using the profiling of users [97]. Data generated by static and mobile sensors implemented in different transportation systems and smartphones allow an understanding of large-scale patterns and habits of citizens [203]. This is used for semantic information ex- traction about user mobility , as well as spatio-temporal variations in travel regulations through transit data [143]. Mobility user profiles can offer valuable information for understanding the disaggregate and aggregate spatiotemporal activity patterns [96] However, the proposed methods are static and do not take into consideration data in motion. Furthermore, performance has not been tested with large datasets.

Profiling is also used in the study of human activities in space and time, which has been an important research topic in recent years. Mobility user profiles can offer valuable information for understanding the disaggregate and aggregate spatio-temporal activity patterns. Ghosh et al. [96] analysed a year of mobility trace data collected by wireless network connections in order to determine users' mobility profiles. A mixture of Bernoulli's distribution is used as the clustering algorithm in order to perform hub-level location predictions. Even if the method is efficient for so-called sociological behaviour analysis, it is not suitable for smart mobility recommendation systems. The presented profiling is static, does not take into consideration data in motion and the performance has not been tested with large datasets.

Data analysis about travel behaviour is often studied using statistical methods, which are useful in analysing aggregated characteristics of individual activities [213], [186]. The same cannot be said about using these methods for the analysis of individual mobility patterns and interactions considering jointly space and time [169]. A number of studies have focused on the visualisation and exploration of individual level activity data in a space/time context, starting from Hägerstrand's time geography conceptual framework for analysing the individual activity patterns with different constraints in space and time [104], until more recent works, which study also the user interaction aspect [122], [127], [187].

One of the most important research questions and objectives in exploring the individual mobility patterns is related to the similarity of spatio-temporal activity patterns. Different measures have been proposed to compare the level of similarity, including dynamic time warping [180], the longest common subsequence [137] or the Fréchet distance [22]. As these measures explore the similarity of mobility traces as spatial shapes without explicitly considering space and time in an integrated manner, they may fall short when not considering variables such as activity duration or time constraints, which are important components of scheduling users' activities.

Other studies consider space and time as an integrated measure with multi-level clustering methods, which represents a method of data visualisation and exploration. Using an individual-level activity diary dataset, [62] presents a geographic information system extension, which is able to cover a set of functions using different methods such as space-time path generation, segmentation and filtering. But these studies do not scale well on big datasets.

Several issues have been identified for further research in order to effectively profile user behaviour in smart mobility systems, including the learning issues for missing values,

data cleansing, dimension reduction, sparse learning, and heterogeneous learning [229]. Massive amounts of raw data collected by nomadic devices (*e.g.,* smartphones) must be cleaned, aggregated and then processed using state-of-the-art methods and algorithms from data science.

Previous research focused on the investigation of ride sharing opportunities [38] showed that through mobility data analysis, efficient solutions for extraction of suitable information from mobility traces can be used to identify ride sharing opportunities. As the ride sharing solutions must provide access to door-to-door transportation, the technological advances and the growing ubiquity of internet enabled mobile devices enable the implementation of dynamic ride sharing. The literature shows there is a need for optimisation of these systems, which are employed to solve different problems related to the required features and characteristics *e.g.,* the dynamic character, automated matching and cost sharing [18]. A suggested solution comes from a good understanding of users' behaviour and preferences, which is an essential feature when designing dynamic shared mobility systems.

In order to make use of collected data for large scale mobility sharing services, users' travel behaviour and preferences must be extracted from raw data. The very first step in this process is the extraction of the duration and location of activities from raw data. A detailed review and comparison of the methodologies from literature is presented in section 2.2.1.3. However, all mentioned methodologies suffer from limitations when applied to dynamic and live profiling on large datasets. More precisely, those methodologies use pre-defined parameters for a one-time extraction of different statistics and analytics. In practice, RSs require the profiling of users in an environment with continuous data generated by dynamic movements of users and means of transport. They need to extract knowledge that can contribute to the mobility services to understand the human travel behaviour and automatically recommend suitable sharing services for each individual. Moreover, the analytics must be done at different levels of aggregation and resolutions, with dynamic precision and scaling *e.g.,* ride sharing requires a higher accuracy than the classification of secondary activities (*e.g.,* shopping, gym or restaurant).

In a next step, using the detected locations from the previous step, special methodologies must be implemented in order to learn user mobility patterns and to perform the knowledge discovery from raw data. An example of knowledge discovery from literature is the trip purpose identification from GPS tracks [155]. The study identified two main groups of trip purpose imputation routines in the literature: rule-based systems based on the position of the activity, timing, and GIS data; and machine learning approaches which focus more on the activity and less on position. Montini et al. [155] used random forests [45], a machine learning algorithm that has been successfully applied in different transport-related classification problems. The input data is represented by the GPS, accelerometer data and a travel diary. The respondents were asked to correct an automatically generated travel diary that was used to extract specific features for semantic interpretation of the data.

A similar application used in the current dissertation is the identification/classification of each activity/visited location (*e.g.,* home, work). The proposed profiling methodology from Chapter 5 uses only the GPS data, specific data science techniques for

indexing, clustering and querying, alongside the training data represented by a set of known location visit patterns for each location type. The key novelty of this approach is that our methodology is able to capture detailed and complex visit patterns of users and locations through the profiling layer, which can be used in a multitude of applications simultaneously. Some usage examples are explained and evaluated in the second part of the thesis which deals with the practical applications from Chapter 6 (*e.g.,* parking sharing, ride sharing, location type and activity classification). Moreover, we provide a complete implementation which is fast, light (uses minimum of resources to the extent that can be deployed even with the resources of a mobile device) and scalable for large datasets.

2.2.2 Data science in the context of smart mobility

This section presents a review of the relevant data science methods, techniques and technologies in the context of smart mobility. First, Section 2.2.2.1 present a background which links this section to the material already presented in this chapter. Section 2.2.2.2 reviews the most popular data science modelling frameworks, followed by Section 2.2.2.3 which examines data analytics platforms and processing frameworks. Finally, Section 2.2.2.4 surveys applications from the domain of smart mobility which use AI technology. A synthesis of the material concludes the section in 2.2.2.5.

2.2.2.1 Background

As we discussed in the previous sections, large-scale data derived from mobile devices, vehicles and traffic information systems have facilitated the understanding of human mobility patterns and similarities. The sharing economy in the transportation domain experienced exponential growth in recent years and the benefits are well known [223]. Private cars, public transportation and parking places must be used in a more resource-efficient manner by ITSs. These systems present the challenge of processing massive amounts of continuously changing data emitted by mobile devices and traffic sensors.

Different sharing solutions have been implemented and their impact is extensively discussed in the literature [92, 152, 121, 157, 58]. Nevertheless, even if all the above services use similar methods and technologies to collect and process the data, each of them has different requirements, challenges and optimisation objectives. For example, while for joint sharing solutions (*e.g.,* carpooling) the objective is to reduce the number of cars used by grouping more people into fewer cars, in the concurrent sharing solutions (*e.g.,* car sharing, shared parking), the objective is to increase the usage of available resources.

Data science and ICT methods and technologies promise great potential for solving these challenges in the transportation domain. A single domain cannot completely solve the problem in a scalable manner and implement it as a service in the real world. Hence the need for an interdisciplinary solution.

Given these realities, data processing in smart mobility becomes one of the most impor-

tant components in a response to the challenge of performing complex processing tasks and rapid scaling in short time. The implementation of Collaborative Mobility (CM) as a service therefore calls for the design of scalable and specialised analytic solutions able to cope with the transportation domain specific complexity and the diversity of underlying data models. Efficient frameworks and platforms, including predictive and prescriptive analytics, complex simulations, ML techniques and AI decision, will need to be merged into a single complex intelligent system.

In summary, the following sections will review the methods, technologies, modelling frameworks, data analytics platform and specialised processing frameworks found in the data science domain which can perform the following operations simultaneously:

1. **Analyse frequently changing data**. Data in motion generated on a very large scale by a wide range of sources, especially sensors embedded in nomadic devices (smartphones, smartwatches), must be processed in near real time in order to deliver a solution at a magnitude of seconds.

2. **Explore many different hypothetical actions**. The future smart mobility recommendation systems will act like a travel advisory which explores all the possible shared mobility alternatives to provide the best solutions for each user.

3. **Reason over distributed data in motion**. The RSs must take in to account context, user preferences and reason in order to provide advice/recommendations regarding the best transportation resources to be used, and to find compatible users able to efficiently share the available transportation resources (*i.e.,* vehicles, parking spaces).

4. **Combine domain knowledge and machine learning at the same time**. On the one hand, the transportation engineering domain knowledge is necessary at the design time to provide all the variables, functions, limitations and to ensure that the entire process flow provides good results. On the other hand, all the relevant information must be extracted autonomously from the raw data using specific methods from the data science domain, coupled with ML techniques and algorithms which can learn and reason about user travel habits and mobility patterns.

In order to perform the above mentioned operations, a complete system is required, composed of a modelling framework, a data analytics platform and specialised processing frameworks which meet the requirements of the CMS case study. The following section presents a review of the modelling frameworks found in the data science domain.

2.2.2.2 Modelling frameworks

The modelling represents an abstraction of a subject one wants to reason about and is a fundamental process in software engineering. The subject can be defined as an entity from the real world which receives a certain purpose.

The next generation of smart mobility recommendation systems must be able to adapt to people's needs and preferences, as opposed to the current mobility scenarios where people must adapt to the services and variables proposed by the system. In order to achieve this, context awareness is a first step needed to trigger system adaptations. Similar examples can be found in the case of ambient intelligent applications which require augmenting the environment with sensing, computing, communicating, and reasoning capabilities [153] or the live analytics requirements for cyber-physical systems in the case of smart grids [106]. This is also the case of ITSs, where their distributed and heterogeneous nature lead to an important complexity when it comes to representing their current context, in order to later reason about it. Therefore, various methods must be implemented to structure and organize the data, in order to process it for a later reasoning.

We will employ contributions made by specific modelling techniques from Model-driven engineering (MDE), which we noted in the second part of this thesis. MDE enables the specification of formal models that express designs in terms of specific application domains (*e.g.,* transportation domain). This technology provides the necessary features to address an increased complexity by combining the **domain-specific modelling languages** (which enable the use of dedicated languages and tools dedicated for each specific domain) and **transformation engines and generators** (*"that analyse certain aspects of models and then synthesise various types of artefacts, such as source code, simulation inputs, Extensible markup language (XML) deployment descriptions, or alternative model representations"* [182]).

Another important requirement for a model able to support a dynamic recommendation system is the enabling and support of self-adaptive systems, which may be used both at design and in runtime. Over the past years, a new paradigm called **models@run.time** [158]has emerged, as a response to the need to equip self-adaptive systems with a model continuously connected to their current state. The model@run.time it is defined in the literature as: *"a causally connected self-representation of the associated system that emphasizes the structure, behavior or goals of the system from a problem space perspective"* [39]. As the models provide a semantic valid way to define the context of the system which uses them to reason about the tasks and processes to be performed, the main advantage of the models@run.time is that any change of the abstraction implies an automatic adaptation of the system at runtime.

One of the most popular frameworks considered as the de facto standard for modelling is the Eclipse Modeling Framework (EMF) [193]. Thanks to its flexibility, EMF is able to define models in multiple ways *e.g.,* annotated Java, MDE, textual modelling. EMF does, however, have a drawback: it has been developed primarily for systems which apply modelling more at design time and less for those using runtime models. Hence, there are significant limitations when the model is used in the context of models@run.time [89].

Another modelling framework developed specifically to meet the models@run.time requirements is GreyCat [110], formerly known as Kevoree Modeling Framework (KMF) [90]. Other features also recommend the GreyCat as optimal for the CMSs as follows [88], [89]:

1. **Reduced memory requirements**. This is a mandatory requirement in the case of CMSs where nomadic devices (smartphones, smartwatches) with limited resources are used by the users of shared mobility applications. Therefore, the model should be compatible with deployment on these remote devices without negatively affecting the user experience due to intensive resource depletion (*e.g.,* using too much memory and power, leading to drained batteries).

2. **Thread distribution**. In complex smart mobility CMSs, the modeling layer should enable concurrent access of multiple applications (*e.g.,* carpooling, car sharing, parking sharing) in order to offer multiple solutions and recommendation to shared mobility users. Moreover, as these systems can have a large scale of users, the processing should be implemented on multi-core/thread nodes, both at the server side and at the end user's device.

3. **Efficient model cloning**. Because of the high data volume and processing demand, the future smart mobility recommendation systems should be able to clone and use the users' nomadic devices to perform fully independent local reason tasks. Consequently, the model should allow efficient cloning and synchronization in the use of remote devices, not only as sensing systems which collect, send/receive data, but also as processing and local reasoning points (*e.g.,* clean/reconstruct/query the data collected).

4. **Lazy loading ability**. The model should be continuously available, but a good strategy for resources optimisation and performance enhancement is to be loaded only when needed (*i.e.,* only the useful parts of the model are loaded and only when needed at the nodes where tasks are processed).

We choose to use the GreyCat model throughout the contributions presented in the second part of the dissertation rather than EMF for several additional reasons. First, GreyCat is a more lightweight framework specifically designed to be deployed at the level of nomadic devices with limited resources. Secondly, GreyCat is more suitable to process the large datasets generated by the sensing systems which continuously collect mobility data (*e.g.,* GPS, motion data, connectivity modules *i.e.,* WiFi, Bluetooth). Thirdly, GreyCat allows the usage of different storage technologies, *e.g.,* key-value stores. This is important for the collaborative mobility ITSs, since the large amount of data generated at the users' device level cannot be stored completely in memory but must be sent and loaded from external servers. Fourthly, GreyCat also offers a complete environment as a processing framework and data analytics platform, which allows the processing the big data in motion using graph processing methods. Moreover, GreyCat provides near real time analytics powered by specialised ML techniques and offers the possibility of performing parallel simulations with *what-if* scenarios for prescriptive analytics.

2.2.2.3 Data analytics platforms and processing frameworks

The challenges of big data analytics platforms in the context of IoT and sensing systems have been extensively identified as an open issue by the industry and academia [117],

[176], [192]. Similar challenges can be found also in the transportation engineering domain and in particular the case study of ITSs which manage the CMSs, as follows:

1. **Real-time analytics**. In the case of CMS, users access the shared mobility services via smartphones applications and perform requests for personalised mobility services, usually on demand, without any advanced planning (*e.g.,* the case of Peer to Peer (P2P) services like Uber). Therefore, the future recommendation systems must find a suitable solution to match multiple compatible users with similar travel behaviour, in order to create a compatible group able to use a shared mobility solution. This requires near real-time analytics to explore the travel habits of all the users which are part of the shared mobility system and to find a solution at a magnitude of a couple of seconds.
2. **Complexity generates high performance requirements**. The paradigm of smart mobility systems has a high complexity, involving the fusion of interdisciplinary methods and technologies. This generates high performance requirements for a number of interconnected distributed systems which must be perfectly synchronised.
3. **Continuously knowledge discovery**. An important challenge resides in the transformation of continuously collected raw data generated by the sensing systems into usable knowledge, using the latest ML methods, algorithms and technologies.

In recent decades we have observed rapid advances in the development of data management systems. These systems were forced to evolve and to push the limits of data analytics by handling ever larger amounts of data. In the rest of this section we present a review of the data analytics platforms and processing frameworks.

Since 1990's, a new type of database processing called Online Analytical Processing (OLAP) [67] addressed the lack of traditional database processing. OLAP was implemented in many commercial database systems. This was one of the attempts to analyse data in multiple dimensions. The dimensions were represented by perspectives in this work, whereas in this dissertation we refer to dimensions as different types of data *i.e.,* time, spatial coordinates, specific attributes of the involved entities. The techniques offered by OLAP were also unsuitable for the case of ITSs, which have complex motion data changing frequently over time.

In a more recent work, Cohen *et al.,* [68] presented a Magnetic, agile, and deep (MAD) data analysis inspired from traditional business intelligence. The work described the best practices for big data analytics called "MAD Skills", in an attempt to better address the large scale of data which analytics platforms must face. However, instead of providing new methods for data analytic platforms, the work presents recommendations instead about the use of already existing technologies, suitable for centralised analytics rather than for a decentralised paradigm as required by the CMSs.

Another popular approach used for data analytic platforms is the Apache Hadoop framework [2]. Featuring distributed storage and processing for big datasets, the framework was built to rapid scale from a low number of computers to big clusters composed

of multiple machines. Using the map-reduce [74] approach, this was successfully implemented across many big companies *e.g.,* Facebook, LinkedIn and Yahoo!. However, the Hadoop stack was essentially designed as a batch processing system. This means that even if the framework is designed to process massive datasets, does not provide native support for all of the four challenges presented in Section 2.2.2.1.

A notable framework among the big data analytic frameworks which become an Apache top project is the Spark stack [4]. Spark's main contribution is in the performance of in-memory computations on big clusters. It is a faster alternative able to replace Hadoop's computation performance. Because Spark provides only the computing core and data structures, it requires additional interconnected components, *e.g.,* Spark Streaming, Spark SQL, GraphX, MLlib, Velox, in order to be used as a complete data analytics framework. Although it is possible to efficiently process mini-batches through components like Spark Streaming, the core of Spark is pipeline-based and is therefore unsuitable for systems requiring immediate reactions *e.g.,* IoT and ITSs.

Other types of big data analytics frameworks are the **stream processing frameworks** designed to address specifically the challenge of near real-time data analytics and are suitable for ITSs *e.g.,* Storm [204], Heron [128], Flink [1], S4 (Simple Scalable Streaming System) [162], Samza [12], Stream Processing Core (SPC) [25]. Hartmann [106] provides an extended review of these frameworks. Even if their design is interesting for the case of ITSs, they lack the features needed for analysing data of complex smart mobility recommendation systems. First, the model's simplistic representation complicates the representation of complex relationships between all entities from the CMS, as well as the exploration of different hypothetical actions. Second, the design does not support natively complex data representation, like graphs, which makes also inhibits the integration of ML algorithms and techniques.

Recently, **graph processing frameworks** gained a lot of interest because of their capabilities for representing various data and complex relationships [185]. Graphs are for representing the context of general cyber-physical system and in similar way of ITSs and have influenced the design of the GreyCat (which is the framework implemented for the contributions from the second part of this thesis). Hartmann [106] provides a detailed review of the most popular frameworks from this category *e.g.,* Pregel [142], Giraph [7], GraphLab [141], PowerGraph [98], GraphChi [129], GRACE [214], Trinity [185], GraphCEP [147]. The results of the comparison between all the above graph processing frameworks shows that none of them are fully compatible with the context of CMSs. First, some require the graph to be completely stored in memory in the case of nomadic devices. In contrast, GreyCat offers the advantage of storing the graphs on secondary storage (*i.e.,* nomadic devices' storage) which is both much cheaper and usually available. Secondly, the graph abstraction does not support the concept of analysing many different hypothetical actions required by the recommendation systems when searching for shared mobility combinations in parallel simulations. Thirdly, none of the reviewed graphs allow the modelling of data collected, of domain knowledge and ML in the same model.

Instead, the GreyCat [110] framework addresses all the above challenges and meets the requirements of the CMS case study. The framework is described in detail in Chapter 4 and practical applications using this framework are presented in Chapter 5 and 6.

2.2.2.4 Artificial Intelligence applications in Smart Mobility

The accelerated adoption of IoT facilitates the implementation of smart mobility analytics, thus providing a wealth of knowledge which may be exploited to improve organizational decision making. The availability of large-scale data from mo- bile devices, vehicles and traffic information systems have facilitated the understanding of human mobility patterns and similarities. People, private cars, public transportation and parking places must be synchronised and informed by intelligent systems that process massive amounts of continuously changing data from mobile devices and traffic sensors. The problem becomes even more complicated when all the above-mentioned solutions must be integrated into a single system which must manage everything at different scales and almost in real time.

The above-mentioned problems from the smart mobility domain are solved in this thesis using a combination of methodologies and techniques from the ICT and data science domain. In this sense, data processing in smart mobility presents the challenge of performance optimisation while executing complex reasoning processes to support exponential scaling in short time *e.g.,* when finding compatible users to share a ride on demand. The massive amount of data used in the transportation economy has called for the next generation of intelligent travel assistants, powered by AI techniques.

Although the AI keyword is abundantly used in industry and business today, an exact definition of AI is surprisingly elusive. A simplistic explanation would describe it as the intelligence demonstrated by machines, in contrast to the natural intelligence displayed by humans. Kaplan and Haenlein provide a more precise definition: *"a system's ability to correctly interpret external data, to learn from such data, and to use those learnings to achieve specific goals and tasks through flexible adaptation"* [120]. In the same work, the authors defines three stages or generations of AI, which adapted to the smart mobility domain can be explained as follows:

1. **Artificial Narrow Intelligence (ANI)** - first generation of AI that can only apply the intelligence to specific designed tasks *e.g.,* self-driving car technology, route planners - already considered near ubiquitous.

2. **Artificial General Intelligence (AGI)** - the second future generation of AI is expected to autonomously identify and solve even problems for which they were never designed. An example (linked with the topic of this dissertation) is the future generation of intelligent travel advisors/recommendation systems, which will autonomously and efficiently organise and plan humans' schedule of activities in order to solve traffic problems, reduce environmental impact and enhance the quality of human life. Such a RS should be able to detect changes in users' travel behaviour, which in turn can generate issues regarding the way transportation system is designed. Consequently, RS should be able to optimise transportation solutions and automatically adapt to the users' needs.

3. **Artificial Super Intelligence (ASI)** - might be possible to see the truly self-aware and conscious AI systems *"that, in a certain way, will make humans redundant"* [120]. Such systems could apply AI to use scientific creativity in order

to design and develop new modes of transportation, services and organise the mobility for an entire society. This is the reason for which some call ASI as the true artificial intelligence.

Over the past few years, smart mobility and sharing mobility RSs increasingly made use of AI technology in various applications. Even if this dissertation does not seek to develop a complete RS, the methods and technologies presented constitute a theoretical and practical foundation which may be used in future research towards the next generation of AI-powered RSs for sharing mobility.

Machine learning makes the AI behind the RSs possible. ML is a topic which combines the statistics domain, optimisation techniques and computer science technologies. It is *"an evolution of pattern recognition and learning theory in artificial intelligence"* [153]. Looking back in the history, ML *"was born as one branch within the major field of artificial Intelligence"* [28] with the objective to perform autonomous data-driven operations (*e.g.,* take decisions, make predictions), using models generated from previous example inputs. The mathematical models can represent any domain notions (*e.g.,* profiles, patterns, correlations) that fit previous observations and are able to extrapolate new observations [153]. A multitude of ML techniques and algorithms are used to solve various types of problems (*i.e.,* hypothesis based, density estimation, classification, clustering, anomaly detection, dimensionality reduction, recommendations), applying both different types of learning (*i.e.,* supervised, unsupervised, semi-supervised, reinforcement learning, meta-learning) and frequency of learning (*i.e.,* online learning, lazy learning, batch learning) [167], [153], [106]. While an extended review of these techniques is out of scope of this thesis, it is important to note that all these advancements in the ML made possible the implementation of AI which is embedded in the RSs.

While an extended review of these techniques is beyond the scope of this thesis, it is important to note that all these advances in the ML made possible the implementation of AI, which is embedded in the RSs.

Recommendation Systems are utilized in a variety of domains including e-commerce (*e.g.,* online shops for products, services, applications), entertainment (*e.g.,* platforms for movies, music, online dating, social networks), research (*e.g.,* research articles, books), search engines, general services (*e.g.,* financial, insurance, restaurants, collaboration) [16], [177], [85]. The general objective of RSs is to suggest/recommend specific items/alternatives/solutions of potential interest for a user [116]. The recommendations relate to a decision-making process, such as what products to purchase, what movie to watch, but may also link to what transportation solution to choose in different contexts. Therefore, RSs are primarily directed toward users who do not have sufficient experience or available time to evaluate the seemingly limitless alternatives which a specific service may offer [179].

Back in 2005, a RSs survey [16] showed that this topic had become an important research area since the mid-1990s, with the introduction of the first collaborative filtering methods [178]. First, the RSs were used mainly for websites to help users dealing with the overload of information and to provide personalized recommendations for products and services. In our own day RSs applications extend to almost every business domain,

using the latest trends and technologies. For example, Felfernig *et al.,* [85] provides an overview of the existing work related to the applications of recommendation technologies in IoT scenarios.

Recommendation systems for Smart Cities. In the smart cities paradigm, the RSs make use of contextualised service delivery methods to provide smart services for solving urban problems [220]. Contextualisation may be defined as the process of identifying relevant data to an entity (*e.g.,* a person or a city) based on the extracted contextual information from various sources [221]. Intelligent RSs then perform an Observation, Orientation, Decision, and Action (OODA) loop [220]. This is also the case of the RSs implemented in the ITSs domain, which involve *Observation* - collecting various information from all the entities using multiple sensing sources, *Orientation* - contextualisation of the extracted information to solve specific needs, *Decision* - autonomously reason and take appropriate decisions, and *Action* - perform specific actions and send recommendations based on the decisions taken in the previous steps.

Location based recommendation systems (LBRSs) represent the first RSs which made use of location-based data, trajectories and geo-tagged media to recommend locations, routes, activities of potential interest to a user. Rehman *et al.,* [177] presents a systematic review of the literature related to the LBRSs and a qualitative comparison of the techniques and algorithms used in the literature (*e.g.,* content-based, collaborative filtering based or hybrid methods). By using the geolocation data and the users' profile information from the social networking services, it was possible to connect the gap between the online services and the physical world. Therefore, the authors classify the LBRSs both in *sequential location recommendation* (which makes use of users' GPS trajectories and geo-tagged social media information) and also *stand-alone location recommendation* (based on the location history, users' trajectories and users' profile).

Travel recommendation systems (TRSs). The first RSs which embedded AI techniques to provide personalized travel recommendations were used in the tourism domain. The tourism/travel recommender system employs AI techniques to generate personalized recommendations regarding touristic information and services. Ravi *et al.,* [175] presents a review of RSs which make use of social network data and geolocation data by considering usage of various recommendation algorithms, functionalities of systems, different types of interfaces, filtering techniques, and artificial intelligence techniques. Cha *et al.,* [57] proposed for the first time real-time RSs that uses the geolocation data from smartphones in order to recommend touristic Points of Interests (POIs) and services in order to create a new type of touristic experiences through the application of user contexts.

Smart Mobility. Increased interest has been directed over the past few years toward context-aware RSs, which use external contextual sources to provide personalised smart mobility recommendations [26], [140], [139], [209]. The main idea is to use external contextual data in addition to the location attributes in order to offer personalised recommendations for potential POIs. Amoretti *et al.,* [26] use a combination of user profiling and context-based data filtering techniques in a smart mobility application that recommends POIs to end users. Logesh *et al.,* [140] propose an user travel behavior based recommendation approach, using basic user profiling techniques from GPS data

and social networking services [139]. This approach is used in a later work to implement an application which can predict personalized list of travel locations by generating a heat map of already visited POIs [209].

The RSs were gradually introduced also in other topics from the smart mobility domain. Thus, the RS was used as environmental travel assistant [183], bike sharing journey advisor [222] or for comfortable public transport recommendations [208]. The PEACOX project [183] proposed a persuasive advisor for CO2-reducing cross-modal trip planning. The aim is to make travellers more aware of the environmental consequences associated with their transport behaviour. The advisor was implemented through a smartphone application which had the objective both to highlight emissions arising from users' existing behaviour and to suggest more sustainable travel alternatives. At the end of the project, however, the research showed that while the provision of emissions may increase awareness of environmental problems, it alone cannot change the users' travel behaviour [44]. Therefore, other projects use the RSs as a mean to incentivise travellers to use existing smart mobility alternatives, providing advice and important information which can increase the rate of usage. Cityride application [222] proposes a personal journey advisor for helping people to navigate the city using the available bike-sharing system. The objective is to minimize the overall ravel time and to maximize the probability of finding available bikes at the stations. The ComfRide application [208] proposes a smartphone based system for comfortable public transport RS, which provides recommendations regarding the most comfortable routes according to users' preference and travel time constraints. The research is supported by existing studies which reveal that different personalized and context dependent factors influence passenger comfort during public transport and consequently has an influence on the rate of usage [91], [43].

Sharing mobility. Similarly, RSs were implemented in the sharing mobility domain for various shared mobility services such as ride sharing, carpooling, parking and multimodal mobility applications, which will be reviewed in the remaining of this section.

- **Ride sharing**. In ride sharing platforms, RSs are used to find compatible users and recommend sharing the ride. Junior *et al.,* [118] proposed a RS to match groups of users with similar preferences and recommend them to share the transportation resources. The RS creates the user profiles using data collected from a questionnaire and information extracted from online social networks. Then the users are matched based on profiling classification techniques, performed through different ML algorithms. Finally, recommendations are used to suggest users with similar hobbies and preference for ridesharing, which improve users' quality of experience in ridesharing services. RSs are used also to provide recommendations for combinations of taxi services and carpooling [226], [227]. Wang *et al.,* [216] proposes a framework which employs a supervised learning model for discovering potential road clusters, which is incorporated into a recommender system for taxi drivers to seek passengers.
- **Carpooling**. In carpooling services, the RSs are used for different objectives. For example, in order to optimise the carpooling routes, the RSs are used to propose optimal carpool driving route options that users may choose from [133]. Recommendations for optimal carpool routes are made by grouping users who share

common trajectories along their trip. Other RSs recommend either vacant/semi-occupied taxicabs in a similar direction for a carpooling service with the minimum detour, without assuming any knowledge of destinations of passengers already in taxicabs [225]. While most of the available approaches and carpooling RSs are route oriented, ComeWithMe application [73] uses the destination as a type of activity instead of a specific location. This novel matching method aims to increase the carpooling rides if passengers accept to perform the same activity (*e.g.,* shopping, eating) in a different location. The authors suggest that this activity-oriented carpooling hugely increases the number of rides, which is not a common phenomenon in traditional carpooling services.

- **Parking recommendation**. Parking space management is a constant issue in big cities [168]. Yavari *et al.,* proposed an approach to contextualise IoT data which is used by a smart parking space recommender application. The RS considers not only the parking information (*e.g.,* available parking spaces) but also takes into account each driver's context and preferences (*e.g.,* type of proffered parking, driving experience, the car's location, the vehicle's properties). Thus, their approach provides a unified solution for extracting knowledge from data obtained from various sensors (*i.e.,* cars' sensors, nomadic devices), and advise each driver accordingly. In order to address overcrowding of parking spaces, Martino and Rossi [77] propose a car-based multimodality RS, which recommend that users leaves their car in a Park-and-Ride infrastructures and reach their destination by public transportation.

2.2.2.5 Synthesis

Whereas much interesting work has been done in the domain of RSs and smart mobility through specialised AI powered applications, existing solutions do not meet the requirements for the next generation of RSs within the AGI paradigm, which will be discussed in the remainder of this section. This thesis provides new methods, technologies and practical applications as solutions and advances in this direction. In the remainder of this section we present a summary of the open issues extracted from the literature in the above sections and their respective contributions.

As we observed in the previous section, the TRSs provide potential POI, activities and services as suggestions according to users' preferences, however the system needs a lot of information to be manually provided by the users and the travel planning needs to be build manually. LBRSs make use of additional data (*e.g.,* trajectories, social network services) to automatically extract user information, however every classical approach (*e.g.,* content-based, collaborative, hybrid) suffers in providing personalised suggestions and recommendations for each user. In order to address this issue, it is important to build the profile of each individual user. We address these issues in Chapter 5, where a new profiling method is proposed for building an individualized user profile. In this way, the RS can offer personalised recommendations regarding sharing mobility services which best match each user's travel behaviour and preferences.

Context-aware RSs use external contextual data to extract additional information in order to build a basic user profile. However, the profiling should provide much more

knowledge (*e.g.,* mobility patterns, travel behaviour) than simple user preferences. Moreover, the profiling should be built dynamically in order to capture travel behaviour's variations and users' mobility patterns extracted from the history of visited locations. Thus, the employed profiling methods should be able to extract important knowledge that can be used by specialised learning techniques to automatically capture the user interactions, travel behaviour and mobility patterns. In order to ensure a good user experience, the entire process of data collection and extraction of the knowledge information should be done with nearly no interaction from the user side. This brings another common issue of the RSs, the so called *cold-start problem* [51]. The problem is that if a new user joins a RSs, the system cannot send recommendations because of lack of previous information about the user profile and preferences. Thus, special profiling techniques and reasoning processes must be employed to automatically extract the relevant user information from raw data with no user intervention. The practical application from Chapter 6 represents a possible solution to these issues. We demonstrate how it is possible automatically to extract the mobility patterns through using the profiling method presented in Chapter 5 and to extract valuable knowledge without any user input. Precisely, the proposed method can automatically classify the type of each visited location (*e.g.,* work, home, restaurant, gym) without any user intervention.

In order to handle the above presented process flow, the next generation of smart mobility RSs should be able to passively collect information from additional data sources (*e.g.,* data collected from the sensing systems of nomadic devices). Therefore, specialised modelling frameworks and data analytics platforms should be able to meet the requirements of analysing complex and frequently changing data from the CMS paradigm. Moreover, because there are multiple smart mobility applications which provide different sharing services, the next generation of smart mobility RSs should offer personalised recommendations from a single intelligent trip advisor. This challenge is addressed by the contribution from Chapter 4 which presents a fast and lightweight data driven modelling framework which can analyse frequently changing data in motion and can be deployed even at the level of the nomadic devices. The framework can combine domain knowledge and ML techniques at the same time and explore many different hypothetical recommendations and actions from multiple services simultaneously. Moreover, because the framework can reason over distributed data in motion, specialised AI techniques can be implemented in order to increase the usage of sharing services but also to increase the user quality of service by offering personalised recommendations, according to the context of each individual user.

Part II

Contributions

3

A data-driven indicator for collaborative mobility

The main objective of this chapter is to derive an indicator for revealing potential collaborative mobility options between individuals using automated data collected from smartphone sensors. This indicator can be used by ITS in order to provide recommendations for all combinations of collaborative mobility sharing systems (e.g., carpooling, parking sharing, car sharing). The proposed indicator must take individual preferences into consideration, schedule the entire chain of activities, and be sensitive to dynamic changes in different scenarios.
This chapter is based on work that has been published in the following paper:

- *Usage of Smartphone Data to Derive an Indicator for Collaborative Mobility between Individuals*
  B Toader, F Sprumont, S Faye, M Popescu, F Viti
  ISPRS International Journal of Geo-Information 6 (3), 62

Contents

3.1 Introduction

The main objective of this chapter is to extract the human mobility patterns from GPS traces in order to derive an indicator for enhancing Collaborative Mobility (CM) between individuals. The first step - extracting activity duration and location - is done using state-of-the-art automated recognition tools. Sensors data are used to reconstruct an individual's activity location and duration across time. For constructing the indicator, in a second step, we defined different variables and methods for specific case studies. Smartphone sensor data was collected from a limited number of individuals for one week. This data was used to evaluate the proposed indicator. Based on the value of the indicator, we analysed the potential for identifying CM among groups of users such as sharing travelling resources (*e.g.,* carpooling, ridesharing, parking sharing) and time (rescheduling and reordering activities).

The proposed collaborative mobility indicator is defined in Section 3.2, followed by the experimentation and results in Section 3.3. Finally we outline the discussion and perspectives in Section 3.4 followed by the conclusions of the study in Section 3.5.

3.2 Methodology

In this chapter we propose an indicator to assess the collaborative mobility between individuals, providing a single score defined hereafter as an index/indicator. The problem consists of generating a score of compatibility for carpooling, car sharing and parking sharing between two or more individuals willing to share the resources. The value of the index indicates if there is compatibility between a group of users for using sharing services, as well as the level of compatibility. This indicator has the following features and objectives:

System optimisation. At the system level the objectives are to provide a compatibility score between individuals for using sharing services in order to reduce carbon emissions, traffic congestion on the roads, and the need for parking spaces, subject to:

a) Minimising the sum of total costs (reduce the cost of shareable resources);

b) Maximising the number of users using the sharing services simultaneously (grouping more people in fewer cars).

Individual optimisation. At the individual level the objective is to minimise the cost of each participant in the collaborative mobility scheme.

Single metric. The index provides a compatibility score combining carpooling, parking sharing and car sharing in a single indicator.

Sensitive. The value of the index reflects any change in the schedule of individuals, the number of users that share the resources, the travel path chosen or the number of shared resources.

Flexible. The indicator can be used to asses the compatibility between individuals for long-term sharing services but also for instant sharing services.

3.2.1 General conceptual model of collaborative mobility indicator

The indicator should evaluate not only the total cost of the system, but also the individual cost of every user, considering different trade-off strategies in some cases. The mathematical problem developed here is challenging in many respects. For instance, implementing the right trade-off rules is far from trivial, given that travellers' characteristics may change over time [210]. It is a multi-objective optimisation problem, involving more than one objective function to be optimized simultaneously.

All variables and cost values used in this study are explained in the following list. The default cost values used in the examples and case studies presented in this study are similar to the values found in specialised literature (see *e.g.,* [210]). Also because the case studies are done in Luxembourg, the parking fee cost is based on the public local rates.

I	Collaborative mobility index value	
C	Individual cost without using sharing services	
C_S	Individual cost when using sharing services	
C_r	Cost ratio between C_S and C	
C_{cp}	Carpooling cost	
C_{ps}	Parking sharing cost	
C_{cs}	Carsharing cost	
F	Distance based costs	0,15 €/km
Tt	Travel time cost	0,17 €/min
Tr	Rescheduling time cost	0,17 €/min
P	Parking cost	17,42 €/day
Tp	Parking time	
O	Other trip related costs *e.g.,* toll, vignette etc.	
$\alpha, \beta, \gamma, \delta, \epsilon$	Weight variable for a specific cost	
n	Total number of users	

Shareable costs are defined as the costs that can be shared between the individuals when using sharing services: F, P, O

Non-shareable costs are the costs that cannot be shared between the individuals when using sharing services: Tt, Tr

Distance in Time (DT) between two activities a and b, is defined as the time difference between the starting (t_s) time of activities:

$$DT(a,b) = t_s(a) - t_s(b) \tag{3.1}$$

Distance in Space (DS) between activities, is represented by the shortest path distance in road network between activities' locations.

$$DS(l_i, l_j) = Distance([x_i, y_i](l_i), [x_j, y_j](l_j)) \tag{3.2}$$

where l_i, l_j representing the location index and x, y are the latitude and longitude of each location.

The general mathematical formulations and constraints for the proposed indicator are defined using the generalised cost as follows:

$$I = \frac{\sum_{i=1}^{n} C_S(i)}{\sum_{i=1}^{n} C(i)} \tag{3.3}$$

where

$$C_S(i) = \alpha_i(C_{cp}(i)) + \beta_i(C_{ps}(i)) + \gamma_i(C_{cs}(i)) \tag{3.4}$$

$$C(i) = \alpha_j(C_{(i,j)}) + \beta_j(Tt_{(i,j)}) + \gamma_j(Tr(i)) + \delta_j(P) + \epsilon_j(O) \tag{3.5}$$

subject to

$$I < 1 \tag{3.6}$$

$$C_r(i) < 1 \tag{3.7}$$

where

$$C_r(i) = \frac{C_S(i)}{C(i)} \tag{3.8}$$

The compatibility between a group of users exists when both (3.6) and (3.7) are met simultaneously. This ensures that the sharing services are efficient both at the individual level but also at the system level. Equation (3.6) ensures that the sum of individual costs when persons are using sharing services is less than the sum of individual costs when they not. Equation (3.7) ensures that the individual cost for each single person when is using sharing services does not exceed the cost when the individual is not using the services. The index value also denotes the level of the compatibility between individuals using the sharing services. The lower the value below 1, the lower the cost resulting in a higher compatibility.

