Treball de fi de màster

Titol: Use of data science techniques for the modelisation of the public bike-sharing system of BIXI in Montreal, Canada

Cognoms: Yvon-Leroux
Nom: Jacob
Titulació: Màster en Ciència i Tecnologia de la Sostenibilitat

Director/a: Karina Gibert

Data de lectura: 12 de juliol 2017
Use of data science techniques for the modelisation of the public bike-sharing system of BIXI in Montreal, Canada

Author: Jacob Yvon-Leroux
Supervisor: Dr. Karina Gibert

A thesis submitted in fulfillment of the requirements for the degree of Master of Sustainability Science and Technologies in the

Institute de Sostenibilitat
Escola Tècnica Superior d’Enginyers de Camins, Canals i Ports

June 16, 2017
Declaration of Authorship

I, Jacob Yvon-Leroux, declare that this thesis titled, “Use of data science techniques for the modelisation of the public bike-sharing system of BIXI in Montreal, Canada” and the work presented in it are my own. I confirm that:

• This work was done wholly or mainly while in candidature for a research degree at this University.

• Where any part of this thesis has previously been submitted for a degree or any other qualification at this University or any other institution, this has been clearly stated.

• Where I have consulted the published work of others, this is always clearly attributed.

• Where I have quoted from the work of others, the source is always given. With the exception of such quotations, this thesis is entirely my own work.

• I have acknowledged all main sources of help.

• Where the thesis is based on work done by myself jointly with others, I have made clear exactly what was done by others and what I have contributed myself.

Signed:

Date: June 16th 2017
“Le cadre du BIXI est inspiré d’un boomerang car le vélo en libre-service revient toujours à bon port.”

Michel Dallaire, BIXI designer

“Le monde du partage devra remplacer le partage du monde.”

Claude Lelouch

“Life is like riding a bicycle. To keep your balance, you must keep moving.”

Albert Einstein

“The miracle is this - the more we share, the more we have.”

Leonard Nimoy
Abstract

Faculty Name
Escola Técnica Superior d’Enginyers de Camins, Canals i Ports

Master of Sustainability Science and Technologies

Use of data science techniques for the modelisation of the public bike-sharing system of BIXI in Montreal, Canada

by Jacob Yvon-Leroux

The final master thesis investigate behavioural patterns of the users of bike-sharing system BIXI in Montreal. A data mining approach is used by applying clustering methods on open data from BIXI, the Government of Canada and the City of Montreal. Hierarchical clustering with Ward and Gower is processed to form clusters using only the BIXI variables at first. Then, because the authors believe that contextual information has a fundamental effect on the user’s behaviour, they investigate the impact of adding such information into the clustering methodology. In the end, a multi-view clustering method is preferred for its quality of preserving the basic characteristics of simpler set of clusters (mixing less variables). Afterwards, local analysis are done over some profiles of the classes. A dynamic balancing analysis of the station using geographical representation offer good insight on the movements of the bikes associated with a cluster. Then, a complex network approach is also used to extract more topological information about the bike-sharing network made from the stations and the trips. All these transdisciplinary approaches used together produce new operational knowledge that can be useful for the operators and the logistics to overcome the redistribution problem as well as improving the quality of the service.
Acknowledgements

I would like to thank first and foremost my director and tutor Karina Gibert without whom this project would not have been possible. She proved the necessary patience and ambitions to elevate a greenhorn at data science and provide the leadership to bring this thesis to success.

I would also like to acknowledge Gabriel Damant-Siroirs and the Local Logic team that help to define the subject and develop the objectives of this project as well as putting me in contact with Nicolas Blain from BIXI Montréal. I am thankful, in fact, of the time and information Nicolas Blain gave me through a phone interview. As well, I am grateful of BIXI for opening its data without which this project would not have been possible.

At the academic level, I would like to thank Martí Rosas-Casals for his contribution of knowledge in complex network theory. I definitely owe a huge heartfelt gratitude to the Stack Exchange community for their constant help and providing of answers to almost all the questions and problems I encountered with LaTex, QGIS, R and even some theoretical concepts. Thank you to the creators of this template Steve R. Gunn, Sunil Patel and Vel.

Finally, I am thankful to Ludovic Alarie and his "Appartement" without which the path to the end would have been way harder and obscure. Above all, I want to express my deep gratitude to Chloé Gérin-Lajoie for her and contributions all along this project.
## Contents

Declaration of Authorship ........................................... i
Abstract .................................................................. iii
Acknowledgements ...................................................... iv

1 Introduction .......................................................... 1
  1.1 Objectives ....................................................... 2

2 Literature Review .................................................. 4
  2.1 History and Actual Context of Public Bike-Sharing Systems .......... 4
    2.1.1 First Generation ......................................... 4
    2.1.2 Second Generation ...................................... 5
    2.1.3 Third Generation ........................................ 5
      The Worldwide Exponential Growth of BSS ................. 6
    2.1.4 Fourth Generation ...................................... 7
    2.1.5 Overview of the Actual Context of Bike-Sharing Systems
      Around the World ............................................. 7
      Environmental Impacts ....................................... 8
      Health and Safety ............................................ 9
      Other social impacts ....................................... 10
      Main Issues and Negative Impacts ......................... 10
    2.1.6 Business Models and Providers .......................... 11
      The Business Models ....................................... 11
      The Key Stakeholders ...................................... 12
      Decision-making and Contracts ......................... 12
    2.1.7 Future Perspectives ..................................... 13
  2.2 The Use of Data in Public Bike-Sharing Systems .................... 14
    2.2.1 User Profile .......................................... 15
    2.2.2 Impact of Weather .................................... 16
    2.2.3 Trip Purpose .......................................... 16
    2.2.4 User Frequency ....................................... 17
    2.2.5 User Preferences ..................................... 18
  2.3 The Main Challenge: Geographical Unbalanced Distribution of the
      Bikes ................................................................ 18
    2.3.1 Network Strategical Design ............................ 19
    2.3.2 Redistribution .......................................... 19
      Operator-Based Redistribution ........................... 20
      User-Based Redistribution .............................. 20
      Static versus Dynamic Redistribution .................. 20
  2.4 Understanding Users Behaviour .................................. 21
3 Presentation of the case study 23

3.1 BIXI Montreal ................................................. 23
  3.1.1 The Story of BIXI ...................................... 23
  3.1.2 The Operating System of BIXI Montreal .............. 25
    The Stations .............................................. 25
    The Bikes ............................................... 25
    The Membership ........................................... 26
  3.1.3 Statistics of Usage ..................................... 27

3.2 Information Sources ........................................ 30
  3.2.1 The Database of BIXI Bike Trips ....................... 30
  3.2.2 The Stations Database ................................ 32
  3.2.3 The Urban Information ................................ 32
  3.2.4 The Weather Database ................................ 34
  3.2.5 Data Selection and Matrix Structure .................. 34

4 Methodology 36

4.1 General Methodology ....................................... 36

4.2 Preprocessing ................................................ 37
  4.2.1 The Database of BIXI Bike Trips ....................... 38
  4.2.2 The Stations Database ................................ 38
  4.2.3 The Geographic Information and Maps ................. 39
  4.2.4 The Weather Database ................................ 39

4.3 Basic Descriptive and Visualization ....................... 40
  4.3.1 Descriptive Statistics ................................ 40
  4.3.2 Presentation of Montreal ............................... 42
  4.3.3 The Dynamic Analysis ................................ 42
    Dynamic Maps of Stations’ Balancing Factors ............. 42
    Map of Trips and Their Frequencies ....................... 43

4.4 A Complex Network Approach ............................... 43
  4.4.1 Centrality measures .................................. 44
    Degree Centrality ....................................... 45
    Shortest Path ............................................ 46
    Closeness Centrality .................................... 47
    Betweenness Centrality .................................. 47
    Eigenvector Centrality .................................. 48

4.5 Clustering ................................................... 49
  4.5.1 The Ward’s Method .................................... 49
  4.5.2 Gower Distance ...................................... 50
  4.5.3 Cutting the Dendrograms ............................... 52

4.6 Robustness, Sampling and Variable Selection ............. 53

4.7 Interpretation of the Clusters ............................. 54

4.8 Integrated Multi-view Clustering .......................... 55
  4.8.1 Local Analysis of Recurrent Patterns ................. 56

4.9 Description of the Tools Used ............................. 56
  4.9.1 R ...................................................... 56
    Presentation of the Software ............................... 56
    Contributions and Packages ................................ 57
  4.9.2 QGIS ............................................... 58
    Presentation of the Software ............................... 58
5  Applications: Presentation of the Results 60
  5.1  Data Descriptive and Visualization ............................. 60
      5.1.1  Descriptive Statistics .................................. 60
      5.1.2  Visualization of BIXI within the city ...................... 66
      5.1.3  Dynamic study of Usage and Users’ Behaviour ............... 70
  5.2  A Complex Network Approach .................................. 74
  5.3  Clustering ..................................................... 78
      5.3.1  Clustering of the Original BIXI Data Variables: H1 ........ 78
            Dendrogram of H1 .......................................... 78
            Profiling and interpretation of the classes .................. 79
      5.3.2  Clustering of All the Variables Including the External Information: H6 ...................................... 81
            Dendrogram .................................................. 81
            Profiling and interpretation of the classes .................. 81
      5.3.3  Clustering of the External Variables Only: H11 ............. 82
            Dendrogram .................................................. 84
            Profiling and interpretation of the classes .................. 85
      5.3.4  Multi-view Clustering .................................... 85
  5.4  Local Analysis ................................................ 88
      5.4.1  Local Analysis of the Mundane & Flat Cloudy Trips ........ 89
      5.4.2  Local Analysis of the Weekend Member & Clear Trips ........ 92
      5.4.3  Local Analysis of the Nightout and the DTe Cloudy trips ... 94
  5.5  Sources of errors ............................................. 98
  6  Conclusion ..................................................... 99
     Future Perspectives ............................................... 101
A  Appendix A ...................................................... 103
    A.1  Metadata Table ............................................. 103
    A.2  Extra Results from the Complex Network Analysis ............. 103
    A.3  Figures accompanying the complementary local analysis .... 103

Bibliography ..................................................... 111
List of Figures

2.1 Growth in cities operating bike-sharing systems from 1998 to 2013. (Fishman, 2015) ................................................. 8
2.2 Distribution of means of transportation substitution by bike-sharing system in selected cities. (Fishman, 2015) .................. 9
2.3 Comparison of travel alternatives based on frequency and length. ................................................................. 17
2.4 Motivation to use a bike-sharing system .................................................................................................................. 18
3.1 The elements of a BIXI station. (Freemark, 2009) ............................................................................................... 25
3.2 A BIXI bike designed by Devito (2017) for the 375th anniversary of the City of Montreal. .......................... 26
3.3 Spatial Dispersion of the BIXI Stations in 2009. ......................................................................................................... 28
3.4 Service Level of the BIXI Stations in 2009 ................................................................................................................ 29
3.5 Redistribution Factor of the BIXI Stations in 2009. .......................................................................................... 29
3.6 Rate of subscription per year ................................................................................................................................ 30
3.7 The sources of information .................................................................................................................................. 31
4.1 Methodology Organizational Chart ......................................................................................................................... 36
4.2 Colour Scale for Balance ........................................................................................................................................ 43
5.1 Descriptive Statistics of the trips in Graphs .............................................................................................................. 61
5.2 Descriptive Statistics of stations in Graphs ............................................................................................................... 64
5.3 Descriptive Statistics of Weather in Graphs ............................................................................................................. 65
5.4 The lower outliers in the trip duration distribution. ............................................................................................... 67
5.5 Impact of bad weather on the use of BIXI. ................................................................................................................ 67
5.6 Topographic and urban features of Montreal. ......................................................................................................... 68
5.7 BIXI Stations vs the Environment .......................................................................................................................... 69
5.8 BIXI Stations vs the Population Density .............................................................................................................. 71
5.9 BIXI Stations vs the Urban Infrastructure .......................................................................................................... 72
5.10 Popularity of each station during the year 2015. ................................................................................................. 73
5.11 Difference between incoming and outgoing trips for each station along 2015. ................................................. 73
5.12 Dynamic Visualization of the Movement ............................................................................................................. 75
5.13 Most popular trips .................................................................................................................................................. 76
5.14 Degree Distribution $\alpha = 0.5$ .......................................................................................................................... 76
5.15 Directed network metrics .................................................................................................................................. 77
5.16 Dendrogram analysis and cut for clustering H1. ................................................................................................. 79
5.17 Interpretation of H1. .............................................................................................................................................. 80
5.18 Dendrogram analysis and cut for clustering H6. ................................................................................................. 82
5.19 Interpretation of H6. .............................................................................................................................................. 83
5.20 Dendrogram analysis and cut for clustering H11. ............................................................................................... 84
5.21 Interpretation of H11. ........................................................................................................................................... 86
5.22 Multi-view Clustering .......................................................................................................................................... 87
5.23 Balance Analysis of the Mundane & Flat Cloudy Trips ..................................................................................... 90
5.24 Most Frequent Mundane & Flat Cloudy Trips .......................... 91
5.25 Degree Distribution of the Mundane & Flat Cloudy Network ........ 91
5.26 Dynamic Analysis of the Mundane Trips During Flat Cloudy Context . 93
5.27 Balance analysis of the Weekend Member & Clear Trips ............... 95
5.28 Most frequent Weekend Member & Clear trips .......................... 96
5.29 Degree Distribution of the Weekend Member & Clear Trips ............ 96
5.30 Dynamic Analysis of the Weekend Member Trips During Clear Context 97

A.1 Centralities calculated using the Barrat et al. (2004) algorithm ....... 108
A.2 Centralities calculated using the Freeman (1978) algorithm .......... 108
A.3 Local Analysis of the Nightout Class and the DTe Cloudy Class independently .................................................. 109
A.4 A dynamic analysis of the DTe Cloudy context class .................... 110
List of Tables

3.1 Pricing scheme of BIXI for the season 2017. All prices are in CAD. (BIXI Montreal, 2017) .................................................. 27
3.2 Extra fees for time extension. (BIXI Montreal, 2017) ................... 27
3.3 Structure of the bike trips database ........................................ 31
3.4 Structure of the network status database ................................. 33
3.5 Structure of the weather data base ......................................... 34
3.6 Structure of the main data frame and description of its variables ... 35
4.1 Details on the variable Period ............................................... 38
4.2 Modifications of the Weather modalities into fewer more generic ones 41
5.1 Descriptive Statistics in Numbers ........................................... 63
5.2 Descriptive Statistics in Numbers ........................................... 66
5.3 Descriptive Statistics in Numbers ........................................... 66
5.4 Dissimilarities ..................................................................... 78
5.5 Characterization of H1 ......................................................... 81
5.6 Characterization of H6 ......................................................... 84
5.7 Characterization of H11 ......................................................... 85
5.8 Number of objects from H11 clusters into the H1 clusters.......... 87
A.1 Metadata Table of the Working Data Frame (Part 1) ................. 104
A.2 Metadata Table of the Working Data Frame (Part 2) ................. 105
A.3 Metadata Table of the Working Data Frame (Part 3) ................. 106
A.4 Metada of the Stations Data Frame ........................................ 107
## List of Abbreviations

<table>
<thead>
<tr>
<th>Abbreviation</th>
<th>Full Form</th>
</tr>
</thead>
<tbody>
<tr>
<td>ANOVA</td>
<td>Analysis of variance</td>
</tr>
<tr>
<td>App</td>
<td>Application for smartphones</td>
</tr>
<tr>
<td>BSS</td>
<td>Bike-Sharing System</td>
</tr>
<tr>
<td>CAD</td>
<td>Canadian Dollar</td>
</tr>
<tr>
<td>EU</td>
<td>European Union</td>
</tr>
<tr>
<td>GIS</td>
<td>Geographic Information System</td>
</tr>
<tr>
<td>IT</td>
<td>Information Technologies</td>
</tr>
<tr>
<td>NA</td>
<td>Not Available</td>
</tr>
<tr>
<td>PBSC</td>
<td>Public Bike System Company</td>
</tr>
<tr>
<td>PPP</td>
<td>Public Private Partnership</td>
</tr>
<tr>
<td>RFID</td>
<td>Radio-Frequency Identification Device</td>
</tr>
<tr>
<td>UK</td>
<td>United Kingdom</td>
</tr>
<tr>
<td>US</td>
<td>United States of America</td>
</tr>
</tbody>
</table>
Dedicated to Montreal, my city, my home.
Chapter 1

Introduction

It has been now more than 50 years that the alarm of the environmental consciousness is ringing (remember the *Silent Spring* from Rachel Carson (1962)). In 2008, the humanity became officially more urban than rural. Almost 30 years ago, the Brundtland report (UNWCED: United Nations World Commission and Development on Environment, 1987) was giving birth to sustainability and to the international agenda to aim for a brighter future. If the trend continues, the future appears as a communal web having to evolve and work collectively to sustain its ecosystem and live in serenity and satisfaction. Cities are winning in power and proactivity while the States are disappearing into the global international politics happening at a too greater scale for the human being.