Therefore, the proposed indicator is sensitive to any individual cost as following:

- **Shareable costs** costs are divided by the number of individuals who share a resource for a specific trip segment/time period.
- **Travel time** is the individual total travel time, which increases in the case of detours and is not shareable.
- **Reschedule time** represents the time cost for each individual user who must reschedule the regular activities in order to synchronise with other travellers.
- **Weight cost variables** represents the weight of the cost for each individual. This offer a realistic cost calculation because *e.g.,* a user might not care about saving shareable costs but has no flexibility for rescheduling some rigid activities (*e.g.,* the working schedule). In this case the weight of shareable costs may be zero.

In the next subsections we present in detail the conceptual model of each of the sharing services i.e. carpooling, parking sharing and car sharing.

3.2.2 Collaborative mobility indicator for assessing carpooling

Carpooling is the sharing of car journeys so that more than one person travels in a car. When travellers are carpooling, they are sharing the cost of the fuel. In our conceptual model, the costs are shared by the segment of the trip where users are sharing the car. The cost is divided by the number of people in the car for that segment.

In this case, the collaborative mobility indicator applied for carpooling is defined as in (3.3), where:

$$C_S(i) = C_{cp}(i) \tag{3.9}$$

$$C_{cp}(i) = \alpha_i(\frac{F_{(i,j)} + O_{(i,j)}}{n}) + \beta_i(Tt_{(i,j)}) + \gamma_i(Tr(i)) \tag{3.10}$$

$$C(i) = \alpha_j(F_{(i,j)}) + \beta_j(Tt_{(i,j)}) + \gamma_j(Tr(i)) + \delta_j(O) \tag{3.11}$$

subject to (3.6) and (3.7).

The value of the index applied to carpooling represents the compatibility between two or more travellers who are carpooling together. The lower the value of the index is, the lower the cost of carpooling, resulting a higher compatibility between users, in accordance with (3.6) and (3.7).

3.2.3 Collaborative mobility indicator for assessing parking sharing

Using the same indicator model it is possible to analyse the compatibility index between individuals for parking sharing services. The conceptual model for parking sharing is defined as in (3.3), where:

$$C_S(i) = C_{ps}(i) \tag{3.12}$$

$$C_{ps}(i) = \alpha_i(\frac{P}{n}) \tag{3.13}$$

$$C(i) = \alpha_j(P) \tag{3.14}$$

subject to (3.6) and (3.7).

In order to demonstrate different applications of parking sharing index, we will consider the following case studies and define the model and conditions for each of them. In the first case study we define the conditions for parking sharing between a group of car-dependent users who are willing to share the parking place in order to reduce the cost. The second case study deals with the usage of parking sharing in combination with carpooling.

3.2.3.1 Parking sharing compatibility index for car-dependent users

Compatibility for parking sharing between a group of car-dependent users is defined as in Section (3.2.3) with the condition that the intervals when they are using the parking place to not overlap. Figure 3.1 presents an example of how the parking is used by two persons during one day. Time of the day in which P_1 and P_2 use the parking. In this example, a_i, b_i are the arrival time in the parking, and $a_{(i+n)}$,$b_{(i+n)}$ are the departure time from the parking.

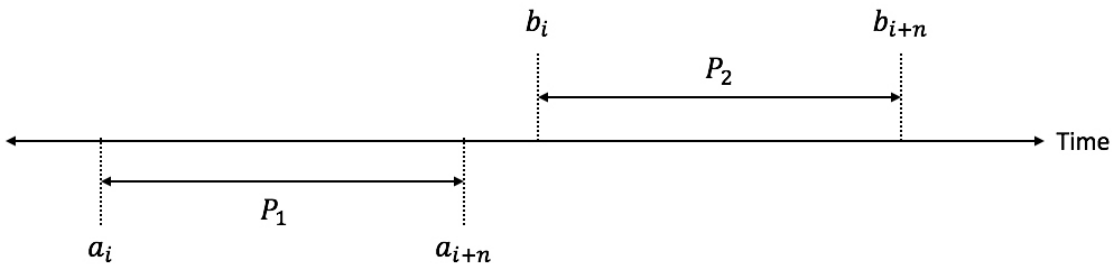


Figure 3.1: Example of parking sharing usage for two users

The intervals when each of them are using the parking place are:

$$P_1 = [a_i,\, a_{(i+n)}] \qquad P_2 = [b_i,\, b_{(i+n)}]$$

In this case, the condition for parking sharing is:

$$a_{i+n} < b_i \tag{3.15}$$

Constraint (3.15) ensures that there will be no overlapping between parking usage periods.

3.2.3.2 Parking sharing index in combination with carpooling

The conceptual model for assessing the parking sharing between individuals in combination with carpooling is defined as in (3.3), where

$$C_S(i) = \alpha_i(\frac{F_{(i,j)} + O_{(i,j)} + P}{n}) + \beta_i(Tt_{(i,j)}) + \gamma_i(Tr(i)) \tag{3.16}$$

$$C(i) = \alpha_j(F_{(i,j)} + O_{(i,j)} + P) + \beta_j(Tt_{(i,j)}) + \gamma_j(Tr(i)) \tag{3.17}$$

subject to (3.6) and (3.7).

3.2.4 Collaborative mobility indicator for assessing carsharing

Carsharing is a resource that has a similar conceptual model and constraints as parking sharing. A vehicle booked by an individual cannot be used simultaneously by other users, as (3.15). There is only one exception: when a group of users needs to use simultaneously the carsharing system and the DS and DT for the origin and destination are zero, then they can instantly carpool using the carsharing system.

The most efficient way of using the carsharing system in a collaborative mobility scheme is in combination with other sharing services, in our case carpooling and parking sharing. In this case, the indicator must be evaluated over a chain of activities, for a longer period *e.g.,* the chain of activities and the related trips between activities over a full day. The conceptual model for assessing the carsharing for a chain of trips and in combination with carpooling is actually defined as in (3.3), (3.4) and (3.5), subject to (3.6) and (3.7).

Basically the index is evaluating the cost of using carpooling for part of the trip chain, of sharing the parking cost and using the carsharing system for the other part of the chain trips, versus the cost of using the private car for the entire trip chain *e.g.,* over a full day.

All the conceptual models defined above will be tested and evaluated with real data, in different scenarios performing various experiments, in the following section.

3.3 Experimentation and results

3.3.1 Data collection and processing

3.3.1.1 Architecture of the Sensing System

The studies conducted as part of this study are based on SWIPE, an open-source platform for sensing, recording and processing human dynamics using smart devices [83] [84]. Figure 3.2 gives an overview of the SWIPE architecture, which consists of two main parts: a local sensing system composed of one or several smartphones, and an online analytics platform where the data from multiple users and devices is aggregated and analysed. Interested readers may refer to [83] to get more information about the data collection parameters (*e.g.,* sampling and recording rates) and existing energy optimization strategies.

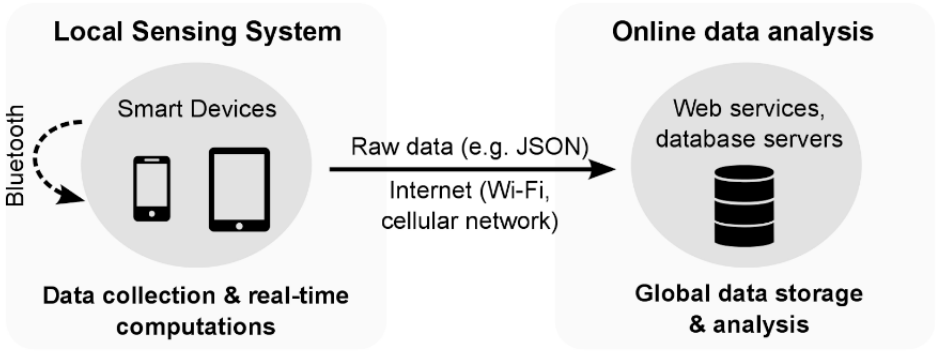


Figure 3.2: Overview of our Sensing System.

3.3.1.2 Data collection

Our study was tested inside the University of Luxembourg, where the data was collected from five employees of the university, both students and teachers, during approximately one week. The participants received a smartwatch and a mobile application that had been installed on their Android smartphone. The application was collecting the data from smartphone and smartwatch sensors in the background, sending the data to a server when the smartphone was connected to the internet.

Even if data was collected from all the sensors, in this study we use only the data regarding the location (GPS latitude and longitude) and the WiFi connections.

Table 3.1 gives a summary of the data collection.

Table 3.1: Data collection summary.

User ID	Covered period	No. of original data points	Frequency of points(seconds)	Frequency of points(meters)	Total distance covered(km)
1	181h,25min	6083	149	204	892.93
2	175h,40min	6588	172	70	256.04
3	191h,57min	6263	117	106	622.18
4	167h,57min	2300	280	211	457
5	149h,48min	2362	492	86	95

As may be observed, each user reported a different number of data points collected and the difference between some users can be quite significant. This is due to the fact that it was not possible to collect geolocalised data all the time. In some cases we observed that users turned off the GPS or switched the smartphone to the flight mode *e.g.,* during the night, when they travel or to save the battery power. In other cases the GPS was turned off accidentally or by the operating system in order to save energy. Once switched off, users must re-activate the GPS in order to relaunch collection of data points. This explains the difference between the different numbers of location data points between *e.g.,* user P2 and user P5.

From the data collection summary we can also observe that users reported a different total distance covered. This is related to the different travel behaviour of different type of users. In our case, students and professors participated in the data collection. From their travel behaviour different hypotheses may be extracted *e.g.,* user P1 reported a short trip outside the country during the week and other trips between campuses. This behaviour is more appropriate to a professor than to a student, with more flexibility and travelling outside the country during the weekdays.

3.3.1.3 Data processing

Data collection methods usually contain errors, resulting in out-of-range values (*e.g.,* vehicle speed: 1500 km/h) or missing values (*e.g.,* data collected without any GPS points). Using this type of data can produce errors if the system is not designed to filter and process this type of data [171].

In order to clean the data we calculated the speed between each consecutive point in order to remove any "supersonic jump" with unusually high speed. Then all the remaining points without GPS location were removed because those points were not only useless for the algorithm employed for the extraction of activities, but could also produce errors because of gaps in the data without any information.

3.3.1.4 Data mining for reconstruction of missing locations

Location data may be collected with errors due to several reasons. First, the GPS may be turned off accidentally by the user during data collection. In such cases, the

system collects data from other sensors but not the location data. Second, due to the urban canyon effect and the smartphone position during travelling or sitting location, the GPS is unable to detect the location or the geospatial coordinates recorded may be erroneous [70].

For the reconstruction of the missing locations, a data mining process has been applied using existing GPS locations and WiFi information in order to reconstruct individual's daily activity locations. Approximately one third of the GPS data has been recovered.

3.3.1.5 Extracting activity duration and location

In the contribution from Chpater 3 we used the methodology developed by [198] using the ArcToolbox [13]. The algorithm requires the definition of a spatial and temporal parameter. The spatial parameter, or bandwidth value, corresponds to the kernel bandwidth (KB). The temporal variable defines the minimal duration of stay that a point must meet in order to qualify as a stop location where an activity is performed or as a trip point. Also, other parameters must be defined like resample frequency or minimum duration for a visit to a location in order to keep as activity. The following values for the parameters were used, following the recommendation and experiments of the authors and some testing done to observe the highest possible accuracy: *KB = 275*, *Resample frequency = 180*, *Minim visit duration = 360*, *minDuration2keepHS = 360*.

In our particular case study, since we are studying the collaborative mobility inside a closed network of respondents having the same workplace, our hypothesis was that the most frequent locations visited by the respondents are the home and work locations. Since the work location is known, we performed some calculations in order to derive an automatic way to extract the home location. First we inspected the frequent activities performed by the respondents in the 9PM - 9AM time interval and preassigned those locations with the home semantic. Another general way to find the home and workplace locations following our hypothesis is by computing the total time spent in each location. In Table 3.2 we can observe that for all the respondents, the location where they are spending most of the time during the week is the home, followed by the workplace. Also we can see that the total time spent in other locations is in general less than the time spent at the workplace.

From Table 3.2 we can observe that there is a significant gap between *e.g.,* time spent at home for P_5 and the rest of the users. This denotes that the algorithm was not able to classify the raw GPS points as a location because either the GPS was turned off or it was not possible to acquire the location from the respective building.

The extracted home locations of each respondent (P_1 to P_5) are plotted in Figure 3.3 with blue and the common workplaces is highlighted in red. Those locations are used in the current study to evaluate the indicator for assessing the compatibility of a group of people for using carpooling, share the parking or car but also scenarios with a solution combining all those options.

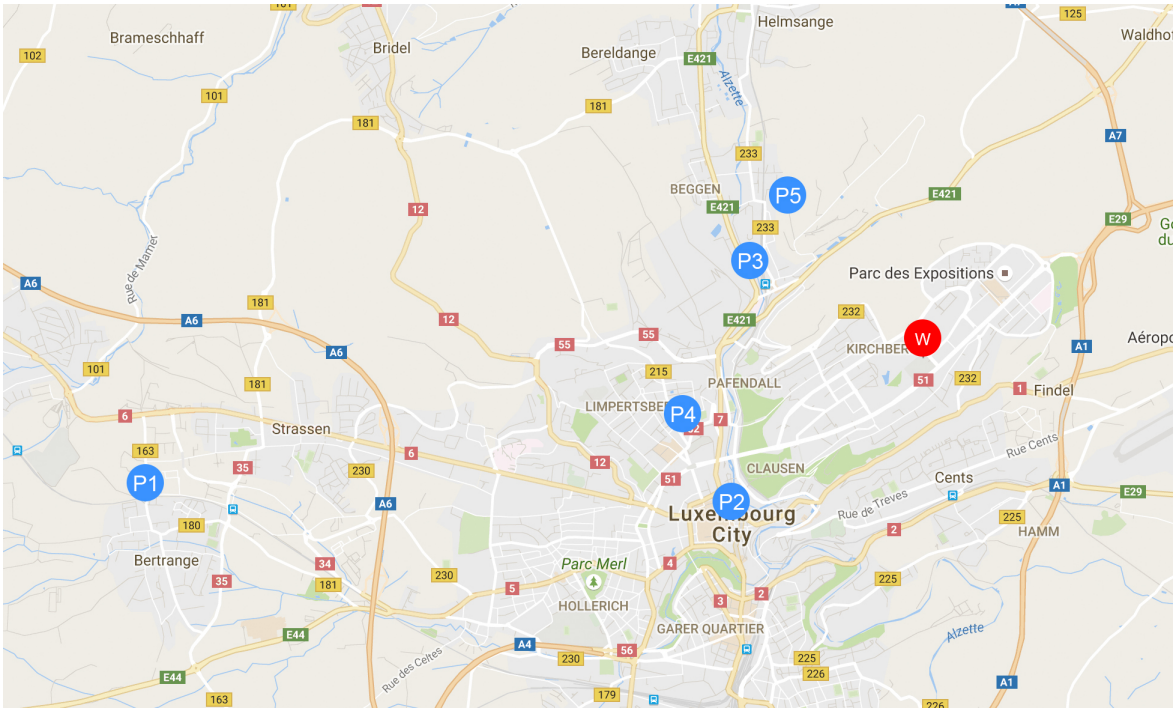


Figure 3.3: Home (P_1 to P_5) and work (W) locations of respondents.

Table 3.2: Time spent in all locations.

User ID	Time spent at home	Time spent at work	Time spent in other locations
1	68h, 32min	19h, 40min	30h, 49min
2	111h, 5min	21h, 23min	12h, 31min
3	117h, 18min	43h, 59min	4h, 9min
4	108h, 34min	20h, 21min	7h, 51min
5	8h, 15min	21h, 3min	1h, 42min

3.3.1.6 Distance in time and space between activities

The distance in the network between the location of activities is an important piece of information used in this study. After the extraction of location and duration for the activity of each user, the distance in network between all the extracted locations was calculated, using the Friendly Batch Routing [148]. Figure 3.4 shows the matrix of distances between all the 28 locations extracted from all five users.

From Figure 3.4 we can observe that some locations are very short, under 100 meters. We refer to those points as common locations between users. Because the respondents from our data collection are co-workers, the common locations can be identified as representing the workplace. Other locations are separated bt more than 20 km distance from others meaning that the user travelled outside the city or country.

The extraction of activity duration and location data followed by the computation

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28
1	0																											
2	2493	0																										
3	2657	1813	0																									
4	174651	174726	173489	0																								
5	370021	370096	368859	210443	0																							
6	360717	360792	359555	201139	10193	0																						
7	365303	365378	364141	205725	8288	3475	0																					
8	365501	365576	364339	205923	8486	3117	7413	0																				
9	365329	365404	364167	205751	8314	3100	1018	1286	0																			
10	366014	366089	364852	206436	9000	2835	872	513	1766	0																		
11	378265	378340	377103	218687	7427	17346	15822	16091	15929	17060	0																	
12	370506	370581	369344	210928	2029	11009	8732	9001	8839	9970	7067	0																
13	369998	370073	368836	210420	23	10501	8224	8493	8331	9462	7723	1904	0															
14	367334	367409	366172	207756	5216	7198	4921	5190	5028	6159	9904	4717	5193	0														
15	3210	870	2377	174662	5639	361134	365438	365707	365545	366676	383669	380178	369614	366675	0													
16	370233	370308	369071	210655	369637	10295	8673	8942	8780	9911	6694	5721	5616	3383	370869	0												
17	369819	369894	368657	210241	5076	9533	7911	8180	8018	9149	7039	5158	5053	2620	370455	623	0											
18	369991	370066	368829	210413	30	10494	8217	8486	8324	9455	7716	1897	7	4400	370627	5920	4920	0										
19	371338	371413	370176	211760	5921	10577	8955	9224	9062	10193	6683	6003	5898	3665	371974	981	1496	5891	0									
20	370230	370305	369068	210652	4572	9469	7847	8116	7954	9085	7926	4654	4549	2557	370866	2346	1504	4542	2335	0								
21	370021	370096	368859	210443	0	10524	8247	8516	8354	9485	7746	1927	23	4430	370657	5950	4950	30	5939	4975	0							
22	368613	368688	367451	209035	2068	9116	6839	7108	6946	8077	8253	2150	2045	3022	369249	4185	3343	2038	4174	3892	2068	0						
23	364825	364900	363663	205247	4573	6296	4362	4631	4469	5600	10300	4955	4550	1611	365461	4720	3649	4543	4709	3674	4573	3291	0					
24	366297	366372	365135	206719	4467	6479	4202	4471	4309	5440	9734	4849	4444	1045	366933	4154	3083	4437	4143	3108	4467	3185	1472	0				
25	373551	373626	372389	213973	5322	7320	5043	5312	5150	6281	11032	7541	5299	1900	374187	5010	3939	5292	4999	3964	5322	4040	2632	1335	0			
26	370013	370088	368851	210435	8	10516	8239	8508	8346	9477	7738	1919	15	4422	370649	5942	4942	22	5931	4967	8	1977	5188	4296	5333	0		
27	368321	368396	367159	208743	2014	8824	6547	6816	6654	7785	8260	2244	1991	2730	368957	4141	3299	1984	4130	3848	2014	93	3496	2604	3641	2006	0	
28	366910	366985	365748	207332	4000	358	5136	5405	5243	6374	9328	4382	3977	1767	367546	3748	2678	3970	3737	2703	4000	2718	2085	1195	2672	3992	2324	0

Distance between 0-100 m Distance between 100 m - 5000 m Distance between 5 km - 20 km Distance more than 20 km

Figure 3.4: Distance matrix between the locations with different main distance groups

of the distances between activities, as described in the previous sections, provide the necessary data which will help us to further analyse the collaborative mobility.

Plotting the *DT* between the extracted activities we obtained the sequence of activities for each individual. Figure 3.5 gives an overview of the extracted activities sequences, for all five users.

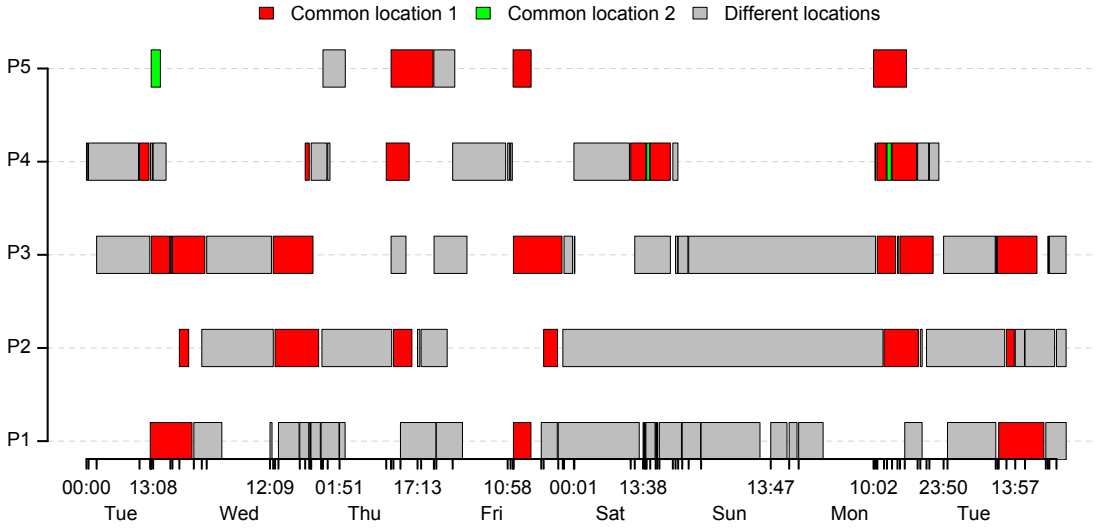


Figure 3.5: Sequence of activities.

The activities performed in a common location are highlighted in red. In our particular case study, the common location for all users is the University of Luxembourg where all the respondents are working. The activities performed in different locations are showed in grey. The gaps between the highlighted activities represents the time periods when

the algorithm has not classified the points as a location. This happens either if the respondents are moving or if it was not possible to acquire the GPS position. Also we can see that user P_4 and user P_5 visited the same location in different time periods, highlighted with green.

The distance in network between home and workplace locations from Figure 3.3 is presented in Table 3.3.

	P_5	P_4	P_3	P_2	P_1	W
P_5	0					
P_4	3,7	0				
P_3	1,6	1,4	0			
P_2	1,7	3,6	2,6	0		
P_1	9,0	8,1	9,5	7,1	0	
W	5,4	5,8	4,3	5,1	10,5	0

Table 3.3: Distance matrix in km between the common workplace and residence of all users.

The topological graph of distances in network between home and work locations computed using the data from Table 3.3 is presented in Figure 3.6.

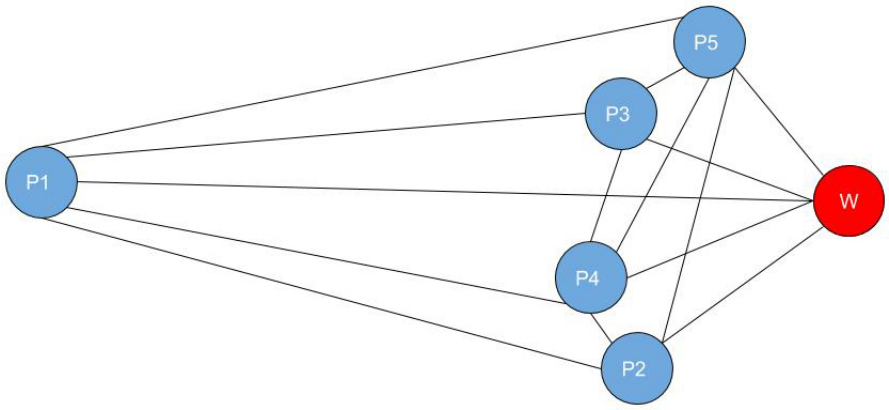


Figure 3.6: Topological graph of distances between home and work locations of respondents.

This will be our input data for assessing the collaborative mobility between individuals in different scenarios in order to test the behaviour of the proposed collaborative index.

3.3.2 Collaborative mobility index for carpooling

3.3.2.1 Carpooling compatibility

In this example we test the compatibility for carpooling between P_3 and P_5 for work commuting. It is assumed that all users in the system have access to a private car and they can drive alone or can offer a ride to other users. Also we consider the weighted cost variable for each user (α, β, γ) equal to 1. This makes the assignment of the trips easier and also leaves the focus on the index computation.

The shortest path between P_3, P_5 and the workplace is $P_5 \rightarrow P_3 \rightarrow W$. Using the equations and constrains from Section 3.2.2, the computations are as following:

$$
\begin{aligned}
C_i(P_5) &= F_{(P_5,W)} + Tt_{(P_5,W)} = 1,91€ \\
C_i(P_3) &= F_{(P_3,W)} + Tt_{(P_3,W)} = 1,52€ \\
C_{cp}(P_5) &= F_{(P_5,P_3)} + Tt_{(P_5,P_3)} + \tfrac{F_{(P_3,W)}}{2} + Tt_{(P_3,W)} = 1,76€ \\
C_{cp}(P_3) &= \tfrac{F_{(P_3,W)}}{2} + Tt_{(P_3,W)} = 1.19€ \\
C_r(P_5) &= \tfrac{C_{cp}(P_5)}{C_i(P_5)} = 0,92 \\
C_r(P_3) &= \tfrac{C_{cp}(P_3)}{C_i(P_3)} = 0,78 \\
I &= \tfrac{C_{cp}(P_5)+C_{cp}(P_3)}{C_i(P_5)+C_i(P_3)} = 0,86
\end{aligned}
\tag{3.18}
$$

As we can observe from the carpooling index, the requirement (3.6) for compatibility is met. Moreover, the condition (3.7) is fulfilled meaning that the carpooling cost for each passenger is less than the cost of travelling alone with the own private car. Also we note that carpooling is even more convenient for user P_3, because it is possible to share the entire cost of the trip with P_5, without any extra time because of the detour involved. For P_5 also the carpooling is more convenient than travelling with the private car even with a detour, because the cost of the second segment trip is shared with P_3.

This example shows how the indicator automatically indicates if the overall carpooling is efficient for any trip, and also provides the benefits for each single individual. This means that based on the individual score, each user can evaluate if the economic benefit is great enough and accept or deny a carpooling proposal. Even if this solution is foreseeable and can seem very simple, this method replicates human thinking and the probability that a user will accept a sharing proposal increases respectively. Of course each individual perceives cost and economic benefit differently. This will be explored in the next examples, where cost weight variables have different values and results.

$C_i(P_1) = 3,17€$ $C_{cp}(P_1) = 5,52€$ $C_r(P_1) = 1,48$
$C_i(P_2) = 1,80€$ $C_{cp}(P_2) = 3,01€$ $C_r(P_2) = 1,66$
$C_i(P_3) = 1,52€$ $C_{cp}(P_3) = 1,65€$ $C_r(P_3) = 1,08$
$C_i(P_4) = 2,05€$ $C_{cp}(P_4) = 2,00€$ $C_r(P_4) = 0,97$
$C_i(P_5) = 1,91€$ $C_{cp}(P_5) = 1,26€$ $C_r(P_5) = 0,66$

3.3.2.2 Carpooling incompatibility

One of the main objectives of collaborative sharing is to group more people in fewer cars. Maximising the number of passengers and minimising both overall and individual costs is a challenge in a dynamic ride sharing problem. Using the provided dataset, we will assess the collaborative mobility for all users together. Similar with Example 3.3.2.1, it is assumed that all users have access to private cars. Moreover $Tr = 0$ meaning that their schedule is synchronised.

Similar with Example 3.3.2.1, the carpooling index computation and the cost variables results between all five users are the following:

In order to carpool together, one driver must pick up all the rest of the passengers. The problem consists of finding the shortest path in the network that passes through all user residences and reach the final destination at the common workplace for all the users. This can be solved with the well-known Dijkstra algorithm, as a solution to find the shortest path from a source to all other nodes in the graph, producing a shortest-path tree. The shortest path presented in Figure 3.7 is the following: $P_1 \rightarrow P_2 \rightarrow P_4 \rightarrow P_3 \rightarrow P_5 \rightarrow W$.

In this particular example, the group formed by all five users is not compatible with carpooling. The total system cost when users are carpooling is higher than if they are each travelling with their private cars. Interestingly, as we can see from the C_r values, user P_4 and P_5 are the only ones who benefit in this case.

$$I_{CP} = \frac{C_{cp}(P_1) + C_{cp}(P_2) + C_{cp}(P_3)}{C_i(P_1) + C_i(P_2) + C_i(P_3)} = 1,22 \tag{3.19}$$

As described in the previous subsections, equation 3.19 represents the ration between sum of individual users' costs when carpooling and when not carpooling.

Figure 3.7: Shortest path between all the residences and workplace.

The question that arises is what is the maximum number of users that can carpool together in this case study? In order to provide an answer we compute the index with different number of users. The results plotted in Figure 3.8 show that in this particular case study, the best index can be obtained by grouping the last three users in order to have the lowest index value below 1 with the maximum number of users.

This experiment demonstrates that the collaborative mobility index can be used at the system level to group more people in fewer cars, but at the same time at an individual level so that all users may benefit from using sharing services. The indicator is sensitive to dynamic changes *e.g.,* the maximum number of users that can join in a carpooling trip from one origin to a destination or the efficiency at the system and individual level if a specific user is joining the carpooling trip. We can conclude that the indicator is flexible and has the potential to be used also in real time carpooling assignment and the system can recommend *e.g.,* to a user which is driving that on the route of the trip can pick-up other compatible users and all the passengers will have an economic benefit. This can be useful for individuals who are commuting or travelling but also to *e.g.,* taxi and carpooling companies which can find new customers ad-hoc in real time without any other intervention. The requirements for this type of sharing services demand that the system to be provided with the origin-destination information for each user.

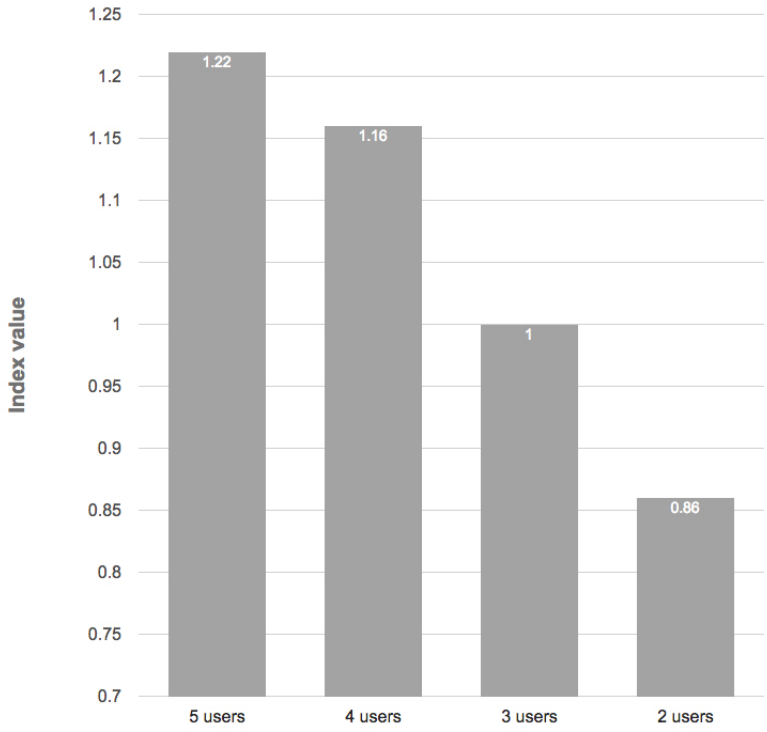


Figure 3.8: Index value for different number of users.

3.3.2.3 Carpooling index with rescheduled activities

In this example we aim to assess the carpooling compatibility between user P_3, P_4 and P_5 in the case when their activities are not synchronised. Tr and weight cost variables are considered, as defined in (3.10). In Figure 3.9 we can observe the DT between activities of all users for a day, extracted from Figure 3.5.

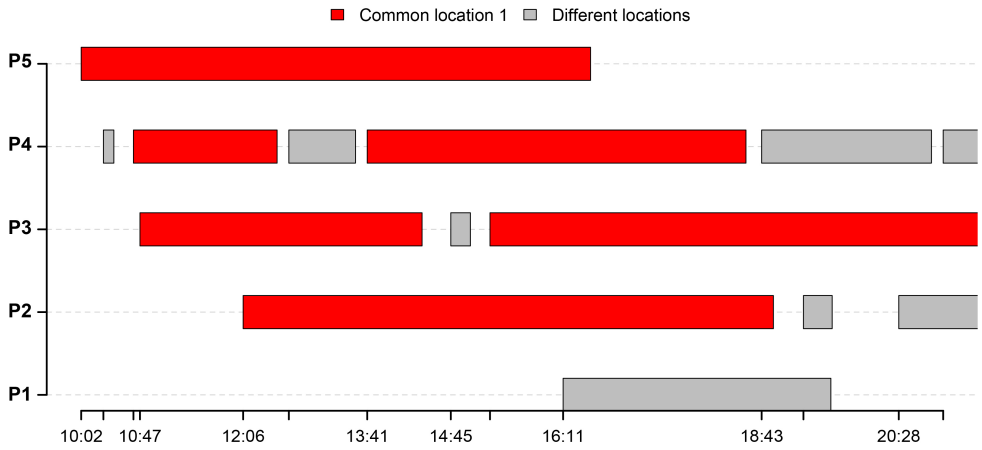


Figure 3.9: Distance in time between activities of all users for one day.

From Figure 3.9, both P_3 and P_4 start work at almost 10:47. Therefore they are well synchronised and the DT between them is very small. The computed carpooling index

values are $I = 0,80$, $C_r(P_3) = 0,78$, $C_r(P_4) = 0,82$ with the result that they are very compatible for carpooling; both (3.6) and (3.7) constraints are met.

It might be possible also for P_5 to join the ride sharing but we can see that P_5 usually arrives at work 45 minutes earlier then P_3 and P_4. In this situation users must reschedule their activities in order to achieve the synchronization and be able to carpool.

We therefore consider the three following scenarios. In the first scenario it is assumed that P_5 with the result that for him $\alpha = 1$, $\beta = 1$ and $\gamma = 0$ as defined in (3.10) and the reschedule is not seen as a cost. Consequently this scenario reshapes as a perfect synchronization between individuals because $T_r(P_5) = 0$. In the second scenario P_5 accepts to reschedule the activity, but α, β and γ are considered by the user (equal to 1) for all users resulting that for P_5 the related reschedule has an extra cost. The third scenario is similar with the second, with the difference that P_3 and P_4 will reschedule their activity, thus both of them will have reschedule time cost.

The comparisons of the index value and the cost for each individual in the proposed scenarios are captured in Figure 3.10.

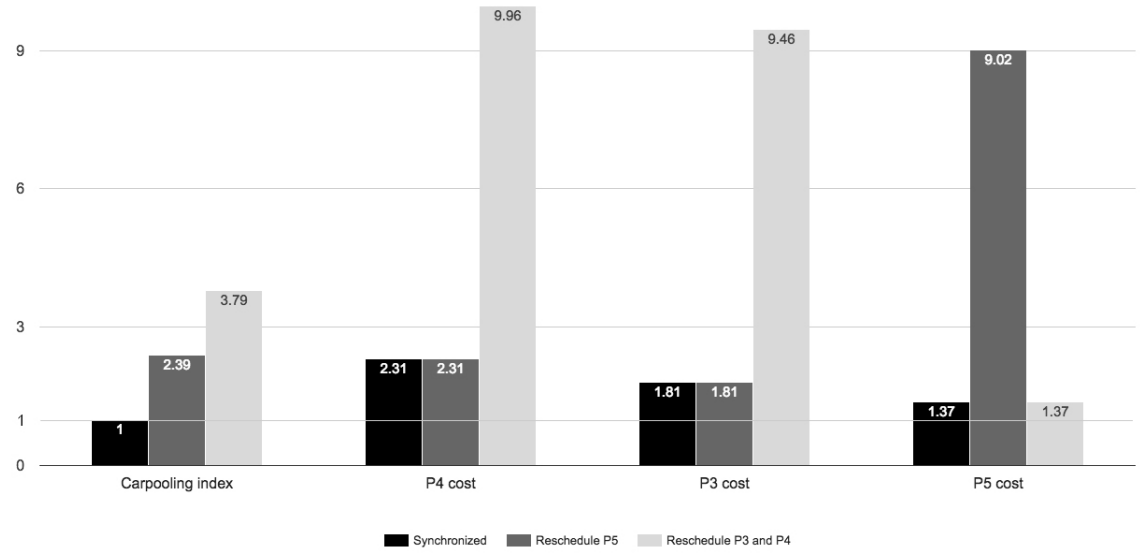


Figure 3.10: Carpooling index and individual user costs for different rescheduled options.

From this case study we can argue that the proposed carpooling index is sensitive to any additional cost in any situation. Based on the index value, a recommendation system that has multiple options for user synchronization can choose the optimal solution for advising users how to reschedule or reorder their activities in order not only to minimise the overall system cost, but also to balance the individual cost for all users. Moreover, the indicator is sensitive to the individual's cost perception, computing the index value considering the weight value for each involved variable. The indicator is sensitive to any change in the weight of cost variables for each user and the final indicator value for a group of individuals can be highly influenced by the weight that each individual assigns to the cost variable. This means that a group of users can be compatible or not just because they have different perceptions of the economic benefit and the trade-

off between comfort, saving money or flexibility e.g., to reschedule the departure or arrival time. Users synchronisation and usage optimisation of existing resources are other objectives where the proposed indicator model can contribute.

3.3.3 Collaborative mobility index for parking sharing

3.3.3.1 Parking sharing compatibility index for car-dependent users

In this scenario it is assumed that there is a daily parking fee at the workplace. This cost can be shared between multiple persons if they use the same parking place. The problem consists of finding car-dependent compatible users who match (3.15). In order to find compatible users, we compute the DS and DT between the parking location and the users activities. Figure 3.11 shows the results for two users during one day.

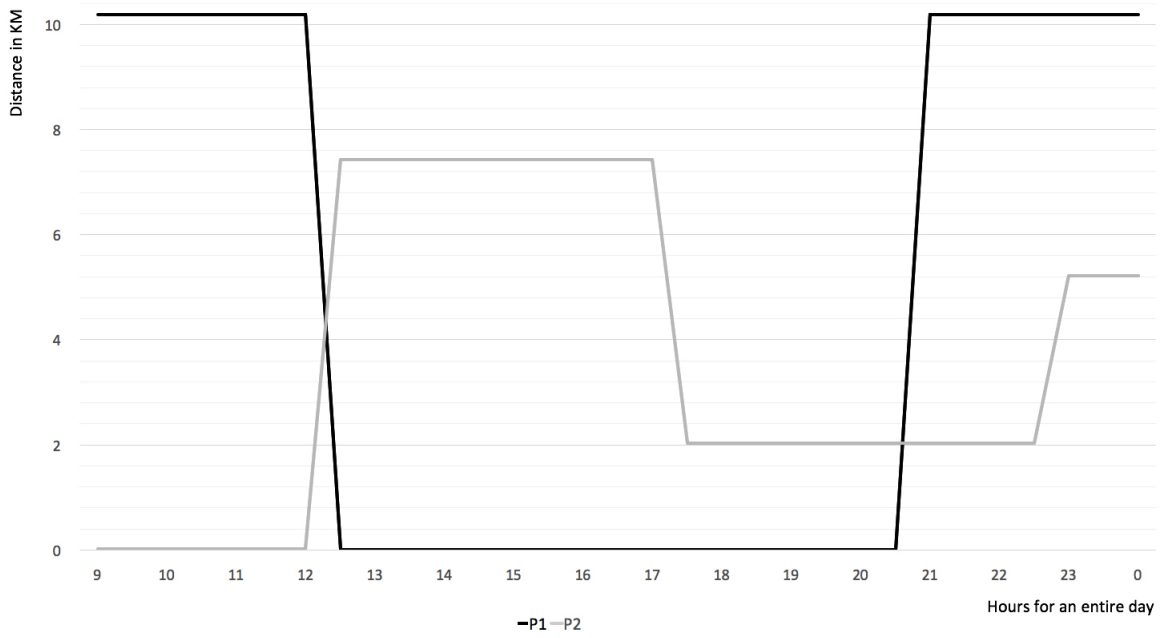


Figure 3.11: DS and DT between parking place and activities of user P_1 and P_2 for full day.

When the distance is zero, users perform an activity in a location near the parking place. In our case study it is the workplace parking spot. From Figure 3.11 we observe that the intervals in which P_1 and P_2 are in the parking place meet the constraint (3.15), with the result that they are compatible for parking sharing. If this repetitive behaviour can be observed over long time periods, they are compatible for long-term parking sharing.

Having travel behaviour data for a long time period and for a higher number of respondents, using the proposed indicator it is possible to compute at large scale the compatibility between all users for long term parking sharing. In this way the parking lot can be used more efficiently and the number of parking places required can be

reduced. This can be useful both to individuals who wish to share the parking cost, but also to organisations which aim to reduce the parking space and the corresponding costs. The indicator can also capture long-term dynamic changes that appear in the travel behaviour of each user and suggest different sharing solutions based on the new knowledge *e.g.,* change of residence or workplace but also new frequent locations visited *e.g.,* a new restaurant, new friends etc. The only requirement is a historical database of locations visited and a learning method used by a recommendation system that detects changes in the behaviour of each user and suggests new solutions.

3.3.3.2 Parking sharing index in combination with carpooling

In this case study we evaluate how the indicator behaves when a scenario is considered with daily parking fee in combination with a carpooling service. For this case study we use the input data from Example 3.3.2.2 from which we have the indicator values for a situation in which all five persons might carpool and in the absence of a parking fee and we compute the index as defined in Section 3.2.3.2. We compare the results from Example 3.3.2.2 with the situation when there is a shareable daily parking fee in order to observe how the parking fee policy can influence the perception and costs of the carpooling service. The results are presented in Figure 3.12.

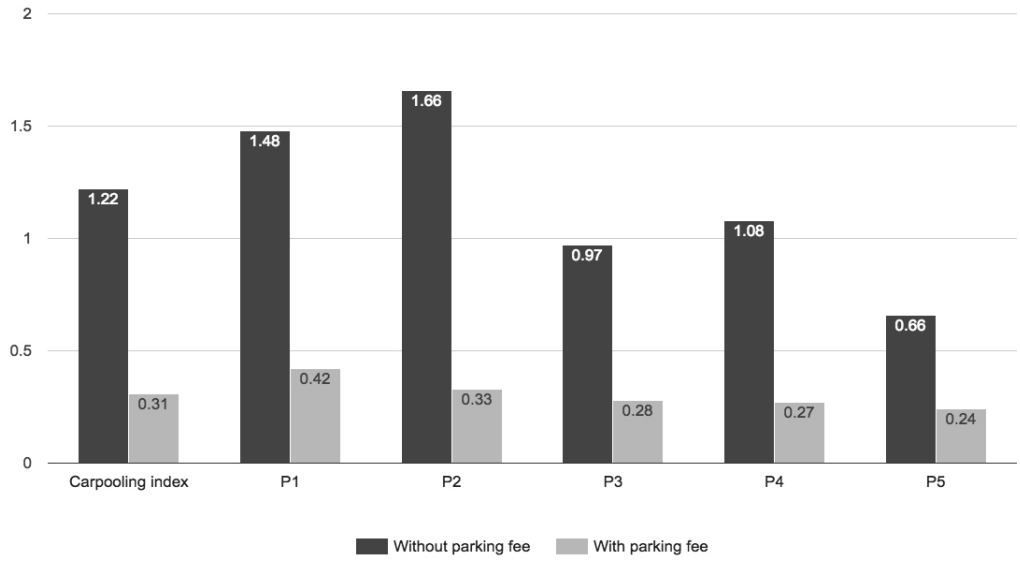


Figure 3.12: Comparison of index value and individual costs ratio C_r with and without parking fee.

From Figure 3.12 we observe that in the situation when there is no parking fee, the index value indicates no compatibility for carpooling, with the index value higher than 1. In this case they will not carpool together meaning more parking places occupied. In the situation when they have to pay a parking fee, the index value indicates an excellent compatibility when sharing the ride and parking cost, both at the system level but also at the individual level. The results from this case study show that the index can be used as an indicator for assessing the impact of different parking fee policies and prices.

The proposed indicator can be part of a recommendation system for advising users for sharing the parking cost with other compatible users in combination with other sharing services, in this case carpooling.

3.3.4 Collaborative mobility index for carsharing

In order to evaluate the indicator for carsharing we consider a case study that combines carsharing, carpooling and parking sharing. Also we consider the entire activity chain of an entire day, as discussed in Section 3.2.4. Figure 3.13 presents a similar situation as in Example 3.3.2.1. From this example we know that P_3 and P_5 are compatible for carpooling when they commute from home to work in the morning and from work to home in the evening, and when they also share the parking cost. In this case study the difference is that P_3 must travel from the initial workplace $(W1)$ to a meeting in $(W2)$ and return back to $(W1)$. In this case the chain of activities for P_3 is: $P3 \rightarrow W1 \rightarrow W2 \rightarrow W1 \rightarrow P3$.

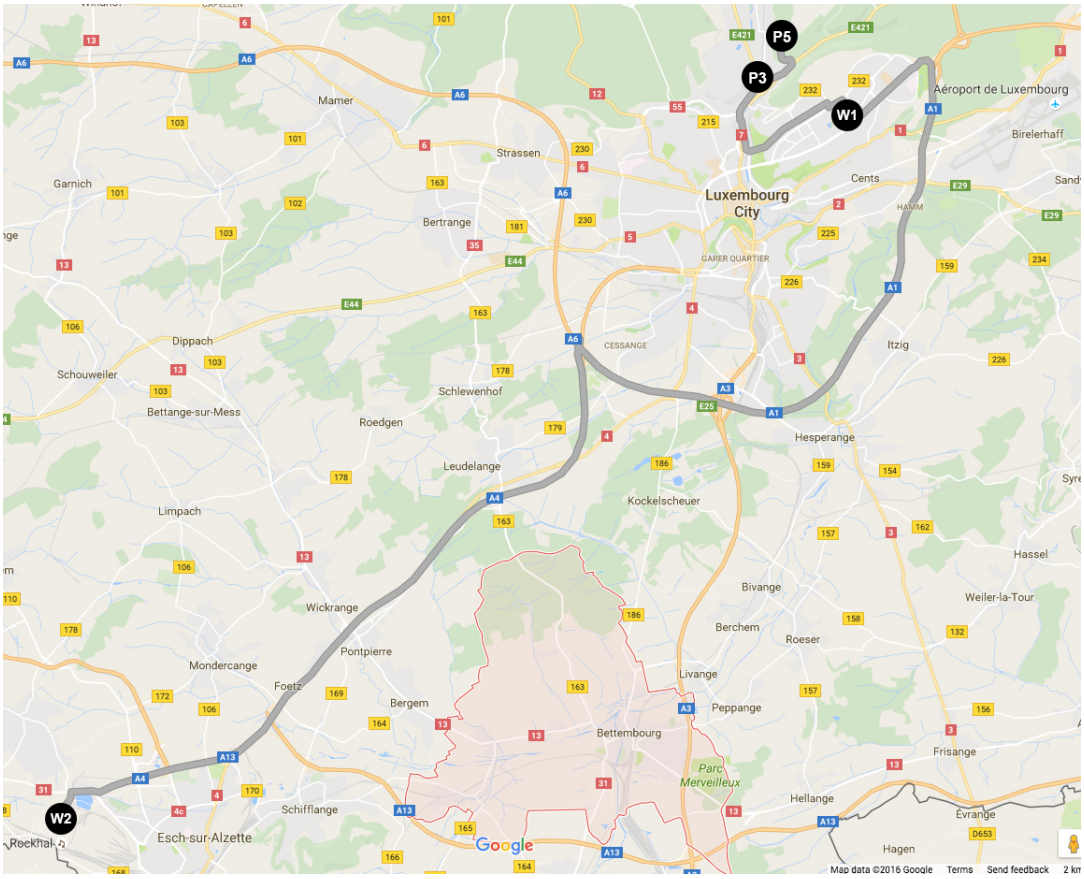


Figure 3.13: Case study with combination of carpooling, parking sharing and car sharing.

For the trips $P3 \rightarrow W1$ and $W1 \rightarrow P3$ the user plans to commute by carpooling and share the parking cost with user P_5. But because of the need to travel an extra $W1 \rightarrow W2 \rightarrow W1$ segment trip and because of the lack of any other user with whom to carpool, the user plans to use the private car for the entire day because of the tight

schedule and the risk of losing precious time. This certainly means extra cost because it is not possible to share the costs in this situation.

In another possible scenario, we assume that the users have access to a carsharing system. This option gives them the flexibility of carpooling for the trips $P3 \rightarrow W1$ and $W1 \rightarrow P3$ with user P_5 and of using the carsharing system for the trip $W1 \rightarrow W2 \rightarrow W1$.

In order to evaluate the indicator behaviour in this case study, we compute the indicator when P_3 use the private car for the entire tip chain, versus the situation when commuting by carpooling and sharing the parking cost with P_5 and using the carsharing system. The computations as defined in (3.3), (3.4) and (3.5) are:

$$C_i(P_3) = F_{(P3,W1)} + F_{(W1,W2)} + F_{(W2,W1)} + F_{(W1,P3)} + P = 26,45 \text{ €}$$

$$C_i(P_5) = 1,91 \text{ €}$$

$$C_s(P_3) = \frac{F_{(P3,W1)}}{2} + F_{(W1,W2)} + F_{(W2,W1)} + \frac{F_{(W1,P3)}}{2} + \frac{P}{2} = 10,51 \text{ €}$$

$$C_s(P_5) = 1,76 \text{ €}$$

$$C_r(P_3) = 0,39$$

$$C_r(P_5) = 0,92$$

$$I = 0,43$$

In this case study the indicator values show that the best option for P_3 is to carpool for the commuting trips and to use the car sharing for the other trips. Also P_3 and P_5 are compatible for using sharing services, both of them saving the related trip and parking costs.

We can argue that the proposed indicator can be used to evaluate collaborative mobility between individuals in a closed environment, considering the entire chain of activities and combinations between all the sharing services. The values obtained can be used by an ITS that acts as a travel advisor which automatically finds the most efficient sharing services and recommends to each individual user in the system, considering not just only one-time-service usage but the entire trips chain for one day. Moreover, the indicator is able to evaluate and optimise combined sharing schemes and modes in a single indicator, quickly and simply. The travel advisor may try different combinations and propose various solutions from which users can choose, depending on their preferences and constraints. The model proposed may be used for both a multi-objective optimisation approach and for more complex problems where different providers, schemes

and solutions can be considered.

3.4 Discussion and Perspectives

In this study we explore the usage of ITS in combination with the actual technologies and with the sharing services as an effective solution for traffic congestion problems. The current study together with others from the literature review emphasised the need for a change in citizen's travel behaviour towards sustainable and efficient collaborative mobility among groups of users. In this case the ITS must pro-actively give recommendations to users and incentivise them to reschedule, reordering and re-routing their activities in order to group as many travellers in fewer cars but also giving advice on how, when and with whom to share the resources and combine the costs involved *e.g.,* parking fees, fuel cost.

This process must be fully automatized by using methodologies similar to those presented in this study. Data must be accurately collected (individual based and available in real-time from mobile devices) and processed. The framework presented in the current study aims to take the full benefits of the data collected and transform it into knowledge. More complex Machine Learning and Data Mining methods and algorithms can be used in order to extract the knowledge from raw data, using a data fusion strategy from all the sensors built in the mobile devices. This means that the ITS must take full benefit from the geospatial Big Data and larger market penetration of portable technologies. With this, the next generation of collaborative ITS platforms would be able pro-actively to give advice in order to change the user's travel behaviour in the interest of more efficient, sustainable and environmentally friendly sharing solutions. ITS must act as an advisor that assists citizens in their daily choices in order to achieve significant travel behaviour change.

The indicator developed in this study aims to fill the gap between the ITS and the user needs and preferences. Using this indicator an ITS can assess collaborative mobility between users in a faster, more flexible and more reliable fashion. As presented with real case studies in the current study, the system can automatically find sharing opportunities and recommend different behavioural changes in order to be compatible with other users for sharing services.

From the case studies presented, we can argue that such a travel advisor is suitable for different types of organisations and communities. In this model, members will be advised by the ITS to reorder or reschedule their activities in order to be compatible with other users inside the organisation for using sharing services in a CM system. This will lead to great benefits at the company level because *e.g.,* , fewer employees will commute with their private cars. This can result in savings for large companies which normally need to rent large parking lots for their employees. The case studies presented in this study suggest that a parking fee can incentivise more people to adopt the shared commuting option. Also the proposed indicator can be used as a tool for assessing the effects of different parking policies at the organisation's level.

At a higher level (*e.g.,* at the city level), fewer people will commute with their pri-

vate cars, resulting in reduced traffic congestion at the peak hours. The indicator's values include also the system cost variable. A general perspective regarding the organisation's system costs can be a good indicator for evaluating the impact of different policies planned by local authorities.

The collaborative mobility scheme presented in this study may also have great benefits at the micro level for each individual user in the system. Cost savings *e.g.,* fuel, parking fees and time lost in lengthy commutes, are strong incentives to attract more and more travellers towards an environmentally friendly and sustainable travel behaviour.

Because user's activities are repetitive, all the computations made for different users and scenarios can be stored and reused by an ITS for speed optimisation. This is mandatory when a massive number of computations must be made for large communities. Finally, the entire process must be fully automatised so that the user will not have to manually perform queries in the system (*e.g.,* , searching for a ride or for a colleague to share the parking fee). The entire human-machine interaction should be minimised until the point that the user will have only to press one button in order to confirm that the system correctly predicts the next destination and the user accepts the system's recommendation.

We can argue that the proposed method has multiple strengths described in the presented proof-of-concept examples. We have demonstrated that the method can evaluate combined sharing schemes, sharing modes and resources in a single indicator. It is sensitive to dynamic changes, flexible and can be used in multi-objective optimisation problems. However, the results and accuracy of the model are dependent upon the quality of the input data. This means that in order to be used, the indicator must be integrated as a component, part of both a complex system where the data is collected, aggregated and processed in an automatic fashion, and a recommendation system which performs different computations based on the model proposed and chooses the best solutions, having taken into account all the objectives, users preferences and constraints. Some variables *e.g.,* the weight of the cost for each user might be very hard to obtain and those are mandatory elements that can have a high influence on the final result. The indicator has been tested only on a small dataset, with a limited period of time and only five users. Even so, because there are many variables, routes, sharing schemes and modes, user preferences and constraints, but also multi-objectives solutions, the computation becomes quickly very intensive and this can require large computing resources and optimisation methods for computation.

3.5 Conclusion and future work

In this study we have proposed an indicator for enhancing collaborative mobility, which can be used by a travel advisor to proactively recommend different actions towards environmentally friendly and sustainable travel behaviour. The research presents different constraints and variables which should be taken into consideration when assessing the collaborative mobility between individuals based on different sharing solutions. Car-pooling, parking sharing and car sharing experiments and case studies were ex-

plored using the real data collected from mobile devices. These experiments confirmed that the proposed indicator is sensitive to dynamic changes, flexible and can be used in multi-objective optimisation problems. We conclude that the behaviour of combined sharing schemes and modes can be assessed in a single indicator. This can be used for evaluating the collaborative mobility for individuals at micro level but also by organisations for simulating the effects of different policies.

There are many directions for extending this work. The methods presented in this study represent the theory that has to be combined with complex ML and artificial intelligence methods in order to develop the engine of an ITS that has a bright future in the next generation traffic systems. The challenge will be to develop the prototype of an elaborated ITS that will be able to execute complex tasks and operations e.g., automatic learning, prediction, optimisation and management of collaborative systems for sharing resources.

In the current research we have used mainly the geospatial data collected from mobile devices. Future exploration of sensing systems will also be done in order to use the data fusion of all the mobile sensors available from the nomadic and wearable devices like smartphones and smartwatches. Each additional sensor brings important data which should be exploited in order to acquire new knowledge.

While we consider the problem of collaborative mobility, our approach is still at a small scale, mainly because of the limited data collection and the focus on the theoretical aspect rather than real-world implementation. Other challenges related to the large scale real data usage collected from smartphone are the data collection high cost and privacy aspect. Authorisations must be obtained from specialised authorities for surveys and data collections that contain sensitive information. Then data anonymization protocols must be implemented using different types of information sanitization with the intent of privacy protection. Encryption protocols, blurring techniques for location privacy or removing personally identifiable information from data sets are some of the protocols that can be used so that the users from whom the data is collected can remain anonymous.

In the following chapter, similar scenarios will be used for larger datasets with a higher number of users and for a long time period. This approach presents complex problems because larger datasets require specialised methods, tools and optimisations for running intensive computations. Also, the knowledge extracted must capture similar travel behaviour of the real users' mobility patterns. In this sense, a new modelling framework that uses temporal graphs and time series in multi-dimensional data is presented in the following chapter.

4

A modeling framework over temporal graphs for big mobility data analytics

In this chapter, we present an innovative data modelling framework that can be used for ITS to deal with big mobility data. We demonstrate that the use of graphs and time series in multi-dimensional data models can satisfy the requirements of descriptive and predictive analytics in real-world case studies with massive amounts of continuously changing data. The features of the framework are explained in a case study of a complex collaborative mobility system that combines carpooling, carsharing and shared parking. The performance of the framework is tested with a large-scale dataset, performing machine learning tasks and interactive real-time data visualization. The outcome is a fast, efficient and complete architecture that can be easily deployed, tested and used for research as well in an industrial environment.
This chapter is based on work that has been published in the following paper:

- *A new modelling framework over temporal graphs for collaborative mobility recommendation systems*
  Bogdan Toader, Assaad Moawad, Francois Fouquet, Thomas Hartmann, Mioara Popescu, Francesco Viti,
  2017 IEEE 20th International Conference on Intelligent Transportation Systems (ITSC)

Contents

.

4.1 Introduction

In this chapter, we propose an innovative data-driven framework that can be used in the transportation domain with large-scale datasets and makes possible the implementation of different data science methodologies, techniques and ML algorithms. The main contribution of this chapter is the implementation of a framework that combines the analytical and statistical results (i.e. from the data science domain) with discrete simulations (i.e. evaluation of alternative actions), adapted for the requirements of the transportation domain and in particular for the case study of collaborative mobility.

The remainder of this chapter is organized as follows. First, Section 4.2 introduce a collaborative mobility case study, which motivates the research behind this contribution. Sections 4.3 introduces the methodology behind the main concepts and features needed in the transportation domain. We thoroughly evaluate our approach in Section 4.4. The related work is discussed in Section 4.5 before concluding in Section 4.6.

4.2 Motivating case study

This study was motivated by the methodology presented in Chapter 3, which describes an indicator that can be used for assessing the CM between individuals. In particular, based on the value of the indicator, the authors analysed the potential for assessing collaborative services among small groups of users for different combinations of sharing services (carpooling, carsharing, shared parking). Our objective is to develop a multi-functional framework that can be used for implementing the CM indicator at a large scale. Moreover, we extend the study's applicability in the dynamic ridesharing domain by the framework features proposed. In the remainder of this section, we present the technical and methodological requirements that can solve the above mentioned problems, extracted from [202].

In large scale ride-sharing systems, sensors from mobile devices and cars positioning systems are sending data continuously. Storage, management and fast access to a temporal unstructured dataset is not trivial [78]. The challenge increases when we take into consideration the complexity of underlying data models and patterns combined with the relationship between entities (people, locations, cars, personal preferences). Analysing such large historical datasets where the context and input values change over time as the entities are constantly moving calls for advanced time series analytics and technologies. While this topic has been explored by the database community [184], it is lately gaining popularity also in the ITS domain.

Time management is a critical component in smart mobility systems. Different services make resources available simultaneously or over time. Therefore, the system must be able to analyse data over time and perform deep and complex temporal search

queries. GPS location data and timestamps are received with different frequencies and sometimes data values are missing (*e.g.,* mobile phones with GPS deactivated). When *e.g.,* a request is made for a carpooling activity, filtering and searching in time and space operations must be performed, returning the possible candidates for carpooling. Often, the exact position of each rider is not exactly known and the system must return either the last known values or an extrapolation of the missing data based on the history and on specific domain rules.

Over the years, data analytics in the ITS domain shifted from *descriptive* data analytics, represented by the understanding of past events, to the emergence of *predictive* analytics, i.e. techniques which make predictions about the future based on learning from historical data. Therefore, different ML techniques can be implemented (*e.g.,* profiling, prediction, clustering, classification) in order to extract knowledge from unstructured data. The ML component must be able to continuously learn and the recommendation system must use the results almost in real-time. The work from Chapter 3.1 [202] supports the idea that the large scale implementation of CM services calls for the next generation of *prescriptive* analytics, capable to take efficient decisions using recommendation systems. In the ridesharing systems the current state of each entity (*e.g.,* location of users) is known from the data received, as well as the desired state expected and results (*e.g.,* grouping more users in fewer cars). Prescriptive analytics can be used to perform an what-if analysis [103] by exploring different alternatives. The final goal is to give recommendations (*e.g.,* with whom an user can carpool) regarding the actions that should be done in order to reach as close as possible from the initial state to the desired state, almost in real-time (*e.g.,* shifting the departure time in order to be compatible with more users).

These are the basic requirements that the next generation of smart mobility systems must meet. This motivates our work on implementing a new modelling framework over temporal graphs that satisfies the requirements of CM recommendation systems. In the next section, the methodology behind each proposed framework features is presented, as long with the implementation details. Although in this study we use mainly a similar case study as [202], the framework can be applied to many other domains [112, 109].

4.3 Features and methodology

This study introduces a new modelling framework that satisfies the requirements of the next generation of recommendation systems, based on the unique combination of the integrated core features. The proposed solution has the foundation in the GreyCat [71] framework, formerly known as KMF [90]. In this section, the methodology behind each core feature is presented, together with a summary of the implementation.

4.3.1 Modelling with graphs

Graph theory has long been used in multiple domains and applications [76]. Figure 4.1 presents the general components and the information processing flow in the case of a

recommendation system that can be implemented for assessing CM using the indicator developed in [202]. The components included in the square with the dashed line are implemented and used in the current study.

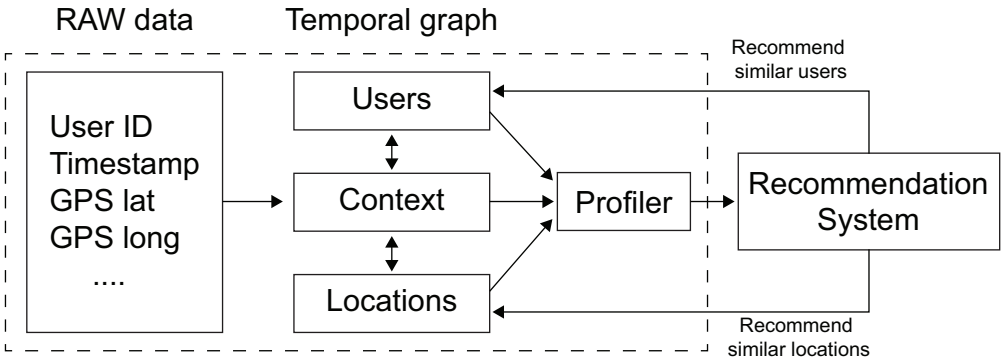


Figure 4.1: Meta model of the graph and system components

The temporal graph is used simultaneously as storage after raw data processing, as an input for the profiler component and as validation for the recommendation system. In this case, the graph stores the data in nodes represented by users, locations and context. Each node has its own attributes according to its type *e.g.,* latitude, longitude, timestamp, sequences of locations, user preferences etc. Between each node, the edges of the graph are represented by the relationships between the nodes.

4.3.2 Temporal aspect

The relationships between nodes (*e.g.,* in our case study users and locations) can evolve over time. Graph structure evolution over time is presented in Figure 4.2.

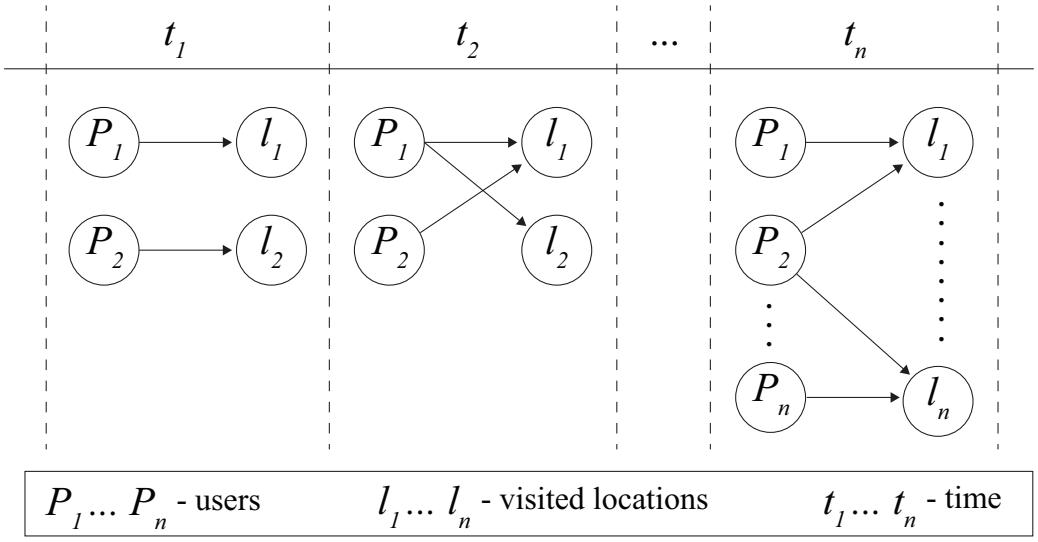


Figure 4.2: Graph structure evolution in time between users and locations

As we can observe, at each timepoint t_n the graph captures the relationships between

users and locations, which constantly change over time. The advantage of this type of structure is that similarities and behavioural patterns between users can be identified by comparing their evolution over time. Moreover, new users can enter the system and others leave, a typical behaviour of users in any ride sharing system.

One major advantage of intelligent systems dealing with time series data is their ability to continuously analyse their context in order to autonomously offer recommendations [111]. In the case of smart mobility systems, the dynamic context of continuously moving of users force the reasoning processes to analyse and compare the current situation with the trend created by the past events. Even if a common approach consists in a temporal discretization, this will lead to large data mining [111]. Figure 4.3 presents the framework's concept of storing only the relevant data (data points marked with a black dot). Data represented by a cross (×) is discarded for compression.

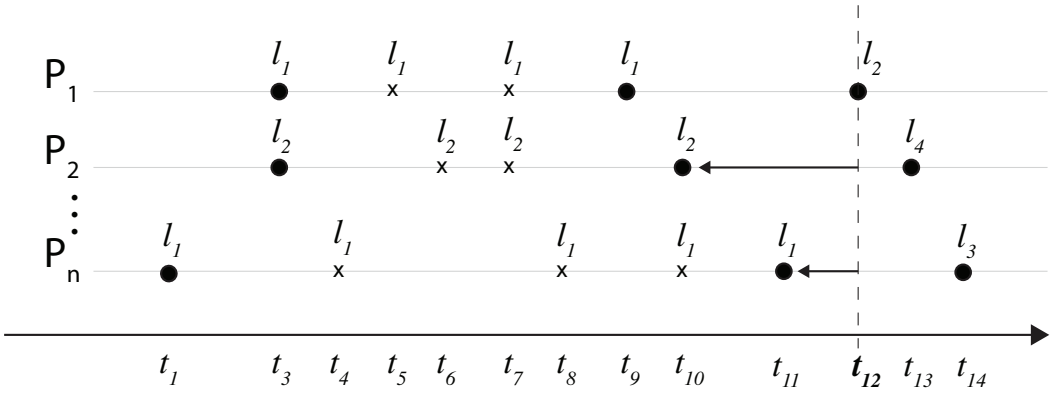


Figure 4.3: Time management and irregular data frequency.

In smart mobility systems, the GPS location data (l_n) is constantly received from P_n users' smartphones. In this case, the system must manage irregular frequency data rates and optimise the storage by discarding redundant data. The proposed model will store new locations (data points marked with black dots) only if the users change their locations. If the users don't change the location, the data points (represented with ×) are discarded. Let us consider the situation when the system is performing a query at the time t_{12}. In this case, only the user P_1 can return the actual location. For the rest of the users, traversing back in time to the last known location is necessary to estimate the current position. In the case of missing data, different extrapolation methods can be used [154]. This technique allows every node to evolve independently of the other in time and greatly contributes to the compression of data storage and access speed.

4.3.3 What-if analysis exploring alternatives

According to [202], in order to asses the CM indicator between a group of n riders, $(n \times (n-1))/2$ calculations of different combinations can be made to group the riders that can share a trip. Using the Quadtree indexing [126] in a graph structure has the advantage of reducing the number of calculations to $nlog(n)$. Similar, the number of steps to traverse a graph is $log(n)$, compared with $n-1$ steps in a classic flat structure.

As [103] stated, *what-if* analysis is a mandatory component in this case. Our proposed framework can generate a large number of *what-if scenarios*, *e.g.,* find the maximum number of riders that can carpool at a certain moment in time. Hartmann [106] introduces the notion of Many World Graphs (MWGs), which can be defined as hypergraphs, which structure and properties can evolve along time and parallel worlds. Each simulation is done in a parallel world, used as an identifier. In Figure 4.4 is presented the general concept of MWG, as a solution for multiple parallel simulations. The worlds that are represented by a continuous line are resolved, while those with dashed lines are discarded as the solution provided is not fitting the objective function.

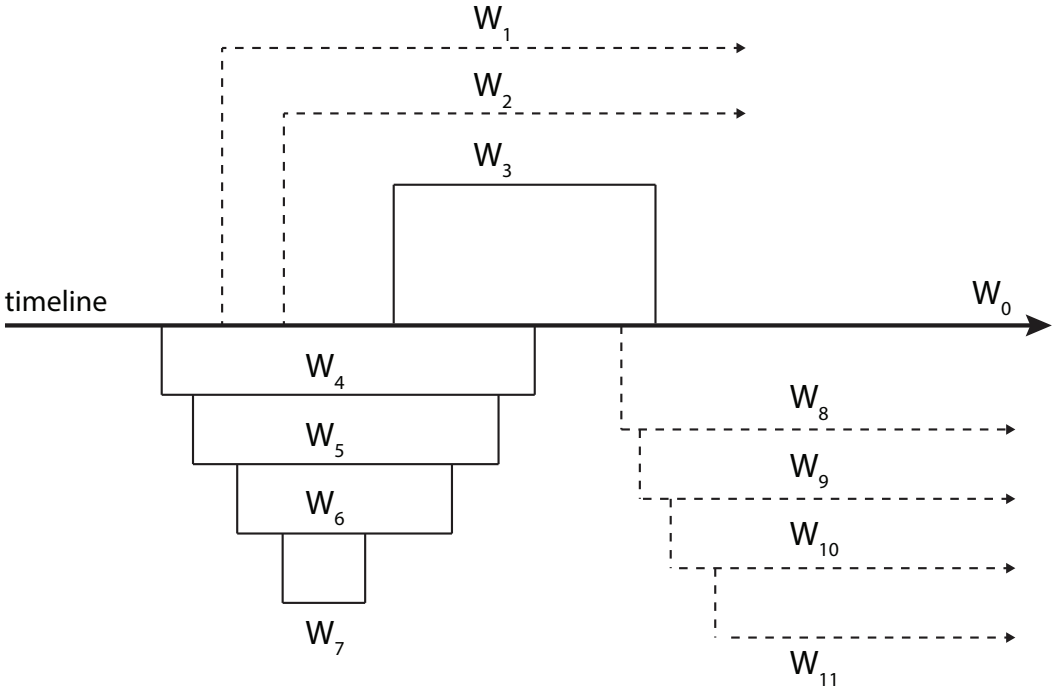


Figure 4.4: Many Worlds Graph as a solution for multiple parallel simulations

We define the first created world (W_0) as the root (parent) world, defined by the collected real data. All other worlds can be created at any time, diverging from the root world (*e.g.,* W_1, W_2, W_3, W_4, W_8). Then, from any other created world, other secondary worlds can occur (*e.g.,* W_5, W_6, W_7, W_9, W_{10}, W_{11}).

In the ridesharing case study, the root world is represented by the real data collected by the riders' mobile devices. When *e.g.,* the driver is asking for riders to share a trip, the framework performs multiple simulations in parallel worlds to test the indicator value for different combinations of grouped riders. If the indicator's computation shows that the riders are compatible for carpooling, the world (*e.g.,* W_3) is solved as represented by a continuous line in Figure 4.4. The scenario is then considered as successful and the recommendation system can inform the driver about the compatible rider(s). A simulation is either successfully completed and merged in W_0, or discarded and represented by a dashed line. If the simulation *e.g.,* W_4 is successful, it is also possible to perform another what-if analysis for finding an additional compatible rider for that trip. The process can be repeated until has reached the maximum capacity of the car. If W_7 has successfully resolved, all the parent worlds are closed and merged.

Regarding the support for prescriptive analytics, [106] showed that MWG is able to handle efficiently hundreds of thousands of independent worlds. Moreover, every node can fork independently and when forked, differential information is stored, resulting in a faster process with data compression.

4.4 Experiments

4.4.1 Dataset and experimental setup

In order to test the proposed framework, the Geolife dataset [151] is used. This publicly available dataset contains over 25 millions GPS points, collected in the Geolife project [233] from 182 users with smartphones and GPS loggers in a period of five years. There are 17,621 trajectories represented by sequences of time-stamped points, which contain the latitude, longitude and altitude with a variety of dense representation sampling rates, *e.g.,* 1-5 seconds. The characteristics of the dataset are suitable for testing the performance of real-time large scale systems and platforms. The dataset was used in many research fields, such as mobility pattern mining [235] and study of human behaviour [234].

Even if the main focus in this study is not related with the performance testing of ICT systems, one of the mandatory conditions of the real-time recommendation systems is the speed of loading, processing and data learning. Experiments are performed using an end-user laptop (Processor: quad-core 2,8 GHz Intel Core i7, RAM: 16 GB 1600 MHz DDR3, Disk: 1TB SSD). LevelDB [134] is used as database management system for its excellent capabilities of storage compression and querying speed.

In order to test the above features of the framework, in the next subsections different experiments are performed. The complete source code of the implemented framework is available from [11].

4.4.2 Scalability

Data received in real-time from different sources must be processed with the following main operations: parsing, learning and saving to the database. The entire dataset was processed in 63,1 seconds with an average of 416.787 values/second. The most consuming operation was as expected the profiling (64,18%), followed by the parsing from CSV (28,68%) and saving in the database (7,13%).

4.4.3 Profiling

The ML component was tested using a Gaussian Mixture Model algorithm [124] for profiling users' location probabilities by the day of the week and hour. In total we created 336 profiles for each user, according to each day of the week at half hour

interval. The results can be visualised live in the browser while the system can perform in parallel computations and update the results in asynchronous mode. In Figure 4.5 is presented a snapshot of the interface which can be accessed online [6].

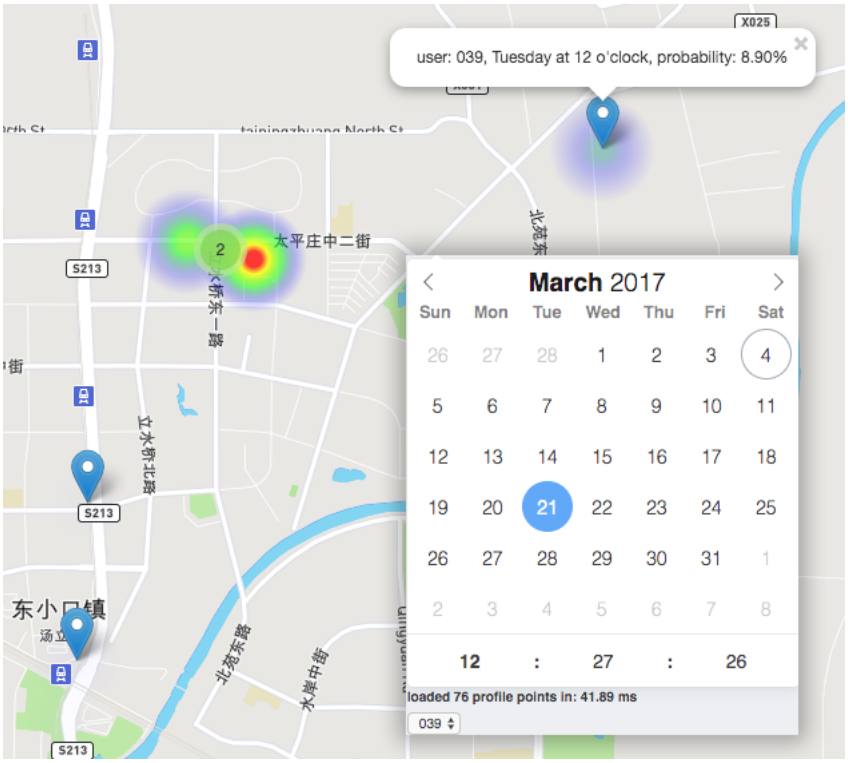


Figure 4.5: User profiling probability map

By selecting any user, a day and hour of the week, the system performs a query in the database and navigates over temporal graphs at the selected timepoint. Then returns the probability map of all the visited locations, learned from the historical data.

4.4.4 Deep search and query capabilities

The very first test was to assess how well the users dynamic movement over the 168 hours of the week was captured.

As mentioned above, the location is stored only when the users move. As a result, the number of users that send new data in each hour of the week over the entire five years represents an indicator of the user's movement dynamics. Performing a deep search query with the average users that are moving for each day of the week, we obtained the results shown in Figure 4.6.

The observed distribution confirmed the hypotheses of the users' typical movements: the working days of the week have a similar pattern with a peak of traffic in the morning and one in the evening (specific to the commuting time period). In the weekend, the patterns are different with a maximum of movement on Saturday evening followed by very low movement on Sunday morning.

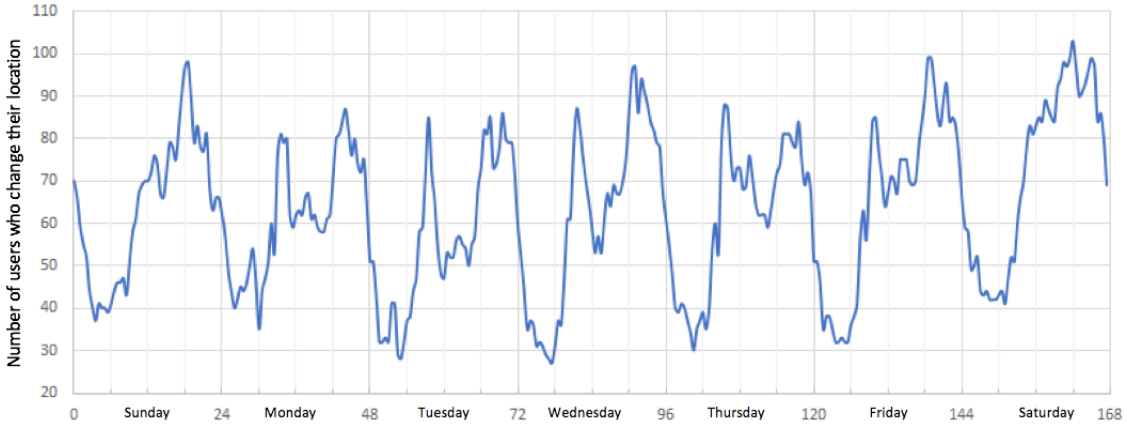


Figure 4.6: User movements over the weekdays

4.4.5 Exploring alternatives for carpooling scenario

In order to perform a what-if analysis experiment, we selected from the users movement distribution over the week presented in Figure 4.6 the day and the hour of the weeks the maximum number of users was moving. There are 112 users travelling around 7 pm on Saturday.