The human scale has to be restrained to the city, its neighbourhoods and its environment. At this scale, mobility is complex, hence the necessity for a mixture of diverse means of transportation. The urban mobility has to be an intricate network of multimodal services. The famous last-mile problem is the new challenge that has to be overcome to take down another segregating factor of the urban society. The advent of the collaborative and sharing economies bear diverse potential solution including the one that is the subject of this master thesis: the bike-sharing system.

Bike-sharing has received increasing attention in recent years with initiatives to increase cycle usage improve the first mile/last mile connection to other modes of transit, and lessen the environmental impacts of our transport activities. The last decade has seen increasing the number of bike-sharing systems in an exponential way. And there are reasons why. On the first hand, this technology offers an array of benefits each as tempting as the next for cities that consider themselves every time smarter and more sustainable. On the other hand, it turns out that the management of such a service is a double-edged sword. Indeed, the logistics and the operations of a bike-sharing system has to face challenges such as theft, vandalism, customer service, marketing, but also, the most important of all, bike redistribution. Actually, the bike-sharing systems have become the sustainable and ecological mean of transportation, while contributing to a healthy lifestyle, but the unbalance between demand and offer has a critical impact on the perception of the service quality from the user perspective as well as on the social penetration of the service. Redistribution of the bikes to satisfy the demand where the demand is happening is even more essential when the network is asymmetric, meaning that a majority of users all go to the same place. The ideal would be to know the user’s intentions before it happens and plan the service consequently. Fortunately, big data collection using implanted technologies into the bikes allows a richness of information that goes in this direction. Thanks to data mining, it is possible to study numerous profiles of usage in function of a plethora of factors. Within the last decade, the academic world has been involved more and more on this domain and many discoveries come slowly to change the rules of the game. Authors like DeMaio (2009), Shaheen, Guzman, and
Zhang (2010), and Fishman (2015) provide professional and extensive overviews of the state of the art, the evolution and the discoveries at different periods and with different approaches of this sector. Other publication like the implementation guide proposed by ITDP (2013) proves again the manifest and growing interest for bike-sharing systems.

This master thesis aims at joining this scientific bibliography by contributing a data mining approach methodology using the open data of the Montreal BIXI system. Contribute to the characterization of the spatio-temporal patterns of BIXI, the behaviour of its users and, thus, increase the global knowledge about its network and functioning.

A clustering approach is used to make emerge behavioural patterns in relation with contextual information about the city or the trips themselves. Clustering allows the flexibility of not knowing beforehand the number or the types of classes we are looking for. It also allows easier and more efficient results of pattern identification in comparison to other classical methods. Once the trips are properly profiled, local analysis on the bikes movements among the network of stations can be done to obtain operational new knowledge to support decision-making at the management level.

This thesis will go through the following structure. First, an exhaustive overview of the bike-sharing system history will precede all the major findings and the state of the art concerning data science and the redistribution problem. Will follow the presentation of the case study. The BIXI system will be presented in detail as well as all the sources of information used to collect the working data. Then, a detailed description of the methodology will also explain the theoretical fundamentals necessary for the understanding of the project. At the end of this section are presented the two principal software used for the completion of this thesis. Afterwards, all the results emerging from the manipulations are exhibited and described. The following section discusses more in details the results used for the knowledge discovery and recommendations. Then, the thesis closes with the future lines of study and the conclusion.

Being intrinsically linked with urban mobility, bike-sharing systems are impregnated by challenging management. One of the biggest problem is to guarantee the availability of bikes and free docks at each station to the users that are looking for service. This directly influence the reliability of the system which indirectly affect the trust of the user into bike-sharing programs and infrastructures. Thus, it directly leads to the willingness of general adoption of this mean of transportation to structure the urban mobility plan. Therefore, it is fundamental to know the behaviour of the users in the network to design strategies that guarantee a full efficiency of the service so that citizens that wish to encourage the bicycle as a sustainable mean of transportation are encouraged to do it and can do it without any barriers.

1.1 Objectives

The primary objective of this research project is to use a data mining approach to process the open data from the BIXI bike-sharing system in Montreal in order to evaluate how data science techniques can provide added value to the company oriented to enhance the management of the service, in particular to avoid over-demand situations of available bikes or free docks. To this purpose, the open data available about trips will be exploited to identify patterns and discover relevant decisional
knowledge about user’s behaviour. To achieve this main goal, it is required to seek the following technical objectives:

- Review the state of the art and the developed data science models for the bike-sharing system services.
- Analyze the relevant data related with bike-sharing system phenomenon.
- Evaluate the availability of this data and find corresponding open data sources where it can be taken from.
- Use data mining methods to understand the user’s behaviour and extract relevant knowledge for the company.
- Use a complex network theory approach to complement the knowledge discovery acquired from the data mining one.
- Make an extensive use of geographic information available over trips to enrich the models.
- Validate the proposed models and extract operative knowledge from the [...] useful for supporting the decision-making and interpret the results. Propose tools to support the management of the bike-sharing service and plan the logistic that will guarantee the availability of bikes and free docks to the users at all time and all stations in the city.
- Acquire and improve personal skills about data science, spatio-temporal approach, geographic mapping, complex network theory and extraction of decisional knowledge from data.
- Evaluate the impacts of additional information on the models. Indeed, the weather, topographical and urban characteristics might have an impact on the user’s behaviour.

In the context of bike-sharing system, the company provides information about trips and stations, but the context of the trips to understand the patterns of mobility is of great importance. For this reason, information external to the company regarding the weather and the urban characteristics of the the city regarding where the trips start or end - like the altitude, the proximity to metro stations, etc. - will be also a matter of interest in this project. This brings into the picture a new objective which becomes essential for the success of this project:

- To evaluate the impact of including external contextual information into the model.
Chapter 2

Literature Review

2.1 History and Actual Context of Public Bike-Sharing Systems

Since the first generation of bike-sharing system in 1965 in Amsterdam, this sector has known a tremendous evolution. Most precisely in the past decade, the number of cities operating a public bike-sharing system (BSS) has increased from 13 in 2004 to 855 in 2014 (Fishman, 2015) and the global bike-sharing fleet recently passed the million cap (Goodyear, 2017). This boom was initiated with the advent of better and cheaper technologies and gave rise to the third generation of BSS such as BIXI in Montreal, Bicing in Barcelona and Vélib’ in Paris. It is now common to speak of four generations of BSS as the fourth generation is on the verge of slowly taking over (Fishman, 2015).

2.1.1 First Generation

The first initiative of public bike-sharing system happened in 1965 in Amsterdam under the name Witte Fietsen (White Bikes). A group of environmental activists called Provo introduced dozens of regular bicycles painted white and left unlocked for anyone to use and leave behind for the next person (Goodyear, 2017; Shaheen, Guzman, and Zhang, 2010). The only rule to use this system was to leave the bike unlock once arrived at destination so someone else could use it. The first goal of this system was to solve circulation problems in downtown Amsterdam. Hence, the first generation was a nonprofit system tackling social and environmental issues (Leduc, 2013).

However, the total absence of security mechanisms led to theft and vandalism, and a rapid demise of Witte Fietsen. Bikes were thrown into the canals or appropriated for private use (DeMaio, 2009). The final nail in the coffin was when the police seized all the unlocked or abandoned bikes under the reason that it was encouraging theft (Shaheen, Guzman, and Zhang, 2010). The program collapsed within days.

Nonetheless, bike-sharing as a concept was born. Two other first generation systems were founded, one in La Rochelle, France, in 1974 and another one on the Cambridge campus in 1993. The Cambridge green bikes initiative died within days as well since the whole fleet got stolen. However, the La Rochelle yellow bikes system surprisingly survived to this day (Midgley, 2009). Still, the first generation experienced very little growth due to its failures and technological advancements turned out to be necessary to reduce the threat of vandalism and theft and bring the second generation to life (Fishman, 2015).

To recapitulate, the first generation of bike-sharing systems is characterized by a lack of regulation, a similar and distinguished colour painting and no stations or parking areas. The main advantages are that free bicycles allow access and ease to a
new latent group of riders and like-coloured bicycles help to identify and promote the program and begin to give uniformity to the fleet. However, the main and many disadvantages are that it is impossible to predict the redistribution (the act of moving bicycles from where they are located to where they are needed, see section 2.3.2), there is no locking mechanism to secure bicycles, there is no way to link identity of user to that of a bicycle and the bicycles can be left anywhere so the supply of bicycles to user is unreliable. (Schroeder, 2014)

2.1.2 Second Generation

It took about 30 years for another major city to attempt an evolved large-scale public bike-sharing program and it is in Copenhagen, in 1995, that the second generation appeared. Copenhagen’s Bycyklen - or City Bikes - allowed users to access sturdy, shared bicycles at specific locations throughout the city via a coin-operated system. The major improvements were way more heavy-duty bikes able to resist vandalism and the introduction of money deposit concept which will remain with BSS for the following generations. (Leduc, 2013)

Basically, this second generation is characterized by standardized bicycles with unique parts, docking stations with mechanized locking systems around city and a coin-deposit system used to check-in and checkout. Its main advantages are the standardization of the bicycles with uniform and proprietary parts making the maintenance easier and more cost-efficient, the stations giving the users a clear idea of where to find bicycles and keep the system tidier, and advertising on bicycles creating revenue stream to help offset operating costs. However, it also has the disadvantages of the anonymity of the user and low deposit leading to thefts and abuses, and the lack of incentive to return the bicycle in a timely manner so trips/bicycle/day is low (a low utility of a public asset). (Schroeder, 2014)

Even though it was still not perfect, its success led to some replications of the system such as the Bycykler in Sandnes, Norway (1996), the City Bikes in Helsinki, Finland (2000), and the Bycykel in Aarhus, Danemark (2005). Just like the first BSS generation, the second one was managed by nonprofit organizations. Hence, the need to be funded by the municipalities since their systems were more expansive to operate. (Leduc, 2013)

2.1.3 Third Generation

The problems experienced by the first two generations of BSS led to the development of the third generation systems. The first system to overcome the anonymity problem was Bikeabout, the small scale BSS funded in 1996 for the Portsmouth University in the U.K. It was limited to students who had to swipe an individualized magnetic-stripe card to borrow a bike, which allowed them to be tracked when they were not returned. (Goodyear, 2017)

The third generation, which is the main one nowadays, brought the use of smart card tracking user’s identity, the bicycles they use, their origin and destination and other data; a fully automated service with real-time info flow and a pricing structure encouraging short trips. It has the advantages of providing accountability and security from theft by tracking identity, the real-time information flow provides useful data to users and operators, thus increasing system optimization, and the automation brings uniformity of service and limits human error. However, it requires a higher initial capital cost investment to cover the hardware and software costs.
These are the elements, in combination with growing public policy interest in cycling, that have enabled the rapid growth of BSS globally and, generating more revenues, the third generation is also accompanied by various business models and types of ownership (Fishman, 2015).

**The Worldwide Exponential Growth of BSS**

To give an overview of the exponential speed at which BSS grew with the advent of the third generation, only a few systems appeared between 1996 and 2005. In 1998, Vélo à la Carte in Rennes, France, was the first city-scale bike-sharing program to use magnetic-stripe cards and a radio-frequency identification device (RFID) technology. It was a partnership between the city of Rennes and Clear Channel, the advertising company, which developed and operated the new "Smart Bike" technology. The program was offered by the city free of charge and launched with 200 bikes at 25 stations. In 2005, the city of Lyon, France, replicated the exact same scheme, but at a greater scale: 1,500 bikes. Inspired by its smaller neighbours, Paris launched the 6,000 bikes Vélib’ system in 2007 then the worldwide movement toward bike-sharing was off and running. By 2015, Paris would have a total of 18,000 bikes. In 2008, Barcelona joined the movement with Bicing and, thus, initiated what quickly became a nationwide boom in Spain. By 2013, 132 Spanish cities would have bike-sharing programs, making Spain a global leader in the phenomenon. (Goodyear, 2017)

Thereupon, North America began to draw interest towards BSS and started to join the movement. It was Montreal the first city to host a city-scale BSS on the continent. In 2009, a government-owned company known as BIXI piloted its own system with innovative, robust bicycles and a modular docking system and quickly went full-scale. The company went on to provide bicycles and docking systems to cities in North America, Australia and London, but its financial situation happened to be precarious around 2013. From 2010 onward, BSS went globally viral. (Goodyear, 2017)

Washington, D.C., launched SmartBike DC, a 10 stations and 120 bikes pilot program that was the first modern BSS in the U.S. and followed Minneapolis’ Nice Ride and Denver’s B-cycle, representing the second of the two major suppliers of bike-sharing technology. Based on the London scheme launched in 2010 where corporate sponsors partially paid for the 6,00 bikes, New York’s bike-sharing system launched also with 6,000 bikes in 2013 and a first-of-its-kind funding model that used no public dollars, fully paid for by corporate sponsorship. The same year, Chicago and the San Francisco Bay Area also launch their own BSS. A bit more south, Mexico City and Buenos Aires launched their EcoBici, one of the first in Latin America. (Goodyear, 2017)

In China, Guanzhou inaugurated a program that was integrated with its Bus-Rapid Transit system, designed to solve the “last mile” problem and Hangzhou, launched its BSS with 2,800 bicycles and is today one of the world’s largest, with more than 78,000 bicycles. Actually, China has more than twice the number of BSS as the next closest country with 237 systems in the country, compared to 114 in Italy (2nd) and 113 in Spain (3rd) (Fishman, 2015). In Australia, Melbourne and Brisbane also started bike-sharing programs, the first in the country, but strict Australian helmet laws hampered adoption of the systems. (Goodyear, 2017)

Globally, 2013 saw a 60% increase in the number of BSS, with 65 new launches in China alone. In 2015, globally, the number of shared bicycles hit an estimated...
Chapter 2. Literature Review

1,000,000. China is far and away the leader in the sheer number of bicycles; three out of four of the world’s shared bikes are in that nation. (Goodyear, 2017)

2.1.4 Fourth Generation

There is still no clear consensus among the academic world about what exactly defines the fourth generation. Most authors describe it differently. Nonetheless, the most credible and accurate definition is the one of Schroeder (2014) that describes the fourth generation as characterized by solar-powered and modular stations that require no excavation, a smart-card integration with other transportation modes and some technological advances like GPS tracking, touch-screen kiosk, electric bikes and dynamic advertising. It decreases the installation cost by requiring no excavation for the stations, the solar power provides environmental benefits and allow greater flexibility in station placement and the technological advances like GPS tracking would provide a real-time mapping and help improve the redistribution.

On the other hand, some authors, like Fishman (2015), also evoke the potential advent of dock-less systems or another a revolutionary feature that still has to come to clearly define the fourth generation of BSS because most of attributes mentioned can be independently added to a third generation system to enhance it. Therefore, many third generation BSS have some, but not all, of the fourth generation attributes. In contrast, in a personal communication, Blain (2017) associated the dock-less systems with a fifth generation. However, he said that these systems encountered several problems recalling the failures of the Witte Fietsen in Amsterdam.

Nevertheless, as Shaheen, Guzman, and Zhang (2010) stress it: the fourth generation is definitely about overcoming the challenges of redistribution and balance between stations and big data and [...] technologies will certainly play a crucial role in the scenario. Emerging fourth-generation systems incorporate technological improvements for bicycle redistribution. For example, some systems use specially designed vehicles with automated technologies that facilitate demand-responsive bike relocation, but using larger, designated vehicles for bike transport increases implementation costs and is not emission free. Fourth-generation bike-sharing models also saw incentivized user-based redistribution (i.e., in which the rider performs bicycle redistribution) by using demand-based pricing in which users receive a price reduction or credit for docking bicycles at empty docking locations.

2.1.5 Overview of the Actual Context of Bike-Sharing Systems Around the World

The promising results from the third generation combined with the slow environmental global awaken and the rise of the cities as critical political players gave recently a massive boom to the BSS worldwide. The figure 2.1 presents the total number of operating bike-sharing systems worldwide as well as the new ones per year between 1998 and 2013. The number of cities operating a BSS has increased of almost 10,000% in a decade.

Each city has made bike-share its own, adapting it to the local context, including the city’s density, topography, weather, infrastructure, and culture. Although other cities’ examples can serve as useful guides, there is no single model of bike-share. This growth is mostly happening in Asia where China has become the world leader in sheer amount of bikes. Europe was a pioneer, but still increase it number of programs. North America is a relatively late comer to BSS, but has known some successful pioneering systems like those of Montreal, Washington, D.C. and New York.
Chapter 2. Literature Review

South America recently began to join the movement and has a promising future with successful implementation in Rio de Janeiro and Buenos Aires. (ITDP, 2013)

The ITDP (2013) report stresses the importance of a strong political support to ensure funding, land use rights, and coordination between various city agencies to achieve a successful implementation of a BSS. It is crucial to ensure the support from all parties to bring life into a sane development and avoid reform the programs every 4 years or so. Nevertheless, it is always easier to convince the political actors when the general impacts of a BSS are positive, although the negative ones are mostly economic.

Most of the literature praise BSS’ impacts highly. Indeed, at a first glance there seems to be only good sides to provide public bikes to a population in order to reduce car use and traffic, and increase general health. Let’s overview these impacts.

Environmental Impacts

Bike-share systems are often implemented as part of a general sustainable transport initiative to reduce pollution and improve mobility options. Indeed, Schroeder (2014) mentions that cycling decreases tailpipe emissions and noise pollution. It also reduces traffic congestion, hence reduce pollution and improve air quality. Interestingly though, Fishman (2015) stresses out that there is an implicit assumption that bike-sharing is used to replace trips previously made by car, yet the data suggest this is not always the case. The figure 2.2 shows the results of a multi-city survey presented by Fishman (2015) that was investigating which means of transportation BSS users substituted for bike-sharing. Bike-sharing primarily replaces trips that would have been taken by, in order, public transportation, walking and driving as his figure indicates. Nonetheless, he adds that a multi-city analysis showed that in all but one city, BSS reduced vehicle kilometres travelled. The exception was London due to its large coverage area and a low substitution rates from cars — probably because people already avoid taking their car to transit to the city centre — which require substantial re-balancing efforts. Fishman (2015) comments another research on the role of topography and weather, and the potential of incentives to encourage users to balance the system themselves, saying: “The results suggest that while price incentives may be sufficient on weekends, usage patterns on weekdays are such that a
Chapter 2. Literature Review

FIGURE 2.2: Distribution of means of transportation substitution by bike-sharing system in selected cities. (Fishman, 2015)

combination of operator and user redistribution is required to maintain an adequate level of service.”

Ricci (2015) confirms this statement and concludes in her paper that there is currently no evidence suggesting that bike sharing produces significant reductions in urban congestion levels and CO$_2$ emissions, or improvements in air quality, at least in the short-medium term. Just like Fishman, she observes that “rather than substituting for car journeys, bike sharing is predominantly used instead of walking and public transport. Moreover, when the effect of using motorized fleets for bike maintenance and re-distribution is accounted for, bike sharing can increase rather than reduce overall motor vehicle usage and emissions, with associated negative environmental and air quality impacts.” However, both nuance the results from the study cited by Fishman (2015) for its inability to include the contribution of casual users, who have been shown to have a different pattern of use than those of members.

Finally, for cities to have a positive environmental impact with their BSS, they must review their maintenance operations use zero-emission vehicles for the redistribution.

Health and Safety

Bike-share offers an active transport choice, providing both physical and mental health benefits. Studies have shown that spending twenty minutes every day on a bike has a significant positive impact on mental health (Obis, 2011). Woodcock et al. (2014) conducted a study showing that physical activity increased considerably at the population level after the implementation of a local BSS. The benefits were shown to differ by gender and age, with men’s major benefit coming from reductions in ischemic heart disease, whereas women were more likely to benefit in terms of reductions in depression. In another study, Fishman, Washington, and Haworth (2015a) found that, “depending on the mode substitution bike-sharing had a positive impact of physical activity, leading to an additional 74 million minutes of physical activity in London and 1.4 million minutes of physical activity in Minneapolis/St. Paul, for 2012. If the user had walked previously, there might be a decrease, but substitution from transit and car use led to increased activity.”

In relation with their safety, bike-sharing users are less likely to be injured in crashes than private bike riders. As an indirect impact, Schroeder (2014) mentions
a rise of motorists’ awareness due to an increase of cyclists in the street, hence leading to cycling safety improvement. It is also link with the so-called ‘safety in numbers’ phenomenon (related with Smeed’s Law (Smeed, 1949)), in which a rise in the amount of cycling does not lead to a proportional rise in the number of injuries. To support this statement, a study of hospital data from ten U.S. cities found that there was a dramatic reduction in the total number of recorded injuries after bike-sharing systems were introduced (Fishman, 2015). In a different multi-city study, Fishman and Schepers (2015b) looked at the reported injuries for various American, European and Australian cities. The results showed that, when compared to the level of risk for general cycling, bike-sharing appeared to be considerably safer. The authors suggest that whilst there is no overwhelmingly obvious explanation for this finding, two factors seem to stand out. First, bike-sharing speeds are lower than general cycling. Second, the bikes are more upright than most bicycles and this may allow the rider to see and be seen more easily. In contrast, Fuller et al. (2012) were unable to find a statistically significant difference in collisions or near misses after two years of assessing the BIXI system in Montreal.

Other social impacts

There are a number of other purported benefits of bike-sharing apart from those brought previously that have direct and indirect impacts on the functioning of a city and its social life. ITDP (2013) makes an exhaustive list in its report. BSS tends to increase accessibility to places that are beyond the reach on foot or public transportation; increase the reach of transit by partly solving the last-mile problem (Shaheen et al., 2012); improve the image of urban cycling culture by projecting a hip and modern image (Goodman, Green, and Woodcock, 2014); improve a city’s image and branding as a “green” or innovative city since cycling is seen as a sustainable transportation option; and generate investment in local industry through demand for hardware and software, as well as provision of the operations.

Schroeder (2014) formulates another exhaustive list which comes here to complete the ITDP’s one. He lists the direct impacts as BSS being a flexible, tailored, point-to-point personal public transport option; being a solution for short trips which eases strain on existing public transport systems, decreases congestion and improves service; being an efficient and organized use of public space by controlling parking; and attracting new or latent users to bicycle use, with consequent benefits associated with increased cycling. Some of the indirect impacts he also states are that BSS solves the "chicken-and-egg" scenario by providing users to justify investment into supporting bicycle infrastructure and increases property values through transit-oriented development and fosters urban revitalization.

Main Issues and Negative Impacts

What always hindered the use and development of BSS were vandalism and theft. They were the two major problems that each new generation was trying to solve. Technology helped to significantly reduce the importance of both these factors, mostly the third generation. The bikes are built in a way which diminishes the risk and the impact of vandalism. They are more solid, massive and heavy and the parts are unique and nontransferable to more traditional private bikes. Nonetheless, these problems still exist and they need to be taken into account if the fourth generation bikes are more fragile, technological or expansive.
Chapter 2. Literature Review

The second most important challenges with the actual systems are the redistribution and the re-balancing of the bikes. Even though the operators recently started to use predicting model to optimize the redistribution, there are still many aspects that can be improve to reach a totally autonomous and efficient system. This problem is not only expensive and generates most of the operating costs, but also pollutes enough to counterbalance the good environmental effects of using bicycle. A badly managed BSS pollutes more than no BSS at all (Ricci, 2015). Since the vandalism and theft problems have been significantly reduced, the redistribution problem should be at the heart of the fourth generation design and implementation.

Another concern that has raised some debates is about the helmet. In most cities, the helmet is not mandatory. However, in the cities were it is mandatory (in Australia), the BSS struggle against low rates of use instead of increasing compliance with the law. The percentage of users not wearing helmet varies a lot depending on the type of user and the city. For example, in New York City, approximately 85% of Citi Bike users do not wear a helmet, and the rate is 45% for Capital bike-sharing in Washington, D.C., and 63% of long-term subscribers did not wear helmets while it rises to 94% for short-term users. According to Fishman (2015), short-term users are more likely to take spontaneous trips, in which they did not have a helmet with them. Besides, some studies (Fishman and Schepers, 2015b) have concluded that the bike-sharing users were less at risk to get involve into accidents, which counterbalance the usual lack of helmet of the users.

However, in one case study, researchers found a negative health impact for women when they applied the general crash risk for all cycling in central London, due to the greater fatality rate among female cyclists there. (Fishman, 2015)

2.1.6 Business Models and Providers

As mentioned previously, with the third generation of BSS also came different business models, different financing methods and different service providers. Since the main objective of this thesis is to provide an insight on the decision making to help the management and operations of the BSS, it is fundamental to understand the internal structure of such a service and how decisions are being made. In its report “The Bike-Sharing Planning Guide”, ITDP (2013) gives an excellent explanation of the BSS business along with one of the best overview of the various business models that are running actually around the world. Therefore, most elements’ explanations are integrally coming from this source.

The Business Models

The business model defines the asset ownership and revenue flow between the municipal government and the operator. The goal is to balance service provision with resource allocation. To do that, the government will need to consider three things: organizational structure, asset ownership and contracting structure. (ITDP, 2013)

As a public transport system, bike-share should be situated similarly to other public transportation systems. Most systems hardly provide a return on investment, which puts the service closer to public transport models than toll roads or parking management models. They are not large profit businesses. (ITDP, 2013)

A bike-share system can be completely public or completely private, but most successful systems are a combination of the two: public-private partnership (PPP). The decision regarding which aspects should be public or private depends on the environment in which the system operates. Different cities need different structures
to meet their specific needs, and this should be analyzed in the feasibility stage. In selecting a business model, the government must weigh the desired utility that bike-share will provide to a user against the necessary resources. (ITDP, 2013)

Zhang et al. (2015) stresses that each BSS is designed and implemented to be unique, bespoke and a one-off. In addition one must recognize that the stakeholder relationships are likely to be interdependent, multi-embedded and, sometimes, intangible.

The Key Stakeholders

The key stakeholders are the implementing agency, the operator and the owner of the assets. The implementing agency is the government entity that oversees the planning, implementation, and operations of the BSS. It is best to consider what a system might look like in five or ten years, and place the agency accordingly. This will streamline decision-making, growth, and general administrative processes. The implementing agency will be responsible for detailed system design, tendering and contracting, developing the financial model, and infrastructure implementation. Once the system has been launched, the implementing agency will need to manage it and evaluate the operator’s performance according to the defined service levels. The implementing agency plays the role of referee, keeping the best interests of the government and the customers in mind, while also focusing on the financial interest of the operator. Because of this, the agency ideally should be independent of the contractor operating the system. (ITDP, 2013)

The operator is the entity that handles the day-to-day operations of the public BSS. The operator’s duties include managing the maintenance and general cleanliness of the fleet of bicycles and stations, as well as the redistribution of the bicycles. Except in special circumstances, the operator also handles the customer service, payment processing, marketing, and general brand management of the system (ITDP, 2013). According to Zhang et al. (2015), BSS operators are those having to deal with most of the challenges discussed in this thesis. For example, they are those in charge of considering the location and size of bike stations, the forecasting and scheduling of customer demands, route choice and development, bike maintenance and bike redistribution. They also need to deal with specific issues such as theft and vandalism and undertake relationship-based marketing to harmonize potential and even direct competition from local taxis, buses, and bike owners.

The ownership of the assets — primarily the stations, terminals, docks, bicycles, and IT system — is usually determined by the implementing agency, as well as the permanency of the assets in the streetscape. The different assets of a system can have varying ownership, and the assets may be shared, transferred, or licensed. Control of the BSS is closely bound to asset ownership: the owner determines the investment, and thus the quality of the system. (ITDP, 2013)

Decision-making and Contracts

Decisions about asset ownership will shape the contracting structure. There may be separate contracts with the suppliers of each of the various components of the BSS, which can include the following: hardware, software, operations, advertising on the BSS, and marketing and public relations. Bundling of the contracts can bring simplicity, with the government having to manage only one contract, thus focusing accountability on a single entity, but in some situations, signing separate contracts can be a better choice since it helps mitigate the risk that accompanies reliance on a
single entity and allow the government to contract with an entity that specializes in the requested service. (ITDP, 2013)

In its report, ITDP (2013) lists three main types of contracting structures, as defined by the ownership of the assets:

- Publicly owned and operated: The government owns the assets and provides the services.
- Publicly owned and privately operated: The government owns the assets but contracts a private entity to run the services.
- Privately owned and operated: The private entity owns the assets and provides the services.

Regardless of the structure, in all cases, the government, through the implementing agency, still oversees the system and is responsible for managing the contracts and monitoring the level of service (ITDP, 2013). Zhang et al. (2015) add that for a bike-sharing system to be successful it is essential to run operations as a not-for-profit charity and to be subsidized by local government or other funding bodies whose ultimate multi-dimensional goal is to reduce environmental impact, lessen traffic congestion, enhance mobile connectivity, and finally improve public health.

In the end, the most crucial aspect of a BSS business model is that it needs to create value for stakeholders — city councils, operators and assets owners, but also advertising agencies, communities and private sector organizations such as bike providers — whilst comprising of a first-rate plan with an economic rationale that increases revenues and lowers costs say Zhang et al. (2015).

2.1.7 Future Perspectives

Apart from the upcoming dock-less BSS, GPS-equipped fleet also raise a lot of attention. As GPS becomes increasingly affordable, they started to appear more in more, mostly in Europe. GPS may reduce the need for physical docks. Apart from the obvious security benefits, GPS may assist bike-sharing operators by detecting when a bicycle has moved outside a given area. Operators may also use GPS to assist with the challenging task of redistributing bicycles across their fleet via the use of real-time tracking. The automated data collection offered through GPS provides new opportunities for data analysis, which may not only be useful for bike-sharing operators to understand how their system is being used, but also from a wider transport planning perspective. Openly available GeoJSON data files may assist governments plan and evaluate bicycle route usage and effectiveness. (Fishman, 2015)

In recent years a number of BSS have launched bikes that offer electric pedal assistance. This technology offers the potential to increase the attractiveness of bike-sharing to those who may not have previously seen it as an option. Longer trips, challenging topography, excessive heat and other factors associated with physical exertion can act as barriers to transport cycling generally (Fishman, 2015). Furthermore, many bike-sharing cities have experienced re-balancing issues associated with the city’s topography. It is typical for users to ride downhill and show a reluctance to return bicycles to stations located at a higher elevation (Fishman, 2015). Electrical bikes may reduce this flow imbalance and be more suitable in hilly, hot or dispersed cities. (Fishman, 2015)

As it has been emphasized many times already, the need to develop more efficient methods of redistribution and stations balancing presents researchers with a complex analytical challenge. Fortunately, the amount of data gathered have already
started to offer researchers a decent platform to embark on this challenge. Transver-
sal collaboration is required since it concern many fields such as behavioural eco-
nomics, data science, engineering, urban planning, demography and more. Con-
cerning the redistribution, more experiments about incentives to use the users as
a redistributing tool could make a notable difference in the BSS to come. Fish-
man (2015) believes that there is a paucity of research with large numbers of people
who are not BSS users. Such studies are of critical importance to bike-sharing user
growth, particularly in underused systems. He also mentions as potential fields of
study the reaction of drivers to bike-sharing users since early investigations indicate
that drivers may react differently to bike-sharing riders than general cyclists. Also
standard and justifiable tools to measure the impacts of bike-sharing, in terms of cli-
mate change, congestion, air and noise quality, as well as health and time savings
are greatly needed. As noted in this report, China has the biggest BSS scale, yet re-
search activity does not reflect this. A much greater focus on Chinese bike-sharing
needs to occur, as the sheer scale of their systems may provide important insights not
just for China but for bike-sharing generally. Finally, Fishman (2015) asks for an in-
ternational platform, supported by governments, universities and the bike-sharing
industry, to share and coordinate research as well as to provide a global framework
for better transportation. (Fishman, 2015)

2.2 The Use of Data in Public Bike-Sharing Systems

Big data presents opportunities to identify problems, analyze and enhance projects
and services. It also provides new opportunities to predict users behaviour before it
happens, using a combination of real time information, historical trends and clever
algorithms. The data is transformed into intelligent information which is then used
by relevant authorities for management. An understanding of origins and destina-
tions of trips – why, who, how and when they are made – are required for short-
and long-term planning or adjust network. Traditionally, it was an expensive and
time consuming task from conducting thousands of surveys, and complementing it
to socioeconomic data analysis. Now, data analytical tools are used to understand
journey patterns from data. This enables a better understanding of service demands
enabling an appropriate response to over-demand or planned maintenance work.
Those data are usually publicly available (open) so that they can be used by applica-
tion developers and researchers as well as the city. Using Big Data does come with
its own set of challenges, including data security and cluster of networks that turn
big data into ‘Smart Information’. The gains from using big data are significant; it
just need to have the right legal structure for safe data sharing.

More precisely, the operator has a particular interest in learning the daily and
weekly cycles of system activity, and the effect of external events such as weather
and transport strikes on movements, because effective redistribution is important
for many of the systems – particularly if they are dominated by asymmetric flows at
certain times of the day (e.g. commuters) but with other user types requiring them
at other times (e.g. tourists) or they are simply too small or too popular for their
city to be able to satisfy demand. To do so, many metrics have been standardized n
the recent years as common indicators to evaluate and compare BSS. The standard
metric for comparing the usage of different bike-sharing systems is the number of
trips per day per bicycle. According to this metric, Barcelona’s Bicing has the most
heavily used BSS across the year and Paris’ Vélib’ reaches the highest peak with
eight trips per day per bike in September (Fishman, 2015). Another example is the
time a docking station remains full or empty. It may actually be desirable to retain or 
even affect such a state, if it is likely that a rapid change in the numbers of bicycles 
(caused by the users themselves rather than operator redistribution activity) can be 
expected shortly (O’Brien, Cheshire, and Batty, 2014).