The experiment consists on finding compatible users for a carpooling activity, using the users profiles from the learning of their five years of data. According to [202] the condition for a group of users to carpool is that the distance in time and space at departure and arrival to be minimal in order to avoid the driver to make a big detour.

The first step was to calculate the distance in time and space between the 6 most probable locations between each user, using the indicator developed in [202]. The goal is to create a score between all the users which will be used as a compatibility classifier between the selected users. In order to use a single measure unit for the distance in time and space, we transformed the geodistance in time, using an average speed of 50 km/h.

The second step is to explore all the possible alternatives for finding users compatible for carpooling, using parallel simulations in MWGs. The computation was done for a hypothetical departure (Saturday at 7pm) and arrival (half an hour later). Our aim is to find users that belongs to the same cluster in both times, which are then compatible for carpooling.

The cluster analysis with the results from the departure time can be visualised using clustered graphs as presented in Figure 4.7.

Compatible users are linked and clustered in the same group colour. For the users that are represented with the grey colour in the centre of Figure 4.7, no compatible users for carpool was found at the selected day and hour.

The same experiment is also performed half an hour later and the results are presented

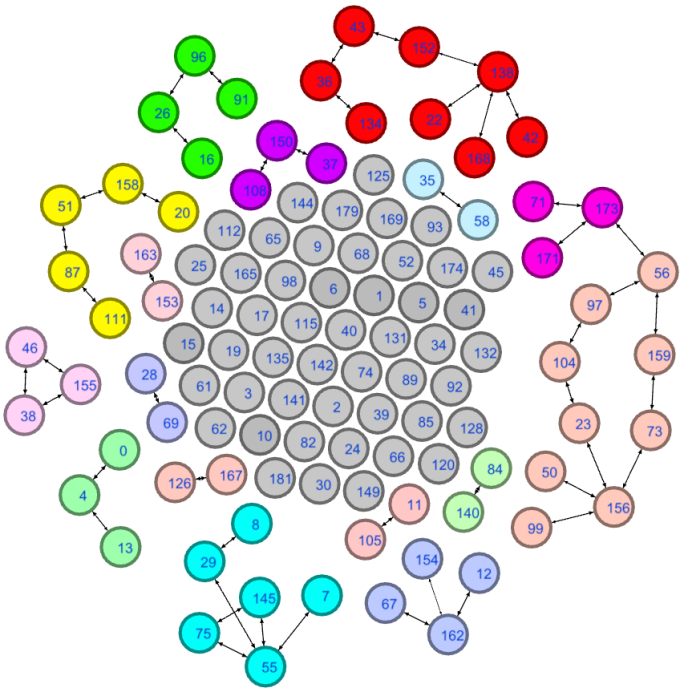


Figure 4.7: Clustering of compatible users for ridesharing departure at 7pm

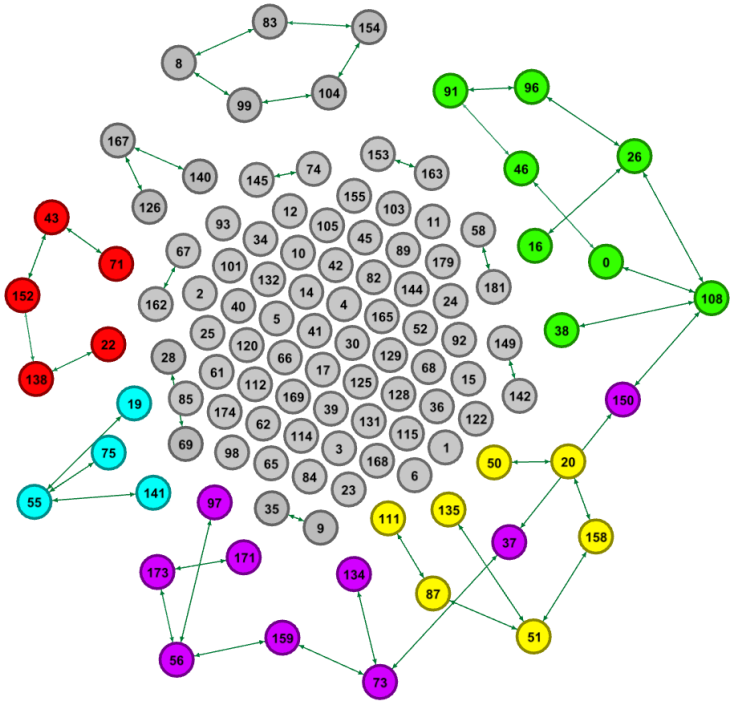


Figure 4.8: Clustering of compatible users for ridesharing at 7.30pm

in Figure 4.8. We can observe five coloured groups, represented by users that are clustered in the same group in both Figure 4.7 and 4.8.

Interestingly, the clusters evolves over time and new users can be added or removed from a group. If we look *e.g.,* at the green cluster in both figures, more users are compatible at the arrival time. This means that further analysis can be done if some of the new added users can be picked up along the trip path.

4.5 Related work

Although ridesharing can provide many benefits, there are still challenges that have restricted its widespread adoption [92]. The research community confirms that there is a need for scalable spatio-temporal stream computing frameworks that can operate big data streams, clustering and queries [93]. There are technologies and methodologies for implementing parts of the features required for the smart mobility systems. For example, Hadoop [3] and Spark [224] have powerful data analytics. Graph representation and processing solutions are offered by Neo4j [8] and GraphLab [141]. What-if analysis combined with hypothetical queries are explored in different database communities [36]. Different methodologies and frameworks have been also proposed for parts of the CM requirements. Mapping and clustering have been explored using data from social media [59]. The existing literature on activity analytics uses synthetic simulations [34]. More recent attention has focused on the human mobility patterns and predictability [173]. Our implementation offers the combination of all the above solutions in a single complete framework that can be used for CM applications.

4.6 Conclusions

We proposed a novel data modelling framework over temporal graphs that can be implemented in the ITS domain. We explained how our implementation can efficiently solve complex scenarios with multiple optimization levels of objectives in the CM applications. The presented implementation introduces, for the first time in the transportation domain, the capability to merge the discrete simulations and statistical results in a single framework.

The results obtained from the performed experiment can be used by a future recommendation system. If the system learns the users' profiles, in the case that an user requests a ride, the search will be performed first on the cluster where the user belongs to. The advantages over classical solutions are the speed of computation and the ability to deal with the missing data which can be replaced by the highest probabilities obtained from the cluster analysis.

The presented framework can be used in any transportation related problem that deals with large-scale datasets and ML algorithms. The algorithms library, the real-time geospatial and timeline visualisation components make the framework a complete

tool for the transportation engineers. Moreover, our implementation is able to unify the functionality development phase with the implementation and final production in an single stage.

Although the results from this chapter presents a framework that can manage large-scale datasets and multiple users in the same time solving efficiently complex tasks, a collaborative mobility recommendation system requires a more user-centric approach. This is important as each user has different preferences when using shared mobility services. The contribution from the following chapter will continue the research in this direction, using the concept of users' profiling. This method is able to extract individual user's travel behaviour from raw data which can be used by a recommendation system to propose additional actions that the user can perform in order to increase the chances to find compatible users, *e.g.,* to reschedule the departure hour or position. This will increase the chances to be included in a cluster and to be compatible with other users for an eventual ride sharing.

5

An user-centric approach for dynamic profiling of travel habits and visited locations

In this chapter, a scalable method for dynamic profiling is introduced, which allows the extraction of users' travel behaviour and valuable knowledge about visited locations, using only geolocation data collected from mobile devices. The methodology makes use of a compact representation of time-evolving graphs that can be used to analyse complex data in motion. In particular, we demonstrate that using a combination of state-of-the art technologies from data science domain coupled with methodologies from the transportation domain, it is possible to implement with the minimum of resources, a demonstration of autonomous sharing mobility services (i.e. long term and on demand parking sharing, combinations of car sharing and ride sharing) and extract from raw data, without any user input and in near real time valuable knowledge (i.e., location labelling and activity classification).

The content of this chapter has been partially submitted to the following journal and its content is unpublished to date:

- *A Data-Driven Scalable Method for Profiling and Dynamic Analysis of Shared Mobility Solutions*
  Bogdan Toader, Assaad Moawad, Thomas Hartmann, Francesco Viti
  IEEE Transactions on Intelligent Transportation Systems (T-ITS)

Contents

.

5.1 Introduction

In this chapter, we propose a method for dynamic and near real time profiling of travel behaviour in time and space, using data in motion. In our case, profiling means the extraction of user habits for visiting a specific location. The real time aspect is a mandatory requirement of some specific applications like safety applications or real time control applications. In our days, the same thing applies also for shared collaborative systems (*e.g.,* car sharing, parking sharing) where a large number of people and goods are moving at high speeds and solutions to combine them in an efficient way must be provided in an order of magnitude of milliseconds (*e.g.,* on the move ride sharing applications which dynamically search for additional passengers while the vehicle is moving). As the status of the involved entities (i.e. car, passengers) change continuously, reasoning processes typically need to analyse and compare the current context with the historical patterns extracted and autonomously provide actions.

In this sense, the proposed methodology makes use of machine learning techniques to automatically build the profile as soon as the data becomes available and proposes efficient techniques to store the results in a temporal index for fast access. Thus, the contribution of this study resides in the dynamic profiling of travel behaviour, which can contribute to multiple transportation problems, such as collaborative services, location classification, prediction, travel habits understanding and changes of habits.

The remainder of this chapter is organised as follows. First, section 5.2 presents a background of this study making the link with previous work and the minimal information needed for a proper understanding of the contribution and challenges involved. We then describe the proposed methodology for activity profiling and classification in section 5.3. The evaluation of the proposed method is presented in section 5.4, alongside with practical usage examples in section 5.5 and future work in section 5.6. We conclude the paper with a discussion of future directions in section 5.7.

5.2 Background

Multi-dimensional and dynamic profiling requires technologies which allow the fast processing, indexing and querying of big data sets. In the remainder of this subsection, we present the data modelling framework which incorporates graphs and time series in multi-dimensional data models, alongside with the major technological implementation challenges.

The GreyCat [71] framework, formerly known as KMF [90], presented in Chapter 4, is a solution for analysing complex data in motion at scale with temporal graphs [107]. There are a number of required features (*e.g.,* modelling with graphs, temporal aspects, what-if analysis exploring different alternatives) presented in [201], which makes the

selected framework suitable for intelligent transportation systems and, to the best of our knowledge, the most efficient solution suitable in the current study.

Another feature that is important in big data systems is the ability to lazy load nodes, meaning to load into main memory only the necessary data that need to be processed rather than to load and query each time the entire dataset. Naturally, many analytic tasks are processing only parts of the dataset. This also counts for the case study of this paper. Therefore, we suggest to load data, i.e. the nodes of our data graph only on-demand, while the graph is traversed. As an example, even if high accuracy datasets are available at an order of a few meters, if the application needs to profile to a maximum of *e.g.,* one kilometre accuracy, there is no need to load and process the data at a higher resolution. This will save both resources and time.

To achieve this, the graph is decomposed into a set of key/value pairs, representing the graph. More specifically, every node value contains its attributes and the relationships to other nodes (in form of sets of IDs). Then, a node value is stored/loaded via an identifier in/from a key/value store. Depending on the requirements of an application, different key/value pairs stores can be used as backend: from simple key/value stores to scalable, replicated and distributed high-frequency stores. To ensure concurrency and fault tolerance, we use a per-node lock policy and we rely on the concrete underlying storage implementation to ensure concurrency and distribution.

If in the previous chapter the proposed framework is used for the first time in the transportation domain to find possible groups of users that can use a ride sharing system using data at rest, the current work proposes a more user centric approach that can handle data in motion at scale for multiple applications (*e.g.,* parking sharing, location classification, non-recurrent trips profiling).

As the data in motion must be indexed and complex searches must be performed in near real time, a map-reduce-like approach is used. The MapReduce [66] paradigm aims at processing and generating big data sets with a parallel, distributed algorithm on a cluster. At the same time it is a flexible data processing tool [75] which simplifies data processing on large clusters [74] and allows the implementation of machine learning technologies on multicore processors [66]. As the name suggests, MapReduce is composed of two distinct methods: Map() and Reduce(). The former has as objectives to filter and sort, the later performs a summary operation. At the infrastructure level, MapReduce manages the communication and data transfer between various parts of the system, which is an essential feature for complex and dynamic ITS systems.

5.3 Methodology

5.3.1 Generic overview

In this chapter, we present a new way of profiling multi-dimensional and temporal data, specifically designed to deal with large quantities of data—in live and with different physical constraints, such as limited memory or processing power (as is the case of

nomadic devices like smartphones).

Our methodology is generic in the sense that it can use any specific profiling algorithm that uses a tree-like structure to divide a parent space into two or more children subspaces. For instance: binary trees, quadtrees, octrees, and K-d trees are all easily implementable and can be integrated in our methodology.

The proposed architecture shown in Figure 5.1 has three independent layers. The base layer is the lowest level which represents the *Raw data layer*, dealing with data management (*e.g.,* collect and store data). The *Processing layer* deals with data processing to produce well structured and fast to query spatio-temporal profiles. Finally the highest layer is the *Application layer* where any specific transportation problem can be translated in high level profile queries. The main advantage of the proposed architecture is that the profile layer is built once and then shared across several transportation applications hence reducing the required infrastructure and resources.

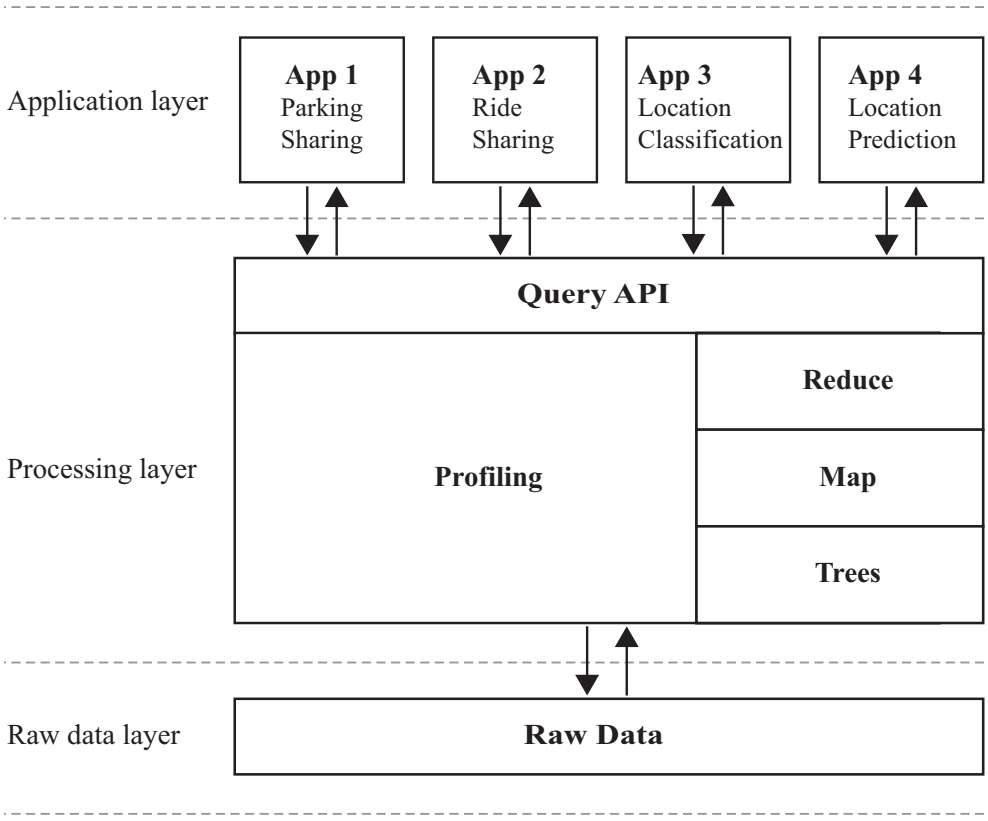


Figure 5.1: Architecture and abstract layers

5.3.2 Terminology

As the current work includes terms from different domains, it is necessary to define the terminology that will be used through the entire work.

- ***Tree***: a directed acylic graph starting from a root node.

- ***Space coverage***: the N-dimensional min-max vectors that define the boundaries of the space covered by the subtree.
- ***The root node*** is the top node in the tree. It covers the widest N-dimensional space of the tree.
- ***Child node***: a node directly connected to another node when moving away from the root node. Child nodes have always a smaller space coverage than their parent nodes.
- ***Parent***: the opposite notion of a child.
- ***Leaf node***: a node without children.
- ***Siblings***: a group of nodes with the same parent.
- ***Degree***: the number of children of a node.
- ***Path***: a sequence of nodes connecting a node with a descendant.
- ***Level***: the number of connections between the root and the node. The root node is of level 0.
- ***Size***: of the tree is the total number of data indexed in the tree.
- ***Height***: the height of a tree is the maximum level reached by its nodes.
- ***Resolution***: the smallest space coverage allowed for the leaf nodes. It is an N-dimensional vector representing the minimum difference allowed between the minimum and the maximum on each of the N dimensions.
- ***Number of dimensions*** represents the number of different features we want to profile (*e.g.,* day of the week, time of the day, geolocation). By default, the proposed architecture support until 32 dimensions that are easily extensible to 64.
- ***Max buffer size*** the maximum size of the data stored in a node before creating a sub-level of child nodes.
- ***Timeline***: a sequence of ordered timepoints.
- ***Temporal resolution***: represents the maximum quota in time for each profiling tree before creating another tree.

5.3.3 Live profiling, indexing and preprocessing

As mentioned in the previous subsection, the live profiling process is composed by several steps and includes several layers, as can be seen in Figure 5.1. In this subsection we describe specifically the preprocessing step, in which the data in motion is indexed as soon as it is received from specific sensing systems (*e.g.,* mobile devices) or databases.

The following example deals with location data represented by points in the geographical space, represented in Figure 5.2. We describe how each tree structure is created based on the space partitioning and indexing of each quadrant, from the root (*Level 0*) to the leaf level (in our example Level 3). It is important to stress that the indexing and profiling methods are completely independent of the final applications that will access and use the data on the application layer.

The profiling methodology can be summarised as the following chronological steps, which are continuously performed as soon as new data is available:

a) Start with an empty timeline.

b) Create the first profiling tree once the data is loaded from a dataset or received through a sensing system, as can be seen in Figure 5.2, at *Level 0*.

c) Once the buffer is full at the root *Level 0* (reached the *max buffer* limit of data stored at the node level, defined at design time), create child nodes of *Level 1* and redistribute the data into the corresponding sub-spaces. Any new data received at *Level 0* will be automatically forwarded to *Level 1* sub-spaces. At this step, the node at *Level 0* is transformed from a node that store data in a *router node*, defined as a node that has no data but acts as a *path* that connects the node with it descendant sub-spaces.

d) Each subspace has its own buffer, and divides the parent dimension boundaries by two or more on each dimension. In the case of geolocation data, Figure 5.2 shows how the space is divided in quadrants. At *Level 1*, there are four subspaces: (1) A, (2) the quadrant formed by B, C, D, and two empty, (3) the quadrant formed by E, F, G and one empty, (4) H.

e) Repeat steps 3 to 5 recursively until one of the following conditions are met:

   (a) The temporal resolution of the current profiling tree has expired. As an example if we set the maximum temporal resolution to one hour, even if the max buffer size was not reached, a new tree will be created.

   (b) The tree reaches the maximum allowed size. This is a requirement to keep the process as fast as possible, as if we allow big trees it's harder to search inside.

   We can observe in Figure 5.2 that the nodes (2) and (3) created at Step 4 on *Level 1* are split again in four children and transformed from nodes with data in router nodes. The same process continues also at *Level 3* until one of the above conditions are met.

f) Once a tree is complete, it is stored and the process continues with the creation of a new tree. As can be observed in Figure 5.4 the new tree will have a new timepoint and the entire process from 3 to 6 will be repeated.

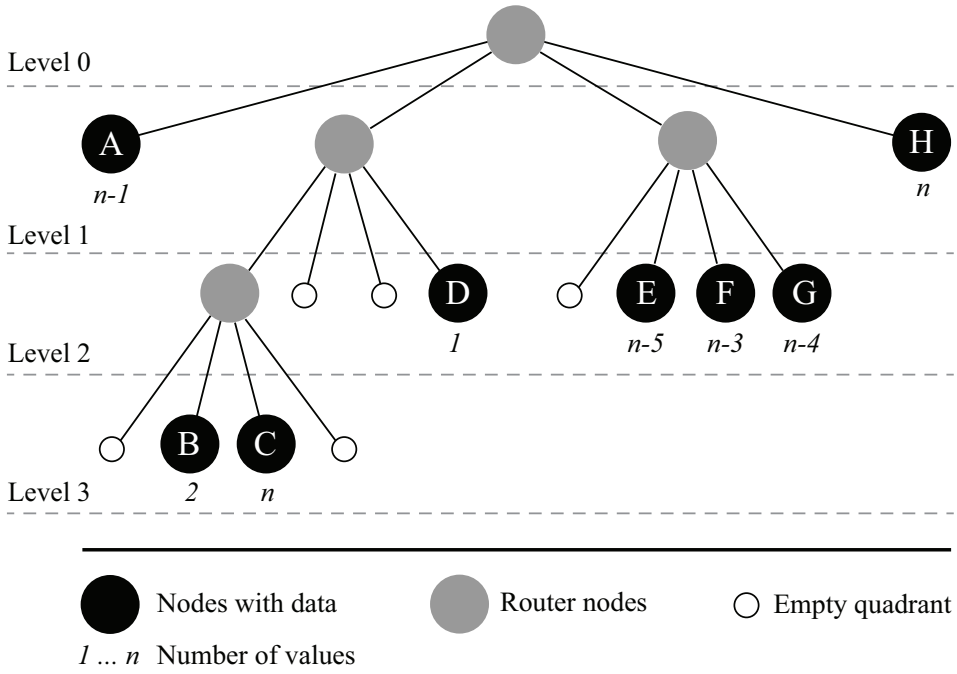


Figure 5.2: Profiling tree structure

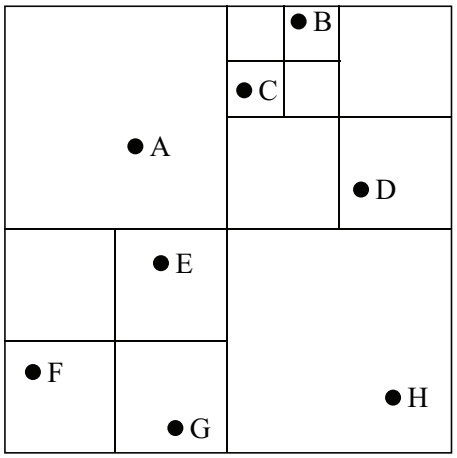


Figure 5.3: Profiling space partitioning for geolocation data

5.3.4 Querying and postprocessing

The multi-dimensional and temporal features of the proposed profile offers several ways to query it in order to allow a wide range of applications to be built on top of it. A query can specify a range in time, specific days and hours of the week, a level of precision in the multi-dimensional space, and can ask either for all the results within a specific range, or the top N results from a specific complex query. The entire flow of the query process is described in Figure 5.4 and will be described in the remaining of this subsection.

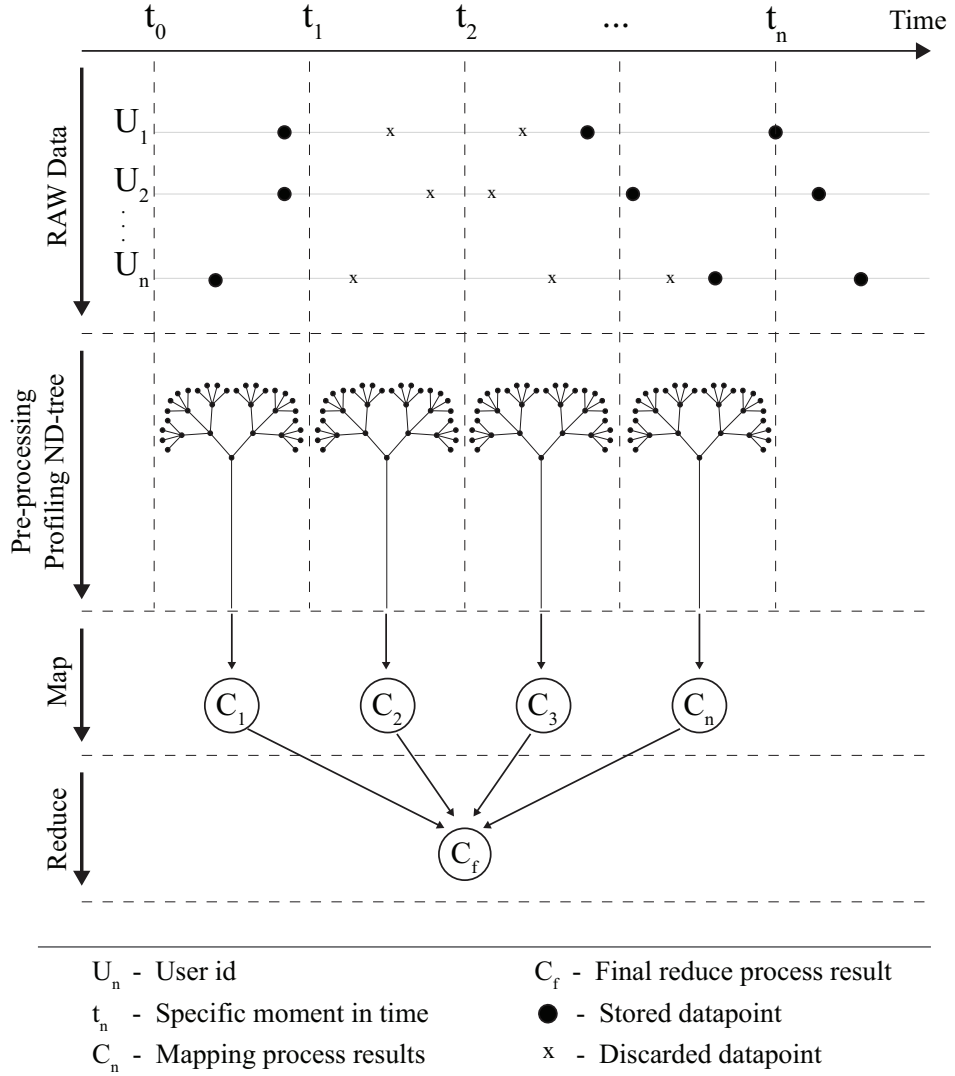


Figure 5.4: Data structure, query and process flow

In order to demonstrate the process flow, in the remaining of this section we present the example of a smart mobility shared system (*e.g.,* car sharing, carpooling, parking sharing) that provides the geolocation data for a number U_n of users through a smartphone application using the integrated sensing system. This type of user centring system is indexing and profiling each user in time and space. The system can then perform specific complex queries which must return a result in an order of milliseconds.

For example, a ride sharing system can perform a query to get all the locations that a user visited in the last two years in a specific geographical region, in specific days of the week and specific time intervals during the day, with a specific geographic resolution. This can be useful to find possible matches with other users that have similar profiles.

Another example can be a location classification application that extracts the user visit pattern for all the visited locations that have been visited at least N times during a specific time period and a specific geographical area. This can be useful to instantly filter trajectories points that don't represent a specific location and to classify the filtered locations based on specific location duration and time of the day.

Moreover, a query can become even more complex and can be used to show the top N locations visited, in a specific day of the week and interval of the day. This query can be used to detect the most visited locations and to quickly detect *e.g.,* home and work location, in order to propose a personalised itinerary end *e.g.,* at the user's home or to be used to match users that are compatible for parking sharing.

It is important to mention that for all of the above examples and independent of any other application, the same process flow is applied. This is described below:

a) The data is captured and processed according to the temporal resolution, in specific time intervals t_n. This temporal resolution must be set in the very beginning and represents the minimum time interval needed to index two specific trees.

   For example, if from the applications domain it is known in advance that no application that uses the profiling data will need a higher time resolution than one minute or a higher geographical resolution than four meters, there is no sense to set this limit lower. A lower time interval than the minimum required will also require more resources or time to process the entire flow, providing also redundant data. There is no upper limit but just the one given by the indexing method (*e.g.,* for the geographical space, a quadrant, no matter of the dimension in measurement units). This dynamic is important to mention as in some applications like carpooling it is possible to perform different queries with different parameters and to increase/decrease the resolution to determine *e.g.,* which specific routes the user uses. This information can be useful to calculate the compatibility for matching different profiles.

   There are some important aspects to mention regarding the data management and temporal resolution. First, if a user is changing the location between two consecutive t_n, t_{n+1} data points, the geolocation is stored in a node. Second, if the user is in the same location for more than two consecutive time intervals, the same data is not replicated through consecutive timepoints but is discarded, represented in Figure 5.4, Raw Data layer with x. Thus, for a visited location, only the arrival and departure timestamp are stored, which helps in cleaning the dataset of duplicate values and reduce the required storage resources. Third, if at any time t_n a query is performed and no points are found at t_n, the data from t_{n-1} will be returned, the process begins to backtrack until a stored point will be found.

b) *Profiling phase*: the trees are created for each time interval, following the methodology from section 5.3.3.

c) *Map phase*: when a query is performed, the query is divided into several sub-queries touching several trees and several sub-spaces (*e.g.,* return the top n points from a specific time interval, specific day of the week and hour, from a specific geospace, with a specific accuracy). The search phase can be distributed among any number of computation units and threads as needed, according to each application and specific domain requirements.

d) The results are then collected, then the *Reduce phase* does the synchronisation, waits for all the running threads to finish, removes the duplicates, sorts the results, and does a final post-filtering if needed.

Another important feature in the context of profile sharing with multiple applications is the ability to have a high level of parallelism. This feature brings important advantages:

a) All queries can run in parallel: this is an important requirement when multiple applications share the same profiling layer, multiple queries can be performed, in the same time, on the same tree indexing.

b) Each query can be mapped to one or several profile trees according to the targeted time range of the query. The search within the targeted trees can be done as well as in parallel.

c) Since a query can involve several sub-spaces within a tree, the search within these sub-spaces can be executed in parallel as well.

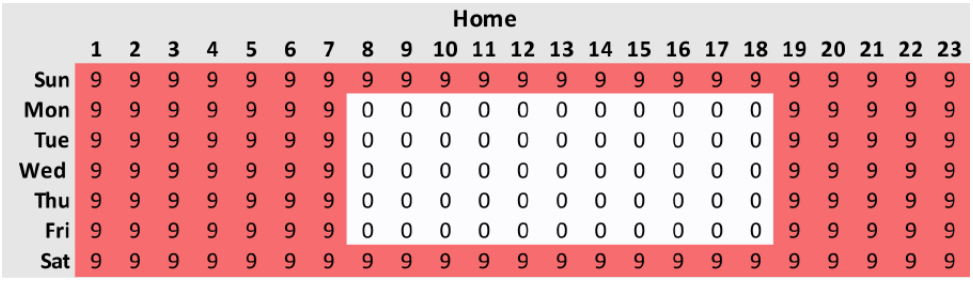
5.3.5 Location visit pattern extraction

In addition to filter in near real time the visited locations based on spatio-temporal complex queries, the final result of the entire process flow (i.e. starting from indexing and preprocessing, live profiling, querying and postprocessing) is used also to extract the weekly activity pattern visit for each location visited. This is done by clustering all the visit records from the time range specified in the query parameters in a matrix with an dimension of 24 hours an the seven days of the week. Each matrix element represents the number of times a person visited a specific location, normalised by 1000, as can be seen *e.g.,* in Figure 5.6.

5.3.6 Location profiling and activity classification

An earlier exploratory study from a previous work [199] presents the extended methodology behind the classification methodology. To summarise, the classification and labelling of each location is done by computing the Euclidean distance (between 0 to 1) from a generated training matrix representing a known location type (*e.g.,* home,

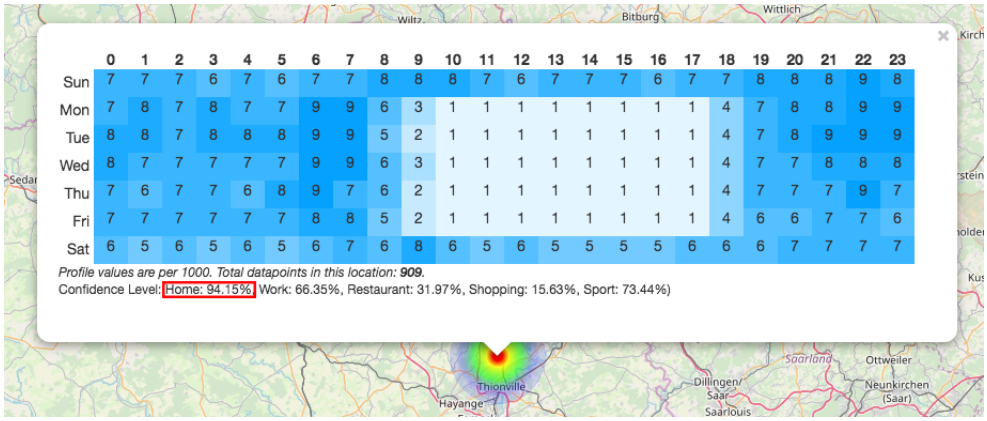
work) and the matrix obtained for each location visit pattern. The smaller Euclidean distance is obtained, the higher probability is that the profile of a specific location can be labelled as a type of a specific location. Figure 5.5 shows an example of the training matrix for a home location, where the total sun of all values must be normalised by 1000.



Home	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23
Sun	9	9	9	9	9	9	9	9	9	9	9	9	9	9	9	9	9	9	9	9	9	9	9
Mon	9	9	9	9	9	9	9	0	0	0	0	0	0	0	0	0	0	0	9	9	9	9	9
Tue	9	9	9	9	9	9	9	0	0	0	0	0	0	0	0	0	0	0	9	9	9	9	9
Wed	9	9	9	9	9	9	9	0	0	0	0	0	0	0	0	0	0	0	9	9	9	9	9
Thu	9	9	9	9	9	9	9	0	0	0	0	0	0	0	0	0	0	0	9	9	9	9	9
Fri	9	9	9	9	9	9	9	0	0	0	0	0	0	0	0	0	0	0	9	9	9	9	9
Sat	9	9	9	9	9	9	9	9	9	9	9	9	9	9	9	9	9	9	9	9	9	9	9

Figure 5.5: Generated matrix for location classification training process

Similar matrix can be generated for other types of locations and activities (*e.g.,* work, restaurant, shopping, sport). Figure 5.6 shows an example of the classification result for a location that has 94.15% confidence level that is a home location.



	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23
Sun	7	7	7	6	7	6	7	7	8	8	8	7	6	7	7	7	6	7	7	8	8	8	9	8
Mon	7	8	7	8	7	7	9	9	6	3	1	1	1	1	1	1	1	1	4	7	8	8	9	9
Tue	8	8	7	8	8	8	9	9	5	2	1	1	1	1	1	1	1	1	4	7	8	9	9	9
Wed	8	7	7	7	7	7	9	9	6	3	1	1	1	1	1	1	1	1	4	7	7	8	8	8
Thu	7	6	7	7	6	8	9	7	6	2	1	1	1	1	1	1	1	1	4	7	7	7	9	7
Fri	7	7	7	7	7	7	8	8	5	2	1	1	1	1	1	1	1	1	4	6	6	7	7	6
Sat	6	5	6	5	6	5	6	7	6	8	6	5	6	5	5	5	5	6	6	6	7	7	7	7

*Profile values are per 1000. Total datapoints in this location: **909**.*
Confidence Level: Home: 94.15% Work: 66.35%, Restaurant: 31.97%, Shopping: 15.63%, Sport: 73.44%)

Figure 5.6: Example of home location classification

Evaluation and usage examples of location profiling and classification are presented in sections 5.4 and 5.5.

5.4 Evaluation

5.4.1 Datasets description

In this study, two types of datasets were used in order to perform different evaluations of profiling results. Each dataset has different purposes and requirements, as follow.

First, a dataset for which we have the ground truth was used in order to evaluate the location classification and labelling accuracy based on the profile extracted for

each location visited by the users. For this type of evaluation, the dataset must be accurate with less GPS errors and the results must be validated by the respondents. As the validation is done on the individual level, we didn't use a huge data set for the validation. The dataset used was based on data of 17 users collected using Google Maps, from the University of Luxembourg, it is individual based and each respondent was able to easily test and validate online and the results [165]. Moreover, the data comes already error-filtered as the data is collected using not only the GPS sensors from the smartphones but also a fusion of sensors like Bluetooth, Wifi, motion sensors, which are used to validate the location even when the GPS signal is poor *e.g.,* inside the buildings.

Second, a larger dataset was used in order to evaluate the computational speed, performance when scaling and accuracy. For this type of evaluation, the most important aspects are the size of the dataset, i.e. the period of time covered and the number of users. This dataset must be very large, covering a long time period and provided for a large number of users in order to test the computational requirements when scaling. The dataset used was the the Geolife dataset [151]. This publicly available dataset contains around 24 millions of GPS points from China, collected in the Geolife project [233] from 182 users with smartphones and GPS loggers in a period of five years.

5.4.2 Location classification accuracy

The evaluation of location classification accuracy was performed by a group of 17 respondents who uploaded their GPS data exported from GoogleMap to the publicly online version of the tool developed in the current research [165]. The respondents were then asked if the home and work location were accurately detected. The results show that 100% of the respondents stated that the home and work location were correctly classified and labelled with the highest confidence level. Of course one can argue that the home and work location are trivial to be used as an example of classification but the scope of our evaluation is only to demonstrate that the proposed framework and methodology are able to automatically classify locations and activities performed in near real time, without any user input and only based on the GPS data. Chapter 6 will extend the evaluation and methodology also for other type of locations and activities (*e.g.,* restaurants, shopping, sport activities).

5.4.3 Computational speed

First, in order to test the computation speed of the proposed profiling method when scaling, multiple tests have been performed with different amounts of data. Figure 5.7 shows the results obtained when processing 12 different amounts of data, from 1 million to 24 million valid points. Moreover, as can be seen in Figure 5.7 the general trend line when scaling is close to logarithmic, something that confirms the complexity reduction explained in Section 5.2.

Second, a speed comparison has been performed between a classical linear computation

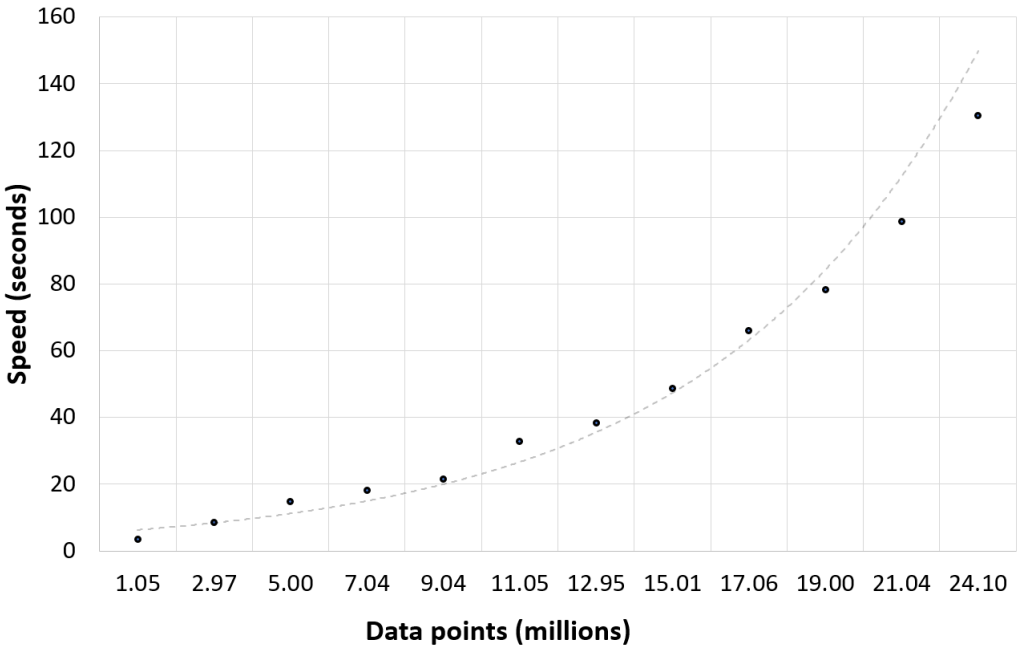


Figure 5.7: Computation speed when scaling

in comparison with a multi-tree profiling method, presented in Figure 5.8. In this experiment, 9 speed tests have been performed, with data from different number of users, from 20 to 180. The results are displayed on a log scale at y axis. The experiment clearly shows that the speed of using a multi-tree architecture is around 10^2 faster compared with a classic linear architecture.