Nonetheless, there are relatively few quantitative studies on BSS employing ac-
tual bicycle usage data according to Imani et al. (2014). Still, the amount of studies 
on the subject are booming since the third generation. For instance, several studies 
have analyzed factors affecting bike flow and usage. To name just a few, Rixey (2013) 
and Daddio (2012) both investigated the effects of demographic and built environ-
ment characteristics on average monthly bicycle usage in three different cities in the 
US (only Washington DC for Daddio) at the station level using a regression analy-
sis. He concluded that population density, job density, income levels, and the share 
of alternative commuters are all critical factors affecting bicycle-sharing ridership. 
Similarly, Wang et al. (2013), in their analysis, considered annual rates for each sta-
tion and examined the effects of nearby business and job densities, sociodemograph-
ics, built environment, and transportation infrastructure variables on annual usage 
flows. They found that locating stations closer to jobs results in higher usage of 
the bicycle-sharing system. Moreover, the presence of food-related businesses near 
stations has a more positive impact on arrivals and departures than non-food com-
mercial businesses. Buck and Buehler (2012) explored the influence of population, 
bicycle lanes, number of households without a car, and retail destinations around the 
stations on bike usage in Washington DC. Shu et al. (2010) used train ridership data 
as demand estimates and developed a network flow model to analyze bike sharing 
systems. Singhvi et al. (2015) used taxi usage as a co-variate in bike usage predic-
tions, finding it particularly useful for predicting pairwise demand, and they pro-
pose a neighbourhood approach in analyzing flows between stations. Krykewycz 
et al. (2010) estimated the demand for a proposed BSS in Philadelphia using ob-
served bicycle flow rates in European cities. Nair et al. (2013) investigated several 
aspects such as the BSS characteristics, utilization patterns and the connection with 
public transit using data from Paris’ Velib’ BSS. Finally, Imani et al. (2014) analyzed 
the BIXI system in Montreal using meteorological data, temporal characteristics and 
built environment attributes.

This short overview of only some researches in the field that represent BSS data 
is just beginning to use the Big Data produced by all the systems and supply the op-
erators with knowledge to enhance the management. Examples of game-changing 
knowledge acquired from data science there is the typical user profile, the impact of 
weather on the usage, the general purpose of use, the frequency of use for each type 
of user, and user preferences.

2.2.1 User Profile

In his global review of the research projects on bike-sharing system, Fishman (2015) 
established the typical user profile based on various studies (principally conducted 
in the US and the UK) as: “bike-sharing uses are on average disproportionately of 
higher education and income, more likely to be male and white. The gender dis-
parity does appear to be smaller, however, than for private bike riding.” In addi-
tion, Ricci (2015) defines the typical user profile as: “male, white, employed and, 
compared to the average population in which BSS are implemented, younger, more 
affluent, more educated and more likely to be already engaged in cycling indepen-
dently of bike sharing.”
2.2.2 Impact of Weather

Weather is a common factor that researchers try to correlate with cycling behaviour. For example, Flynn et al. (2012) affirmed that precipitation and temperature appeared to be strong influences on the odds of commuting to work by bicycle. The odds of bicycle commuting nearly doubled when no precipitation was recorded for the morning commuting hours. Bicycle commuting decisions similarly appeared to be sensitive to temperature. Finally, increased wind speed diminished the odds of bicycle commuting modestly.

In a bike-sharing system context, Imani et al. (2014) carried out an elaborated study using real usage data from BIXI BSS in Montreal. Their study examines the influence of meteorological data, temporal characteristics, bicycle infrastructure, land use and built environment attributes on arrival and departure flows at the station level using a multilevel approach to statistical modelling. They identified the following key correlates to bicycle flows: weather conditions, with users more likely to bike-sharing under good weather conditions; time of day/week: during the weekends the bicycle usage reduced, however Friday and Saturday nights were positively correlated to arrival and departure rates; the provision of cycle infrastructure, with bicycle flows and usage of the BSS increasing with cycle lanes nearby a BIXI station; and the characteristics of the built environment around the stations, with bicycle flows decreasing further away from the core business district. Accessibility indicators appeared to be correlated to bicycle usage for every BIXI station. Restaurants, other commercial enterprises and universities in the vicinity of a station significantly influenced the arrival and departure rates of the BIXI station.

Comparatively, Corcoran et al. (2014) suggested that inclement weather conditions are significant detractors for both recreational and commuting bicycle trips. Temperature was found not to be a significant independent factor influencing the number of trips taken, but when modelled simultaneously with other factors returned a significant and small positive influence on usage. both rain and wind (specifically stronger winds) reduce the frequency of trips, particularly the longer trips. As well, they showed that calendar events (in particular public holidays) exerted some subtle variation in the spatial distribution of trips despite showing no effect at the system-wide level.

2.2.3 Trip Purpose

According to Fishman (2015) the typical purpose of a bike-sharing user in North America depends on the type of user. The long-term members will use the BSS to commute to work or school whereas the short-term and casual users will use it for leisure or tourism. Consequently, the purpose vary considerably depending if it is during the weekend or not, or a public holiday. Members create peaks of usage during working rush hours. Usage between systems can vary significantly, but generally peaks between 7:00 and 9:00 and 16:00 to 18:00 weekdays; on weekends, usage is often strongest in the middle of the day and in the afternoon. The average ride duration is between 16 and 22 minutes, with casual users typically taking longer trips than annual members. (Fishman, 2015)

As well, Fishman (2015) mentions a study claiming that trip purpose can vary by residential location, age, gender, ethnicity and whether the member has a car available for their use. Women were found to be more likely to report making errands by bike-sharing, whereas men were more likely to report commute trips by bike-sharing.
These patterns of use are so consistent that they form the basis for most predicting models as it will be discussed in the further section 2.4.

2.2.4 User Frequency

Various studies concluded that BSS has the ability to drive up use and frequency of cycling, i.e., for expanding the market, with an expansive effect on demand similar to that recorded in other transportation sectors, such as the effect of low cost airlines on air transport, the high-speed train on rail passenger transport, and container-carriers on maritime goods transport (Castillo-Manzano, Castro-Nuño, and López-Valpuesta, 2015). Surprisingly, even though BSS increase the use and frequency of cycling, bike-sharing members are not particularly frequent bike-sharing users based on North American and Australian data (Fishman, 2015). It appears that many bike-sharing subscribers may view the service as an occasional complement to their primary or secondary transport modes. In focus group discussions with bike-sharing members, a commonly reported motivation for signing up was a desire to show support for the government decision to initiate a BSS. This may help explain why around half of members report no usage in the previous month concludes Fishman (2015).

Mátrai and Tóth (2016) explain that BSS are a complementary mean of transportation between walking and public transportation. Its main advantage is that the cost of ownership is not a burden since users don’t have to deal with bike storage, protection and maintenance as oppose to owning its private bike. Figure 2.3 shows the preferred mean of transportation in function of the frequency and the length of the trip. Mátrai and Tóth (2016) highlight the niche that occupy BSS among other means of transportation. They say that “it is mainly used for short distance, occasional travel, therefore it usually does not provide alternatives for commuters. However, these systems can be an option for last mile, provide further transport alternatives, promote cycling in general, boost up safe infrastructure investments and raise awareness.”

![Figure 2.3: Comparison of travel alternatives based on frequency and length. (Mátrai and Tóth, 2016)](image-url)
Bachand-Marleau, Lee, and El-Geneidy (2012) studied the factors influencing the use of shared bicycle systems and those impacting the frequency of use in Montreal. They brought to light four interesting conclusions about the location of the stations, the transportation habits of current and potential users, the fear of bicycle theft and the status and perceptions associated with shared bicycles. First, a greater number of docking stations close to origins of potential users in residential neighborhoods has a greater impact on the increase in the number of system users than the proximity of docking stations to destinations, yet both are incentives. Second, transit users, people combining cycling and transit for their trips, and those who have a driver’s license are more likely to use BSS. Special multimodal offers would encourage individuals to adopt shared bicycles by making the integration into their current travel habits as seamless as possible. Third, individuals recognize shared bicycles as an interesting active travel option in minimizing bicycle theft. Finally, individuals who like the design of shared bicycles tend to use the system more often. BSS that are considered “trendy” increase their frequency of usage.

2.2.5 User Preferences

In 2013, Washington, D.C.’s Capital bike-sharing surveyed some 11,100 of its members. The results are reported at figure 2.4 taken from Fishman (2015). Of those responding, 69% said that getting around fast and easily was a “very important” part in their motivation. Therefore, convenience is a significant factor in bike-sharing use.

Lower-income users cited the importance of saving money compared with other transportation options. In a study of London’s system, members who resided in poorer areas had higher trip rates than those living in more affluent suburbs. (Fishman, 2015)

The distance between users’ homes and the nearest docking station is an important predictor for bike-sharing membership. For example, Bachand-Marleau, Lee, and El-Geneidy (2012) found that Montreal residents living within 500 meters of a docking station were 3.2 times more likely to have used the BSS. Consequently, system density is a key factor in overall usage.

2.3 The Main Challenge: Geographical Unbalanced Distribution of the Bikes

According to Médard de Chardon and Caruso (2015), a large proportion of the existing bicycle sharing literature can be grouped into two fields, the mathematical models focusing on re-balancing and those characterizing BSS through analysis. This
section will introduce the main challenges that require data science and computer modelling, as well as the famous re-balancing problem and the proposed solutions.

2.3.1 Network Strategical Design

One important subject of prediction modelling is about the development of a BSS network and positioning of the stations. Briefly, Zhang et al. (2016) define it as the design work about various aspects of the BSS—like the number and locations of bicycle stations, the creation of bicycle lanes, the selection of paths, etc.—and requires comprehensive analysis and modelisations. They mention works about methods for maximizing the throughput of a mobility-on-demand urban transportation system; a re-balancing policy that minimizes the number of vehicles performing re-balancing trips and considered the solution to a linear program effectively in the proposed model; and a strategic design problem for BSS incorporating bicycle stock considerations as a hub location inventory model.

For BSS that exhibit well-defined and constant patterns of use like it is the case in North America, the UK and Australia (Fishman, 2015), the modelling efforts are more about network strategical design than redistribution [personal communication with Blain, 2017]. Mostly, when the patterns of usage present some simplifying characteristics—for example, all trips going to the city centre in the morning rush hours—once this factor is well managed, the challenge is to predict who will be the next potential subscriber and how to develop the network to convince and sustain him as well as make evolve the whole BSS with the evolving users and contexts. (Blain, 2017)

Although this field is of great interest, it is not the focus of this research study.

2.3.2 Redistribution

Redistribution is broadly defined by ITDP (2013) as “the re-balancing of bicycles from stations that are near or at capacity to stations that are close to empty.” When either occurs, half of the functionality of a BSS station is lost, either no bicycle can be returned or taken precise Médard de Chardon and Caruso (2015). Well managed systems aim to avoid these situations by predicting when they are likely to occur based on travel patterns throughout the day or the week. Successful redistribution is critical to the viability of the system from the customer’s perspective, and redistribution is one of the greatest challenges of operating a bike-share system, accounting for as much as 30% of operating costs in European systems (Obis, 2011). Bacciu et al. (2016) found that users identified “finding an available bike and a parking slot” as the two most important problems encountered. These problems should therefore be addressed in the best possible way within the obvious budget constraints, while keeping the number of unavoidable kilometres made by the redistribution fleet as small as possible. If an operator has an adequate IT system, redistribution becomes predictive, and is better thought of as pre-distribution—the movement of bicycles to stations where users will need them and away from stations where users will be dropping them off.

Redistribution indicators according to ITDP (2013) are:

- Percentage of the time that high-priority stations are empty during peak hours (7:00-10:00 and 16:00-19:00)
- Percentage of the time that high-priority stations are empty during off-peak hours
• Percentage of the time that low-priority stations are empty during peak hours (7:00-10:00 and 16:00-19:00)
• Percentage of the time that low-priority stations are empty during off-peak hours
• Minimum percentage of total cycle fleet available at 6:00

The redistribution of bikes is closely related to the quality of service from the customer perspective. However, the quality of the service provided by BSS operators is not easily verifiable. Some methodologies use spatial and temporal frequencies of re-balancing completed by providers which allow a municipal oversight of service quality. Although, it cannot be stated that systems with greater re-balancing quantities are better managed as individual BSS may have non-symmetric spatial tendencies, such as elevation differences, that accentuate the need for re-balancing. (Médard de Chardon and Caruso, 2015)

This is why the analysis of statistical bicycle usage patterns is primordial (Han, Come, and Oukhellou, 2014). Real-time monitoring and forecasting stations’ state—both from the aggregate and local perspective—is of high value to transport operators, who may re-balance the system in a dynamic manner (Lathia, Ahmed, and Capra, 2012).

Redistribution scheme can be mainly divided into two standard categories: “user based” and “operator based” (Mátrai and Tóth, 2016).

**Operator-Based Redistribution**

For operator-based redistribution, the relocation of the bikes is assigned to the service staff which generates operating costs. Common operating costs are expressed in an annual-per-bike amount and can range drastically depending on redistribution mechanisms and needs, labour costs and service level delivery. Redistribution vehicles are usually flatbed trucks or trailers carried behind vans or pick-up trucks and they represent a significant investment (ITDP, 2013). For example, to manage its 20,600 bicycles, Vélib’ uses 20 natural gas powered vehicles to transport bicycles from one station to another (Shaheen, Guzman, and Zhang, 2010).

The analysis of statistical bicycle usage patterns (Han, Come, and Oukhellou, 2014). Real-time monitoring and forecasting stations’ state—both from the aggregate and local perspective—is of high value to transport operators, who may re-balance the system in a dynamic manner (Lathia, Ahmed, and Capra, 2012).

**User-Based Redistribution**

In this case, the users are stimulated to return to an empty station in order to balance the distribution of the bikes among the stations. This incentive usually take the form of a bonus-malus structure system. These tariff structures provide bonuses for users who use the system in the opposite direction of the regular flow in a given time (e.g. out-ward from the city centre in the morning peak, upward to the hill, etc.). (Mátrai and Tóth, 2016)

**Static versus Dynamic Redistribution**

The BSS reallocation can be carried out either during the night when the bikes demand is negligible (static re-positioning, as defined in literature), or during the day
when the bikes distribution among the stations rapidly changes due to the high demand level (dynamic re-positioning). According to ITDP (2013), most system were trying to do most of the redistribution at night, when there is less traffic and it is more efficient. However, most of the trips occur during the day between 7:00 and 9:00, and 16:00 and 18:00. During those periods, redistribution is necessary to provide a satisfying service, especially for stations that experience high peak-demand. For example, BIXI is redesigning redistribution trucks to include on-board computers that can provide drivers with real-time information on bicycle stations to facilitate a speedier and more efficient response to bicycle shortages and station overcrowding during the day (Shaheen, Guzman, and Zhang, 2010).

Mathematical approaches and modelisations emerge more and more to solve the large problem of the redistribution of bikes due to the inequalities of demand (Mátrai and Tóth, 2016). With these mathematical processes the redistribution can be more cost-effective; the distribution of the bikes can reflect more on the demand. However, according to Caggiani and Ottomanelli (2013) and Lathia, Ahmed, and Capra (2012), most of the literature attacks only the static bike-sharing pickup and delivery problem: optimizing how a fleet of trucks should move throughout the city given that the station contents are, for the most part, not themselves on the move (i.e., at night time). Caggiani and Ottomanelli (2013) stress that the dynamic case is generally investigated without focusing at redistribution patterns and time periods. Some works propose a fixed re-positioning time interval, a variable re-positioning time interval or assume redistribution vehicles moving at random from saturated stations to empty ones. Therefore, there is definitely a need for more research on the dynamic redistribution scheme and, in deed, it is a burgeoning sub-topic within bike-sharing research. On this matter, Fishman (2015) dresses an extensive list of recent findings in this field. Some researchers (e.g. Imani et al. (2014)) have examined the factors associated with higher and lower levels of docking station activity, finding that weather and the presence of restaurants have a predictable impact of station activity. Other researchers have also identified a relationship between weather and station activity; however, inclement weather is much more likely to impact on casual users than members with a commuting function (Fishman, 2015). Other work (like Frade and Ribeiro (2014)) has examined the impact of topography on station activity. Parkes et al. (2013) suggest altering the price to achieve re-balancing objectives may increasingly be employed as an option to resolve fleet distribution issues. Fricker and Gast (2016) have investigated the effectiveness of providing users with incentives to redistribute bikes, using complex mathematical modelling. Pfommer et al. (2014) used historical data on the London BSS to model the effectiveness of employing trucks for redistribution as well as the impact of introducing price incentives to the user to mitigate fleet imbalance. The results suggest that while price incentives may be sufficient on weekends, usage patterns on working days are such that a combination of operator and user redistribution is required to maintain an adequate level of service. (Fishman, 2015)

2.4 Understanding Users Behaviour

It is obvious now that redistribution is at the core of a well-operated bike-sharing system. One way to deal with this topic is to deepen our knowledge of the user’s behaviour and classify it. Actually, the use of a clustering approach to identify patterns in the user’s behaviour is also a burgeoning field of research.
Bike-sharing systems can be grouped into behaviourally similar categories based upon their size, but according to Sarkar, Lathia, and Mascolo (2015) cluster analysis shows that larger systems display greater behavioural heterogeneity among their stations, and smaller systems generally have stations which all behave similarly in terms of their daily utilization patterns. Froehlich, Neumann, and Pontes (2009) were the first to apply clustering techniques and forecasting models to identify patterns of behaviour in stations. They used principally stations’ location and time of day when the trips occurred. Lathia, Ahmed, and Capra (2012) used a similar methodology to assess the effect of policy changes in London. Côme and Latifa (2014) introduced a model-based method to group stations with similar bike usage patterns - stations near restaurants and train stations - and predicted their bike usage pattern in different temporal settings.