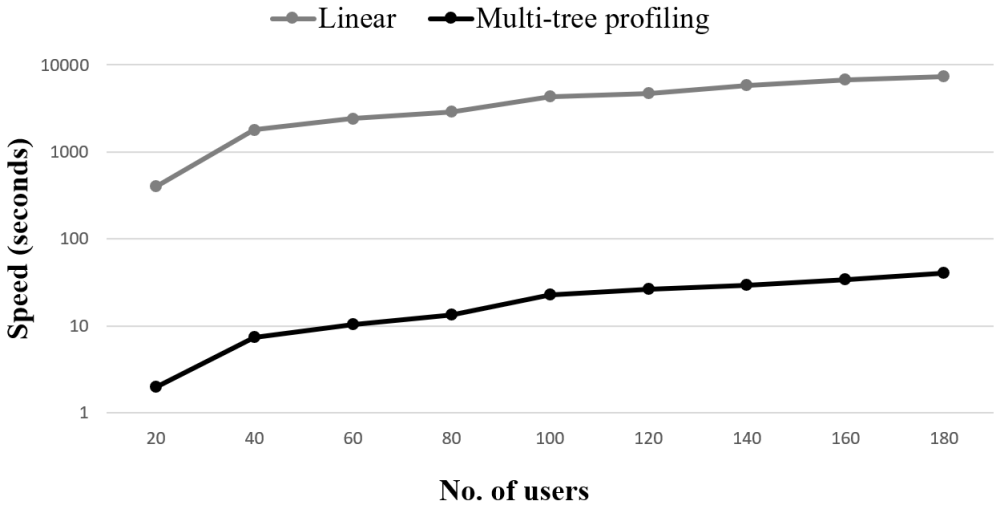


Figure 5.8: Linear profiling vs multi-tree profiling speed

Another important aspect in practical implementation of fast processing is the amount of resources needed. Most of the time, the bigger the size of the dataset is, more processing resources are needed in order to have the highest speed. The database research community identified Graphics Processing Units (GPUs) as the most effective co-processors for parallel data processing [46] mainly because the dataset is processed by hundreds or thousands of small CPUs nodes.

Table 5.1: Resources and performance comparison

Speed (ms)	Cost(EUR)	Hardware	Scaling	Method	Summary
0.5 - 20	40000	8x Nvidia Tesla K80	Linear in hardware	Load all x Tesla K80	BIG Data on BIG Hardware
36 - 51	2000	User PC	Log in time or hardware	Multi-tree profiling	BIG Data on Small Hardware + Smart Models
6895 - 7579	2000	User PC	Linear on time	Linear parsing or hardware	BIG Data on Small Hardware

To the best of our knowledge, one of the fastest massive parallel architecture is MapD [159]. Recent experiments show that massive datasets with billions of geolocation routes can be processed and visualised in milliseconds [145]. But everything comes at a cost. Table 5.1 shows a comparison of speed and resources needed to process the GeoLife dataset using MapD with a linear in hardware scaling method and a very powerful but costly hardware, compared with a multi-tree profiling performed by a user PC which is a logarithmic in time or hardware method.

5.4.4 Accuracy

The dynamic aspect of the profiling is given not only by the capability to load different amount of data asynchronously but also using different precisions (resolutions) from a minimum of $4,77$ x $4,77$ m (the minimum precision value of a regular smartphone) to a maximum of 5000 x 5000 km. The list of all the used resolutions can be seen in Table 5.2.

The profiling accuracy of any point $P(x, y)$ depends on the distance from the centre of the smallest sub-space resolution unit needed L, presented in Figure 5.9.

The error of any point $P(x, y)$ in the L_xxL_y space is represented by the Pythagorean distance between the point and the centre of the sub-space surface. Thus, the general mathematical error can be expressed as shown in equation 5.1:

$$\sqrt{(\frac{L_x}{2})^2 + (\frac{L_y}{2})^2} \tag{5.1}$$

The highest accuracy can be obtained if the point $P(x, y)$ is located at centre of the sub-space surface $O(L_x/2, L_y/2)$ while the highest error can be obtained if the point is located at the edge of the surface.

The average error is the double integration of all possible points within the L_xxL_y

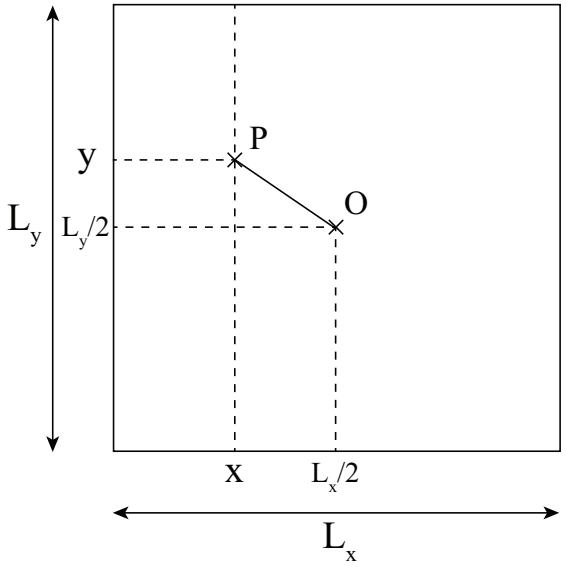


Figure 5.9: Error computation at the leaf level

space, divided by the area of the rectangle:

$$\frac{1}{L_x L_y} \int_0^{L_x} \int_0^{L_y} \sqrt{(x - \frac{L_x}{2})^2 + (y - \frac{L_y}{2})^2} dx dy \tag{5.2}$$

where L_x and L_y represent the width and height of the sub-space, x and y the coordinates of any point $P(x, y)$ from Figure 5.9.

Since the profiling is done for different sub-spaces sizes (resolution), the accuracy is highly correlated with the resolution (the smaller resolution the better accuracy) and the number of levels that need to be queried to reach from the root to the leaf level (more levels means higher details, but longer time to search).

In our specific case, the accuracy depends on the size (width and height) of each resolution. Using the GeoLife dataset, accuracy tests have been performed and the results are presented in Table 5.2 from a minimumof *4,77 x 4,77 m* to a maximum of *5000 x 5000 km*.

The experiments confirm that the dataset average error is very close to the mathematical predicted error, computed using Equation 5.1. The provided results can be used as a guideline for choosing the minimum resolution for each transportation application, based on the average and maximum error of each resolution. Thus, for any type of application it must be assessed if the average and maximum error are acceptable and tolerated in the domain. Different transportation applications require different precision and maximum errors. Table 5.3 presents an example of comparison for different applications with the precision and amount of data needed.

For some application like ride sharing, the accuracy is important as *e.g.,* the meeting point of different users that can share the same car cannot have large error. A study

Table 5.2: Accuracy table

Lx	Ly	Unit	Worst err.	Dataset avg. err.	Math. err.
4.77	4.77	m	3.372899	1.824923	1.824993
38.2	19.1	m	21.354449	11.330335	11.330761
153	153	m	108.187337	58.537250	58.537314
1.22	0.61	km	0.682000	0.361813	0.361873
4.89	4.89	km	3.457752	1.870842	1.870902
39.1	19.5	km	21.846398	11.589896	11.590412
156	156	km	110.308657	59.691685	59.685239
1250	625	km	698.771243	370.746174	370.771200
5000	5000	km	3535.533906	1913.060314	1912.992000

Table 5.3: Usage examples

Type	Precision needed	Maximum error	Amount of data needed
Ride sharing	High	< 50 m	Medium
Location/activity classification classification	High	< 50 m	Small
Parking sharing	Medium	< 150 m	High
Non-recurrent trips	Low	< 50 km	Medium

[31] shows that on average only about 60% of the passengers will accept to walk 150 meters for transit to another bus stop and 90% of the passengers will accept to walk 50 meters. The same strict requirement has also the location classification (*e.g.,* home, work, shops, restaurants) or activity classification (*e.g.,* sport, shopping) systems. In order to keep a higher quality of service and a higher rate of user retention, the maximum error must be lower than the acceptable distance that a passenger has to walk if *e.g.,* the suggested location/meeting point is not precisely in the location designated by *e.g.,* a recommendation system/trip planner.

For other transportation problems like parking sharing the error can be a bit higher as is not unusual to park the car and walk for a decent distance until the destination, but again under the limit of the maximum user's tolerated error [206]. There are also other applications where the error can be bigger, there is no need of very detailed profile and the error can be much higher, of an order of a dozens of kilometres. This is the case of non-recurrent trips analysis *e.g.,* holidays or business trips where anyway the clusters and visualisation are much bigger than the above examples.

The capability to be versatile in order to handle various application requirements in the same system represents a requirement in a shared architecture. In the next section will be presented practical usage examples of all the applications compared in Table 5.3.

5.5 Usage examples

The proposed profiling methodology can be applied in various topics within the ITS domain, due to the features presented: first the compatibility with various data types and secondly the dynamic characteristics of modelling the analytic parameters. The former feature allows the usage of any mobility data type, provided as a data collection (*e.g.,* databases of entity attributes) or as data input which is sent from sensor systems (*e.g.,* geolocation, additional real time information like traffic condition). The latter feature provides the possibility to analyse and process the data using dynamic parameters, which can be modified at any moment, without the need of re-processing the entire dataset. This is an important feature, especially applied in the analysis of smart mobility services, which will be exemplified in the rest of this section.

Collaborative mobility services (*e.g.,* ride sharing, car sharing, parking sharing) represent one of the best case studies of the methodology presented in Section 5.3. The dynamic profiling provided by the proposed methodology can be used to asses the profile similarity of different users in order to find users that can share existing mobility resources and which can collaborate on various combinations of sharing services. This is a NP-hard problem as different types of data collected from users' smartphones represent a new dimension that adds an additional grade of complexity on finding efficient solutions for combining users and transportation resources in complex collaborative systems *e.g.,* finding compatible users to share a car or a parking place, taking into account the user's schedule and personal preferences - everything in a tolerated time interval of milliseconds. Each dimension is represented by the properties of those entities, combined with the types of queries which are performed, *e.g.,* day, hour, location, age, sex etc. Some of them have sub-dimensions, *e.g.,* the location where users perform activities can have as sub-dimensions the starting and ending hour of an activity, the geographic coordinates of the location (represented by the latitude and longitude), and the radius of the geographic space that represents a location. In this study, the multi-dimensional profiling refers to the concept of profiling on all of these dimensions, where each entity has specific properties.

5.5.1 Parking sharing

A group of two or more users can share the same parking place if they use it at different times of the day. In other words, the more dissimilar the users' profiles are for the same location, the higher compatibility for parking sharing there is. The dynamic character of the proposed profiling methodology makes possible the assessment of parking sharing both for (a) planned long term parking sharing and (b) short term or ad hoc parking

sharing. In order to demonstrate the usefulness of a flexible, dynamic and fast profiling framework we present the following case study.

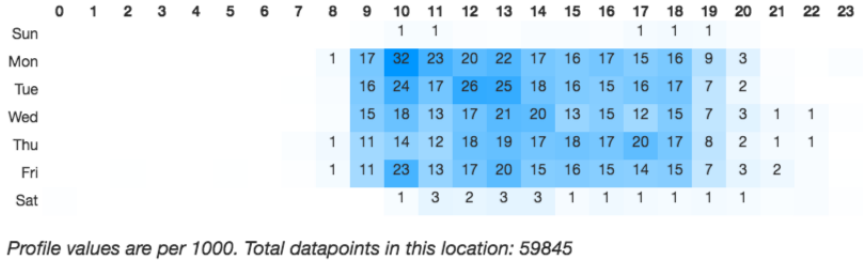


Figure 5.10: Profile of User 1 for long term parking sharing

Figure 5.10 shows the profile of *User 1* which works in the proximity of the home of *User 2*. *User 2* is part of a Peer-to-peer (P2P) parking sharing application. As we can observe, for *User 1* the highest probability to be in the parking location is from *9 AM* to *7 PM*, from *Monday* to *Friday*. Similar, Figure 5.11 shows the profile of *User 2*, which has the lowest probability to be in the same location when *User 1* is in the same location.

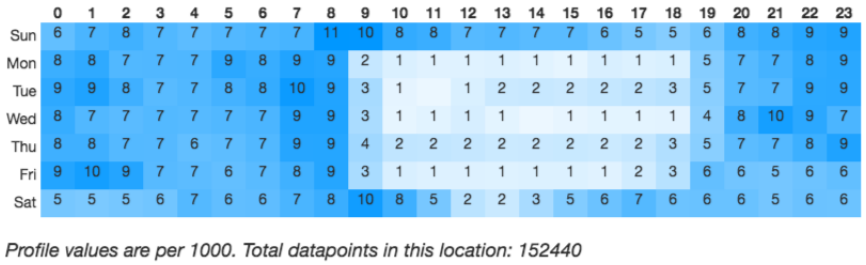


Figure 5.11: Profile of User 2 for long term parking sharing

With this information, a system can match the two profiles with highest compatibility index (*i.e.,* the highest Euclidean distance between the profiling matrices) to share the same parking location as they partly overlap. Moreover, this is done without asking the users any prior information but profiling their behaviour, extracting their pattern to visit the location, classify the location and match profiles that are synchronised for specific sharing services.

The results denote a good example of how the presented profiling methodology can be used to asses the compatibility for long term parking sharing of two or more users using *e.g.,* specific indicators for collaborative mobility between individuals presented in Chapter 3. More precisely, the profiling can be used to search in a specific region, users that have profiles that match other users for specific applications/sharing services. Nevertheless, there are some conditions that must be met to have an accurate long term profiling.

First, there should be enough time data in order to have an accurate profiling. This will ensure that the location profile has a specific pattern over time and is not just a location that is visited randomly. *Home* and *work* locations are typical examples of locations that have a specific pattern over time.

Second, it is not required for the location accuracy to be extremely precise, as in the case of long parking duration, people are likely willing to walk for a decent location from the parking lot to destination [206].

In the same way, a P2P parking sharing service can be used also for ad hoc or instant parking sharing, presented in the following example. Particularly, if an user is part of the P2P parking sharing application and during a trip notifies the application that must perform a stop for a specific time period in a specific place, the application can instantly search for other users in the system which has free parking slots during that specific time interval. In order to test this case study, using the Geolife dataset described in section 5.4.1, we took a random user and a random visited location and perform a search for compatible users to simulate a match of an ad hoc P2P parking sharing request. Table 5.4 presents the results of different searches, at different resolutions.

Table 5.4: Ad hoc p2p parking sharing matching

Lx	Ly	Unit	Users' matched	Max error
4.77	4.77	m	0	3.372899
38.2	19.1	m	0	21.354449
153	153	m	3	108.187337
1220	610	m	7	682.00000
4890	4890	m	11	3457.752000
39100	19500	m	24	21846.398000

As can be observed, at very small resolutions (i.e. $4.77x4.77m$ and $38.2x19.1m$) no compatible user has been found, as the search is too detailed. When increasing the resolution to $153x153m$, three compatible users were found and the maximum error can be $108.187337m$ which is acceptable. If we increase the resolution, more compatible users are found but also the maximum possible error increases, to the extent that some results are not relevant, as the walking distance will be then too long and most likely not acceptable for the user to walk. In this case, we can argue that the best resolution would be in this case $153x153m$, which can give the best results regardless the maximum possible error.

This concrete example demonstrates the capabilities that the proposed profiling methodology offers: fast processing large quantity of data, on demand and in near real time, coupled with the ability to extract user insights, behaviour and travel pattern with minimum of computation, storage resources and user input. This is of great importance for the next generation of AI autonomous travel planners and sharing services, as in most of the cases, the data collection and processing are done through the passenger's mobile device. In a matter of seconds, the system must process years of geolocation data, extract the insights, user habits, preferences and provide reliable services. The fact that now it is possible to use the online tool [165] on any browser without installing any software and all the computation is done locally on the device is inline with the requirements of mobile devices which have limited autonomy and

computation resources. In the same time fits also the mobile applications' user preferences because asking continuously user input information is no more applicable and sustainable in our days.

5.5.2 Ride sharing

The profiling of users' mobility for the days and hours of the week is an important information that can be used for a recommendation systems in order to analyse which users can match for ride sharing/carpooling. There are some conditions that should be met in order to organise a ride sharing between two or more users, such as the departure and arrival position to be suitable for all the participants and that the departure and arrival time to synchronise matching at best their schedule. The latter condition can be assessed by analysing the probability to be in a specific location, by the days and hours of the week and used in a collaborative mobility system [202].

In order to exemplify this case study, we searched in the small database presented in section 5.4.1 for compatible respondents that can match a carpooling service. Figure 5.12 presents the extracted weekly heatmap of time spent in the residence for two neighbours (User 1 and User 2) that are working also in the same area, as can be seen in Figure 5.14. This is a typical situation where users can participate in a long term ride sharing, as their schedule is pretty fixed in most cases.

As can be observed from the heatmap, they are at home typically outside of the working hours. Moreover they both leave the house during the week around *9 AM* and they return at home around *7 PM*, resulting in a good synchronisation.

Figure 5.13 shows the schedule that they have for the work location, make it suitable for a long term carpooling as they can share the same car for commuting to work.

In the same time, we noticed also that also User 3 works close to User 1 and User 2 and when available, User 3 can join the ride sharing. Analysing Figure 5.14 we can observe that in order to pick-up User 3, the trip must be rerouted and will be four minutes longer, but another passenger will be picked-up in the same car and there will be a car less on the road. The same principle can be applied to the entire community, which will result more people in less cars and less traffic congestion.

This concrete example shows how the profiling can be used both for long term carpooling and short term ride sharing as combined services. It is important to observe that using the proposed profiling methodology, all the necessary steps for matching people and sharing services (i.e. location visit pattern extraction, search of compatible users and trip planner) can be done automatically and dynamically, without any user input but only the access of history GPS data, which in our days can be easily obtained via mobile devices.

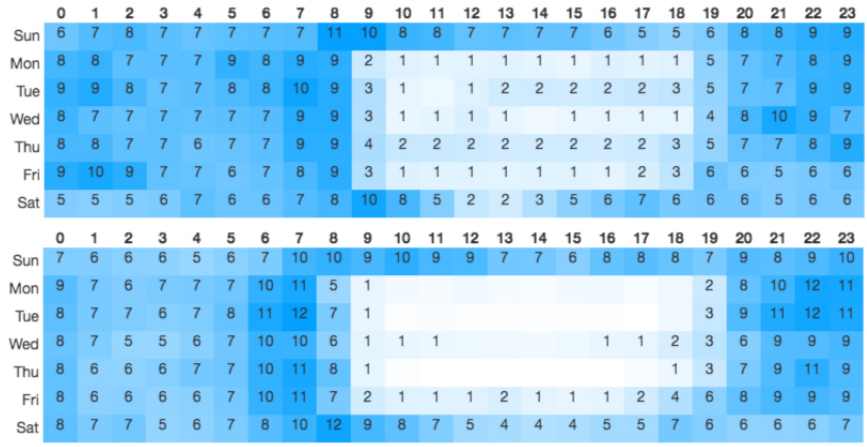


Figure 5.12: Home location profile for User 1 and User 2

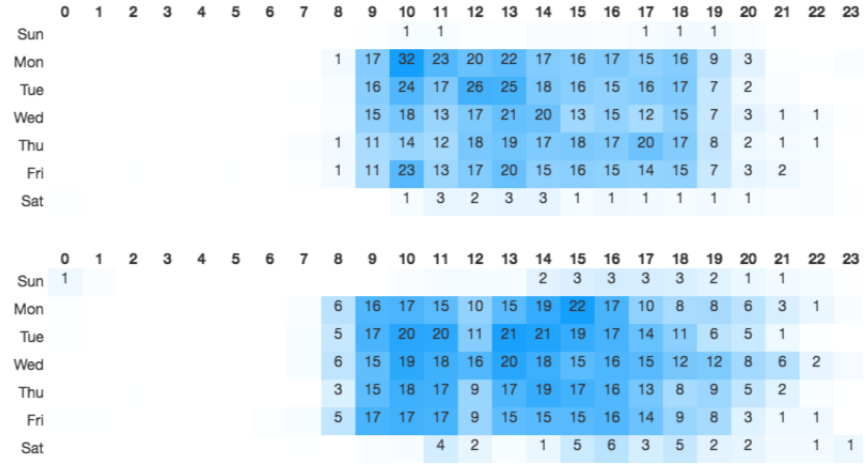


Figure 5.13: Profile of visit work location for User 1 and User 2

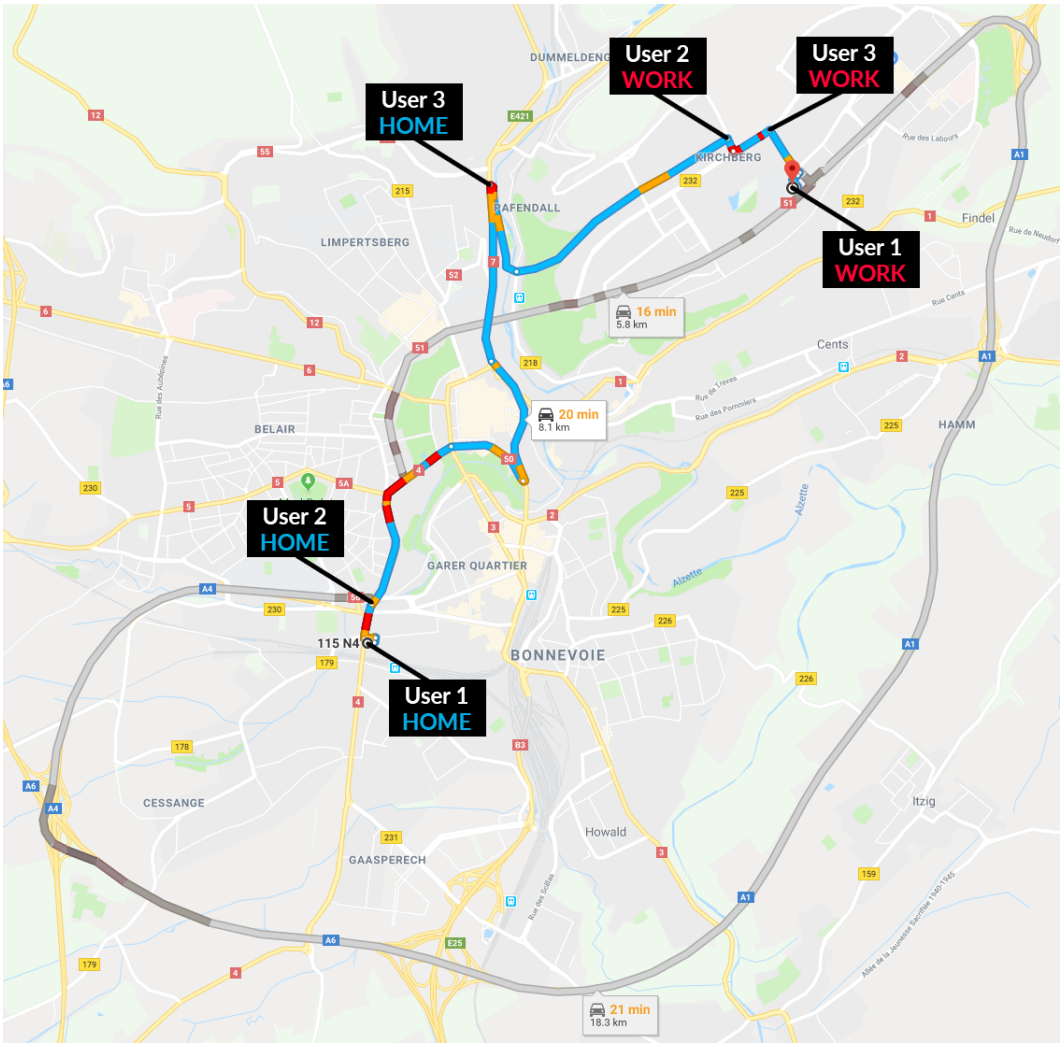


Figure 5.14: Ridesharing example three users example

5.5.3 Location type and activity classification

Profiling the pattern for visiting specific locations gives also a possibility to automatically classify the location to a specific category. As we can observe in Figures 5.10, 5.11, 5.12 the location visit patterns obtained can be clearly identified as *Home* and *Work* locations and dynamically displayed as in Figure 5.15.

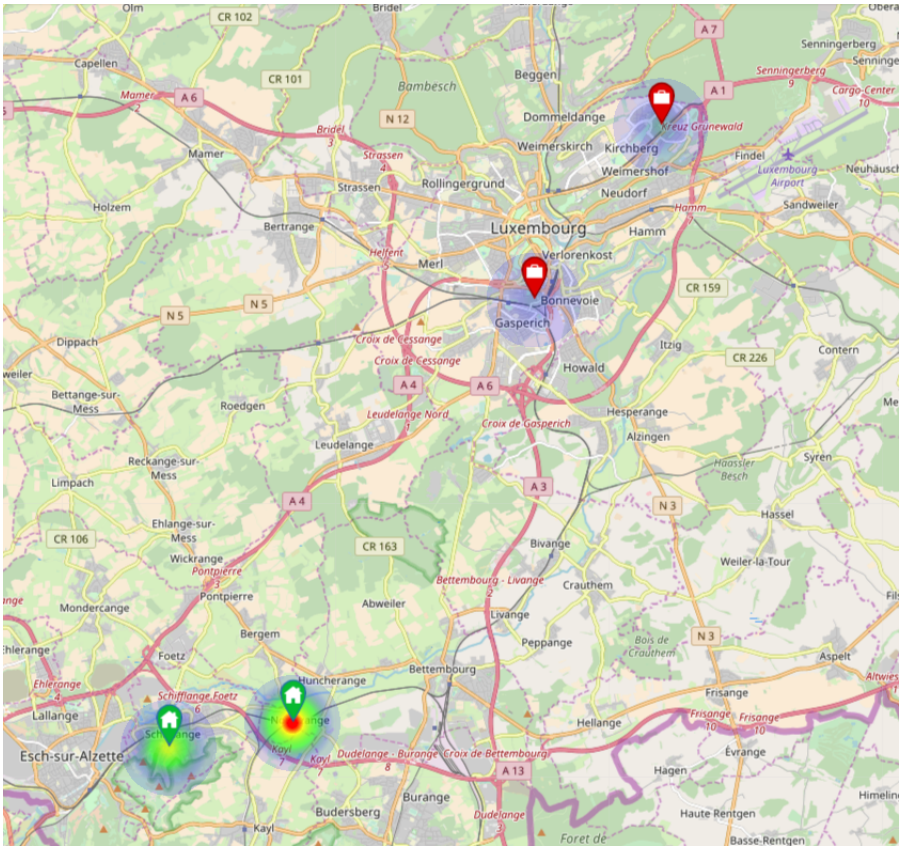


Figure 5.15: Location type classification: Home and Work classification. Residence and workplace change detection.

The same can be done with any type of location for which there are defined patterns which can be identified and the classification can be done automatically. In an extended exploratory study [199] we showed that using the combination of the proposed profiling method combined with observed user habits learned from surveys and extracted as activities matrix was possible to automatically classify the type of activity performed in a specific location.

Moreover, as the profiling can take a large time period into consideration, it is now possible to detect changes of user travel habits by detecting changes of regular visit to specific locations. In Figure 5.15 we can see that for the same user two homes and workplaces are detected, as the profiling detects recursive similar patterns in different time periods. This means that the proposed profiling method can not only detect recurrent habits to secondary activities, but also change of habits, something that is not easy to detect with a static method. The insights obtained based on these changes can be used to adapt and personalise transportation services to match the passenger's

habits, which will result in a better service quality.

5.5.4 Non-recurrent trips profiling

Another application of the proposed methodology is the profiling of non-recurrent trips i.e. holidays and business trips. For this type of profiling the accuracy can be very low, to an order of dozens or hundreds of kilometres. Figure 5.16 shows a typical non-recurrent trips profiling, with the main countries and parts of the world where the user travelled.

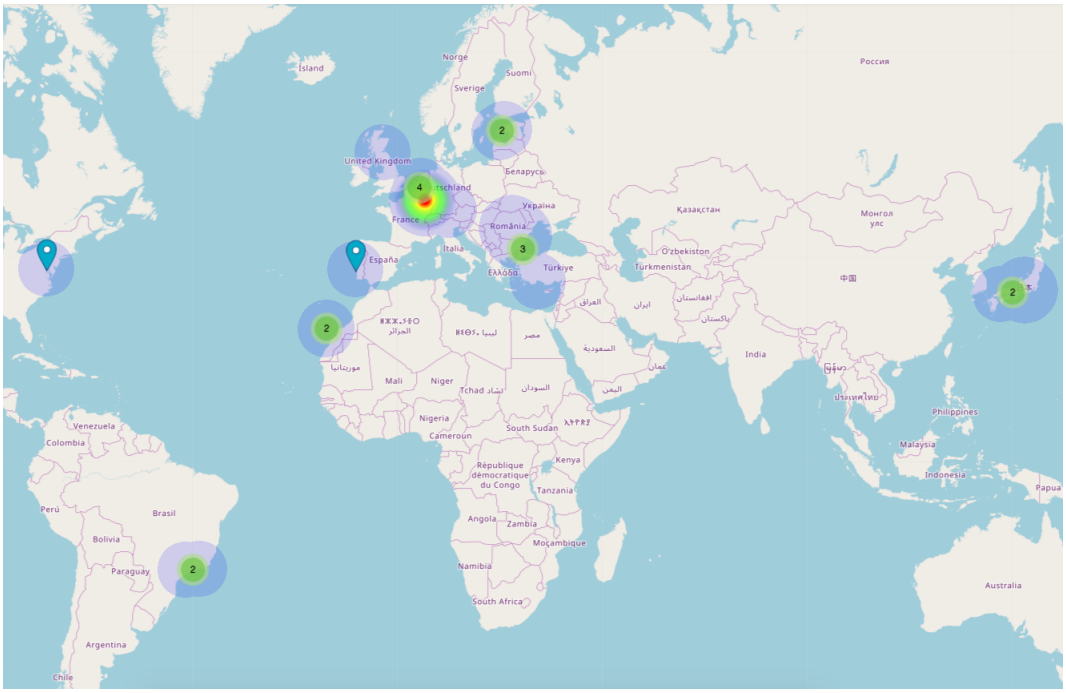


Figure 5.16: Non-recurrent trips profiling

This knowledge is very important for tourism application systems as in an order of milliseconds, the tourist's habits and preferences for visit certain places or parts of the world can be learned. This knowledge is extremely valuable for the next generation of AI trip advisers which will be able to offer support and personalised recommendations for tourists before, during and after their trip. The proposed profiling methodology is suitable for this, as in most of the cases the tourists use a mobile application which they download and in a matter of seconds they need reliable services. The ability to extract very fast insights from historical data without the need of the user input and perform all the computations locally on the device, minimising the resources and quantity of information that need to be transmitted over internet are of a great importance.

5.6 Future work

Future work includes potential optimisation on the technological part but also on the addition of new features and capabilities. Some technological optimisation can be included

Finding common sub-queries across several requests and execute them once it is something that can reduce the number of operations and tasks executed. In the case when multiple queries are performed simultaneous, where it is possible these can be combined *e.g.,* if two simultaneous queries are performed with a temporal range from 8 to 18 and 10 to 17, this might be combined. In the same time, caching techniques can be useful in the future to store temporally the results of the most asked queries on the most recent trees *e.g.,* if there is a regular query that is searching for how many points are in a particular region. Another optimisation would be the implementation of a subscription system which will perform automatically live updates of latest queries and trees.

The presented usage examples show that only using the geolocation data it is already possible to support some sharing services (*e.g.,* parking sharing, car sharing - as for those services only presence/absence of users/resources in time and space is needed), detect travel habits and identify/classify location and activities. In the future, an implementation of a route planner can offer the possibility to have a complete ride sharing service which can match people and vehicles. On the other hand, the usage of semantic external data of visited locations (*e.g.,* type of facilities from existing maps) can better infer secondary activity types and reduce the identification and classification errors.

5.7 Conclusion

The contribution of this chapter is twofold: on the one hand, we present a novel methodology that provides a dynamic profiling of users' mobility and locations visit pattern. The proposed profiling method can be used in many applications and even in a simultaneous manner. The usage examples explained and evaluated throughout the current paper (i.e. parking sharing, ride sharing, location type and activity classification) provides the first directions on how the profiling can be used for a dynamic analysis of sharing mobility users and solutions.

On the other hand, using state-of-the art technologies from data science and computer science we provide a complete implementation of the proposed methodology which can be tested through an online demonstrative prototype. The demo application demonstrates how is possible to load the data and extract complex profiles from geolocation data (i.e. location data from Google Maps), with different accuracy levels and spatio-temporal scales, in an order of magnitude of milliseconds. Moreover, for any visited location, a classification is dynamically performed, which demonstrates that different actions and computations can be performed in motion, at scale and in near real time. Different evaluations were performed in order to assess the speed, scalability and to

evaluate the required resources for implementation, which demonstrates that the proposed profiling can be implemented in a distributed way at the smallest hardware level (*e.g.,* micro computers or mobile devices).

In the following chapter, we present an enhancement of the estimation/learning of complex mobility patterns by using a combination of user data, GIS and specific rules derived from utility theory.

6

A hybrid model and data-driven approach to learn complex mobility habits

The core contribution of this chapter is related to the extraction of locations of individuals' daily and weekly activity-travel patterns, based on the historical visit patterns. Using raw GPS data, special indexing techniques, and a set of aggregate statistics about activity scheduling and preferences, the proposed methodology provides the probability to classify the activity performed in each location. Then, a heuristic rule improves this estimation by considering the value of the information over time, coupled with GIS data about the number of facilities located in a certain area to further improve the overall estimation. Results of this exploratory study support the idea that the proposed approach can reconstruct complex mobility patterns while minimizing the number of inputs from the respondent.

This chapter is based on work that has been published in the following papers:

- Using Passive Data Collection Methods to Learn Complex Mobility Patterns: An Exploratory Analysis
  B Toader, G Cantelmo, M Popescu, F Viti
  2018 21st International Conference on Intelligent Transportation Systems (ITSC)
- Inferring Urban Mobility and Habits from user location history
  Guido Cantelmo, Bogdan Toader, Constantinos Antoniou, Francesco Viti
  22nd EURO Working Group on Transportation Meeting, EWGT 2019, 18 - 20 September 2019, Barcelona, Spain (submitted in 2018, unpublished to date)

Contents

6.1 Introduction

In this chapter, we propose a method to automatically detect the respondents' activity performed in each location, using only the GPS data collected by the smartphones, without the need of any user reported information. The core contribution is related to the extraction of locations' weekly visit pattern, based on the number of visits, time of the day, and day of the week. Using raw GPS data, special indexing techniques, and a set of aggregate statistics about activity scheduling and preferences, the proposed methodology provide the probability to classify the activity performed in each location.

The remainder of this chapter is organised as follows. We describe the methodology for activity classification and extraction of mobility pattern in section 6.2. The evaluation of the proposed method and results are presented in section 6.3. We conclude the study with a discussion of future directions in section 6.4.

6.2 Methodology

The proposed theoretical foundation and methodology from this study consists of several analyses and methods which will be presented in the following sections.

First, in section 6.2.1 we empirical analyse data on the relationship between different mobility patterns and set the foundation for the automatic classification of the location/activity performed in each visited location. The methods used for each stage of this process are presented in the rest of this section.

Section 6.2.3 presents the automatic classification process which begins with the extraction of the weekly visit pattern for each location where respondents spent some time. The clustering technique used to perform the extraction of the visit pattern matrix, it is introduced in the following sections.

Section 6.2.4 presents the methodology behind the automatic location/activity type classification, using only the user's historical mobility patterns collected through the sensing systems of nomadic devices. Based on the visit patterns of each location, we calculate the degree of similarity among different activities with observed visit patterns from travel surveys (which represents the training data used in our classification process).

Then, section 6.2.5 presents an improvement of the location classification method's accuracy from the previous section, using a heuristic rule which adds a weight on the probability/confidence level, according to the importance of the mobility pattern of different times of the day.

Finally, section 6.2.6 presents a Bayesian updating rule of the classification method described in the previous sections, enriching the classification model with external GIS contextual data.

6.2.1 Individual and aggregate mobility patterns

In this section, we investigate the relationship between individual and aggregate mobility patterns. The first step consists in identifying the right number of activities to profile. For instance, if many similar options are considered, the model would probably fail to distinguishing events like "*gym*" and "*swimming pool*". On the other hand, to considering only a few activities would result in an oversimplified problem and pattern. Thus, in this section, we perform a cluster analysis in order to identify the correct number of activities to consider in the model.

Specifically, we assume that activities can be classified in two main groups: *rigid activities e.g.,* work, and *flexible activities*, *e.g.,* daily shopping. It is intuitive to realize that rigid activities are easier to identify, as typical user behaviour is highly repetitive by definition. People sleep almost always in the same location - home - and spend most of their time during working days in their office. On the other hand, flexible activities are more difficult to detect, as they are influenced by different factors, including traffic conditions and household composition. Cantelmo [54] defines in an extended work specific categorisations of activities. Based on these considerations, we identify at least three groups of activities as follows:

a) *Within-Day-Systematic Activities (DSA)*: Activities in which activity scheduling is not flexible.

b) *Within-Week-Systematic Activities (WSA)*: Activities that are not systematic within the day but recur regularly, e.g. every week (i.e. swimming pool, weekly shopping).

c) *Not-Systematic Activities (NSA)*: Flexible activities that represent extraordinary events with respect to the usual user activity scheduling (i.e. visiting the doctor).

We then propose to perform a cluster analysis to classify activities based on two variables:

- N_a^T represents how many people joined activity *"a"* during the reference time period *"T"*.
- N_a^d represents how many people joined activity *"a"* at day *"d"*.

For instance, if we observe two users going to work from Monday to Friday for one week, then we will have that $T = 7$ (i.e. one week of observations), $N_{work}^T = 10$ and $N_{work}^{Monday} = 2$ (i.e. we observed 10 times activity work over one week, two of which on Monday). Given N_a^T and N_a^d, we can compute N_a^t by sorting N_a^d in ascending order. This means that for $t = 1$, N_a^t represents the day with the lowest number of observations for activity *"a"*, while for $t = T$ we will have the highest participation ratio:

$$N_a^{t=1} \leq N_a^{t=2} \leq ... \leq N_a^{t=T-1} \leq N_a^{t=T} \tag{6.1}$$

The cumulative of the frequency for each value of t can then be calculated as

$$P_a^t = \frac{N_a^t}{N_a^T} \tag{6.2}$$

We refer to P_a^t as *cumulative probability* in the rest of this study. Fig. 6.1 represents how P_a^t looks like for activities *work* and *Home*, given fourteen weeks of observations for one hundred users, all systematically commuting five days a week.