These researches conducted on individual cities help to characterize a city’s bike share spatio-temporal patterns, mostly when experiments are repeated with different approaches for the same city. A recurring conclusion across analyses is that spatio-temporal system usage patterns are tied to, and reflect, city-specific characteristics. By focusing on single cities’ systems, Sarkar, Lathia, and Mascolo (2015) believe that these works indicate that each city has a unique pattern, and that forecasting algorithms applied to each one may not be generalized across the world. This is why they decided to conduct a multi-city study of stations clustering. Sarkar, Lathia, and Mascolo (2015) used a hierarchical clustering with an agglomerative strategy on various BSS (but only on the stations unlike this current project that clusters the trips). They discovered naturally-occurring behavioural classes for stations across all systems. They concluded that by studying the behaviour over many different systems, there were significant transferable knowledge between stations and between systems and that certain behavioural classes were system-independent.

The next step further is to predict the required redistribution to come based on the behavioural knowledge acquired. García-Palomares, Gutiérrez, and Latorre (2012) mentions that knowing the distribution of the potential demand and distinguishing areas that are trip generators from those that are trip attractors makes it possible to anticipate the asymmetric travel demands. At the user level, Bacciu et al. (2016) proposes to inform her/him on the status of the stations at run time. “Most of the available systems currently provide the information in terms of number of bicycles parked in each docking stations by means of services available via web. However, when the departure station is empty, a BSS user could be happy to know how the situation will evolve in the next few minutes and if a bike is going to arrive [and vice-versa.]” However, this knowledge given to the user might greatly affect the prediction per se. To mention only one example of demand prediction, Li et al. (2015b) propose a hierarchical prediction model to predict the number of bikes that will be rent from/returned in a future period for BSS, which focus more on the macroscopic bike traffic flow in the bicycle-sharing system and is different from the microscopic trip destination and duration prediction problem of a specific trip.

Nevertheless, this thesis will not elaborate more on this topic. Still, an important aspect stressed by Chen et al. (2016) is that the prediction results of static clusters may not yield consistent accuracy across different contexts, since the bike usage patterns of stations might be affected by various contextual factors such as weather condition and social events. Therefore, even though this study does not go as far as developing a prediction model, there is an attempt at forming dynamic classes.
Chapter 3

Presentation of the case study

3.1 BIXI Montreal

Since this project is conducted using the open data of the public bike-sharing system of BIXI in Montreal, it is essential to introduce this system, its origin, its operations, its context and its characteristics. The data per se will also be introduced and described.

3.1.1 The Story of BIXI

BIXI is a portmanteau combining the words “bicycle” and “taxi” to underline the concept that bicycles can be used just like taxis within the city. Launched in 2009, BIXI was the first large scale public bike-sharing system in North America. The project was included in the transportation plan for the City of Montreal, which aimed at encouraging active means of transportation such as bicycle. The initiator of the project was Stationnement de Montréal, the city’s parking authority, with the Public Bike System Company (PBSC). The later spread the original BIXI brand of systems internationally in 2010 by exporting its technology to London in the UK, Minneapolis, Washington D.C. in the USA, and Melbourne in Australia. Within the 3 following years, it also conquered Boston, Toronto, Chattanooga, New York City, Aspen, San Francisco, Chicago, Columbus and the Stony Brook University Campus. In total, it provided close to 47,000 bikes and 3,800 stations for 15 cities and two university campuses (Gerbet, 2016).

Despite its quick and global growth, PBSC suffered financially and was forced to bankrupt at the beginning of 2014. The bankruptcy was the result of a line of events that led to a critical financial situation. PBSC had some major liabilities, and it was owed money, too. A number of cities had delayed payments to PBSC over complaints about its bike sharing software programs and it had more than CAD 50 million in debt that had to be restructured (Maynard, 2014). Based on Riga (2012), five big debts led to the PBSC bankruptcy:

- New York and Chicago: They owed PBSC US 5.3 million, withheld by because of delays in implementing software across the Citibike and Divvy Bikes systems respectively.

- Damages: Alta Bike Share, which operates Citibike for New York, wanted US 11 million from PBSC because of those software delays.

- Loans: Bixi owed Montreal and its taxpayers a minimum of CAD 38 million. About CAD 31.6 million was owed on a CAD 37 million loan from the city. Another CAD 6.4 million was owed on a line of credit guaranteed by Montreal taxpayers.
• Suppliers: They were owned CAD 9 million by PBSC.

• Montreal: Beyond the loan and line of credit, it could cost CAD 1.5 million more for Montreal to run the BIXI system in 2014.

On top of that, the Montreal’s auditor general released a scathing report on PBSC in which he pointed to administrative problems, an illegal organizational structure, inadequate planning and an absence of oversight and accountability. Among other things, he noted a city is not permitted to be involved in a commercial activity. Consequently, the Quebec government ordered the City of Montreal to sell off PBSC’s international operations, which were supposed to bankroll the Montreal service BIXI. (Riga, 2012)

The bankruptcy of the biggest name in North American BSS arrived at the time when experts predicted a boom of the U.S. bike sharing fleet, supposed to double between 2012 and 2013 (Maynard, 2014). So, the bankruptcy sent some shudders across bike sharing cities. Luckily, this unfortunate event has served more as a lesson than an hindrance or discouragement to pursue the BSS expansion in North America (Yu, 2014).

After the bankruptcy, Bruno Rodi purchased the international division in April 2014, renamed it PBSC Urban Solution and left the Montreal’s on-the-ground BIXI operations to the City of Montreal. Then, he sold the majority share of PBSC Urban Solutions to Luc Sabbatini in January 2015 who became the CEO (Décarie, 2016). Since then, PBSC Urban Solutions has extended its activity in Toluca, Mexico, and Honolulu, Hawaii. The separation of the international division and the local operations has proved to be successful since both divisions have continue to grow and show positive results. (Gerbet, 2016; Magder, 2015)

On the basis of its commitment to promote the active means of transport in Montreal, and following its acquisition of PBSC after its bankruptcy, the City of Montreal created in 2014 BIXI Montreal, a nonprofit organization, in order to operate the BSS in Montreal for the 2014 season. This way, the Montrealers were able to benefit for one more year from this urban and alternative means of transport. The City of Montreal entrusted to BIXI Montreal the mission to manage properly the 6th season of BIXI and make it a success after the difficult previous years while it ensured the necessary material and financial resources. On one side, the property of the assets belongs to the City of Montreal and BIXI Montreal ensure the management of all operations in a totally autonomous, accountable and transparent way. (BIXI Montréal, 2015)

The BIXI system won several awards, including being named on the best innovations of the year (2008) by Time Magazine, an Edison Award (Gold) for best new product, and the 2009 Transportation Association Canada’s Sustainable Urban Transportation Award, among others (Blain, 2015).

BIXI Montreal counts about 70 employees managing a CAD 7.5 million budget of which CAD 4.3 million come from the City of Montreal which also provides goods, services and rights on the intellectual property of the equipment. It is the social economy company Cyclochrome which is responsible of the maintenance and the repairs of the bikes since 2009.

Being a nonprofit organization, BIXI Montreal is under a volunteer board of directors made of seven administrators and two observers from the City of Montreal. On the basis of its strong mid-season performance in 2014, the board of directors recommended to follow through with another season in 2015. (BIXI Montréal, 2015)
3.1.2 The Operating System of BIXI Montreal

The BIXI system was installed in Montreal during the summer 2009. It went from 20 stations to 215 in a month. The BIXI system met such a success that the second phase - which was planned for 2010 - was forestalled and new stations were installed in August 2009 (Leduc, 2013). As of June 2016, BIXI counted 452 permanent stations and 5,220 active bikes. However, one should note that, according to Fishman (2015), the numbers reported by the operators often differ from those observed and he states that BIXI had 3,594 observed bicycles which made it the 8th biggest BSS in the world in 2015. For the 2017 season, BIXI added almost 1,000 bikes and 100 stations to reach 6,200 bikes and 540 stations in Montreal (Lévesque, 2017).

The Stations

Each station is composed of a terminal or pay station with a touchscreen, the bikes and the bike docks which are joined into a modular platform as the figure 3.1 is illustrating. Each dock is equipped with a radio-frequency identification (RFID) scanner which transmits the information to the terminal. Due to the flexibility of its station and their ease of use, BIXI is said to be among the first fourth generation BSS. Indeed, the stations are taking their energy through solar panels which enhance the flexibility of the stations and no excavation or preparatory work is required for the instalment. They are simply added to on-street parking and removed for winter. According to Leduc (2013), a station needs up to 40 minutes to be installed.

The location of a station is determined by several parameters, including population density, points of interest and activities (universities, bike paths, other transportation networks, and data on travel patterns of the general public gathered from origin-destination studies). Moreover, the flexibility of the BIXI stations allow quick on-the-fly adaptations to punctual events such as festivals or transportation problems, such as metro delays and traffic accidents.

![Diagram of a BIXI station](image)

FIGURE 3.1: The elements of a BIXI station. (Freemark, 2009)

The Bikes
The BIXI bikes are utility bicycles with a unisex step-through frame which makes them comfy, secure, manageable and mostly heavy-duty to survive vandalism and prevent theft. The bikes weight approximately 18 kg. Their tires are designed to be puncture-resistant and are filled with nitrogen gas to maintain proper pressure for longer. They have three gear speeds through an efficient and hidden gearing system. They also have a little basket at the front for convenience, strong self-sufficient flashing LED lights (light-emitting diodes) at the front and the back for security and night time or fuzzy weather conditions. They communicate with the dock through an RFID chip as well (Leduc, 2013). All the bikes are designed and built in the Province of Quebec by Cyclochrome using local materials (ITDP, 2013; BIXI Montreal, 2015). Every docking spot has a button to notify if a bike needs maintenance. For instance, if a flat tire occurs, the user checks-in the bike, presses the button and the bike is then locked until a maintenance truck comes pick it up shortly after.

The Membership

There are two kinds of users: members and casual users. Therefore, the pricing scheme in date of March 2017 is elaborated. There are four options for members and three for casual users. The options mainly differ in length and price. The seven options are detailed in table 3.1 below. All memberships grant a BIXI RFID key which is link with a user ID. The casual members don’t have this key, so they must use their credit cards to operate the system. Therefore, the members produce richer data by enhancing all their trip information with their personal data such as name, gender, age, address and language. However, all this extra information is private, hence not use in this project. The BIXI key allows the user to quickly rent a bike by directly checking it out where the bike is by using its respective RFID slot at the dock instead of going to the pay station. A casual user can rent a maximum of four bike per credit card. (BIXI Montreal, 2017)

The one year, 90 days and 30 days contracts give a free 45 minutes every time bike is borrowed. All other options allow only a free 30 minutes. When these time are passed, additional fees are charged. They are CAD 1.75 for the first extra 15 minutes (a trip between 45 and 60 minutes or 30 and 45 minutes), then CAD 3 for every additional 15 minutes (see figure 3.2. (BIXI Montreal, 2017) New to the 2017 season, it will be possible to buy a book of 10 single-trips of which the price will be known later on during the season. (Lévesque, 2017)

A membership or a casual rental is not transferable between individuals, i.e. only one person is allowed to use a BIXI key or a rental from a credit card. To balance this aspect, BIXI offers multi-user keys and a 20% reduction for the purchase of at least 20 one-year memberships. These offers are aimed at enterprises which want to enhance their reputations and provide sustainable transport to its employees. (BIXI Montreal, 2017)

Finally, members can combine many different offers between several transportation services of Montreal such as STM, AMT, Communauto and Vélo Québec and
benefit from discounts. In 2017, Car2Go and Téo Taxi will also join the list of collaborators (Lévesque, 2017).

### 3.1.3 Statistics of Usage

The BIXI usage is also following the general trends detailed in section 2.2. Indeed, many studies about BIXI contributed to the literature as early as 2009 during the implementation of the service. Some have highlighted interesting features like Fuller et al. (2013) who conducted a cross-sectional telephone survey with some 2,500 individuals before and after the implementation of BIXI in Montreal, to determine the potential mode shift and health benefit of the programme and they concluded that BIXI was associated with a shift towards active transport. In a previous study, Fuller et al. (2011) demonstrated that newly implemented BSS attracted a substantial fraction of the population and was more likely to attract younger and more educated people who previously used cycling as a primary transportation mode.

Imani et al. (2014) conducted a thorough study on the BIXI BSS that highlighted various interesting features from the system. They confirmed that users are more likely to cycle under good weather conditions as well as that the usage reduces during the weekends. Friday and Saturday nights are positively related with bike flow. They found out that the bike flow decreases with the distance from the city centre. The proximity of public infrastructure like university, shops, bars and restaurants affect significantly the usage of the nearby stations. They conclude that their most interesting discovery is about the nitricate relationship between the number of stations and their capacity, and their impacts on usage. They found out that “that adding additional stations (either by relocating existing capacity from large stations or adding new bicycle slots) is more beneficial in terms of arrival and departure flows compared to adding capacity to existing stations.”

Bernatchez et al. (2015) added interesting information about novel information about the processes leading to inequalities in awareness of facilities available in the built environment. To do so, they studied the association between the proximity of BIXI docking stations to home, educational achievement, and awareness of the BSS across time. They highlight in their paper that the lack of awareness is associated with three factors: time, availability of docking stations in the neighbourhood, and education. Also, they show that lower education is associated with lower awareness of BSS and that a lack of awareness decreases more slowly across time despite the availability of BIXI stations in the neighbourhood.

---

**Table 3.1: Pricing scheme of BIXI for the season 2017. All prices are in CAD. (BIXI Montreal, 2017)**

<table>
<thead>
<tr>
<th>Type of user</th>
<th>Member</th>
<th>Casual</th>
</tr>
</thead>
<tbody>
<tr>
<td>Term duration</td>
<td>1 year 90 days 30 days 24 hours</td>
<td>3 days 24 hours 1 single trip</td>
</tr>
<tr>
<td>Registration cost</td>
<td>$89 $55 $30 $5</td>
<td>$14 $5 $2.95</td>
</tr>
<tr>
<td>Allocated trip period</td>
<td>45 min 45 min 45 min 30 min</td>
<td>30 min 30 min 30 min</td>
</tr>
</tbody>
</table>

**Table 3.2: Extra fees for time extension. (BIXI Montreal, 2017)**

<table>
<thead>
<tr>
<th>Extra time</th>
<th>Rental cost</th>
</tr>
</thead>
<tbody>
<tr>
<td>1st 15 minutes</td>
<td>+$1.75</td>
</tr>
<tr>
<td>Subsequent 15 minutes</td>
<td>+$3/(15 minutes)</td>
</tr>
</tbody>
</table>
On her part, Leduc (2013) realized her final master thesis about optimizing the operations of the BIXI system and its redistribution scheme. She developed an operational tool to be used in a real context in order to optimize the static and dynamic redistribution activities. She mentioned that no previous authors offered a concrete and practical solution despite the frequent mention and discussion of the problem in the literature. She also presents all the following information from an unavailable source (Morency, Trepanier, and Godefroy, 2011).

Morency, Trepanier, and Godefroy (2011) (Leduc, 2013) list various statistics based on the months of July, August and September of the first season of BIXI in 2009. These statistics have clearly evolved since then, but they can still bring interesting insights about the understanding of the context in which the BIXI system evolves and the factors influencing the redistribution. Members were representing 50.3% and 72.8% of the daily users at that time with lower proportions during the weekends. Men are 1.73 more likely to use the service than women. Users are 34 years old on average and the authors mention that there is an over representation of people within the 25-34 years old branch within the members in comparison with the Montreal population.

Morency, Trepanier, and Godefroy (2011) (Leduc, 2013) developed three indicators characterizing various parameters of the stations. The first indicator is about the balancing factor representing the ratio between the number of departures and arrivals. A station with a high balancing factor is a popular starting station (or a repulsor or a gregarious station further in the text) and a station with a low balancing factor is a popular ending station (or an attractor or a popular station further in the text). The figure 3.3 allows to visualize the results from their analysis over a map of Montreal.

The second indicator developed by Morency, Trepanier, and Godefroy (2011) (Leduc, 2013) concerns the service level and is split into two ratios (see figure 3.4).
Chapter 3. Presentation of the case study

Figure 3.4: Spatial dispersion of the stations according to their service level. (Leduc, 2013; Morency, Trepanier, and Godefroy, 2011)

Figure 3.5: Spatial dispersion of the stations according to their redistribution factor. (Leduc, 2013; Morency, Trepanier, and Godefroy, 2011)

The left one indicates the time proportion when a station is full, thus users cannot check-in more bikes at this station. The right one indicates the time proportion when a station is empty, thus users cannot check out bikes at this station.

The third indicator developed by Morency, Trepanier, and Godefroy (2011) (Leduc, 2013) is about the redistribution and is also split into two ratios (see figure 3.5). The left frame represents the ratio of bikes checked out at a station also including both operators and users. The right frame represents the ratio of bikes checked-in at a station including those from the operators and from the users.

In addition, Shaheen et al. (2012) reports more statistical information like that the majority of BIXI users (56%) use the system to transit to work or school, that 81% of the users are totally in favour with the fact that BIXI has improved public transportation in Montreal, and that the majority of BIXI users never wear helmets when they use the service.

Finally, since its bankruptcy, BIXI has always known a growth in the amount of users (members and casuals) every year as shows figure 3.6.
Chapter 3. Presentation of the case study

Figure 3.6: Amount of subscribers through memberships or casually per year.

A personal communication with Blain, 2017 revealed that the current system of BIXI uses geographical clusters of the stations that they call zones. They benefit from their public nonprofit organisation attached to the City of Montreal by having access to plenty of rich data sources like the sociodemographic data about the users. Therefore, they are in position to develop powerful tool to enhance their service. Since the usage patterns are quite known and stable, a dynamic redistribution is good enough to provide a good balance among the stations, hence a good service. The punctual events that might disturb the patterns are managed thanks to the modularity of the stations which offers a great flexibility and can quickly adjust to the demand. For example, Riga (2012) reported that there would be two “bike depots” downtown in 2012. These depots are locations where a user can return a bike to a BIXI employee without the need to find an empty slot at a bike station. This kind of initiative helps solving rebalancing problems. Since then, bike depots are a common strategy for BIXI.

In consequence, since the redistribution is under control, this allows BIXI to put a greater emphasis on the member profiles and try to predict the best network topology to attract more members while keeping the actual ones. According to the users and the knowledge of the general population, an elaborated strategy needs to be put in place to define where to put the stations as well as their capacity in order to provide the best service possible while still growing like they did in 2017.

3.2 Information Sources

In this section, the data used for this project are presented. Four different databases were used: one gathering the bike trips between stations, one identifying the geolocation of the stations in Montreal, one to add climate information for each entry and a fourth one to elaborate sociodemographic, geological and urban analysis and combine them into the stations characterization database. First, the sources where these data were extracted will be presented. Then, a portray of the relevant variables from each database kept for the main data frame is given. Finally, the manipulations and the preprocessing done over the data are described. An illustration of the combination of each database for the final working database is shown at figure 3.7.

3.2.1 The Database of BIXI Bike Trips

The bike trips information was extracted from the open data web page of BIXI Montréal (2017) on March 16th 2017. This information from the 2015 season and on includes all the bike trips taken by users since then. BIXI offers open information on
As mentioned earlier, CSV files of all the bike trips taken during the 2015 season were downloaded from the BIXI open data web page. Globally, it includes 3,503,355 trips between 460 stations (in 2015) spread during the 7 months of service of BIXI, from mid-April to mid-November. Basically, this database gathers all the movements from the users, either they are registered members or simply casual users. It gives the exact moment when a bike is checked out and check in, and at which stations it is done so. Moreover, the type of user is specified. The exhaustive list of all the variables included in this database is available below in table 3.3.

<table>
<thead>
<tr>
<th>Variables</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Start.date</td>
<td>Date and time of the beginning of the transit</td>
</tr>
<tr>
<td>Start.station.number</td>
<td>Starting station's terminal identifier</td>
</tr>
<tr>
<td>Start.station</td>
<td>Starting station’s name</td>
</tr>
<tr>
<td>End.date</td>
<td>Date and time of the arrival</td>
</tr>
<tr>
<td>End.station.number</td>
<td>Ending station’s terminal identifier</td>
</tr>
<tr>
<td>End.station</td>
<td>Ending station’s name</td>
</tr>
<tr>
<td>Account.type</td>
<td>Type of user</td>
</tr>
<tr>
<td>Total.duration</td>
<td>Transit’s total duration</td>
</tr>
</tbody>
</table>
3.2.2 The Stations Database

A database was constructed to gather all the information collected about the stations. This database was first built around the geolocation information coming from the minute-by-minute BIXI bicycle availability data for all stations in service. These data are updated every 5 minutes on the web page of BIXI Montreal (2017) and describe the current state of every station as oppose to the archives (see section 3.2.1) that detail the movement of the users. Hence, both databases are complementary. The other variable collected from the minute-by-minute data is the total docking capacity of each station. If the geolocation does not change from season to season, it is not the case with the capacity that can be adjusted according to the needs since the stations are modular. Therefore, let’s keep in mind that the capacities are those for the 2017 season and might differ from the actual ones in 2015.

Most of the studies realized using BIXI’s data have been done using the live feed data provided by the bike-sharing operator, such as Leduc (2013) and Imani et al. (2014). However, this method requires a previous recording of the data in order to use it because the minute-by-minute availability data are not stored anywhere. Thus, it extend the length of the project studies greatly as it is impossible to record data between mid-November to mid-April since the service is out for the winter season. Therefore, this database was mostly used to obtain the geolocation of each station to be able to geographically locate and visualize the bike trips as well as apply a visual dynamic analysis. Hence, only the latitudes and longitudes of each station were extracted. Knowing the structure of this database is essential in order to implement and operate the prediction model dynamically. Two formats are available to withdraw the data: JSON and XML. They are very similar in the sense that both JSON and XML can be used to receive data from a web server, they are “self describing” (human readable), they are hierarchical (values within values), they can be parsed and used by lots of programming languages and they can be fetched with an XMLHttpRequest. However, JSON is more practical because it doesn’t use end tag, it is shorter, it is quicker to read and write, it can use arrays and mostly because it is easier to parse using a standard JavaScript function. Nonetheless, the format used in the present case was the XML for practical reasons and simplicity with R, the software used. The table 3.4 presents thoroughly the variables and their descriptions for both formats of the database.

3.2.3 The Urban Information

To enhance the characterization of the stations, urban, demographic and geological information were collected, manipulated and added to the stations’ database. These data were downloaded from open governmental platforms such as the Government of Canada (Natural Resources Canada, 1983; Statistics Canada, 2016) and the City of Montreal (2017). These data include information about the environment (topographic elevation, green areas, water streams, parks, etc), the public transportation system of Montreal (metro and buses), infrastructure (cycle paths, streets, one-ways, etc), the urban use of the space (skyscrapers, schools, universities, cultural heritage, commercial and residential zone, religious buildings, vacant lots, parking lots, etc), sociodemographic information (population density), administrative frontiers, and toponymy. Most of this information was use to enhance the rendering of the maps presented later on. No table resume this information as they are mostly shapefiles treated with a geographic information system software.
### Table 3.4: Structure of the network status database

#### JSON

<table>
<thead>
<tr>
<th>Variables</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>id</td>
<td>Station’s unique identifier</td>
</tr>
<tr>
<td>s</td>
<td>Station’s name</td>
</tr>
<tr>
<td>n</td>
<td>Station’s terminal identifier</td>
</tr>
<tr>
<td>st</td>
<td>Station’s state</td>
</tr>
<tr>
<td>b</td>
<td>Boolean value specifying if the station is blocked</td>
</tr>
<tr>
<td>su</td>
<td>Boolean value specifying if the station is suspended</td>
</tr>
<tr>
<td>m</td>
<td>Boolean value specifying if the station is out of order</td>
</tr>
<tr>
<td>lu</td>
<td>Timestamp of the last update of the station’s data</td>
</tr>
<tr>
<td>lc</td>
<td>Timestamp of last communication between the station’s terminal and the central server</td>
</tr>
<tr>
<td>bk</td>
<td>(for future use)</td>
</tr>
<tr>
<td>bl</td>
<td>(for future use)</td>
</tr>
<tr>
<td>la</td>
<td>Station’s latitude according to the CSR WGS84</td>
</tr>
<tr>
<td>lo</td>
<td>Station’s longitude according to the CSR WGS84</td>
</tr>
<tr>
<td>da</td>
<td>Number of empty docks available at this station</td>
</tr>
<tr>
<td>dx</td>
<td>Number of empty docks unavailable at this station</td>
</tr>
<tr>
<td>ba</td>
<td>Number of bikes available at this station</td>
</tr>
<tr>
<td>bx</td>
<td>Number of bikes unavailable at this station</td>
</tr>
</tbody>
</table>

#### XML

<table>
<thead>
<tr>
<th>Variables</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>id</td>
<td>Station’s unique identifier</td>
</tr>
<tr>
<td>name</td>
<td>Station’s name</td>
</tr>
<tr>
<td>terminalName</td>
<td>Station’s terminal identifier</td>
</tr>
<tr>
<td>lastCommWithServer</td>
<td>Timestamp of last communication between the station’s terminal and the central server</td>
</tr>
<tr>
<td>b</td>
<td>Station’s latitude according to the CSR WGS84</td>
</tr>
<tr>
<td>su</td>
<td>Station’s longitude according to the CSR WGS84</td>
</tr>
<tr>
<td>m</td>
<td>Boolean value specifying if the station is installed</td>
</tr>
<tr>
<td>lu</td>
<td>Boolean value specifying if the station is temporarily deactivated</td>
</tr>
<tr>
<td>lc</td>
<td>Installation date of the station</td>
</tr>
<tr>
<td>bk</td>
<td>Removal date of the station</td>
</tr>
<tr>
<td>bl</td>
<td>Temporary Boolean value specifying if the station is temporary</td>
</tr>
<tr>
<td>la</td>
<td>Public Boolean value specifying if the record is public</td>
</tr>
<tr>
<td>lo</td>
<td>NbBikes Number of bikes available at this station</td>
</tr>
<tr>
<td>da</td>
<td>NbEmptyDocks Number of empty docks available at this station</td>
</tr>
<tr>
<td>dx</td>
<td>LastUpdateTime Timestamp of the last update of the station’s data</td>
</tr>
<tr>
<td>ba</td>
<td>Number of bikes available at this station</td>
</tr>
<tr>
<td>bx</td>
<td>Number of bikes unavailable at this station</td>
</tr>
</tbody>
</table>
Chapter 3. Presentation of the case study

TABLE 3.5: Structure of the weather data base

<table>
<thead>
<tr>
<th>Variables</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Date.Heure</td>
<td>Date and time of the measurement</td>
</tr>
<tr>
<td>Année</td>
<td>Year of the measurement</td>
</tr>
<tr>
<td>Mois</td>
<td>Month of the measurement</td>
</tr>
<tr>
<td>Jour</td>
<td>Day of the measurement</td>
</tr>
<tr>
<td>Heure</td>
<td>Hour of the measurement</td>
</tr>
<tr>
<td>Temp (°C)</td>
<td>Air temperature</td>
</tr>
<tr>
<td>Point de rosée (°C)</td>
<td>Dew point temperature</td>
</tr>
<tr>
<td>Hum. rel (%)</td>
<td>Air relative humidity</td>
</tr>
<tr>
<td>Dir. du vent (10s deg)</td>
<td>Wind direction</td>
</tr>
<tr>
<td>Vit. du vent (km/h)</td>
<td>Wind speed</td>
</tr>
<tr>
<td>Visibilité (km)</td>
<td>Visibility</td>
</tr>
<tr>
<td>Pression à la station (kPa)</td>
<td>Atmospheric pressure</td>
</tr>
<tr>
<td>Hmdx</td>
<td>Humidity index</td>
</tr>
<tr>
<td>Refroid. éolien</td>
<td>Wind chill factor</td>
</tr>
<tr>
<td>Temps</td>
<td>Weather description</td>
</tr>
</tbody>
</table>

3.2.4 The Weather Database

Since cycling is an activity highly related with the weather conditions, meteorological information is considered to enrich the BIXI database to better understand the user behaviour. The weather data were extracted from the historical climate data of the Government of Canada on February 2\textsuperscript{nd} 2017 (\textit{Historical Climate Data - Climate - Environment Canada}). The data were collected by the Pierre-Elliot Trudeau Montreal International Airport (YUL) weather station and they offer hourly information of all the weather variables throughout the whole year.

From the whole year 2015 archive, only the data from April to November 2015 inclusively were downloaded to match the BIXI season analyzed in this project. The resulting weather database is composed of 5,856 observations. The exhaustive detailing of the variables this database provides is put in table 3.5. The relevant variables that were used are: the moment (Date and time) of the measure, the temperature, the relative humidity, the wind direction and speed, the atmospheric pressure and the main qualifier of the weather.

3.2.5 Data Selection and Matrix Structure

Finally, the main working data frame to be used further on in the project is a combination of the variables of the previously described databases (see table 3.6). This data frame is composed of 39 variables and 3,503,355 entries. The metadata is available in appendix A.
<table>
<thead>
<tr>
<th>Variable</th>
<th>Meaning</th>
</tr>
</thead>
<tbody>
<tr>
<td>Start.date</td>
<td>Date and time of the beginning of the transit</td>
</tr>
<tr>
<td>End.date</td>
<td>Date and time of the arrival</td>
</tr>
<tr>
<td>Total.duration</td>
<td>Transit’s total duration</td>
</tr>
<tr>
<td>Year</td>
<td>Year of the trip</td>
</tr>
<tr>
<td>Month</td>
<td>Month of departure of the trip</td>
</tr>
<tr>
<td>Day</td>
<td>Day of departure of the trip</td>
</tr>
<tr>
<td>Start.weekday</td>
<td>Day of the week where the transit begins</td>
</tr>
<tr>
<td>Weekend</td>
<td>True if the trip begins during the weekend (Saturday or Sunday)</td>
</tr>
<tr>
<td>Nightout</td>
<td>True if the trip begins Friday or Saturday night between 20:00 and 5:00</td>
</tr>
<tr>
<td>Period</td>
<td>Period of the day when the trip begins</td>
</tr>
<tr>
<td>Start.time</td>
<td>Time when the trip starts</td>
</tr>
<tr>
<td>Hour</td>
<td>Time when the trip starts rounded down to the hour</td>
</tr>
<tr>
<td>TenS</td>
<td>Time when the trip starts rounded down to the ten of minutes</td>
</tr>
<tr>
<td>Start.station.number</td>
<td>Starting station’s terminal identifier</td>
</tr>
<tr>
<td>End.station.number</td>
<td>Ending station’s terminal identifier</td>
</tr>
<tr>
<td>Capacity.S</td>
<td>Bike capacity of the origin station</td>
</tr>
<tr>
<td>Capacity.E</td>
<td>Bike capacity of the destination station</td>
</tr>
<tr>
<td>Metro.station.S</td>
<td>True if the station of origin lies within 250m of radius from a Metro station</td>
</tr>
<tr>
<td>Metro.station.E</td>
<td>True if the station of arrival lies within 250m of radius from a Metro station</td>
</tr>
<tr>
<td>Metro.sORe</td>
<td>True if the station of origin OR arrival lies within 250m of radius from a Metro station</td>
</tr>
<tr>
<td>Metro.sANDe</td>
<td>True if the station of origin AND arrival lies within 250m of radius from a Metro station</td>
</tr>
<tr>
<td>Downtown.S</td>
<td>True if the station of origin is within the determined downtown area</td>
</tr>
<tr>
<td>Downtown.E</td>
<td>True if the station of arrival is within the determined downtown area</td>
</tr>
<tr>
<td>Altitude.S</td>
<td>Altitude of the station of origin</td>
</tr>
<tr>
<td>Altitude.E</td>
<td>Altitude of the station of arrival</td>
</tr>
<tr>
<td>Elevation</td>
<td>Difference in altitude between the stations of arrival and origin</td>
</tr>
<tr>
<td>Loop</td>
<td>Is true if the trip ends at the same station where it started</td>
</tr>
<tr>
<td>Start.lat</td>
<td>Starting station’s latitude in the CSR WGS84</td>
</tr>
<tr>
<td>Start.long</td>
<td>Starting station’s longitude in the CSR WGS84</td>
</tr>
<tr>
<td>End.lat</td>
<td>Ending station’s latitude in the CSR WGS84</td>
</tr>
<tr>
<td>End.long</td>
<td>Ending station’s longitude in the CSR WGS84</td>
</tr>
<tr>
<td>Account.type</td>
<td>Type of user</td>
</tr>
<tr>
<td>Temperature</td>
<td>Air temperature</td>
</tr>
<tr>
<td>Rel. Humidity</td>
<td>Air relative humidity</td>
</tr>
<tr>
<td>Wind speed</td>
<td>Wind speed</td>
</tr>
<tr>
<td>Wind.cardinal</td>
<td>Cardinal wind direction</td>
</tr>
<tr>
<td>Atm. Pressure</td>
<td>Atmospheric pressure</td>
</tr>
<tr>
<td>Main weather</td>
<td>Simplified weather description</td>
</tr>
<tr>
<td>Proper.conditions</td>
<td>True if the weather conditions are proper to cycle (Clear, Cloudy or Fog)</td>
</tr>
</tbody>
</table>
Chapter 4

Methodology

This chapter presents the methodology, the theoretical concepts and the tools used to exploit data for extracting relevant decisional knowledge useful for a better management of the BIXI system. First of all, the general steps are described. Then, the preprocessing of the data, a basic data analysis, the complex network approach and the whole clustering process are described in details. Finally, a brief overview of the software used and their input in the development of the model is done.