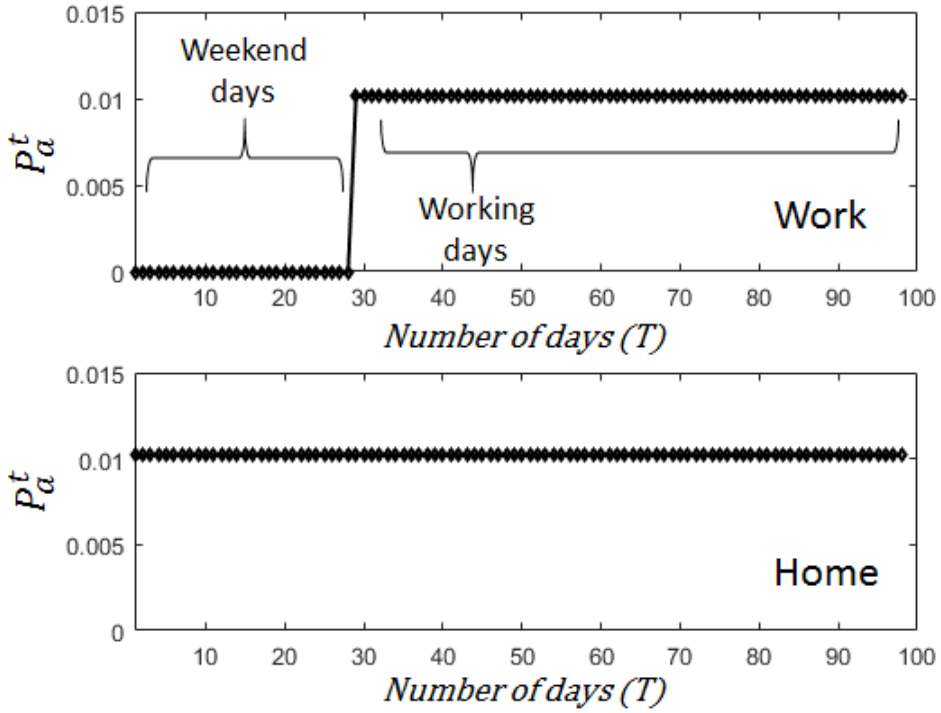


Figure 6.1: Cumulative representation of the probability of the activities *Work* and *Home*

For the activity *work* it is intuitive to see that there is a probability equal to zero during weekends, while it is uniform during the working days. Similarly, the probability to go back home is constant regardless the day of the week, as people always return back home. Clearly, this is a rather naive example where we assume extremely regular travellers, who work five days/week and always return home after work. In reality, these sharp and regular relations are smoothed by factors like part-time jobs, out-of-office work days, vacations, etc. While this is an illustrative example, as we will show in the empirical data analysis, similar patterns can be identified for all activities.

Moreover, using the probability to visit a specific location combined with the visit pattern of each location, in the following sections we demonstrate how it is possible to classify each location and/or the activity performed (*e.g.,* work, home, restaurant, shopping, gym). The first step in this process is to extract the visit pattern of each visited location, which is described in the following section.

6.2.2 Cluster Analysis

In this section, we present present the cluster analysis method and the databases used for this.

a) The first one is the *"Behaviour and Mobility Within the Week" (BMW)*[55]. This travel survey, collected in the region of Ghent (Belgium), contains information from 717 different individuals in the form of Travel Diaries, covering a period of three months (from 08 September 2008 to 07 December 2008).

b) The second database is the *"Multiday travel Survey" (MS)* collected at the University of Luxembourg in 2015 [190]. In this work, we use travel surveys from 52 users observed for a two weeks period (15-30 of June).

Both databases have the same structure, and provide the following information: departure time, arrival time, the origin of the trip, the destination of the trip, activity type, sequence of activities and mode of transport.

In both cases, 12 different types of activity have been reported. Fig. 6.2 shows the list of activities together with the cumulative probability P_a^t for each activity in the case of the *BMW* database. Similar results have also been obtained for the *MS* database.

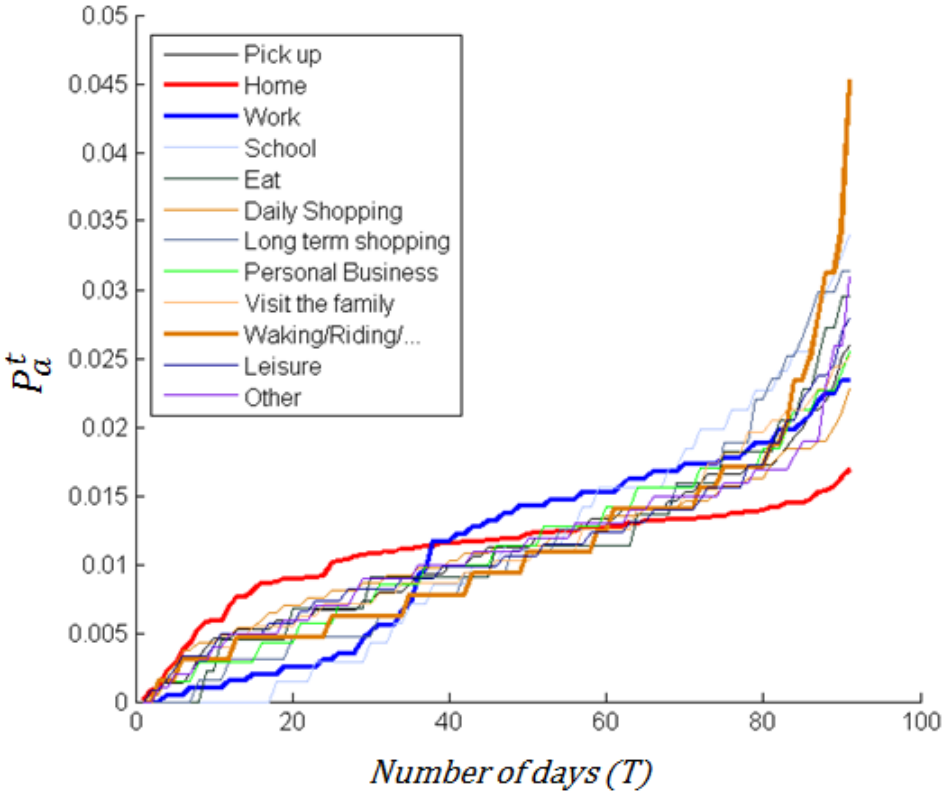


Figure 6.2: Cumulative probability and list of activities for the *BMW* database

Fig. 6.2 shows that, although different from those showed in Fig. 6.1, activities *work* and *home* show the expected trend. The probability for *home* is almost constant over

time, while for *work* we see two trends, one for the weekends and one for the working days. It is also interesting to see that *Not-Systematic Activities*, such as *Walking/Riding*, have an exponential trend, meaning that observations are concentrated over a few days. Finally, *within-Week-Systematic Activities*, such as *Personal Business*, have a more linear trend, indicating a more variable frequency of visits across the observed population.

By taking into account these four fundamental shapes, we created four artificial functions, named [DSA-Home, DSA-Work, WSA, DSA] which represent the typical shape for each activity according to the classification proposed in section 6.2.1. We then performed the hierarchical cluster analysis using the euclidean distance as index of similarity among activities. Fig. 6.3 and 6.4 shows the results for both the *BMW* and *MS* databases.

Results support two interesting points. First, based on the *"Dissimilarity"*, in both cases it is possible to cluster all 12 activities in four clusters, without losing too much information. Second, clusters are very similar in both cases. This was expected, as typical habits are similar in Belgium and Luxembourg. However, we can also see some major differences. Activities *"eat"*, *"Visit family/friends"* and *"other"* belong to the *WSA* cluster in the *BMW*, while to the *NSA* in the *MS*.

There are two possible explanations for this. First, even though similar, we are still analysing two different populations with different habits. The *BMW* is a less biased group of respondents being selected from an entire region and without any specific selection criteria while the *MS* represents only the University staff, including *Ph.D* students. It is thus possible that *e.g.,* this population eats more often at the canteen of the University rather than going somewhere else. Another possible explanation is that, since the *MS* database is relatively small, there are not enough observations to properly represent these three activities. However, the cluster analysis provides in our opinion a satisfactory result in both cases, showing that activities can be aggregated and differences among different populations can be observed. In the rest of this study, we will adopt the data from the *BMW* travel survey to profile users, as these data are statistically more representative.

The results of cluster analysis are represented in Fig. 6.5, where the activities performed by respondents have been clustered in five main groups (*i.e.,* Home, Work, Restaurant, Daily Shopping, Sport). Using the surveys' aggregate results, a visit pattern matrix has been generated for each group of activities.

As discussed in the Section 6.2.2, since the probability for *home* is almost constant, the matrix has been generated manually. We choose this also to reduce the respondents error on recording the *home* location time arrival and departure, since the arrival should be recorded in the previous day of departure.

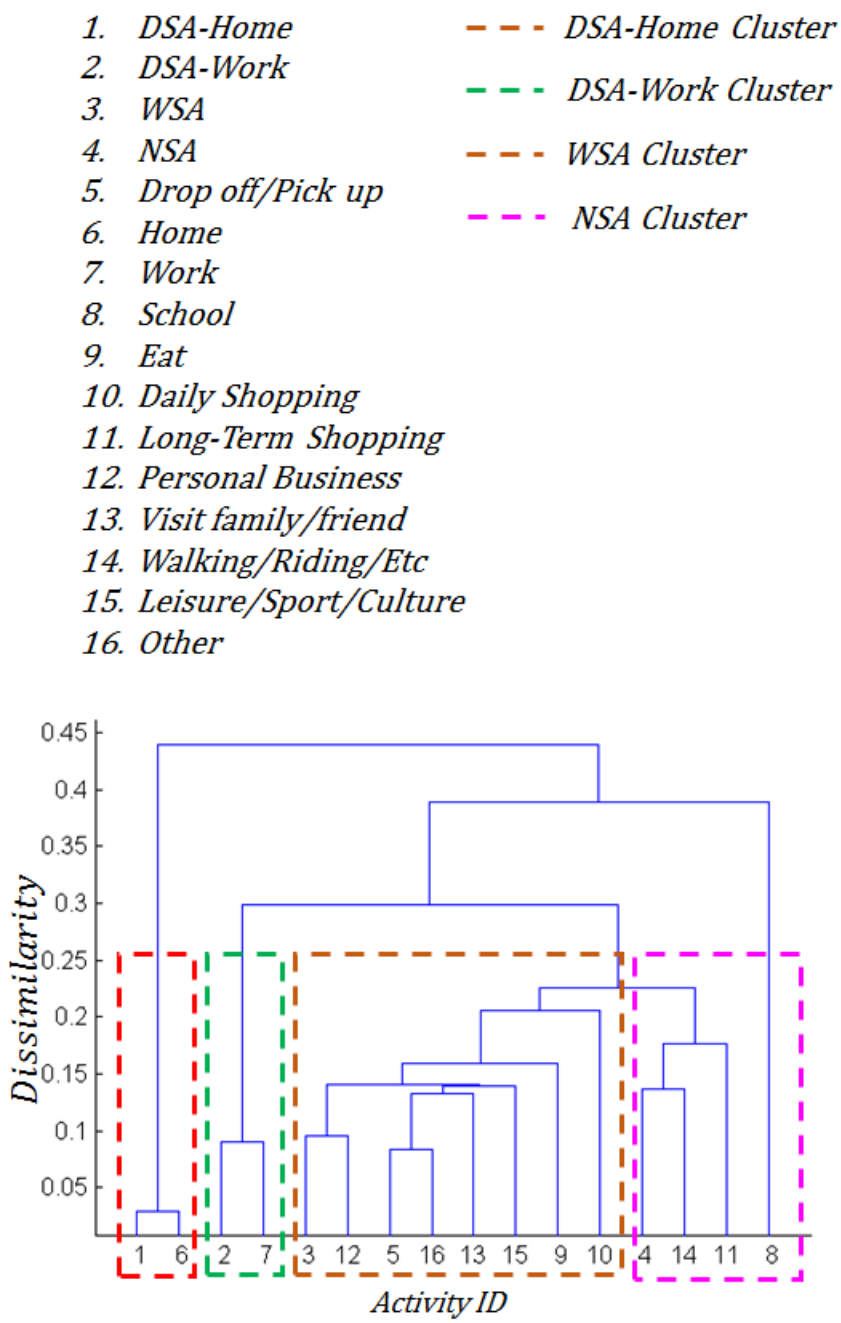


Figure 6.3: Hierarchical Cluster Analysis for the *BMW* database

6.2.3 Mobility patterns extraction from raw data

The weekly activity pattern for each location visited is generated by clustering all the visit records from the beginning of data collection. The clustering technique is based on specific indexing techniques using temporal graphs (such as ND-tree). More details about this technique can be found in [125], [172].

The cluster unit of analysis is one hour, resulting in a matrix with an dimension of 24 hours an seven days of the week. Each matrix element represents the number of times a person visited a specific location, normalised by 1000. In figure 6.9 we can observe

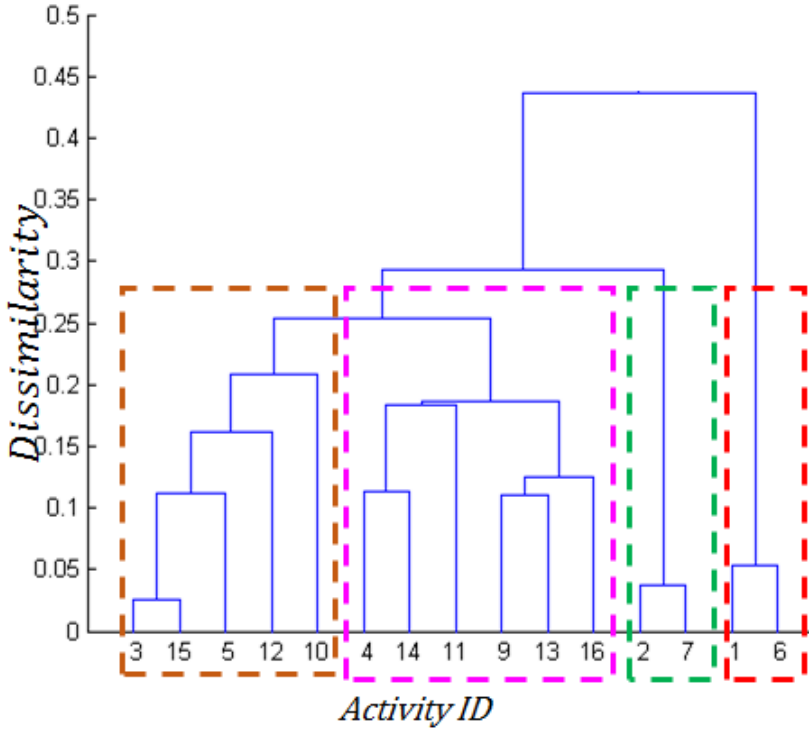


Figure 6.4: Hierarchical Cluster Analysis for the MS database

an example of matrix obtained using this method.

The location clustering performs the extraction based on temporal and spatial parameters. A demo location profiling tool can be accessed and tested online [10] with a provided demo dataset or by uploading individual user data which can be downloaded from Google Map. The application has different settings which control the temporal and profiler parameters, as seen in Fig. 6.6:

a) *Date range*: The time interval on which the clustering is performed. If nothing selected, the entire dataset time interval will be taken by default.

b) *Specific days and hours of the week*: Can be selected specific days and hours of the week. Locations that are visited outside the selected range will not be excluded.

c) *Precision*: Clustering size groups the visited points in the selected range. The precision varies from a few meters to thousands of kilometres, which can be used to profile activities which do not require high precision e.g. holiday profiling.

Using the visit pattern obtained using the method described in this section and a training dataset obtained from travel surveys, the following section presents the methodology used for classification of any visited location and consequently the probable activity performed in each location.

Home

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23
Sun	9	9	9	9	9	9	9	9	9	9	9	9	9	9	9	9	9	9	9	9	9	9	9
Mon	9	9	9	9	9	9	9	0	0	0	0	0	0	0	0	0	0	0	9	9	9	9	9
Tue	9	9	9	9	9	9	9	0	0	0	0	0	0	0	0	0	0	0	9	9	9	9	9
Wed	9	9	9	9	9	9	9	0	0	0	0	0	0	0	0	0	0	0	9	9	9	9	9
Thu	9	9	9	9	9	9	9	0	0	0	0	0	0	0	0	0	0	0	9	9	9	9	9
Fri	9	9	9	9	9	9	9	0	0	0	0	0	0	0	0	0	0	0	9	9	9	9	9
Sat	9	9	9	9	9	9	9	9	9	9	9	9	9	9	9	9	9	9	9	9	9	9	9

Work

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23
Sun	0	0	0	0	0	0	0	0	0	0	0	0	0	0	2	2	2	3	3	2	2	1	1
Mon	0	0	0	0	0	0	12	47	71	82	89	88	79	84	84	81	65	38	17	6	3	2	1
Tue	0	0	0	0	0	0	6	23	47	54	58	59	56	59	58	54	45	29	9	2	2	1	1
Wed	0	0	0	0	0	0	6	53	84	90	96	97	97	98	91	84	66	37	10	6	4	0	0
Thu	0	0	0	0	0	0	10	42	66	77	80	78	75	80	76	73	57	31	12	4	1	1	1
Fri	0	0	0	0	0	0	7	33	51	60	61	64	63	63	63	53	43	30	17	5	5	0	0
Sat	0	0	0	0	0	0	0	2	2	2	2	2	3	3	3	3	2	1	0	0	0	0	0

Restaurant

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23
Sun	0	0	0	0	0	0	0	0	0	6	9	15	21	15	9	3	6	12	12	15	12	9	0
Mon	0	0	0	0	0	0	3	0	0	0	9	24	24	15	6	0	3	3	9	15	15	3	0
Tue	0	0	0	0	0	0	0	0	3	3	0	12	21	15	9	0	6	9	18	12	9	3	3
Wed	0	0	0	0	0	0	0	3	0	0	3	27	29	3	0	0	0	12	18	15	15	9	3
Thu	0	0	0	0	0	0	0	0	0	0	0	9	9	3	0	0	9	9	6	9	9	3	0
Fri	0	0	0	0	0	0	0	0	0	0	0	6	12	6	3	3	0	6	18	18	12	0	0
Sat	0	0	0	0	0	0	3	3	12	12	24	35	41	18	9	6	6	12	32	47	41	18	9

Daily Shopping

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23
Sun	0	0	0	0	0	0	4	13	9	0	22	26	13	4	0	0	4	4	0	0	0	0	0
Mon	0	0	0	0	0	0	4	9	13	4	4	9	4	9	9	13	26	34	26	9	4	0	0
Tue	0	0	0	0	0	0	0	0	4	4	13	22	9	0	4	4	17	30	26	13	4	0	0
Wed	0	0	0	0	0	0	4	13	4	4	9	4	0	0	0	13	13	34	4	4	4	0	0
Thu	0	0	0	0	0	0	0	4	4	9	13	9	9	9	0	4	30	47	30	9	4	0	0
Fri	0	0	0	0	0	0	0	0	0	0	0	0	0	13	17	9	13	17	13	4	0	0	0
Sat	0	0	0	0	0	0	0	9	13	22	26	22	4	4	22	22	13	13	13	13	0	0	0

Sport

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23
Sun	0	0	0	0	0	0	1	1	10	17	23	25	23	27	24	25	22	19	6	6	7	5	4
Mon	0	0	0	0	0	0	1	2	5	2	2	10	10	6	5	4	5	7	6	16	16	11	5
Tue	0	0	0	0	0	2	2	1	1	2	2	10	8	11	7	11	10	11	11	14	12	5	1
Wed	0	0	0	0	0	1	2	4	1	0	0	1	2	2	1	2	4	13	18	19	17	11	5
Thu	0	0	0	0	0	1	1	1	1	0	0	2	4	5	2	2	2	4	7	12	11	7	5
Fri	0	0	0	0	0	0	1	4	4	1	0	0	0	0	4	7	11	13	14	13	10	6	2
Sat	0	0	0	0	0	0	0	1	11	23	24	22	18	24	23	25	24	23	17	8	6	5	2

Figure 6.5: Matrix of activities used as training data, obtained from aggregated travel surveys

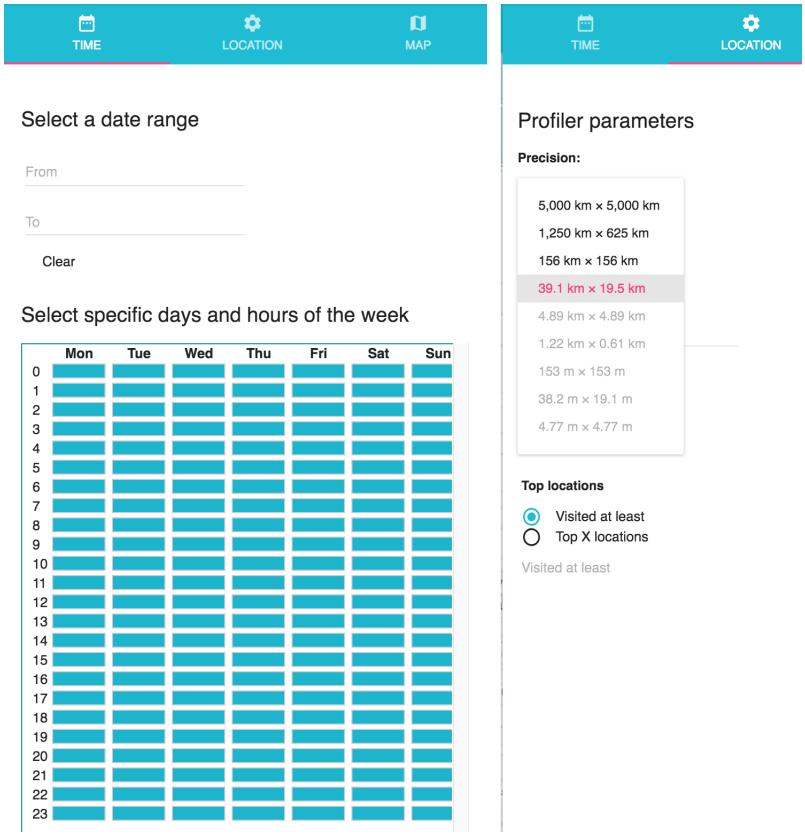


Figure 6.6: Profiling parameters interface

6.2.4 Location/activity type classification

The classification of each location based on the profiling results is done by computing the Euclidean distances between the vectors obtained from two types of matrices:

a) The real data visit pattern matrix which is generated with the pattern extraction method from the previous section. This matrix is extracted for each visited location point, from user's data collected through nomadic devices (smartphones, smartwatches). An example of this matrix can be seen in figure 6.9.

b) The training visit pattern matrix samples from travel surveys. In the following sections, the datasets and the methods used to generate these matrix will be described in details. Thus, for each location type (work, home, restaurant, shopping, gym) a visit pattern matrix is generated. Examples of such matrix can be seen in Figure 6.5.

The classification process consists of computing the Euclidean distances between the matrices described at point (a) with the set of matrices described at point (b). The process can be described as follows:

1) For any selected location, extract the visit pattern matrix *(a)* from the data collected using the nomadic devices.

2) Compute the Euclidean distance between the extracted matrix *(a)* and each of the training sample matrices *(b)*.

3) Compute the probability that the selected location can be classified as one of the training sample matrix *(b)*.

The computing final value belongs to the $(0, 1)$ interval. As a result, the smaller Euclidean distance is obtained, the higher probability that a specific location is one of the five activities represented in Fig. 6.5 (*home, work, restaurant, daily shopping, sport*). In other words, if the matrices are almost identical, the distance is close to zero and the probability that a location can be classified as one of the pre-specified categories is higher. Consequently, the bigger the distance, the more dissimilar the two matrices, resulting in a kind of *penalty* (a method used often for solving constrained optimization problems), which increases the distance and the final result goes close to one.

In order to increase the accuracy of the proposed model, in the following section we present a heuristic rule which aims to improve the final estimation.

6.2.5 Home - work location classification: a heuristic rule

The pattern extraction phase (described in section 6.2.3) provides a first input for a coarse estimation of the probability to perform an activity in a certain location (presented in section 6.2.4). Specifically, the clustering phase calculates the similarity between mobility data collected using the nomadic devices and observations extracted from travel surveys. However, human behaviour is not explicitly modelled within this framework, which is purely data driven.

For instance, the method does not take into account that the *utility* derived by performing a certain activity changes over time. This also means that some time intervals carry more information than other intervals. For instance, it is very likely that users will be at home between 3-4AM on a working day (which can help us to reason that a location where this behaviour is seen very often can be classified as home), whereas not the same can be said about a location visited after work (which can be home, restaurant, shopping or gym). Hence, we introduce a weight that modifies the classification method by stressing that information during some specific time intervals is more relevant. In order to better identify *work* and *home* location, in the rest of this section we introduce some weights within the location classification phase.

An example is shown in Figure 6.7. By focusing on activity *home*, Figure 2 shows that the information at late night (3-6AM) is considered up to three times more important than during the previous time intervals. Similarly, we consider up to three times more valuable the information of being *home* or not during the working hours. In other words, this means that we expect users to be *home* at night and not to be there during working hours. If one of the two conditions is not satisfied, then we strongly decrease the probability of performing activity *"home"* in that location. Similarly, for *"work"* we consider the information at late night more important because people are

not supposed to work in those hours. Similarly, we increased the weight at 11AM and 4PM, as users are very likely to be at work during those hours. Finally, we decreased the relevance of the information during the early morning, lunch-time, and late afternoon to capture both the elasticity of the demand (*i.e.,* not everybody arrives at the same time) and the lunch break (people might leave the working place).

In other words, in order to obtain a better classification, during the time periods of the day when users most likely are in a certain type of location, a bigger *penalty* is applied if the user is not in that location. This will reduce the final probability and confidence that the location can be classified as one of the defined types of location/activity. Consequently, if the user visits the location during the peak time intervals when the *penalty* is higher, the final probability and the total confidence level will increase.

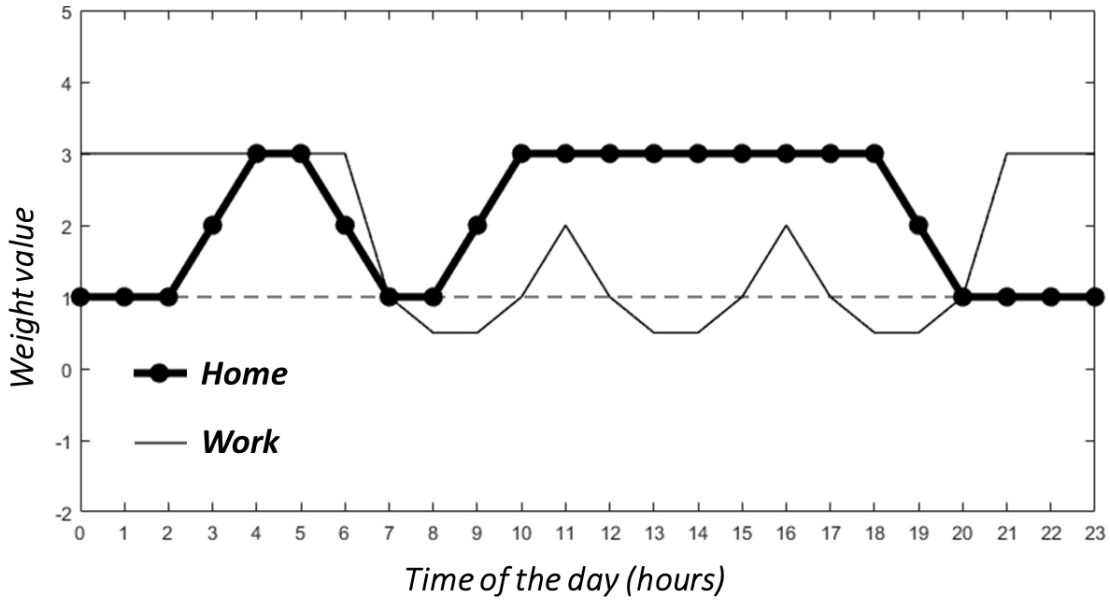


Figure 6.7: Weights of the information over time for activities "Home" and "Work".

While Figure 6.7 represents a very simplistic way of defining the weight for a working day, in general, we derive a more sophisticated heuristic rule derived from utility theory. By defining W_a^t the weight of activity a at time t, and assuming that each activity has a utility function $U(t)$, as discussed in [82], the heuristic rule proposed in this paper can be written as:

a) If $U(t)$ is close to a local minimum or $U(t) = 0$, then $W_a^t > 1$

b) If $U(t)$ is not a local minimum, then $W_a^t = 1$

c) If $U(t) \neq 0$ and it is next to the beginning/end of the activity, then $W_a^t \leq 1$

Rule (a) is a *must-be* (since $U(t)$ is a local minimum the information is maximum), rule (c) takes into account the flexibility of the demand (it is more likely to have larger errors next to the beginning/end of the activity), and rule (b) takes into account all other scenarios.

Even if the methodology presented in this section has the objective to improve the classification process, the method described takes into consideration only the user's historical mobility patterns. In the following section, we present an extension of this methodology which takes into consideration external contextual GIS data which can additionally improve the identification and classification accuracy of visited locations.

6.2.6 Bayesian updating rule

While the simple heuristic rule discussed in subsection 6.2.5 could represent any activity, we prefer limiting its application to model *"home"* and *"work"*. The obvious reason is that it is easy to implement such a rule to capture a highly repetitive behaviour, while it would probably fail in capturing dynamics that are more complex (*e.g.,* to identify locations like restaurants, shopping centres or gym facilities). Thus, in order to increase the reliability of our results, we introduce a probabilistic approach based on GIS data. Specifically, we exploited the read-only "Overpass" API to retrieve online data through OpenStreetMap [217].

In this study, we focused on three specific secondary activities: shopping, sport and food. We adopted the following process to extract how many services are located in a certain location:

a) Provide a location (x, y), where x and y are the coordinates of the point.

b) Identify edges of the surrounding area $[(x + r, y + r), (x - r, y - r)]$, where r is the radius of the area we want to consider.

c) For each element e within the area $[(x + r, y + r), (x - r, y - r)]$:

- Draw *activity_type* from tag {*amenity, shop, leisure, sport*}
- Assign *activity_type* to one of the categories { *shopping, sport, food* }

d) Count the number of locations n_a for each *activity_type* a.

For more details about *Overpass* and *OpenStreetMap*, we refer to their official documentation [217]. We then use a Bayesian updating rule to combine this information with the probability previously calculated in subsection 6.2.4. Specifically, the indexing techniques provides the posterior probability $P_a(U|L)$ of user U doing an activity a given a location L, meaning that we can write the Bayesian updating rule as:

$$P^a(L|U) = \frac{P^a(U|L)P_a(L)}{P^a(U)} \tag{6.3}$$

With:

- $P^a(U)$ the probability of user U doing activity a;

- $P^a(L)$ the probability of performing activity a in location L;

- $P^a(L|U)$ the posterior probability of performing activity a in location L, given a certain user U and the location history.

In order to calculate $P^a(L|U)$, we define $P^a(U)$ and $P^a(L)$ as:

$$P^a(U) = \frac{1}{Number\_Activities} \tag{6.4}$$

$$P^a(L) = \frac{e^{\frac{n_a+\varepsilon}{\theta}}}{\sum_a e^{\frac{n_a+\varepsilon}{\theta}}} \tag{6.5}$$

Equation 6.4 implies that the prior probability $P^a(U)$ follows a uniform distribution (in this case we consider 5 activities, thus $P^a(U) = 0.2$). To calculate $P^a(L)$, we leverage the GIS data and the number of activity location n_a. In essence, if there are two restaurants and no sport centers, the probability for activity *"food"* will be higher. The error term ϵ in equation 6.5 takes into account that, even though *OpenStreetMap* has a rich database, our information could be incomplete (*i.e.,* unreported activities).

Moreover, *home* and *work* location are clearly not in the database, so the probability of performing work related (out-of-office) activities is penalized. To overcome this issue, we define as *"shadow locations"* those locations that are not described within our database. Then, we can introduce the number of *shadow locations* - n_S - which represents the trust we have in our database. A high value of n_S means that the available database poorly represents activities in that location. We can now calculate the error term as:

$$\epsilon = \frac{n_S}{Number\_Activities} \tag{6.6}$$

Equation 6.6 assumes that *shadow locations* are uniformly distributed with respect to the different activities we are considering. This also means that for $\epsilon \to \infty$ equation 6.5 becomes also a uniform distribution (which is expected, since we do not have information about activity distribution in location L).

6.3 Model testing, evaluation and results

This section presents the analysis performed based on the provided methodologies from section 6.2, together with the evaluation of each method and the results obtained.

6.3.1 Location profiling and classification results

In order to test the profiling method presented in section 6.2.4, we used the data of five respondents from University of Luxembourg. Data is collected from history of

locations recorded by Google Map from respondents' smartphones. The time for each user varies from four to eight years. The tool can be accessed online [150] and tests can be done using the demo data provided or by loading any Google Map dataset, following the instructions provided.

The process of profiling and classification employed in this evaluation is presented in section 6.2.4, which begins with *(1)* the extraction of the visit pattern matrix (presented in section 6.2.3) from the raw data collected using the Google Map application, followed by *(2)* the computation of Euclidean distances between the extracted matrices for each location and the matrix from Figure 6.5, and finally *(3)* the computation of the classification probability for each of the main five selected activities.

A visual example with some classified activities can be seen in Figure 6.8 which displays side by side *(a)* the matrix of aggregate activities from the data obtained from travel surveys and *(b)* the matrix extracted from the historical visits of a specific location.

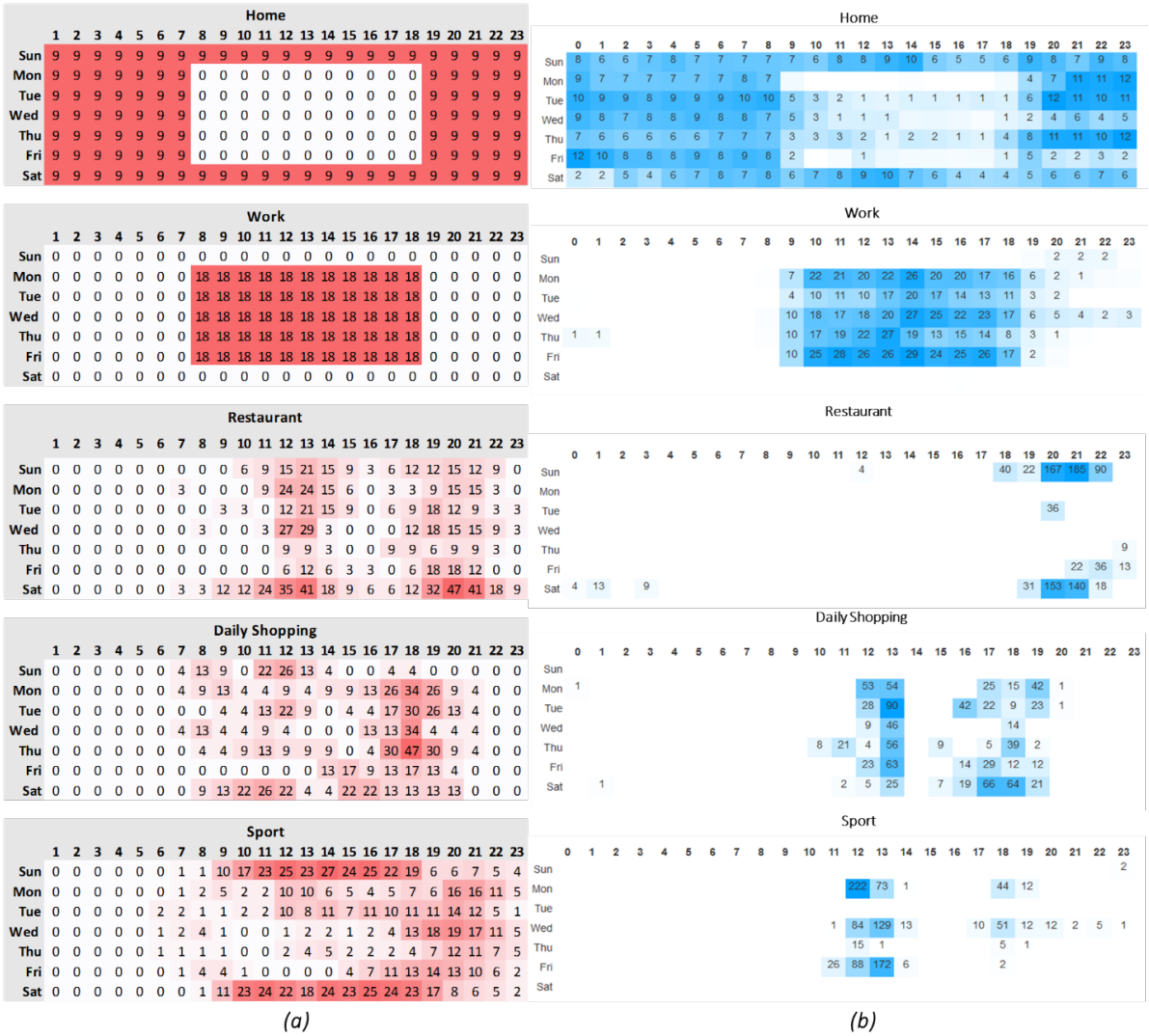


Figure 6.8: (a) Matrix of aggregate activities from the data; (b) Matrix of historical visit pattern for a user.

An example of a detailed result can be seen in Figure 6.9, which shows the matrix of visit patterns of an user to a specific location and the probabilities obtained. In this example, the selected location has the highest probability (91.66% Confidence level) to be a home location. Similar, in Figure 6.10 shows the detected *work* location, for the same user.

Figure 6.9: Example of home location

Figure 6.10: Example of work location

Interesting to note, in Fig. 6.10, *work* location has 83,49% confidence level, followed by *sport* with 73,51%. When the respondent was interviewed, we received the confirmation that around the *work* there is a gym location that is frequently visited by the same user. Moreover, the *restaurant* activity seems to be also correctly identified (with 34,1%), and we received the confirmation that the workplace canteen is a popular place for lunch. Thus, the profiling method captured through this methodology captures multiple activities which are done in a specific range due to the profiler precision parameter (presented in section 6.2.3 and shown in Figure 6.6). In this case, since a larger area has been selected, multiple activities have been performed in the selected area, which effectively contributes to the probability for each group of activities.

Another interesting feature that the proposed profiling method provides is the ability to capture the activity location changing over time. In Figure 6.11 three different *home* locations are represented, since the user changed the home location three times during the data collection. Similar, in Figure 6.12 two *work* locations can be observed, as the respondent confirmed the changing of the *work* location during the collection period.

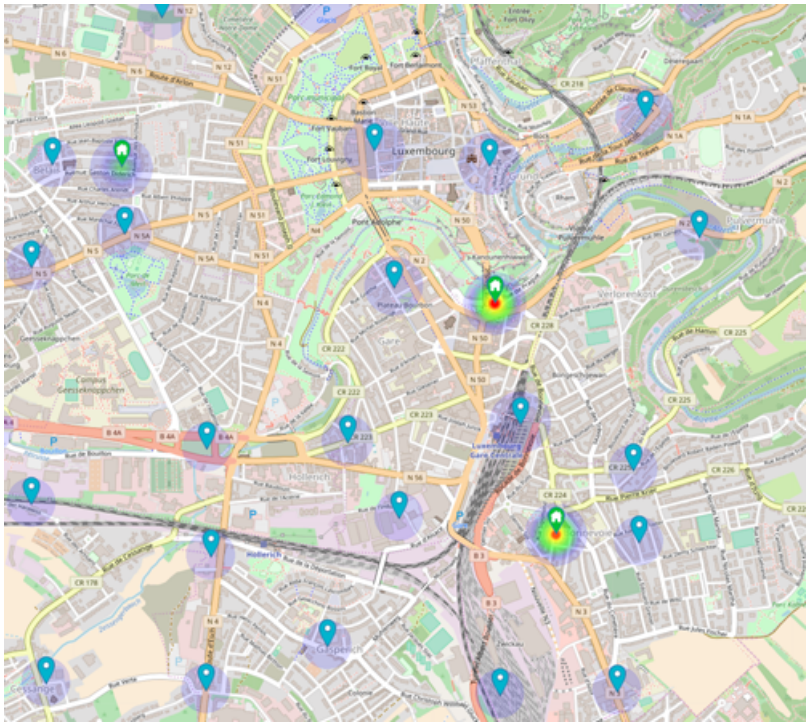


Figure 6.11: Home location change detection

The changing of activity location is a very important information that cannot be captured by the classical travel surveys. The only possibility is that the surveys are repeated with the same respondents at regular periods of time, something which is very expensive and with a very low rate of success because of the burden that a travel survey involves.