4.1 General Methodology

In this section, the general methodology used to develop the model is detailed. The following figure 4.1 shows the different steps describing the present methodology.

The research began with a literature review that has been covered in the previous chapter 2. This review allowed to highlight the different approaches that have been used to develop prediction model for public bike-sharing system. As well, the literature review guided towards the modelisations and clustering techniques to be use in this project.
The second step of the methodology is the data examination as it was introduced in section 3.2. This step is based upon the open data offered by the BSS BIXI in Montreal. All the trips registered during the 2015 season have been downloaded and used. This database gave the spatio-temporal information about the origin and destination of a trip as well as the user type (member versus casual). To these data were added the additional information about the stations such as its specific geolocation and its capacity (based on the 2017 season) coming from the live open data offered by BIXI as well. This live feed of data gives information about the state of each stations updated every 5 minutes. To these values were added the weather information at the moment of the trip. These data were downloaded from the Environment Canada archive web page. Finally, some sociogeographic data about Montreal were acquired from the open data web page of the City of Montreal (2017), Statistics Canada (2016) and Natural Resources Canada (1983). The four last databases added were mostly used to better understand the context of each trip taken by a user.

All these data were manipulated and thoroughly analyzed to better understand the complexity emerging from the BSS and identify its weaknesses and opportunities. The analysis were performed using the software R (R Core Team, 2017) and QGIS (QGIS Development Team, 2016) described bellow at the section 4.9. The results from the analysis are presented and described in the chapter 5.

The third step of the methodology is the choice of the modelisation technique. As previously elaborated in section 2.4, many different methods have been used to learn about BSS users’ behaviour. In the case of this thesis, a hierarchical clustering using Ward and Gower have been performed using R to characterize different types of trips. Then, an integrated multi-view clustering technique was used to obtained the selected set of clusters. Local analysis were performed on some of the classes to express the utility of the developed model.

The last step consists in using the interpretation of the patterns emerging from the clustered trips to elaborate guidelines for decision makers in order to enhance the service of BIXI. The following steps are those leading to the development of the decisional knowledge.

Will follow a detailed description of each step and manipulation from the data preprocessing and on, including the clustering, the interpretations of the classes’ profiles, the multi-view analysis and the local analysis of two cases. The results are presented in the following chapter 5.

The very first steps of the project have been to overview the academic literature on the topic while collecting the data for the project as it has been discussed in chapters 2 and 3. So, this section begins with the preprocessing of the data.

### 4.2 Preprocessing

All the databases used in this project have been introduced previously in their raw form. All of them had to undergo modifications in accordance with the needs of this project. Some general modifications were applied to all databases. All the text values and the variables’ names were translated from French to English and the final adjustment of the variables’ formats in R was done such that the R-classes of all variables would be adequate to be properly processed and manipulated afterwards. The types of the variables are detailed in the metadata table A.3 in appendix A. (Gibert, Sánchez-Marrè, and Izquierdo, 2016)
4.2.1 The Database of BIXI Bike Trips

Being the spinal core of the main working dataset, the bike trips data did not have to go through any major changes or manipulations. The only minor one was to convert the trip duration from milliseconds to seconds - a more understandable unit - or to a time format of the form Days:Hours:Minutes:Seconds using the R package `lubridate` (Grolemund and Wickham, 2011). However, the time variables were duplicated into a new dataset called "Datetime" because it led to many complications with the software R. At first, the date and time variable (Datetime) was divided into various variables more specific such as Year, Month, Day, Weekday, Time, Hour and TenS which is the time segregated into intervals of 10 minutes like "8:10", "8:20", etc. Then, the complications were about their formats in R which is limited when dealing with time variables. Therefore, some of them are strings and others are factors. R cannot deal properly with a variable that contains both the date and the time. For example, it is not possible to study all the trips taken between 8:00 and 9:00 from the variable Start.date which is in the POSIXlt format, the best format to treat date and time in the same variable. This is why this variable was subdivided into 9 temporal variables: Year, Month, Day, Weekday, Hour, TenS (tens of minutes), Weekend, Nightout and Period. However, all these new variables cannot stay in the POSIXlt format. Year, Month, Day, Weekday and Period are thus converted in factors, Weekend and Nightout are Boolean, and Hour and TenS have to be character.

The variable Period is a variable dividing the day into 5 periods: morning, noon, afternoon, evening and night. The hours included into each period are as follow in table 4.1.

<table>
<thead>
<tr>
<th>Period</th>
<th>from</th>
<th>until</th>
</tr>
</thead>
<tbody>
<tr>
<td>Morning</td>
<td>6:00</td>
<td>11:00</td>
</tr>
<tr>
<td>Noon</td>
<td>11:00</td>
<td>14:00</td>
</tr>
<tr>
<td>Afternoon</td>
<td>14:00</td>
<td>18:00</td>
</tr>
<tr>
<td>Evening</td>
<td>18:00</td>
<td>22:00</td>
</tr>
<tr>
<td>Night</td>
<td>22:00</td>
<td>6:00</td>
</tr>
</tbody>
</table>

TABLE 4.1: Details on the variable Period.

After all the manipulations and fine tuning was completed in the "Datetime" dataset, all the variables were added back to the main working dataset. At first, there were efforts to keep the working data frame as small as possible for random access memory reasons, but in the end, all time variables happened to be necessary.

For some manipulations, the Start.date, Hour and TenS variables had to be converted into numbers. The Start.date was then in seconds since January 1st 1970, and the other two simply had the ":" removed so 8:00 would appear as 800.

4.2.2 The Stations Database

From the live BIXI data was created the "idlocation" database used to characterize the stations. This database contains all the information about the stations individually such as its identification number, its name (made from the streets intersecting), its longitude and latitude, and its docking capacity. During the course of the project, BIXI applied some changes on its minute-by-minute database in early March 2017 for the upcoming season. It relocated five stations and added 20 more. This caused some association problems when merging the geolocations to the stations.
in the bike trips database. Five stations used in 2015 were not in the live database anymore. Consequently, a new database gathering all the previous and new stations with their respective identification numbers, names, latitudes and longitudes was created. This database contains 544 stations even though only 460 were in activity in 2015. This is because it also includes the new stations added after the 2015 season and because many stations are temporary (only for events, to compensate disturbance by construction, etc). Also, a new variable named Capacity was created summing the variables nbBikes (number of bikes present at the station) and nbEmptyDocks (number of empty docks available at the station) together. Therefore, this variable also changed considerably as BIXI was preparing its 2017 season in early April. Consequently, an ultimate update for the capacities has been made on April 25th 2017. Note that the capacity values are those for the 2017 season and not the 2015 one, but it can still give an idea of the distribution and importance of each station.

4.2.3 The Geographic Information and Maps

To complete the characterization database of the stations, the geographic information from the City of Montreal (City of Montreal, 2017) and the Government of Canada (Natural Resources Canada, 1983; Statistics Canada, 2016) was used to create some additional variables. First, a Boolean variable was added according to the proximity of a BIXI station to a Metro station. If a BIXI station falls within 250 meters of radius from a Metro station, it will be TRUE, and FALSE otherwise. The 250 meters buffer regions around the metro stations was automatically computed by QGIS. Based on this variable, 2 other Boolean variables were calculated. The first being TRUE if at least one of the origin or destination station is TRUE and the second being TRUE if both of the origin and destination stations are TRUE. Then, another Boolean variable was added according to if a BIXI station is within the downtown area. This area was delimited according to the commercial zone determined by the urban use and limited to the lower part, south of Sherbrooke street in the neighbourhood Ville-Marie (see maps in section 5.1.2 for reference). Finally, the altitude of the stations was added using the topographical map of the greater Montreal area. This feature is quite relevant due to the presence of a mountain in the middle of the city, thus creating a tendency by the users to go down more often than up with the bikes.

4.2.4 The Weather Database

The weather variables were also treated in a separated dataset before being merged with the main working dataset. Several manipulations had to be done with this database. In its original format, the wind direction is presented in tens of degrees with values ranging from 0 to 35 (0° being North). To simplify this variable and give it an understandable meaning, they were converted into the classic cardinal values such as North (N), Northeast (NE), East (E), Southeast (SE), South (S), Southwest (SW), West (W) and Northwest (NW). As well, a new cardinal value named “No wind” had to be created to fill in the missing values related to when there is no wind, of which there were only 5 occurrences in 2015.

The variable Main.Weather had some undefined values (“ND”), but following a pattern. When the weather was either clear, cloudy, generally clear or generally cloudy, the status of the weather was only specified once every three hours. We are in front of systematic missing values. These do not require such intense monitoring because the intrinsic dynamic of meteorological conditions does not evolve fast
enough so that its missing values are real lack of information. In this scenario, the implicit imputation is: the conditions change slowly so there is no need of measuring every 5 minutes, but every 3 hours is enough. The conditions in between the measurements are reasonably assumable as constant. According to that, the "ND" were all imputed with their respective previously specified value.

Then, a simplification of the modalities was needed since there were too many (36 modalities) and some were irrelevant. For instance, it was difficult to deal efficiently with the multiple combined values such as "Snow,Rain,Fog". Thus, the simplification consisted in reducing the multiple factors into the most influential one of the combination. For example, "Snow,Rain,Fog" was converted to only "Rain" since it is the most influential factor of the three affecting the use of BIXI (Imani et al., 2014; Flynn et al., 2012; Corcoran et al., 2014). The conversions made are gathered into table 4.2.

Finally, another greater simplified variable was built to facilitate the consideration of the weather conditions during the clustering. To do so, a Boolean variable labelled Proper.Cond indicates if the weather conditions are proper or not to cycle. Based on the Main.Weather variable, the the Proper.Cond variable will be TRUE if the weather is either "Clear", "Cloudy" or "Fog", and FALSE in all the other cases.

Five entries of the Weather database had missing values for the measurements, but not for the time. So, a linear interpolation was applied to fill in the holes. Whereas in the case of the Main.Weather variable, it was subjectively decided to look at the 3 previous and 3 following hours information and the other corresponding variables information such as temperature, wind and atmospheric pressure to approximate the weather description.

4.3 Basic Descriptive and Visualization

4.3.1 Descriptive Statistics

After having built the working data frame by combining the different original databases, a statistical description of the variables was conducted. Descriptive statistics is used to describe and summarize the variables of a dataset. They are not meant to extract knowledge from the target population but to provide an overview of how the data is and orient preprocessing and analysis. Descriptive statistics are a collection of measurements and graphical representations showing some basic characteristics of the variables’ underlying distribution. For numerical variables: central trend, variability, symmetry, existence of outliers, bi-modalities, etc. The distributions are visualized through histograms and box-plot. Central trend tells the central value of the variable (the mean or the median are used depending on the symmetry of the distribution. Variability refers to the spread of the data around the central trend (i.e. variance, standard deviation, interquartile range). The dataset being composed of mixed variables, the numerical variables were summarized into graphs and tables of basic numerical statistics whereas logical and qualitative variables were summarized into bar-plot including the conditional distribution table of the variables.

As a basic comparison, the summary of the numerical original weather variables were joined in the same table with the weather variables from the trips. This way, the reader can observe the differences between the statistical values and understand better the underlying information about the users’ behaviour.
Table 4.2: Modifications of the Weather modalities into fewer more generic ones

<table>
<thead>
<tr>
<th>Simplified variables</th>
<th>Original Variables</th>
</tr>
</thead>
<tbody>
<tr>
<td>NA</td>
<td>NA, ND</td>
</tr>
<tr>
<td>Snow flurries</td>
<td>Averses de neige</td>
</tr>
<tr>
<td>Showers</td>
<td>Averses de pluie, Averses de neige, Averses de pluie, Brouillard, Averses de pluie forte, Averses de pluie modérées, Brouillard</td>
</tr>
<tr>
<td>Fog</td>
<td>Brouillard</td>
</tr>
<tr>
<td>Freezing fog</td>
<td>Brouillard verglaçant</td>
</tr>
<tr>
<td>Mist</td>
<td>Bruine, Brouillard</td>
</tr>
<tr>
<td>Clear</td>
<td>Dégagé, Généralement dégagé</td>
</tr>
<tr>
<td>Cloudy</td>
<td>Généralement nuageux, Nuageux</td>
</tr>
<tr>
<td>Snow</td>
<td>Neige, Brouillard</td>
</tr>
<tr>
<td>Blowing snow</td>
<td>Neige, Brouillard, Poudrerie élevée, Neige, Poudrerie élevée</td>
</tr>
<tr>
<td>Storms</td>
<td>Orages, Averses de pluie, Averses de pluie, Brouillard, Averses de pluie forte, Averses de pluie modérées, Orages, Pluie modérée</td>
</tr>
<tr>
<td>Rain</td>
<td>Pluie, Brouillard, Pluie, Averses de pluie, Pluie modérée, Pluie forte, Pluie, Neige, Pluie modérée, Brouillard</td>
</tr>
<tr>
<td>Freezing rain</td>
<td>Pluie verglaçante, Neige, Brouillard</td>
</tr>
</tbody>
</table>
Chapter 4. Methodology

4.3.2 Presentation of Montreal

The data treated in this project have a spatial dimension; bike trips link stations dispersed in a city. Consequently, spatial representation of the trips, their environment and their context is essential. To attain this, a visual and urban presentation of Montreal is done in addition to the descriptive and the data are also geographically represented to help make connections. The contextual information was rendered on maps to visually express the relation the environment can have with the operations of the BIXI system. A series of maps made using a geographic information system software show the topography of the city, its urban composition, its administrative features, the natural environment and the demography, and how all of these characteristics are connected and might influence - and be influenced by - the metabolism of the BIXI system.

In all cases, the data and the maps were simply downloaded from the web either in CSV or shapefiles formats. All the data were incorporated as layers into the GIS software and only some adjustments were made to enhance the rendering as wanted.

The topographic map indicates the elevation of the territory of the city. The administrative map shows the major streets and the central neighbourhoods. The environmental map locates the principal green and blue infrastructures (parks, woods, rivers, streams, lakes, etc) as well as all the BIXI stations and their capacity in dates of the 2017 season which has been taken from the BIXI live database, incorporated to the stations CSV file and directly uploaded into the GIS software. The sociodemographic map presents the density of population of each neighbourhood, the location of the BIXI stations as well as those of the Metro. A buffer zone of 250 meters around the metro station automatically computed by QGIS helps grasping the connectivity between the Metro and BIXI systems. The population density has been calculated simply by taking the total population of the neighbourhood and dividing it by its area. Therefore, the margin of error is quite high since it does not take into account the presence of parks, rivers or empty spaces. The urban composition map puts in relation the BIXI system with the urban use of the city. On this map are shown residential and commercial zones, skyscrapers, Universities, institutions, job sectors and cultural or religious heritage like cemeteries and religious buildings.

4.3.3 The Dynamic Analysis

Once the profile of the city is known, a dynamic analysis of the stations evolution reveals way more information. Two kinds of geographical visualization are done: one is a dynamic map of the stations balance at every hour on an annual average, and the other is a representation of the trips and their frequencies between the stations.

Dynamic Maps of Stations’ Balancing Factors

The balancing factor (as defined by Leduc (2013)) of a station corresponds to a new variable associated to the BIXI stations that is computed in the following way:

1. The total amount of trips that started and ended during working days was calculated for each station.
2. They were separated according to the hour of the day when they happened.
3. They were divided by the number of working days included during the analyzed period.
4. The average number of leaving trips was subtracted to the average number of arriving trips to give the balancing factor.

The resulting variable takes negative values when the station knows more departures than arrivals during a specific hour on average in the total period analyzed, thus reducing its general bike availability. Conversely, the values are positive if there are more arrivals than departures at a station, thus increasing its general bike availability. In the project, this analysis has been focused to users activities during working days because the goal was to understand and represent a pattern that is stable and well-known from the literature: on working days, the service is mostly used to transit to work in the morning (around 8:00) and back in the afternoon (around 17:00). Thus, the analysis is done by selecting trips made by users between April 15th and November 13th 2015, which accounts for 153 working days. The results were mapped with a colour scale representing the balance as shown in figure 4.2. For each hour, a picture of the state of the BIXI system illustrates which stations are attractors (in blue) and which ones are repulsors (in red). This kind of visual rendering of the BIXI system enables the operational manager to identify which stations encounter an excess of arrivals and might need a redistribution of their bikes and when they would need it. Conversely, the manager is also in position to identify which stations might need more bikes when, or even before, they turn red. His gave 24 images corresponding to the average hourly balance of the stations for the 24 hours of the day. The 24 images were grouped and animated under a graphics interchange format, or commonly called an animated GIF so the daily evolution of departures and arrivals around the city are viewed.

\[
\text{Balance} \\
\bullet [-20, -4] \\
\bullet ]-4, -0.5] \\
\bullet ]-0.5, 0.5] \\
\bullet ]0.5, 4] \\
\bullet ]4, 55]
\]

**Figure 4.2:** The colour scale used to characterize the stations balance.

### Map of Trips and Their Frequencies

To complete the dynamic analysis, another map was made representing the most frequently done trips. Basically, the number of occurrences of a trip from station A to station B was calculated and converted as a weight of the edge (trip) of the graph when each station represent a node. This graph is visualized by considering the geolocation of the nodes. Then, over the map of Montreal, all the trips were plotted; a gradient of colour highlighting the weight (or the frequency) of a trip such that darker edges are more frequent trips.