The same considerations do not hold for leisure activities. The reason is that the reference data (Figure 6.5) are calculated as the average behaviour for a reference population. For instance, no user goes to the restaurant or to the gym every day. In this sense, the clustering technique still provides a reasonable result, but it will always underestimate the probability of performing leisure activities. For this reason, it is extremely important to consider both the proposed heuristic rule and the GIS data. The heuristic improves the estimation for activities home and work, which also means reducing the probability for these activities in all other cases. Similarly, the GIS data introduces a prior probability that compensates the underestimation related to the data aggregation phase. Both improvements are discussed and evaluated in the following section.

6.3.2 Improving estimation by leveraging GIS data

In this section we aim to improve the estimation of the classification by implementing the heuristic rule presented in section 6.2.5 and the Bayesian rule presented in section 6.2.6.

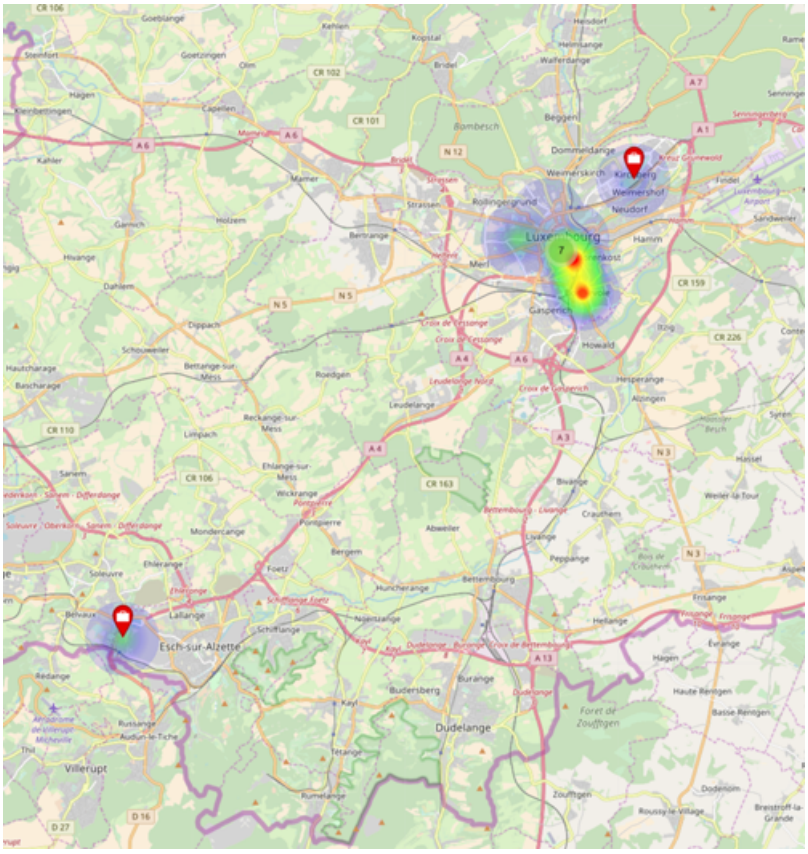


Figure 6.12: Work location change detection

Table 6.1 shows the improvement related to the heuristic rule proposed in 6.2.5. Specifically, we present the results for three users who accepted to collaborate during the validation phase.

For each user, we calculated the probability of performing activity home (P_home) and work (P_work) in four different cases. All users have at least two work locations (WP1 and WP2). In one case this is related to the relocation of their office, showed in Figure 6.12, while in the other two cases the reason is that users have multiple working locations. In this case, we can observe that, when the proposed heuristic is adopted, for all users the value of P_work increases while P_home decreases.

A *Leisure* location has also been analysed. In this case, as expected, both P_home and P_work decreases when the proposed rule is implemented.

Finally, location *Home* has also been analysed. In this case, we would expect P_work to decrease and P_home to increase. However, as shown in Table 6.1, the estimated probability for location home is slightly lower when the heuristic rule is applied. However, we do not consider this to be an issue, as in all cases the estimated probability is close to 90%, thus the model properly identifies the activity performed in that location. If it is true that, in this case, we reduce the probability, while looking at all analysed locations, we clearly see that results are more consistent when the heuristic rule is implemented.

Table 6.1: Experiment results

	User 1				*User 2*				*User 3*			
	With Heuristic		*No Heuristic*		*With Heuristic*		*No Heuristic*		*With Heuristic*		*No Heuristic*	
	P_work	*P_home*	*P_work*	*P_home*	*P_work*	*P_home*	*P_work*	*P_home*	*P_work*	*P_home*	*P_work*	*P_home*
WP 1	90.59	50.69	89.13	67.8	88.51	48.77	86.85	66.91	78.9	50.93	77.53	68.13
WP2	88.05	49.3	86.67	67.05	90	49.62	88.36	67.32	91.26	51.17	90.03	68.07
Leisure	70.67	38.2	73.97	61.4	67.6	77.03	73.83	80.94	75.03	51.09	79.07	68.2
Home	65.46	92.65	71.87	93.57	65.11	92.61	71.75	92.92	72,30	87.25	77.27	90.38

Table 6.2: Experiment results

Results without GIS data:					
	$P^{Work}(U\|L)$	$P^{Home}(U\|L)$	$P^{eat}(U\|L)$	$P^{sport}(U\|L)$	$P^{Shop}(U\|L)$
	50.12%	*0%*	*23.20%*	*41.12%*	*0.13%*
Results with GIS data:					
	Number of Shadow Locations $n_S = 0$				
	$P^{Work}(U\|L)$	$P^{Home}(U\|L)$	$P^{eat}(U\|L)$	$P^{sport}(U\|L)$	$P^{Shop}(U\|L)$
$\theta = 0.3$	*18.44%*	*0%*	*45.19%*	*80.10%*	*0.40%*
$\theta = 0.6$	*32.96%*	*0%*	*35.11%*	*62.23%*	*0.80%*
$\theta = 1$	*39.79%*	*0%*	*30.36%*	*53.82%*	*1.03%*
	Number of Shadow Locations $n_S = 1$				
	$P^{Work}(U\|L)$	$P^{Home}(U\|L)$	$P^{eat}(U\|L)$	$P^{sport}(U\|L)$	$P^{Shop}(U\|L)$
$\theta = 0.3$	*27.61%*	*0%*	*38.82%*	*68.81%*	*0.70%*
$\theta = 0.6$	*38.63%*	*0%*	*31.17%*	*55.25%*	*1%*
$\theta = 1$	*43.27%*	*0%*	*27.95%*	*49.54%*	*1.12%*
	Number of Shadow Locations $n_S = 10$				
	$P^{Work}(U\|L)$	$P^{Home}(U\|L)$	$P^{eat}(U\|L)$	$P^{sport}(U\|L)$	$P^{Shop}(U\|L)$
$\theta = 0.3$	*44.42%*	*0%*	*27.15%*	*48.12%*	*1.10%*
$\theta = 0.6$	*47.30%*	*0%*	*25.15%*	*44.58%*	*1.20%*
$\theta = 1$	*48.43%*	*0%*	*24.36%*	*43.19%*	*1.25%*

Finally, we boost the prediction for leisure activities. To verify the result, we studied the activity "*sport*". The profiling phase showed us that, during lunch time, one of the respondents was often visiting a location close to the University Campus. Based on OpenStreetMap, the area has both a sport center and a restaurant, thus both options have to be considered. The respondent confirmed that performs sport activity in that location three times a week during the lunch break. Table 6.2 shows the probability we estimated through the Bayesian updating process, for different values of n_S and θ.

The results from Table 6.2 shows that the probability obtained through the Bayesian updating rule (equation 6.3) is more reliable than the one derived only from the location history. For n_S equal to 0 or 1, it always identifies the right solution (*i.e.,* sport is the most likely option in the given location), while for large values of n_S and θ the model collapses to the original estimation.

It is also interesting to highlight that, for average parameter values ($n_S = 1$ and $\theta = 0.6$), the Bayesian Approach returns similar values of $P^{Work}(U|L)$ and $P^{eat}(U|L)$, which is realistic considering that *(i)* the user is recurrently visiting that location during the lunch break and *(ii)* probability $P^{Work}(U|L)$ for that user to perform working activities at location L is greater than 50%.

6.4 Conclusions and future work

The main purpose of this study was to explore the possibility to automatically detect the activity performed in frequent visited locations, without any user report or additional information.

Different from most of the state-of-the-art approaches (frameworks to collect travel information) the proposed framework leverages existing location history to infer activity location, meaning that there is no need of collecting additional data (*i.e.,* to ask users which activity performed in each location).

Our integrated framework uses three interconnected components to learn user behaviour:

i) A clustering technique to identify the most likely activity performed in a location;

ii) A heuristic rule to explicitly account for user behaviour while estimating the *Home* and *Work* location;

iii) GIS data to include external contextual data from land use and properly estimate leisure activities.

Results reveal the tool's capabilities to automatically compute the probability that activity performed in each location is either at *home*, *work*, *restaurant*, *daily shopping*, *sport*, using only the location data.

Moreover, this study brings the following scientific and practical contributions:

a) The proposed framework enables the process and analysis of travel data over a long time period (several years);

b) Because of the efficient clustering and data extraction techniques, GIS information can be downloaded in real time and new user data can be processed in a few seconds;

c) Since the model estimates a probability for each activity and each location, there is no need for users to validate their information. While this is still possible, to decrease the error, by further reducing respondent's efforts, larger samples of users can be involved in the process;

d) The tool properly identifies dynamics such as activity relocation, which is an essential information difficult to retrieve with traditional or digital travel surveys.

Future research will explore the possibility to include from external sources additional information regarding the visited locations (such as opening hours) and to improve the clustering approach to provide estimations that are more accurate. Moreover, the duration of the visits can be also a good direction which should be explored. Finally, the authors stress that working with existing data is not always feasible. With the increasing concern about privacy, users are becoming more aware of their rights. On one hand, this has a huge potential, as users can freely access their information and decide to share with the community. On the other hand, many users systematically delete, which is a clear limit for methodologies such as the one proposed in this study. Given these assumptions, the authors aim at working with anonymize data and validating this work with a test case on a larger number of users.

Part III

Conclusion

7

Conclusion

This chapter concludes the dissertation and presents a summary and future research directions.

Contents

This chapter is organised as follows. Section 7.1 summarises the contributions of this dissertation. Section 7.2 discusses potential directions for future work and a short-term outlook concludes the work in Section 7.3.

7.1 Summary

In today's complex mobility domain, data is generated at large scale by different entities that are part of the transportation network (*i.e.,* people, vehicles, goods) and from different sources, such as traffic sensors, nomadic and wearable devices [144]. One of the biggest challenges is to extract insights and knowledge from raw data and consequently to extract data driven models that can describe the multi-layered relationships between all the involved entities. Moreover, all the entities must be matched and synchronised in a seamless way without to compromise the transportation quality service. In order to address these challenges, a data driven approach for mobility analysis was developed and used in this thesis, which can solve some specific issues and challenges presented in each of the contributions from Part II.

Interdisciplinary research that takes into account the different facets of ITSs [149] is in high demand. Beyond this, a deep knowledge of the transportation domain is equally essential in order to better understand all the problems and challenges begging for solutions. A deep knowledge of the data science domain is also essential to apply advanced methods and technologies. Lastly, industry is warning of a lack of data scientists in the transportation domain who can implement the technologies and methodologies for efficient big data management [115]. Bridging the gap between transportation and data science is clearly not a trivial task!

Given these challenges, the studies presented in this dissertation combine the methodologies and technologies from data science with the knowledge domain of transportation engineering. The main objective was to research how data science-driven methods can be developed to dynamically analyse, profile and match people in order to efficiently use collaborative mobility services and exploit shared mobility solutions.

The first part of this dissertation presented the context and motivation behind this research, the objectives and the challenges we were facing when implementing data science methods and techniques in the transportation domain. The objectives of this thesis were translated into specific research questions in Section 1.2. The challenges were discussed in Section 1.3.

In Chapter 2, we presented background information and discussed the state of the art. The state of the art section introduced important terms, methods and techniques from the domains of smart mobility and data science. We then presented the relevant topics and the related literature review in order to correctly frame the contributions presented in the second part of the thesis.

The second part of the thesis focused on practical contributions. Each challenge of this dissertation is addressed by a contribution. Figure 7.1 presents a flowchart summary with the research questions, related challenges and the corresponding chapters/contri-

butions were addressed, which will be discussed in the remainder of this section.

Research Questions	Challenges	Addressed by contributions
RQ1. How can sensing systems contribute to improving and automating collaborative mobility solutions?	**Challenge #1** Universal metric that gives an indication of compatibility for groups of shared mobility users.	**Chapter 3** **A data-driven indicator for collaborative mobility solutions**
RQ2. How to manage big data and perform data analytics in complex mobility scenarios?	**Challenge #2** A framework which can efficiently analyze multi-dimensional large-scale data in motion.	**Chapter 4** **A modeling framework over temporal graphs for big mobility data analytics**
RQ3. How profiling analysis can give new insights and offer new perspective in the study of people behaviour?	**Challenge #3** Develop a method for dynamically profiling users' travel habits and classify visited locations.	**Chapter 5** **An user-centric approach for dynamic profiling of travel habits and visited locations**
RQ4. What is the impact of the proposed contributions in practical applications?	**Challenge #4** Automatically learn users' travel habits and mobility patterns without users' intervention.	**Chapter 6** **Learn complex mobility patterns and habits using external contextual data**

Figure 7.1: Summary of research questions, challenges and contributions

Chapter 3. The main objective of this chapter was to study how the data collected by the sensing systems can contribute to matching people and transportation resources in CMSs. When addressing this objective based upon data collected by nomadic devices, we faced the challenge (Challenge #1) of developing a generic methodology which could compute a metric/index to assess the compatibility of people and transportation resources for different shared mobility solutions. This challenge was addressed in Chapter 3, which proposed an indicator to signal if a user group is compatible for potential collaborative mobility solutions. The indicator can assess if (1) a user group is compatible for a shared mobility solution (*i.e.,* carpooling, car sharing, parking sharing) and (2) if it is economically beneficial. Economic benefit must be both at the system level (by reducing the overall cost of travel and/or travel time and/or transportation resources used) and also at the individual level (by assuring that each user will reduce the total cost of travel when choosing a shared mobility solution). The indicator is designed to take into consideration all variables and possible costs at both the system and individual levels. The indicator also provides an implementation strategy which facilitates use by future ML based recommendation systems.

The experiments and results presented in Section 3.3.1 showed the effectiveness of the proposed method when applied to different sharing mobility services. First, Section 3.3.2 presented an example in which the proposed indicator can reveal whether is

beneficial or not for users to employ a carpooling service and to share trip costs instead of travelling individually. By testing scenarios in which some users will reschedule activities in order to be part of a compatible group of users, experiments revealed that the proposed indicator can identify compatible users who share a ride. Secondly, the proposed method was used to check the compatibility of a group of users for parking sharing services. Section 3.3.3 described a scenario where a group of users who commute by car are compatible to share the same parking place without overlapping. Moreover, the findings revealed that a combination of carpooling and parking sharing services coupled with a parking fee policy could incentivise users to opt for the CMSs instead of commuting in their own vehicle. Finally, we tested the proposed method in a case study that combines carsharing, carpooling and parking sharing services. The results showed that the proposed indicator can be applied to evaluate the CM between individuals, taking into consideration the entire chain of activities and combinations of sharing services. This study was conducted using a small dataset and a reduced number of respondents. It was apparent that factoring in large datasets and high numbers of users would have been far more challenging. Given that activity location and duration were extracted manually for demonstration purposes, implementation in real case scenarios with large scale datasets would have presented many new challenges as well.

Chapter 4. The remaining challenges from Chapter 3 were solved in Chapter 4. The chapter aimed to connect data science methodologies with compter science technologies which could be implemented to handle mobility data at scale. The literature emphasized that ITSs can dramatically improve urban mobility [27]. These systems are expected to perform data analysis in near real time and to react to rapid changes and non-recurrent behaviours *e.g.,* change of the residence. Because sensing systems used for collecting and processing data (*i.e.,* nomadic and wearable devices, traffic sensors, vehicle counting units) usually have only limited computational capabilities; they must handle big clusters and batch processing for their analytic tasks. We pointed out the need for appropriate data-driven models, which are able to extract crucial knowledge and to represent the appropriate context (personal preferences and surrounding environment) able to be deployed at the level of nomadic devices.

Finding a technological solution able to analyse the data in motion (*i.e.,* frequently changing data) was the main challenge that we faced in Chapter 4. Data in motion is continuously collected from all sensing systems involved in the CMSs paradigm (*e.g.,* nomadic devices, traffic sensors, GIS data). Given that the main objective of a smart mobility RS is to provide sustainable travel advice and solutions, hypothetical recommended actions must be explored. In order to address the Challenge #2, we proposed a complete solution that combines a modelling framework and a data analytics platform. The proposed framework makes use of dynamic multi-dimensional data models, temporal graphs and time series. This enables near real time analytics, parallel simulations and deep search capabilities, meeting the requirements of smart mobility complex scenarios. The practical usage and benefits are explained in a case study of collaborative mobility with a large dataset, performing complex tasks and providing interactive real-time data visualization. In Section 4.4 we presented different experiments which demonstrated that the proposed framework is indeed able to manage big data. We also tested deep search and query capabilities. Furthermore, a practical experiment that

made use of the framework's MWGs feature was presented. The experiment consisted of finding compatible groups of users for a carpooling activity drawn from a big dataset of GPS data. The results demonstrated that the proposed framework was fast enough to find clusters of compatible users for a ride sharing service on demand. Moreover, the proposed method revealed its ability to merge discrete simulations and statistical results into a single framework. However, this approach assumed that all users had the same travel behavior. Therefore it could not give all the time satisfactory results when providing personalised shared mobility solutions for each individual. This represents a challenge that was addressed in the following chapter, where a user-centric approach for dynamic profiling was proposed.

Chapter 5. In order to address Challenge #3, we dedicated Chapter 5 to the development of a method capable of dynamically profiling users and all visited locations in order to extract knowledge from raw data *i.e.,* travel habits and mobility patterns. Using special indexing techniques through temporal graphs and deep search capabilities, the provided methodology made dynamic profiling in time and space possible. The profiling was used to extract users' travel behavior and visited locations. It also provided an indication of the type of activities performed in each visited location (*e.g.,* leisure, sport, work). This study also answered RQ3 by presenting case studies in which profiling of people's travel habits could give new insights and offer new perspectives in the study of travel behaviour. More precisely, the profiling method was used to extract insights, travel patterns and habits, and to perform automatic classification of visited locations and activities by using only GPS data collected from nomadic and wearable devices. The main contribution of this study was the demonstration of the possibility of automatically extracting valuable knowledge regarding travel habits without any user input or intervention. The usage examples presented in Section 5.5 demonstrated that the proposed profiling methodology could be applied to various issues within the smart mobility domain. First of all, by using the information provided by the profiling, we demonstrated that it was possible to confirm the compatibility of a group of users for sharing the same parking place. The highest level of compatibility was when their profiles were completely dissimilar *i.e.,* they use the same parking space but at different times. Secondly, the profiling was used in the ride sharing applications to determine if a group of users were synchronised with respect to departure/arrival times. A similar visit profile of the same location indicated a possible ride sharing opportunity. Thirdly, the profiling was used to automatically classify locations types. This was done by extracting the location visit pattern and by computing the similarity distance between the extracted profile and a range of training data relative to different types of locations *i.e.,* home, work, restaurant, shopping centre, gym facility. Finally, we demonstrated that the provided profiling method could be also used for non-recurrent trips profiling *e.g.,* holidays and business trips. By using lower precision we were able to cover wider areas. Overall, the proposed profiling method could be used by an RS to automatically extract travel behaviour and mobility patterns, which in turn could match people and shared mobility services autonomously, quickly and dynamically.

Chapter 6. In this chapter we presented an enhancement of the estimation/learning of complex mobility patterns from Chapter 5 by using a combination of user data, GIS and specific rules derived from utility theory. This answered the RQ4 by offer-

ing a concrete demonstration of the proposed contributions' impact through practical applications. The main challenge was to seamlessly integrate data science and behavioural methods to learn complex mobility patterns and travel habits while performing advanced analytics without any respondent's input. This requirement was derived from mobile users' unforgiving expectations of faultless performance by their mobile apps.Specialised reports revealed these user demands: *"key to this is having the necessary depth of application intelligence in real time so that any problems can be anticipated or rapidly solved"* [29]. This means that knowledge discovery and information extraction methods employed should be able to use passive data collection techniques while offering a high response time and performance. The sensing systems embedded in the nomadic devices should passively collect meaningful information, while reducing the energy consumption and the resources used. This was done by using the data-driven modelling framework presented in Chapter 4 and the dynamic profiling method from Chapter 5. The evaluation and results presented in Section 6.3.1 demonstrated that it is possible to learn users' travel habits and perform classification of visited location types by using only data provided by the sensing systems. Moreover, we showed that the proposed method can use external contextual data from GIS information. When coupled with different behavioural modelling rules, the method can improve the overall accuracy of the proposed model.

7.2 Future research directions

7.2.1 Future smart mobility recommendation systems

The proposed methods and technologies have major implications for future travel RSs, the study of travel behaviour and activity-based modelling. Even if throughout this dissertation we often discussed the next generation of RSs, it was not possible to offer a completely implemented RS. This was due primarily to the time and resources required for the development and rigorous experimental testing of a new fully functional RS. However, the methods and technologies presented, evaluated and tested in this dissertation represent a good foundation able to accelerate future research in this direction.

The next generation of RSs must be able to automatically detect and semantically interpret the activities performed by users in all the visited locations. The main task of an RS is to provide travel advice for individuals or user groups to implement sustainable collaborative services *i.e.,* carpooling, dynamic ride sharing, car sharing, parking sharing. The methodologies and technologies presented in this dissertation have made a contribution in this direction. The RS must not become dependent by waiting for users to supply additional information. The development of an autonomous RS would solve the cold start problem, in which new users who join the system and for whom the RS has no information except historical GPS data. However, the experiments performed (*e.g.,* location classification) use general training datasets extracted from previous surveys and well-known travel behaviour models. Further research should be undertaken to investigate the usage of specific self-learning techniques (*e.g.,* unsuper-

vised learning, reinforcement learning).

The contributions presented illustrate that different what-if scenarios, simulations, optimisations and combinations of people and services may be done at any time. Thus, future RSs could have the capability to autonomously perform complex analytics tasks and to anticipate users' behaviour. By using these methods it would be possible to automatically perform complex analytic tasks without user intervention *e.g.,* the prediction of the next location to be visited or the classification of the activity performed at each visited location. Consequently these systems could classify all activities performed by each user as "daily systematic" (*e.g.,* home, work, daily shopping), "weekly systematic" (*e.g.,* restaurant, sporting activities), or "non-systematic" (*e.g.,* holiday, visiting the doctor), simply by using the location data alone. The most common types of data used in this thesis are the GPS data and GIS information. There is abundant room for future research to explore the potential of adding contextual information and semantics by tapping into external sources (*e.g.,* different IoT devices). This can improve the accuracy of classification and transportation services optimisations.

The evaluations performed and results obtained demonstrate that future RSs will be able to perform detailed analytics of each of heterogeneous locations visited by an individual and uncover valuable information from this analysis. The proposed tool will likely also capture changes in activity location (such as home and workplace relocation). In that case more investigation can be done by exploring other changes in travel behaviour. Indeed, capturing changes in travel behaviour will become a key point for future optimisations of public and private transportation services. With respect to activity based modelling, the information extracted could be used to improve the accuracy and reliability of the modelling methods. In future, it might be possible to use completely different methods to enhance accuracy. This could mean that future RSs could recommend not only how to travel more efficiently, but also how to organise and individual's schedule and sequence of activities. The objective would be to provide sustainable travel solutions (*e.g.,* using shared mobility solutions) that not only provide an economic benefit but also enhanced comfort (*e.g.,* by avoiding time lost in traffic jams).

7.2.2 Machine learning and artificial intelligence

A very promising future research direction is the evaluation and usage of other ML techniques and algorithms in our approach. We are confident the proposed data-driven modelling framework and analytics methodologies can work with any ML and data science method. Of course, in this thesis we tested only basic methods of clustering and learning from training data. Future directions can include the evaluation and testing of more complex ML algorithms and methods in order to improve results and to extract more in-depth knowledge from raw data.

For example, reinforcement learning algorithms are used in other domains to help software agents to take autonomous action to maximise rewards. Inspired by behavioural psychology, a similar approach may be used in shared mobility systems. The strategy may incentivise users increasingly to use soft and shared mobility solutions and to

adopt gamification principles. The technology we have discussed can enable mobility systems to become autonomous intelligent systems which recommend appropriate actions because they have analysed previously learned behaviour and user response to offered incentives. Further studies, which take these variables into account, will need to be undertaken so that RSs can harness advanced methods from the artificial intelligence domain. RSs can become self-learning agents which autonomously take actions, offer recommendations, organise people and manage transportation resources in order to globally optimise the transportation systems.

7.3 Outlook

The approach we have presented in this dissertation has pursued the concept of data-driven transportation engineering fused with the data science analytics. In one way, we might view this concept simply as the inevitable and necessary advance of ITSs in step with the increasing requirements of technology, coupled with the need for efficient smart mobility. The research summarized in this thesis aims to close the gap between knowledge about the transportation engineering domain and data-driven intelligent learning, which can drive autonomous analytic processes. That is why we have combined various research areas and rubrics, including transportation engineering, smart mobility, shared mobility, software engineering, machine learning, data-driven model engineering, database management, and big data analytics.

The sequence of contributions from Part II has offered a complete methodological workflow to support the next generation of smart mobility RS. We began by presenting the theoretical foundation of sharing mobility requirements for matching people and mobility services. An RS can use our proposed method to automatically assess the compatibility of a user group potentially interested in different sharing solutions. Secondly, we proposed a specific data driven modelling framework which could handle and process data at scale. The fact that this framework may be deployed even at the level of low resources devices (*e.g.,* nomadic devices) makes it suitable for use by a RS in multiple smart mobility applications. Thirdly, the proposed profiling method can enable a RS to perform advanced data driven mobility analysis by extracting individual complex travel behaviour from raw data, with nearly no user intervention. Finally, the entire process flow can enable a RS to autonomously execute complex tasks such as visited locations' classification and activities' identification. Moreover, because it is possible to make use of external data, the entire process flow enables future RSs autonomously to take decisions and to provide advice about concrete action to be taken by the final user.

Finally, this thesis has focused on developing data-driven analytics for collaborative mobility solutions and on providing a basic contribution for the next generation of smart mobility RSs. At the same time, we believe the concepts, methods and technologies we have presented may be also applied in domains other than transportation engineering. For example, autonomous driving requires the combination of raw data, domain knowledge, and ML in a single model able to drive near real time analytic processes. Analysing dynamic data in motion that changes frequently and at different

paces is challenging. The data-driven approach, methods and technologies presented in this thesis can constitute an efficient solution. Freight distribution and logistics also require sustainable decisions which factor in the impact of certain actions. Given that every alternative has its own impact, the high combinatorial complexity of alternatives is very hard to analyse. The graph data model used throughout the experiments performed in this thesis could enable an efficient representation and analysis of many different alternatives even in near real time. Other domains such as traffic engineering use different distributed systems, which typically need to collect and share their context information before taking decisions aimed at reducing traffic problems. In order to achieve this objective traffic monitoring systems could reason over distributed data using a similar multi-dimensional graph data model able to handle frequent changes. Last but not least, many transportation systems need to become increasingly intelligent. To make smart decisions, these systems must continuously extract and use behavioural models generated only by learning from live data. ML algorithms can help to extract certain behaviour using the general profiling concepts and methods developed in this thesis. The profiling can then be used in almost any domain requiring a deep understanding of the individual behaviour of each entity involved. This strategy opens up new pathways in research and opportunities to explore innovative AI techniques, which could resolve - perhaps unexpectedly - many of the major as-yet unsolved problems of hi-tech societies.

List of Papers and Tools

Published papers included in the dissertation

- Usage of Smartphone Data to Derive an Indicator for Collaborative Mobility between Individuals, B Toader, F Sprumont, S Faye, M Popescu, F Viti, ISPRS International Journal of Geo-Information 6 (3), 62 [202]
- A new modelling framework over temporal graphs for collaborative mobility recommendation systems, Bogdan Toader, Assaad Moawad, Francois Fouquet, Thomas Hartmann, Mioara Popescu, Francesco Viti, 2017 IEEE 20th International Conference on Intelligent Transportation Systems (ITSC) [200]
- Using Passive Data Collection Methods to Learn Complex Mobility Patterns: An Exploratory Analysis, B Toader, G Cantelmo, M Popescu, F Viti, 2018 21st International Conference on Intelligent Transportation Systems (ITSC) [199]

Paper currently under Submission

- A Data-Driven Scalable Method for Profiling and Dynamic Analysis of Shared Mobility Solutions, Bogdan Toader, Assaad Moawad, Thomas Hartmann, Francesco Viti, IEEE Transactions on Intelligent Transportation Systems (submitted in 2018, under review) [40]
- Inferring Urban Mobility and Habits from user location history, Guido Cantelmo, Bogdan Toader, Constantinos Antoniou, Francesco Viti, 22nd EURO Working Group on Transportation Meeting, EWGT 2019, 18 - 20 September 2019, Barcelona, Spain (unpublished to date) [102]

Tools developed during the thesis

- Pofiler demo: `https://mobilab.lu/profiler-demo`
- Location probability demo: `https://github.com/bogdan-xplode/playmobel`

Bibliography

[1] Apache flink. https://flink.apache.org/. Accessed: January 2019.

[2] Apache hadoop. https://hadoop.apache.org. Accessed: January 2019.

[3] Apache hadoop. https://hadoop.apache.org.

[4] Apache spark. http://spark.apache.org/. Accessed: January 2019.

[5] Autonomous travel suite concept. http://www.aprilli.com/autonomous-travel-suite/.

[6] Demo tool:. https://www.playmobel.bogdantoader.com.

[7] Giraph. http://giraph.apache.org/. Accessed: January 2019.

[8] neo4j. http://neo4j.com.

[9] Occupancy rates of passenger vehicles, European Environment Agency, http://www.webcitation.org/6kduojoag.

[10] Online demo. https://mobilab.lu/profiler-demo/. Accessed: 2018-29-01.

[11] Playmobel source code. https://github.com/bogdan-xplode/playmobel.

[12] Samza. http://samza.apache.org/. Accessed: January 2019.

[13] SPHERE Lab, http://www.spherelab.org/, webcite: http://www.webcitation.org/6krwbre9m.

[14] Toyota launches new mobility ecosystem and concept vehicle at 2018 ces. https://newsroom.toyota.co.jp/en/corporate/20546438.html.

[15] Tamer Abdulazim, Hossam Abdelgawad, Khandker Habib, and Baher Abdulhai. Using smartphones and sensor technologies to automate collection of travel data. *Transportation Research Record: Journal of the Transportation Research Board*, (2383):44–52, 2013.

[16] Gediminas Adomavicius and Alexander Tuzhilin. Toward the next generation of recommender systems: A survey of the state-of-the-art and possible extensions. *IEEE Transactions on Knowledge & Data Engineering*, (6):734–749, 2005.

[17] G. Agamennoni, J. Nieto, and E. Nebot. Mining GPS data for extracting significant places. In *IEEE International Conference on Robotics and Automation, 2009. ICRA '09*, pages 855–862, may 2009.

[18] Niels Agatz, Alan Erera, Martin Savelsbergh, and Xing Wang. Optimization for dynamic ride-sharing: A review. *European Journal of Operational Research*, 223(2):295–303, 2012.

[19] MaaS Alliance. Guidelines and recommendations to create the foundations for a thriving maas ecosystem. Technical report, 2017.

[20] Andreas Allström, Ida Kristoffersson, and Yusak Susilo. Smartphone based travel diary collection: experiences from a field trial in stockholm. *Transportation Research Procedia*, 26:32–38, 2017.

[21] Javier Alonso-Mora, Samitha Samaranayake, Alex Wallar, Emilio Frazzoli, and Daniela Rus. On-demand high-capacity ride-sharing via dynamic trip-vehicle assignment. *Proceedings of the National Academy of Sciences*, 114(3):462–467, 2017.

[22] Helmut Alt and Michael Godau. Computing the fréchet distance between two polygonal curves. *International Journal of Computational Geometry & Applications*, 5(01n02):75–91, 1995.

[23] Andrew Amey. A proposed methodology for estimating rideshare viability within an organization, applied to the mit community. In *TRB Annual Meeting Procediings*, pages 1–16, 2011.

[24] Andrew Amey, John Attanucci, and Rabi Mishalani. Real-time ridesharing: opportunities and challenges in using mobile phone technology to improve rideshare services. *Transportation Research Record: Journal of the Transportation Research Board*, (2217):103–110, 2011.

[25] Lisa Amini, Henrique Andrade, Ranjita Bhagwan, Frank Eskesen, Richard King, Philippe Selo, Yoonho Park, and Chitra Venkatramani. Spc: A distributed, scalable platform for data mining. In *Proceedings of the 4th international workshop on Data mining standards, services and platforms*, pages 27–37. ACM, 2006.

[26] Michele Amoretti, Laura Belli, and Francesco Zanichelli. Utravel: Smart mobility with a novel user profiling and recommendation approach. *Pervasive and Mobile Computing*, 38:474–489, 2017.

[27] Sheng-hai An, Byung-Hyug Lee, and Dong-Ryeol Shin. A survey of intelligent transportation systems. In *Computational Intelligence, Communication Systems and Networks (CICSyN), 2011 Third International Conference on*, pages 332–337. IEEE, 2011.

[28] Constantinos Antoniou, Loukas Dimitriou, and Francisco Pereira. *Mobility Patterns, Big Data and Transport Analytics: Tools and Applications for Modeling*. Elsevier, 2018.

[29] Inc. AppDynamics. The app attention span. Technical report, 2014.

[30] Daniel Ashbrook and Thad Starner. 6. Using GPS to learn significant locations and predict movement across multiple users. *Personal and Ubiquitous Computing*, 7(5):275–286, oct 2003.

[31] Kittelson & Associates, United States. Federal Transit Administration, Transit Cooperative Research Program, and Transit Development Corporation. *Transit capacity and quality of service manual*. Number 100. Transportation Research Board, 2003.

[32] Joshua Auld and Abolfazl Mohammadian. Framework for the development of the agent-based dynamic activity planning and travel scheduling (adapts) model. *Transportation Letters*, 1(3):245–255, 2009.

[33] Kay W. Axhausen, Andrea Zimmermann, Stefan Schönfelder, Guido Rindsfüser, and Thomas Haupt. Observing the rhythms of daily life: A six-week travel diary. *Transportation*, 29(2):95–124.

[34] Milos Balać and Kay W Axhausen. Activity rescheduling within a multi-agent transport simulation framework (matsim). *Arbeitsberichte Verkehrs-und Raumplanung*, 1180, 2016.

[35] Roberto Baldacci, Vittorio Maniezzo, and Aristide Mingozzi. An exact method for the car pooling problem based on lagrangean column generation. *Operations Research*, 52(3):422–439, 2004.

[36] Andrey Balmin, Thanos Papadimitriou, and Yannis Papakonstantinou. Hypothetical queries in an olap environment. In *VLDB*, volume 220, page 231, 2000.

[37] Martin Berger and Mario Platzer. Field evaluation of the smartphone-based travel behaviour data collection app "smartmo". *Transportation Research Procedia*, 11:263–279, 2015.

[38] Nicola Bicocchi and Marco Mamei. Investigating ride sharing opportunities through mobility data analysis. *Pervasive and Mobile Computing*, 14:83–94, 2014.

[39] Gordon Blair, Nelly Bencomo, and Robert B France. Models@ run. time. *Computer*, 42(10), 2009.

[40] Thomas Hartmann Francesco Viti Bogdan Toader, Assaad Moawad. A data-driven scalable method for profiling and dynamic analysis of shared mobility solutions. Submitted in 2018 to IEEE Transactions on Intelligent Transportation Systems (T-ITS), unpublished to date.

[41] Wendy Bohte and Kees Maat. Deriving and validating trip purposes and travel modes for multi-day gps-based travel surveys: A large-scale application in the netherlands. *Transportation Research Part C: Emerging Technologies*, 17(3):285–297, 2009.

[42] Wendy Bohte and Kees Maat. Deriving and validating trip purposes and travel modes for multi-day gps-based travel surveys: A large-scale application in the netherlands. *Transportation Research Part C: Emerging Technologies*, 17(3):285–297, 2009.

[43] Maria Bordagaray, Luigi dell'Olio, Angel Ibeas, and Patricia Cecín. Modelling user perception of bus transit quality considering user and service heterogeneity. *Transportmetrica A: Transport Science*, 10(8):705–721, 2014.

[44] William Brazil, Brian Caulfield, Efthimios Bothos, and ICCS Athens. Transport emissions information: lessons from the peacox project, 2015.

[45] Leo Breiman. Random forests. *Machine learning*, 45(1):5–32, 2001.

[46] Sebastian Breß, Max Heimel, Norbert Siegmund, Ladjel Bellatreche, and Gunter Saake. Gpu-accelerated database systems: Survey and open challenges. In *Transactions on Large-Scale Data-and Knowledge-Centered Systems XV*, pages 1–35. Springer, 2014.

[47] Stacey Bricka and Chandra Bhat. Comparative analysis of global positioning system-based and travel survey-based data. *Transportation Research Record: Journal of the Transportation Research Board*, (1972):9–20, 2006.

[48] Mark W Burris and Justin R Winn. Slugging in houston—casual carpool passenger characteristics. *Journal of Public Transportation*, 9(5):2, 2006.

[49] Young-Ji Byon, Baher Abdulhai, and Amer Shalaby. Real-time transportation mode detection via tracking global positioning system mobile devices. *Journal of Intelligent Transportation Systems*, 13(4):161–170, 2009.

[50] Francesco Calabrese, Giusy Di Lorenzo, Liang Liu, and Carlo Ratti. Estimating origin-destination flows using opportunistically collected mobile phone location data from one million users in boston metropolitan area.

[51] Lesly Alejandra Gonzalez Camacho and Solange Nice Alves-Souza. Social network data to alleviate cold-start in recommender system: A systematic review. *Information Processing & Management*, 54(4):529–544, 2018.

[52] Paolo Campigotto, Christian Rudloff, Maximilian Leodolter, and Dietmar Bauer. Personalized and situation-aware multimodal route recommendations: the favour algorithm. *IEEE Transactions on Intelligent Transportation Systems*, 18(1):92–102, 2017.

[53] S Canale, A Di Giorgio, F Lisi, M Panfili, L Ricciardi Celsi, V Suraci, and F Delli Priscoli. A future internet oriented user centric extended intelligent transportation system. In *Control and Automation (MED), 2016 24th Mediterranean Conference on*, pages 1133–1139. IEEE, 2016.

[54] Guido Cantelmo. *Dynamic Origin-Destination Matrix Estimation with Interacting Demand Patterns*. PhD thesis, University of Luxembourg, 2018.

[55] Marie Castaigne. *Behaviour and Mobility Within the Week:" BMW"*. Belgian Science Policy, 2011.

[56] German Castignani, Thierry Derrmann, Raphaël Frank, and Thomas Engel. Driver behavior profiling using smartphones: A low-cost platform for driver monitoring. *IEEE Intelligent Transportation Systems Magazine*, 7(1):91–102, 2015.

[57] Sangwhan Cha, Marta Padilla Ruiz, Monica Wachowicz, Loc Hoang Tran, Hung Cao, and Ikechukwu Maduako. The role of an iot platform in the design of real-time recommender systems. In *Internet of Things (WF-IoT), 2016 IEEE 3rd World Forum on*, pages 448–453. IEEE, 2016.

[58] Nelson D Chan and Susan A Shaheen. Ridesharing in north america: Past, present, and future. *Transport Reviews*, 32(1):93–112, 2012.

[59] Emmanouil Chaniotakis, Constantinos Antoniou, and Francisco Pereira. Mapping social media for transportation studies. *IEEE Intelligent Systems*, 31(6):64–70, 2016.

[60] Vineeta Chaube, Andrea L Kavanaugh, and Manuel A Perez-Quinones. Leveraging social networks to embed trust in rideshare programs. In *System Sciences (HICSS), 2010 43rd Hawaii International Conference on*, pages 1–8. IEEE, 2010.

[61] Cynthia Chen, Hongmian Gong, Catherine Lawson, and Evan Bialostozky. Evaluating the feasibility of a passive travel survey collection in a complex urban environment: Lessons learned from the new york city case study. *Transportation Research Part A: Policy and Practice*, 44(10):830–840, 2010.

[62] Jie Chen, Shih-Lung Shaw, Hongbo Yu, Feng Lu, Yanwei Chai, and Qinglei Jia. Exploratory data analysis of activity diary data: a space–time gis approach. *Journal of Transport Geography*, 19(3):394–404, 2011.

[63] Sheng-Tzong Cheng, Gwo-Jiun Horng, and Chih-Lun Chou. Using cellular automata to form car society in vehicular ad hoc networks. *IEEE Transactions on Intelligent Transportation Systems*, 12(4):1374–1384, 2011.

[64] Driss Choujaa and Naranker Dulay. Predicting Human Behaviour from Selected Mobile Phone Data Points. In *Proceedings of the 12th ACM International Conference on Ubiquitous Computing*, UbiComp '10, pages 105–108, New York, NY, USA, 2010. ACM.

[65] Driss Choujaa and Naranker Dulay. Predicting human behaviour from selected mobile phone data points. In *Proceedings of the 12th ACM international conference on Ubiquitous computing*, pages 105–108. ACM, 2010.

[66] Cheng-Tao Chu, Sang K Kim, Yi-An Lin, YuanYuan Yu, Gary Bradski, Kunle Olukotun, and Andrew Y Ng. Map-reduce for machine learning on multicore. In *Advances in neural information processing systems*, pages 281–288, 2007.

[67] Edgar F Codd, Sharon B Codd, and Clynch T Salley. Providing olap (on-line analytical processing) to user-analysts: An it mandate. *Codd and Date*, 32, 1993.

[68] Jeffrey Cohen, Brian Dolan, Mark Dunlap, Joseph M Hellerstein, and Caleb Welton. Mad skills: new analysis practices for big data. *Proceedings of the VLDB Endowment*, 2(2):1481–1492, 2009.

[69] European Commission. Eu transport in figures. statistical pocketbook 2015., 2015.

[70] Youjing Cui and Shuzhi Sam Ge. Autonomous vehicle positioning with GPS in urban canyon environments. *IEEE Transactions on Robotics and Automation*, 19(1):15–25, feb 2003.

[71] DataThings. Greycat framework. `https://github.com/datathings/greycat`. Accessed: 2017-10-10.

[72] Nicholas Davis. What is the fourth industrial revolution. In *World Economic Forum*, volume 1, page 2016, 2016.

[73] Vinícius Monteiro de Lira, Valeria Cesario Times, Chiara Renso, and Salvatore Rinzivillo. Comewithme: An activity-oriented carpooling approach. In *Intelligent Transportation Systems (ITSC), 2015 IEEE 18th International Conference on*, pages 2574–2579. IEEE, 2015.

[74] Jeffrey Dean and Sanjay Ghemawat. Mapreduce: simplified data processing on large clusters. *Communications of the ACM*, 51(1):107–113, 2008.

[75] Jeffrey Dean and Sanjay Ghemawat. Mapreduce: a flexible data processing tool. *Communications of the ACM*, 53(1):72–77, 2010.

[76] Narsingh Deo. *Graph theory with applications to engineering and computer science*. Courier Dover Publications, 2017.

[77] Sergio Di Martino and Silvia Rossi. An architecture for a mobility recommender system in smart cities. *Procedia Computer Science*, 98:425–430, 2016.

[78] Janet E Dickinson, Tom Cherrett, Julia F Hibbert, Chris Winstanley, Duncan Shingleton, Nigel Davies, Sarah Norgate, and Chris Speed. Fundamental challenges in designing a collaborative travel app. *Transport Policy*, 44:28–36, 2015.

[79] Florian Drews and Dennis Luxen. Multi-hop ride sharing. In *Sixth Annual Symposium on Combinatorial Search*, 2013.

[80] Ministère du Development Durable et des Infrastructures (2012). Modu: Stratégie globale pour une mobilité durable pour les résidents et les frontaliers., 2012.

[81] Ministère du Development Durable et des Infrastructures (2018). Modu: Stratégie globale pour une mobilité durable pour les résidents et les frontaliers., 2018.

[82] Dick Ettema, Fabian Bastin, John Polak, and Olu Ashiru. Modelling the joint choice of activity timing and duration. *Transportation Research Part A: Policy and Practice*, 41(9):827–841, 2007.

[83] Sébastien Faye, Raphael Frank, and Thomas Engel. Adaptive Activity and Context Recognition Using Multimodal Sensors in Smart Devices. In Stephan Sigg, Petteri Nurmi, and Flora Salim, editors, *Mobile Computing, Applications, and Services*, number 162 in Lecture Notes of the Institute for Computer Sciences, Social Informatics and Telecommunications Engineering, pages 33–50. Springer International Publishing, nov 2015. DOI: 10.1007/978-3-319-29003-4_3.

[84] Sébastien Faye, Nicolas Louveton, Gabriela Gheorghe, and Thomas Engel. A Two-Level Approach to Characterizing Human Activities from Wearable Sensor Data. *Journal of Wireless Mobile Networks, Ubiquitous Computing, and Dependable Applications*, 7(3), sep 2016.

[85] Alexander Felfernig, Seda Polat-Erdeniz, Christoph Uran, Stefan Reiterer, Muesluem Atas, Thi Ngoc Trang Tran, Paolo Azzoni, Csaba Kiraly, and Koustabh Dolui. An overview of recommender systems in the internet of things. *Journal of Intelligent Information Systems*, pages 1–25, 2018.

[86] Erik Ferguson. The rise and fall of the american carpool: 1970–1990. *Transportation*, 24(4):349–376, 1997.

[87] International Transport Forum. *ITF Transport Outlook 2017*. OECD Publishing, 2017.

[88] François Fouquet, Erwan Daubert, Noel Plouzeau, Olivier Barais, Johann Bourcier, and Arnaud Blouin. Kevoree: une approche model@ runtime pour les systèmes ubiquitaires. In *UbiMob2012*, 2012.

[89] François Fouquet, Grégory Nain, Brice Morin, Erwan Daubert, Olivier Barais, Noël Plouzeau, and Jean-Marc Jézéquel. An eclipse modelling framework alternative to meet the models@ runtime requirements. In *International Conference on Model Driven Engineering Languages and Systems*, pages 87–101. Springer, 2012.

[90] Fouquet Francois, Grégory Nain, Brice Morin, Erwan Daubert, Olivier Barais, Noël Plouzeau, and Jean-Marc Jézéquel. Kevoree modeling framework (kmf): Efficient modeling techniques for runtime use. *arXiv preprint arXiv:1405.6817*, 2014.

[91] André Luís Policani Freitas. Assessing the quality of intercity road transportation of passengers: An exploratory study in brazil. *Transportation Research Part A: Policy and Practice*, 49:379–392, 2013.

[92] Masabumi Furuhata, Maged Dessouky, Fernando Ordóñez, Marc-Etienne Brunet, Xiaoqing Wang, and Sven Koenig. Ridesharing: The state-of-the-art and future directions. *Transportation Research Part B: Methodological*, 57:28–46, 2013.

[93] Zdravko Galić. *Spatio-temporal data streams*. Springer, 2016.

[94] Robert Geisberger, Dennis Luxen, Sabine Neubauer, Peter Sanders, and Lars Volker. Fast detour computation for ride sharing. *arXiv preprint arXiv:0907.5269*, 2009.

[95] Keivan Ghoseiri, Ali Ebadollahzadeh Haghani, Masoud Hamedi, and MAUT Center. *Real-time rideshare matching problem.* Mid-Atlantic Universities Transportation Center, 2011.

[96] Joy Ghosh, Matthew J Beal, Hung Q Ngo, and Chunming Qiao. On profiling mobility and predicting locations of wireless users. In *Proceedings of the 2nd international workshop on Multi-hop ad hoc networks: from theory to reality*, pages 55–62. ACM, 2006.

[97] Fosca Giannotti, Lorenzo Gabrielli, Dino Pedreschi, and Salvatore Rinzivillo. Understanding human mobility with big data. In *Solving Large Scale Learning Tasks. Challenges and Algorithms*, pages 208–220. Springer, 2016.

[98] Joseph E Gonzalez, Yucheng Low, Haijie Gu, Danny Bickson, and Carlos Guestrin. Powergraph: distributed graph-parallel computation on natural graphs. In *OSDI*, volume 12, page 2, 2012.

[99] Elizabeth Greene, Leah Flake, Kevin Hathaway, and Michael Geilich. A seven-day smartphone-based gps household travel survey in indiana. In *Transportation Research Board 95th Annual Meeting (16-6274)*, 2016.

[100] Elizabeth Greene, Leah Flake, Kevin Hathaway, and Michael Geilich. A seven-day smartphone-based gps household travel survey in indiana 2. In *95th Annual Meeting of the Transportation Research Board, Washington, DC*, 2016.

[101] Philip A Gruebele. Interactive system for real time dynamic multi-hop carpooling. *Global Transport Knowledge Partnership*, 2008.

[102] Constantinos Antoniou Francesco Viti Guido Cantelmo, Bogdan Toader. Inferring urban mobility and habits from user location history. Submitted in 2018 to 22nd EURO Working Group on Transportation Meeting, EWGT 2019, 18 - 20 September 2019, Barcelona, Spain, unpublished to date.

[103] Peter J Haas, Paul P Maglio, Patricia G Selinger, and Wang Chiew Tan. Data is dead... without what-if models. *PVLDB*, 4(12):1486–1489, 2011.

[104] Torsten Hägerstrand. Survival and arena: on the life-history of individuals in relation to their geographical environment. *Timing space and spacing time*, 2:122–45, 1978.

[105] Ramaswamy Hariharan and Kentaro Toyama. 4. Project Lachesis: parsing and modeling location histories. In *Geographic Information Science*, pages 106–124. Springer, 2004.

[106] Thomas Hartmann. *Enabling model-driven live analytics for cyber-physical systems: The case of smart grids.* PhD thesis, University of Luxembourg, Luxembourg, 2016.

[107] Thomas Hartmann, Francois Fouquet, Matthieu Jimenez, Romain Rouvoy, and Yves Le Traon. Analyzing complex data in motion at scale with temporal graphs. In *The 29th International Conference on Software Engineering & Knowledge Engineering (SEKE'17)*, page 6. KSI Research, 2017.

[108] Thomas Hartmann, Francois Fouquet, Jacques Klein, Yves Le Traon, Alexander Pelov, Laurent Toutain, and Tanguy Ropitault. Generating realistic smart grid communication topologies based on real-data. In *Smart Grid Communications (SmartGridComm), 2014 IEEE International Conference on*, pages 428–433. IEEE, 2014.

[109] Thomas Hartmann, François Fouquet, Jacques Klein, Grégory Nain, and Yves Le Traon. Reactive security for smart grids using models@ run. time-based simulation and reasoning. In *International Workshop on Smart Grid Security*, pages 139–153. Springer, 2014.

[110] Thomas Hartmann, Francois Fouquet, Assaad Moawad, Romain Rouvoy, and Yves Le Traon. Greycat: Efficient what-if analytics for data in motion at scale. *arXiv preprint arXiv:1803.09627*, 2018.

[111] Thomas Hartmann, François Fouquet, Grégory Nain, Brice Morin, Jacques Klein, and Yves Le Traon. Reasoning at runtime using time-distorted contexts: A models@ run. time based approach. In *Proceedings of the 26th International Conference on Software Engineering and Knowledge Engineering*, pages 586–591. Knowledge Systems Institute Graduate School, USA, 2014.

[112] Thomas Hartmann, Assaad Moawad, François Fouquet, Yves Reckinger, Jacques Klein, and Yves Le Traon. Near real-time electric load approximation in low voltage cables of smart grids with models@ run. time. In *Proceedings of the 31st Annual ACM Symposium on Applied Computing*, pages 2119–2126. ACM, 2016.

[113] Wesam Herbawi and Michael Weber. Evolutionary multiobjective route planning in dynamic multi-hop ridesharing. In *European Conference on Evolutionary Computation in Combinatorial Optimization*, pages 84–95. Springer, 2011.

[114] Peter Hidas and Shalendra Ram. Changing Travel Behaviour. *Transport Engineering in Australia*, 10(1):1, 2006.

[115] Inmarsat Research Programme report. The Future of IoT in Enterprise – 2017. Accessed: 2007-10-01.

[116] Dietmar Jannach, Markus Zanker, Alexander Felfernig, and Gerhard Friedrich. *Recommender systems: an introduction*. Cambridge University Press, 2010.

[117] Antonio J Jara, Dominique Genoud, and Yann Bocchi. Big data for cyber physical systems: an analysis of challenges, solutions and opportunities. In *Innovative Mobile and Internet Services in Ubiquitous Computing (IMIS), 2014 Eighth International Conference on*, pages 376–380. IEEE, 2014.

[118] Eduardo Lucio Lasmar Junior, Renata Lopes Rosa, and Demostenes Zegarra Rodriguez. A recommendation system for shared-use mobility service. In *2018 26th International Conference on Software, Telecommunications and Computer Networks (SoftCOM)*, pages 1–6. IEEE, 2018.

[119] Jair Ferreira Júnior, Eduardo Carvalho, Bruno V Ferreira, Cleidson de Souza, Yoshihiko Suhara, Alex Pentland, and Gustavo Pessin. Driver behavior profiling: An investigation with different smartphone sensors and machine learning. *PLoS one*, 12(4):e0174959, 2017.

[120] Andreas Kaplan and Michael Haenlein. Siri, siri, in my hand: Who's the fairest in the land? on the interpretations, illustrations, and implications of artificial intelligence. *Business Horizons*, 62(1):15–25, 2019.

[121] Kalon L Kelly. Casual carpooling-enhanced. *Journal of Public Transportation*, 10(4):6, 2007.

[122] Ryuichi Kitamura. An evaluation of activity-based travel analysis. *Transportation*, 15(1):9–34, 1988.

[123] Alexander Kleiner, Bernhard Nebel, and V Ziparo. A mechanism for dynamic ride sharing based on parallel auctions. 2011.

[124] Joonho Ko and Randall L Guensler. Characterization of congestion based on speed distribution: a statistical approach using gaussian mixture model. In *Transportation Research Board Annual Meeting*, 2005.

[125] Dashiell Kolbe, Qiang Zhu, and Sakti Pramanik. Efficient k-nearest neighbor searching in nonordered discrete data spaces. *ACM Transactions on Information Systems (TOIS)*, 28(2):7, 2010.

[126] Ravi Kanth V Kothuri, Siva Ravada, and Daniel Abugov. Quadtree and r-tree indexes in oracle spatial: a comparison using gis data. In *Proceedings of the 2002 ACM SIGMOD international conference on Management of data*, pages 546–557. ACM, 2002.

[127] Menno-Jan Kraak and Otto Huisman. 17 beyond exploratory visualization of space–time paths. *Geographic data mining and knowledge discovery*, page 431, 2009.

[128] Sanjeev Kulkarni, Nikunj Bhagat, Maosong Fu, Vikas Kedigehalli, Christopher Kellogg, Sailesh Mittal, Jignesh M Patel, Karthik Ramasamy, and Siddarth Taneja. Twitter heron: Stream processing at scale. In *Proceedings of the 2015 ACM SIGMOD International Conference on Management of Data*, pages 239–250. ACM, 2015.

[129] Aapo Kyrola, Guy E Blelloch, and Carlos Guestrin. Graphchi: Large-scale graph computation on just a pc. USENIX, 2012.

[130] N. D. Lane, E. Miluzzo, H. Lu, D. Peebles, T. Choudhury, and A. T. Campbell. A survey of mobile phone sensing. *IEEE Communications Magazine*, 48(9):140–150, sep 2010.

[131] Nicholas D Lane, Emiliano Miluzzo, Hong Lu, Daniel Peebles, Tanzeem Choudhury, and Andrew T Campbell. A survey of mobile phone sensing. *IEEE Communications magazine*, 48(9), 2010.

[132] David E LeBlanc. *Slugging: The commuting alternative for Washington, DC*. Forel Pub., 1999.

[133] DongWoo Lee and Steve HL Liang. Crowd-sourced carpool recommendation based on simple and efficient trajectory grouping. In *Proceedings of the 4th ACM SIGSPATIAL International Workshop on Computational Transportation Science*, pages 12–17. ACM, 2011.

[134] LevelDB. Database leveldb. `http://leveldb.org/`.

[135] Jianling Li, Patrick Embry, Stephen Mattingly, Kaveh Sadabadi, Isaradatta Rasmidatta, and Mark Burris. Who Chooses to Carpool and Why?: Examination of Texas Carpoolers. *Transportation Research Record: Journal of the Transportation Research Board*, 2021:110–117, dec 2007.

[136] Zhigang Jason Li and Amer S Shalaby. Web-based gis system for prompted recall of gps-assisted personal travel surveys: System development and experimental study. Technical report, 2008.

[137] Rake& Agrawal King-lp Lin and Harpreet S Sawhney Kyuseok Shim. Fast similarity search in the presence of noise, scaling, and translation in time-series databases. In *Proceeding of the 21th International Conference on Very Large Data Bases*, pages 490–501, 1995.

[138] TIR Consulting Group LLC. The third industrial revolution startegy study, 2016.

[139] R Logesh, V Subramaniyaswamy, and V Vijayakumar. A personalised travel recommender system utilising social network profile and accurate gps data. *Electronic Government, an International Journal*, 14(1):90–113, 2018.

[140] R Logesh, V Subramaniyaswamy, V Vijayakumar, and Xiong Li. Efficient user profiling based intelligent travel recommender system for individual and group of users. *Mobile Networks and Applications*, pages 1–16, 2018.

[141] Yucheng Low, Joseph E Gonzalez, Aapo Kyrola, Danny Bickson, Carlos E Guestrin, and Joseph Hellerstein. Graphlab: A new framework for parallel machine learning. *arXiv preprint arXiv:1408.2041*, 2014.

[142] Grzegorz Malewicz, Matthew H Austern, Aart JC Bik, James C Dehnert, Ilan Horn, Naty Leiser, and Grzegorz Czajkowski. Pregel: a system for large-scale graph processing. In *Proceedings of the 2010 ACM SIGMOD International Conference on Management of data*, pages 135–146. ACM, 2010.

[143] Ed Manley, Chen Zhong, and Michael Batty. Spatiotemporal variation in travel regularity through transit user profiling. *Transportation*, pages 1–30, 2016.

[144] James Manyika, Michael Chui, Brad Brown, Jacques Bughin, Richard Dobbs, Charles Roxburgh, and Angela H Byers. Big data: The next frontier for innovation, competition, and productivity. 2011.

[145] Mark Litwintschik. 1.1 billion taxi rides with mapd and 8 nvidia tesla k80s. Accessed: 2018-10-10.

[146] Bernard Marr. Why everyone must get ready for the 4th industrial revolution. *The Forbes*, 2016.

[147] Ruben Mayer, Christian Mayer, Muhammad Adnan Tariq, and Kurt Rothermel. Graphcep: Real-time data analytics using parallel complex event and graph processing. In *Proceedings of the 10th ACM International Conference on Distributed and Event-based Systems*, pages 309–316. ACM, 2016.

[148] Cyrille Medard de Chardon and Geoffrey Caruso. Friendly Batch Routing (FBR), http://geow.uni.lu/apps/fbr/, webcite: http://www.webcitation.org/6krwis2tf. 2012.

[149] Andrea Melis, Marco Prandini, Laura Sartori, and Franco Callegati. Public transportation, iot, trust and urban habits. In *International Conference on Internet Science*, pages 318–325. Springer, 2016.

[150] Microsoft. geolife-dataset. `https://www.microsoft.com/en-us/download/details.aspx?id=52367`.

[151] Microsoft Research Asia. Geolife gps trajectories. Accessed: 2018-12-04.

[152] Harvey J. Millera. Collaborative mobility: using geographic information science to cultivate cooperative transportation systems. *Procedia - Social and Behavioral Sciences*, 21:24–28, 2011.

[153] Assaad Moawad. *Towards ambient intelligent applications using models@ run. time and machine learning for context-awareness*. PhD thesis, University of Luxembourg, 2016.

[154] Assaad Moawad, Thomas Hartmann, François Fouquet, Grégory Nain, Jacques Klein, and Johann Bourcier. Polymer: A model-driven approach for simpler, safer, and evolutive multi-objective optimization development. In *MODELSWARD 2015-Proceedings of the 3rd International Conference on Model-Driven Engineering and Software Development*, pages 286–293. SCITEPRESS, 2015.

[155] Lara Montini, Nadine Rieser-Schüssler, Andreas Horni, and Kay W Axhausen. Trip purpose identification from gps tracks. *Transportation Research Record*, 2405(1):16–23, 2014.

[156] Raul Montoliu, Jan Blom, and Daniel Gatica-Perez. 1. Discovering places of interest in everyday life from smartphone data. *Multimedia Tools and Applications*, 62(1):179–207, jan 2013.

[157] Catherine Morency. The ambivalence of ridesharing. *Transportation*, 34(2):239–253, 2007.

[158] Brice Morin, Olivier Barais, Jean-Marc Jezequel, Franck Fleurey, and Arnor Solberg. Models@ run. time to support dynamic adaptation. *Computer*, 42(10), 2009.

[159] Todd Mostak. An overview of mapd (massively parallel database). *White paper. Massachusetts Institute of Technology*, 2013.

[160] Bat-hen Nahmias-Biran, Yafei Han, Shlomo Bekhor, Fang Zhao, Christopher Zegras, and Moshe Ben-Akiva. Enriching activity-based models using smartphone-based travel surveys. *Transportation Research Record*, page 0361198118798475, 2018.

[161] Bat-hen Nahmias-Biren, Yafei Han, Shlomo Bekhor, Fang Zhao, Christopher Zegras, and Moshe Ben-Akiva. Enriching activity based models using smartphone-based travel surveys. Technical report, 2018.

[162] Leonardo Neumeyer, Bruce Robbins, Anish Nair, and Anand Kesari. S4: Distributed stream computing platform. In *Data Mining Workshops (ICDMW), 2010 IEEE International Conference on*, pages 170–177. IEEE, 2010.

[163] Ben Nham, Kanya Siangliulue, and Serena Yeung. Predicting mode of transport from iphone accelerometer data. *Machine Learning Final Projects, Stanford University*, 2008.

[164] Hiroki Ohashi, Phong Xuan Nguyen, Takayuki Akiyama, Masaaki Yamamoto, and Akiko Sato. Trip-Extraction Method Based on Characteristics of Sensors and Human-Travel Behavior for Sensor-Based Travel Survey. *Journal of Information Processing*, 24(1):39–48, 2016.

[165] Online profiler tool link:. Accessed: 2018-10-01.

[166] Engin Ozatay, Simona Onori, James Wollaeger, Umit Ozguner, Giorgio Rizzoni, Dimitar Filev, John Michelini, and Stefano Di Cairano. Cloud-based velocity profile optimization for everyday driving: A dynamic-programming-based solution. *IEEE Transactions on Intelligent Transportation Systems*, 15(6):2491–2505, 2014.

[167] Francisco Câmara Pereira and Stanislav S Borysov. Machine learning fundamentals. In *Mobility Patterns, Big Data and Transport Analytics*, pages 9–29. Elsevier, 2019.

[168] Elena Polycarpou, Lambros Lambrinos, and Eftychios Protopapadakis. Smart parking solutions for urban areas. In *2013 IEEE 14th International Symposium on*, pages 1–6. IEEE, 2013.

[169] Allan Pred. The choreography of existence: comments on hägerstrand's time-geography and its usefulness. *Economic geography*, 53(2):207–221, 1977.

[170] PwC. The sharing economy, 2015.

[171] Dorian Pyle. *Data Preparation for Data Mining*. Morgan Kaufmann, 1999.

[172] Gang Qian, Qiang Zhu, Qiang Xue, and Sakti Pramanik. A space-partitioning-based indexing method for multidimensional non-ordered discrete data spaces. *ACM Transactions on Information Systems (TOIS)*, 24(1):79–110, 2006.

[173] Weicheng Qian, Kevin G Stanley, and Nathaniel D Osgood. The impact of spatial resolution and representation on human mobility predictability. In *International Symposium on Web and Wireless Geographical Information Systems*, pages 25–40. Springer, 2013.

[174] Ragunathan Rajkumar, Insup Lee, Lui Sha, and John Stankovic. Cyber-physical systems: the next computing revolution. In *Design Automation Conference (DAC), 2010 47th ACM/IEEE*, pages 731–736. IEEE, 2010.

[175] Logesh Ravi and Subramaniyaswamy Vairavasundaram. A collaborative location based travel recommendation system through enhanced rating prediction for the group of users. *Computational intelligence and neuroscience*, 2016:7, 2016.

[176] Asok Ray. Autonomous perception and decision-making in cyber-physical systems. In *Computer Science & Education (ICCSE), 2013 8th International Conference on*, pages 1–10. IEEE, 2013.

[177] Faisal Rehman, Osman Khalid, and Sajjad Ahmad Madani. A comparative study of location-based recommendation systems. *The Knowledge Engineering Review*, 32, 2017.

[178] Paul Resnick, Neophytos Iacovou, Mitesh Suchak, Peter Bergstrom, and John Riedl. Grouplens: an open architecture for collaborative filtering of netnews. In *Proceedings of the 1994 ACM conference on Computer supported cooperative work*, pages 175–186. ACM, 1994.

[179] Paul Resnick and Hal R. Varian. Recommender systems. *Commun. ACM*, 40(3):56–58, March 1997.

[180] Yasushi Sakurai, Masatoshi Yoshikawa, and Christos Faloutsos. Ftw: fast similarity search under the time warping distance. In *Proceedings of the twenty-fourth ACM SIGMOD-SIGACT-SIGART symposium on Principles of database systems*, pages 326–337. ACM, 2005.

[181] Andrea Sassi and Franco Zambonelli. Coordination infrastructures for future smart social mobility services. *IEEE Intelligent Systems*, 29(5):78–82, 2014.

[182] Douglas C Schmidt. Model-driven engineering. *COMPUTER-IEEE COMPUTER SOCIETY-*, 39(2):25, 2006.

[183] Johann Schrammel, Marc Busch, and Manfred Tscheligi. Peacox-persuasive advisor for co2-reducing cross-modal trip planning. In *PERSUASIVE (Adjunct Proceedings)*, 2013.

[184] Arie Segev and Arie Shoshani. Logical modeling of temporal data. In *ACM Sigmod Record*, volume 16, pages 454–466. ACM, 1987.

[185] Bin Shao, Haixun Wang, and Yatao Li. Trinity: A distributed graph engine on a memory cloud. In *Proceedings of the 2013 ACM SIGMOD International Conference on Management of Data*, pages 505–516. ACM, 2013.

[186] Mary Shapcott and Phillip Steadman. Rhythms of urban activity. *Human activity and time geography*, pages 49–74, 1978.

[187] Shih-Lung Shaw and Hongbo Yu. A gis-based time-geographic approach of studying individual activities and interactions in a hybrid physical–virtual space. *Journal of Transport Geography*, 17(2):141–149, 2009.

[188] Sharon Shewmake. Can Carpooling Clear the Road and Clean the Air? Evidence from the Literature on the Impact of HOV Lanes on VMT and Air Pollution. *Journal of Planning Literature*, page 0885412212451028, jul 2012.

[189] Frank Spielberg and Phillip Shapiro. Mating habits of slugs: Dynamic carpool formation in the i-95/i-395 corridor of northern virginia. *Transportation Research Record: Journal of the Transportation Research Board*, (1711):31–38, 2000.

[190] François Sprumont, Paola Astegiano, and Francesco Viti. On the consistency between commuting satisfaction and traveling utility: the case of the university of luxembourg. *European Journal of Transport and Infrastructure Research*, 17(2):248–262, 2017.

[191] François Sprumont, Geoffrey Caruso, Francesco Viti, and Eric Cornelis. Considering activity pattern to achieve a more sustainable commuting behavior. 2016.

[192] John A Stankovic. Research directions for the internet of things. *IEEE Internet of Things Journal*, 1(1):3–9, 2014.

[193] Dave Steinberg, Frank Budinsky, Ed Merks, and Marcelo Paternostro. *EMF: eclipse modeling framework*. Pearson Education, 2008.

[194] P Stopher and L Wargelin. Conducting a household travel survey with gps: Reports on a pilot study. In *12th World Conference on Transport Research*, pages 11–15, 2010.

[195] Peter Stopher, Eoin Clifford, Jun Zhang, and Camden FitzGerald. Deducing mode and purpose from gps data. *Institute of Transport and Logistics Studies*, pages 1–13, 2008.

[196] Peter Stopher, Camden FitzGerald, and Min Xu. Assessing the accuracy of the sydney household travel survey with gps. *Transportation*, 34(6):723–741, 2007.

[197] Yusak O. Susilo and Kay W. Axhausen. Repetitions in individual daily activity–travel–location patterns: a study using the Herfindahl–Hirschman Index. *Transportation*, 41(5):995–1011, sep 2014.

[198] Benoit Thierry, Basile Chaix, and Yan Kestens. 2. Detecting activity locations from raw GPS data: a novel kernel-based algorithm. *International journal of health geographics*, 12(1):1, 2013.

[199] Bogdan Toader, Guido Cantelmo, Mioara Popescu, and Francesco Viti. Using passive data collection methods to learn complex mobility patterns: An exploratory analysis. *IEEE 21th International Conference on Intelligent Transportation Systems, November 4-7, 2018, Maui, Hawaii, USA*, 2018. accepted paper pending to be published.

[200] Bogdan Toader, Assaad Moawad, François Fouquet, Thomas Hartmann, Mioara Popescu, and Francesco Viti. A new modelling framework over temporal graphs for collaborative mobility recommendation systems. In *Intelligent Transportation Systems (ITSC), 2017 IEEE 20th International Conference on*, pages 1–6. IEEE, 2017.

[201] Bogdan Toader, Assaad Moawad, François Fouquet, Thomas Hartmann, Mioara Popescu, and Francesco Viti. A new modelling framework over temporal graphs for collaborative mobility recommendation systems. *IEEE 20th International Conference on Intelligent Transportation Systems, Yokohama, JAPAN, October 16 - 19*, (in-press), 2017.

[202] Bogdan Toader, François Sprumont, Sébastien Faye, Mioara Popescu, and Francesco Viti. Usage of smartphone data to derive an indicator for collaborative mobility between individuals. *ISPRS International Journal of Geo-Information*, 6(3):62, 2017.

[203] Emeric Tonnelier, Nicolas Baskiotis, Vincent Guigue, and Patrick Gallinari. Smart card in public transportation: Designing a analysis system at the human scale. In *Intelligent Transportation Systems (ITSC), 2016 IEEE 19th International Conference on*, pages 1336–1341. IEEE, 2016.

[204] Ankit Toshniwal, Siddarth Taneja, Amit Shukla, Karthik Ramasamy, Jignesh M Patel, Sanjeev Kulkarni, Jason Jackson, Krishna Gade, Maosong Fu, Jake Donham, et al. Storm@ twitter. In *Proceedings of the 2014 ACM SIGMOD international conference on Management of data*, pages 147–156. ACM, 2014.

[205] Sheung Yuen Amy Tsui and Amer S Shalaby. Enhanced system for link and mode identification for personal travel surveys based on global positioning systems. *Transportation Research Record*, 1972(1):38–45, 2006.

[206] Peter van der Waerden, Harry Timmermans, and Marloes de Bruin-Verhoeven. Car drivers' characteristics and the maximum walking distance between parking facility and final destination. *Journal of Transport and Land Use*, 10(1):1–11, 2017.

[207] David A Vautin and Joan L Walker. Transportation impacts of information provision & data collection via smartphones. Technical report, 2011.

[208] Rohit Verma, Surjya Ghosh, Mahankali Saketh, Niloy Ganguly, Bivas Mitra, and Sandip Chakraborty. Comfride: a smartphone based system for comfortable public transport recommendation. In *Proceedings of the 12th ACM Conference on Recommender Systems*, pages 181–189. ACM, 2018.

[209] V Vijayakumar, Subramaniyaswamy Vairavasundaram, R Logesh, and A Sivapathi. Effective knowledge based recommender system for tailored multiple point of interest recommendation. *International Journal of Web Portals (IJWP)*, 11(1):1–18, 2019.

[210] F. Viti and F. Corman. Equilibrium and sensitivity analysis of dynamic ridesharing. In *16th International IEEE Conference on Intelligent Transportation Systems (ITSC 2013)*, pages 2409–2414, oct 2013.

[211] Francesco Viti and Francesco Corman. Equilibrium and sensitivity analysis of dynamic ridesharing. In *Intelligent Transportation Systems-(ITSC), 2013 16th International IEEE Conference on*, pages 2409–2414. IEEE, 2013.

[212] Francesco Viti, C Tampere, Rodric Frederix, Marie Castaigne, Eric Cornelis, and Fabien Walle. Analyzing weekly activity–travel behavior from behavioral survey and traffic data. In *World Conference on Transport Research*, 2010.

[213] Katerina Vrotsou, Kajsa Ellegard, and Matthew Cooper. Everyday life discoveries: Mining and visualizing activity patterns in social science diary data. In *Information Visualization, 2007. IV'07. 11th International Conference*, pages 130–138. IEEE, 2007.

[214] Guozhang Wang, Wenlei Xie, Alan J Demers, and Johannes Gehrke. Asynchronous large-scale graph processing made easy. In *CIDR*, volume 13, pages 3–6, 2013.

[215] Lei Wang, Wanjing Ma, Yingling Fan, and Zhongyi Zuo. Trip chain extraction using smartphone-collected trajectory data. *Transportmetrica B: Transport Dynamics*, pages 1–20, 2017.

[216] Ran Wang, Chi-Yin Chow, Yan Lyu, Victor CS Lee, Sam Kwong, Yanhua Li, and Jia Zeng. Taxirec: recommending road clusters to taxi drivers using ranking-based extreme learning machines. *IEEE Transactions on Knowledge and Data Engineering*, 30(3):585–598, 2018.

[217] Overpass API OpenStreetMap Wiki. Overpass api - openstreetmap wiki. `https://wiki.openstreetmap.org/wiki/Overpass_API`.

[218] Jizhe Xia, Kevin M. Curtin, Weihong Li, and Yonglong Zhao. A New Model for a Carpool Matching Service. *PloS one*, 10(6):e0129257, 2015.

[219] Longgang Xiang, Meng Gao, and Tao Wu. 7. Extracting Stops from Noisy Trajectories: A Sequence Oriented Clustering Approach. *ISPRS International Journal of Geo-Information*, 5(3):29, mar 2016.

[220] Ali Yavari, Prem Prakash Jayaraman, and Dimitrios Georgakopoulos. Contextualised service delivery in the internet of things: Parking recommender for smart cities. In *Internet of Things (WF-IoT), 2016 IEEE 3rd World Forum on*, pages 454–459. IEEE, 2016.

[221] Ali Yavari, Prem Prakash Jayaraman, Dimitrios Georgakopoulos, and Surya Nepal. Contaas: An approach to internet-scale contextualisation for developing efficient internet of things applications. 2017.

[222] J. W. Yoon, F. Pinelli, and F. Calabrese. Cityride: A predictive bike sharing journey advisor. In *2012 IEEE 13th International Conference on Mobile Data Management*, pages 306–311, July 2012.

[223] Biying Yu, Ye Ma, Meimei Xue, Baojun Tang, Bin Wang, Jinyue Yan, and Yi-Ming Wei. Environmental benefits from ridesharing: A case of beijing. *Applied energy*, 191:141–152, 2017.

[224] Matei Zaharia, Mosharaf Chowdhury, Tathagata Das, Ankur Dave, Justin Ma, Murphy McCauley, Michael J Franklin, Scott Shenker, and Ion Stoica. Resilient distributed datasets: A fault-tolerant abstraction for in-memory cluster computing. In *Proceedings of the 9th USENIX conference on Networked Systems Design and Implementation*, pages 2–2. USENIX Association, 2012.

[225] Desheng Zhang, Tian He, Yunhuai Liu, Shan Lin, and John A Stankovic. A carpooling recommendation system for taxicab services. *IEEE Transactions on Emerging Topics in Computing*, 2(3):254–266, 2014.

[226] Desheng Zhang, Tian He, Yunhuai Liu, and John A Stankovic. Callcab: A unified recommendation system for carpooling and regular taxicab services. In *Big Data, 2013 IEEE International Conference on*, pages 439–447. IEEE, 2013.

[227] Desheng Zhang, Ye Li, Fan Zhang, Mingming Lu, Yunhuai Liu, and Tian He. coride: carpool service with a win-win fare model for large-scale taxicab networks. In *Proceedings of the 11th ACM Conference on Embedded Networked Sensor Systems*, page 9. ACM, 2013.

[228] J. Zhang, F. Y. Wang, K. Wang, W. H. Lin, X. Xu, and C. Chen. Data-Driven Intelligent Transportation Systems: A Survey. *IEEE Transactions on Intelligent Transportation Systems*, 12(4):1624–1639, dec 2011.

[229] Junping Zhang, Fei-Yue Wang, Kunfeng Wang, Wei-Hua Lin, Xin Xu, Cheng Chen, et al. Data-driven intelligent transportation systems: A survey. *IEEE Transactions on Intelligent Transportation Systems*, 12(4):1624–1639, 2011.

[230] Fang Zhao, Francisco Câmara Pereira, Rudi Ball, Youngsung Kim, Yafei Han, Christopher Zegras, and Moshe Ben-Akiva. Exploratory analysis of a smartphone-based travel survey in singapore. *Transportation Research Record: Journal of the Transportation Research Board*, 2(2494):45–56, 2015.

[231] Fang Zhao, Francisco Câmara Pereira, Rudi Ball, Youngsung Kim, Yafei Han, Christopher Zegras, and Moshe Ben-Akiva. Exploratory analysis of a smartphone-based travel survey in singapore. *Transportation Research Record: Journal of the Transportation Research Board*, 2(2494):45–56, 2015.

[232] Yilin Zhao. Mobile phone location determination and its impact on intelligent transportation systems. *IEEE Transactions on Intelligent Transportation Systems*, 1(1):55–64, mar 2000.

[233] Yu Zheng, Yukun Chen, Xing Xie, and Wei-Ying Ma. Geolife2. 0: a location-based social networking service. In *Mobile Data Management: Systems, Services and Middleware, 2009. MDM'09. Tenth International Conference on*, pages 357–358. IEEE, 2009.

[234] Yu Zheng, Quannan Li, Yukun Chen, Xing Xie, and Wei-Ying Ma. Understanding mobility based on gps data. In *Proceedings of the 10th international conference on Ubiquitous computing*, pages 312–321. ACM, 2008.

[235] Yu Zheng, Lizhu Zhang, Xing Xie, and Wei-Ying Ma. Mining interesting locations and travel sequences from gps trajectories. In *Proceedings of the 18th international conference on World wide web*, pages 791–800. ACM, 2009.

[236] Jianyu Jack Zhou and Reginald Golledge. Real-time tracking of activity scheduling/schedule execution within a unified data collection framework. *Transportation Research Part A: Policy and Practice*, 41(5):444–463, 2007.