### 4.4 A Complex Network Approach

In order to shed some light on more unexpected features of the BIXI system, a complex network approach was used to study its topology and its structure. Most networks - such as the ones created by the operations of a BSS - display substantial non-trivial topological features, with patterns of connections between their elements that
are neither purely regular nor purely random. Such features can be observed, for example, through the degree distribution which can be used to characterize the type of network analyzed. As well, it is possible to quantify the community structure of a network through its modularity (the strength of division of a network into modules) and its clustering coefficient. For example, Li et al. (2015) recently identified bottle-necks in city traffic using network theory. As a matter of fact, the present case is also about identifying common patterns between similar stations in order to find some community structure and, thus, predict their behaviour.

In the studied case, the network is made out of the trips taken between stations. Each station corresponds to a node (or a vertex) in the network and every trip taken by a user will form an edge (or lattice or tie) between the stations of origin and destination. Since the trips (edges) have an origin and a destination, the resulting network is directed. Directness in a network is more likely to involve reciprocity and triad significance profile which are both characteristic of a high clustering potential. The network is also weighted by the frequency of a specific trip between two stations; i.e. if the trip from station A to station B has occurred 3 times, the corresponding edge will have a weight of 3. Thus, the whole system is a weighted directed network.

It is absolutely essential to not consider the network in a geospatial way. The network does not (in the present case) take the geolocation into account. Therefore it is the network topology and structure that is studied here. When distances are mentioned, they are topological distances based on connections (edges) and not real euclidean distances.

The construction of the network started by simply counting the amount of incoming and outgoing trips per stations to first have a look at the in and out degrees of each one. Then, the stations were paired and the directed trips between each pair were summed. Directed pair means that they are not commutative; a trip from station A to station B is distinguished from a trip from B to A. The number of trips is the weight of the respective pair. These pairs and their weights are the resulting adjacency matrix of the system and this matrix was used to build a complex network. Each station represents a node, each pair of stations (or each trip) forms an edge between two nodes, and the weight of the pair is the weight of the edge. Once the network is built, it is easy to calculate various parameters.

4.4.1 Centrality measures

The centrality of nodes, or the identification of which nodes are more “central” than others, has been a key issue in network analysis (Opsahl, Agneessens, and Skvoretz, 2010). Research on network theory led to the observation of three significant features: having an advantaging position among the network such as having more edges, being able to reach all the others more quickly and controlling the flow between the others. Based on these three features, Freeman (1978) formalized three different measures of node centrality: degree, closeness, and betweenness respectively.

Dealing with a weighted and directed network has some particularities. The degree, the betweenness and the closeness centralities have been previously generalized to weighted networks. In a first set of generalizations, Barrat et al. (2004) calculated degree by taking the sum of weights instead of the number of ties, while Newman (2000) and Brandes (2001) used Dijkstra (1959) algorithm of shortest paths for generalizing closeness and betweenness to weighted networks, respectively. However, these generalizations focused solely on tie weights and ignored the original
feature of the measures: the number of ties. As such, a second set of generaliza-
tion were proposed by Opsahl, Agneessens, and Skvoretz (2010) that incorporates
both the number of ties and the tie weights by using a tuning parameter $\alpha$. (Opsahl,
Agneessens, and Skvoretz, 2010)

**Degree Centrality**

The degree is the simplest of the node centrality measures because it uses the local
structure around nodes only. Basically, the degree of a node is determined by the
number of edges connected to it. In a directed network, a node may have a different
number of outgoing and incoming edges, hence degree is split into out-degree and
in-degree, respectively. The degree measures the involvement of the node in the
network. Its simplicity is an advantage. However, there are limitations: the measure
does not take into consideration the global structure of the network. For example,
although a node might be connected to many others, it might not be in a position
to reach further ones quickly to access resources, such as information or knowledge
(Opsahl, Agneessens, and Skvoretz, 2010). For a BSS, a station with high in-degree
will be a popular one where a lot of trips are arriving. Conversely, a station with
high out-degree is a gregarious one where many trips begin. In both cases, stations
located in the business downtown area, for example, are more likely to have high in
and out degrees on weekdays.

Degree has generally been extended to the sum of weights when analyzing
weighted networks and labelled node strength (Barrat et al., 2004; Newman, 2004).
It is equal to the traditional definition of degree if the network is binary (i.e., each
tie has a weight of 1). Conversely, in weighted networks, the outcomes of these two
measures are different. Since node strength takes into consideration the weights of
ties, this has been the preferred measure for analyzing weighted networks proposed
by Barrat et al. (2004) (formula 4.2). However, node strength is a blunt measure as it
only takes into consideration a node’s total level of involvement in the network, and
fail to take into account the main feature of the original measures formalized by Free-
man (1978): the number of ties (formula 4.1). (Opsahl, Agneessens, and Skvoretz,
2010)

\[
k_i = C_D(i) = \sum_{j} x_{ij}
\]  

(4.1)

where $i$ is the focal node, $j$ represents all other nodes, $N$ is the total number of
nodes, and $x$ is the adjacency matrix in which the cell $x_{ij}$ is worth 1 if nodes $i$ and $j$
are connected, and 0 otherwise.

\[
s_i = C^{w}_{D}(i) = \sum_{j} w_{ij}
\]  

(4.2)

where $s$ is the strength of the node $i$ and $w$ is the weighted adjacency matrix in which
$w_{ij}$ is $\geq 0$ if nodes $i$ and $j$ are connected, and the value indicates the weight of the
edge.

In an attempt to combine both degree and strength, Opsahl, Agneessens, and
Skvoretz (2010) used a tuning parameter $\alpha$ to set the relative importance of the num-
ber of ties compared to tie weights. Specifically, the proposed degree centrality mea-
sure proposed was the product of the number of nodes that a focal node is connected
to, and the average weight to these nodes adjusted by $\alpha$ (formula 4.3). There are two
benchmark values for the tuning parameter (0 and 1), and if the parameter is set to either of these values, the existing measures are reproduced as follow. If the parameter is set to the benchmark value of 0, the outcomes of the measures are solely based on the number of ties, and are equal to the one found when applying Freeman (1978) measure to a binary version of a network where all the ties with a weight greater than 0 are set to present. In doing so, the tie weights are completely ignored. Conversely, if the value of the parameter is 1, the measure is based on tie weights only, and are identical to the already proposed generalization by Barrat et al. (2004). This implies that the number of ties is disregarded. (Opsahl, Agneessens, and Skvoretz, 2010)

Mathematically, this measure is expressed as:

$$C_w^{\alpha}(i) = k_i \times \left( \frac{s_i}{k_i} \right)^\alpha = k_i^{(1-\alpha)} \times s_i^\alpha$$  \hspace{1cm} (4.3)

where $\alpha$ is a positive tuning parameter that can set according to the research setting and data. High degrees are favoured by $0 \leq \alpha \leq 1$ and low degrees are favoured by $\alpha > 1$.

**Shortest Path**

While the shortest path often is not of interest in itself, it is the key component of a number of measures such as the closeness and the betweenness centralities. The shortest distance among nodes in a network is easy to calculate if the edges are single weighted: you just sum them up. Two directly connected nodes are at a distance of 1 of each other. Two indirectly connected nodes separated by intermediaries are at a distance equal to the lowest number of edges to cross to connect both nodes. In the context of a BSS, the shortest path does not consider the geographical distance, but the connectivity by the trips between the stations. Therefore, it is possible be that the two geographically furthest stations are "closer" than two stations next to each other that have experienced no trips between them.

Difficulty occurs when ties are differentiated, as they are in a weighted network. If we are looking at the diffusion of information or diseases in a network, then the speed that it travels, and routes that it takes, are clearly affected by the weights. Based on the original binary shortest distance formula (equation 4.4, Dijkstra (1959) proposed an algorithm that sums the cost of connections and finds the path of least resistance. For example, GPS devices uses this algorithm by assigning a time-cost to each leg of the road, and then find the route that cost least in terms of time. This algorithm can also be used in social network analysis. Newman (2000) applied it to a collaboration network of scientists by inverting the tie weights (dividing 1 by the weight, see equation 4.5). This implies that a stronger tie gets a lower cost than a weaker tie.

This application of Dijkstra’s algorithm is fine to use if we wish to identify the shortest paths; however, what does an average distance mean? To this end, Opsahl, Agneessens, and Skvoretz (2010) suggests to “normalize” the weights by the average weight in the network (see equation 4.6). A unit of distance then refers to one step with the average weight in the network. (Opsahl, Agneessens, and Skvoretz, 2010)

$$d(i, j) = \min (x_{ih} + ... + x_{hj})$$  \hspace{1cm} (4.4)

where $h$ represents the intermediary nodes on the path between $i$ and $j$.

$$d^w(i, j) = \min \left( \frac{1}{w_{ih}} + ... + \frac{1}{w_{hj}} \right)$$  \hspace{1cm} (4.5)
where $\phi$ is the average weight of the edges in the network and $\alpha$ is a positive tuning parameter. Similarly with the degree algorithm, when $\alpha = 0$, then Opsahl’s equation 4.6 is equivalent to the original equation 4.4; and when $\alpha = 0$, it is equivalent to Newman’s equation 4.5.

### Closeness Centrality

To capture the access to information or knowledge and consider the global structure of the network, closeness centrality was defined as the inverse sum of shortest distances to all other nodes from a focal node; i.e. as the inverse of farness. The intent behind this measure is to identify the nodes which can reach others quickly. A main limitation of closeness is the lack of applicability to networks with disconnected components (Opsahl, Agneessens, and Skvoretz, 2010). As for a BSS, a station will have a high closeness centrality if people from all over the city are going there and coming back so this station is at a distance 1 from all the other stations. Therefore, this station is more likely to be located in a zone fulfilling many functions: different job sectors, leisure, education, bars, restaurants, social activities, etc.

Closeness has been generalized to weighted networks by Newman (2000) who used Dijkstra (1959) algorithm for shortest paths as explained previously. Similarly to Barrat et al. (2004) generalization of degree, Newman (2000) generalized algorithm solely focuses on the sum of tie weights, and fails to consider the number of ties on paths, which Opsahl, Agneessens, and Skvoretz (2010) solve by determining the length of the paths using their generalization of shortest paths. Therefore, the resulting equations are respectively:

\[
C^C(i) = \left[ \sum_{j} d(i, j) \right]^{-1}
\]

\[
C^\omega_C(i) = \left[ \sum_{j} d^\omega(i, j) \right]^{-1}
\]

### Betweenness Centrality

The extent to which a node is part of transactions among other nodes can be studied using Freeman (1978) betweenness measure. This centrality measure assesses the degree to which a node lies on the shortest path between two other nodes, and are able to funnel the flow in the network. In so doing, a node can assert control over the flow. Although this measure takes the global network structure into consideration and can be applied to networks with disconnected components, it is not without limitations. For example, a great proportion of nodes in a network generally does not lie on a shortest path between any two other nodes, and therefore receives the same score of 0 (Opsahl, Agneessens, and Skvoretz, 2010). This measure is quite abstract in the context of a BSS. Indeed, ones should not see a station with high betweenness as a station where people “go through”, but more as a hub probably connecting different neighbourhoods in the city. For example, a station located close to a metro station far from the city centre will be used by the people using the metro
and living in the surroundings, but also by the determined BSS users that travel to
the city centre and back by bike. Therefore, because of the later, this station becomes
the funnel between the neighbourhood and the centre of the network. This example
was fictive in order to better illustrate the betweenness.

Brandes (2001) proposed a new algorithm for calculating betweenness faster. In
addition to reducing the time, this algorithm also relaxed the assumption that ties
had to be either present or absent (i.e. a binary network), and allowed betweenness
to be calculated on weighted networks. This generalization takes into account, that
in weighted networks, the transaction between two nodes might be quicker along
paths with more intermediate nodes that are strongly connected than paths with
fewer weakly-connected intermediate nodes. This is due to the fact that the strongly
connected intermediate nodes have, for example, more frequent contact than the
weakly connected ones. If we are studying the nodes that are most likely to be fun-
elling information or diseases in a network, then the speed at which it travels, and
routes that it takes, are clearly affected by the weights. If we assume that transactions
in a weighted network follow the shortest paths identified by Dijkstra’s algorithm
instead of the one with the least number of intermediate nodes, then the number of
shortest paths that pass through a node might change.

Similarly to Newman (2000) generalization of closeness, Brandes (2001) gener-
eralized algorithm solely focuses on the sum of tie weights, and fails to consider the
number of ties on paths\(^1\). Opsahl, Agneessens, and Skvoretz (2010) generalization
of shortest paths can also applied to identifying them (equation 4.10). (Opsahl, Ag-
neessens, and Skvoretz, 2010)

\[
C_B(i) = \frac{g_{jk}(i)}{g_{jk}}
\]  

(4.9)

where \(g_{jk}\) is the number of shortest paths \((d(i, j))\) between two nodes, and \(g_{jk}(i)\) is
the number of those paths that go through node \(i\).

\[
C_{B}^{wα}(i) = \frac{g_{jk}^{wα}(i)}{g_{jk}^{wα}}
\]  

(4.10)

The generalizations proposed by Opsahl, Agneessens, and Skvoretz (2010) can
also be applied to directed networks. The main difference being an added constraint:
a path from one node to another can only follow the direction of present edges.

**Eigenvector Centrality**

The eigenvector centrality (also called eigencentrality) is a measure of the influence
of a node in a network. It assigns relative scores to all nodes in the network based
on the concept that connections to high-scoring nodes contribute more to the score
of the node in question than equal connections to low-scoring nodes. It considers
nodes connected to other high degree nodes as highly central. It can be considered
as an extended form of degree centrality and closely related to centrality measures
used in web search engines. For instance, Google’s PageRank (Brin and Page, 1998)

\(^1\)Complementary information: The implementation of Brandes’ algorithm finds multiple paths if
they have exactly the same distance. For example, if one path is found over the direct tie with a weight
of 1 (distance = 1/1 = 1) and a second path is through an intermediary node with two ties with weights
of 2 (distance = 1/2 + 1/2 = 1), the two paths have exactly the same distance. However, if there is a
third path through two intermediaries with three ties with weights of 3 (distance = 1/3 + 1/3 + 1/3),
it does not exactly equal 1 as computers read these values as 0.3333333 and the sum of these values is
0.9999999. Therefore, this path is considered shorter than the other two paths (distance = 1).
and the Katz centrality (Katz, 1953) are variants of the eigenvector centrality. The eigenvector centrality of a vertex in an unweighted network is defined to be proportional to the sum of the centralities of the vertex's neighbours, so that a vertex can acquire high centrality either by being connected to a lot of others (as with simple degree centrality) or by being connected to others that themselves are highly connected. (Newman, 2004)

Tore Opsahl does not provide an alternative to the eigencentrality. However, Newman (2004) demonstrates that the eigencentrality is resilient to the addition of the weights since it is still the eigenvector of the adjacency matrix. He explains that network neighbours that are connected to a vertex with twice the weight now contribute twice as much to the vertex’s eigenvector centrality. As a result, the correct generalization of eigenvector centrality to a weighted network is still the leading eigenvector of the adjacency matrix, with the elements of the matrix being equal to the edge weights, as before.

After having calculated the principal network measures, the top ten stations with the highest values for each measure were represented on a map for a deeper analysis. The analysis of these maps is an attempt at better understanding some hidden features of the BIXI BSS brought to light by the complex network approach.

4.5 Clustering

So many variables and parameters can affect the behaviour of a BSS user. Also, the topology of the information concerning the users and their trips makes it more suitable to use a clustering approach in order to classify the users’ behaviour and the associated trips. By classifying the behaviour of the users, it is easier and more efficient to predict the usage based on the behaviour than by trying to predict the demand for a specific station at a specific time by extrapolating from the past. It also has the advantage of being open to include contextual information and leave the interpretations opened. Therefore, it is not mandatory to know exactly the number of classes beforehand.

Since the working database is composed of several different types of variables, the agglomerative hierarchical clustering method is preferred over, for example, the k-means one which only works for numerical values. Hierarchical clustering does not require to pre-specify the number of clusters and is generally deterministic. It also has an embedded flexibility regarding the level of granularity. It provides an easy handling of any distances and it can be applied to any kind of variable. These advantages of hierarchical clustering come at the cost of lower efficiency.

4.5.1 The Ward’s Method

The Ward’s method (Ward, 1963) is usually preferred over the others because it works with inertia and variability. As stated by Benzécri (1977), the more information is inside a group, the greater is its inertia. This is referring to information in the sense of the information theory from Shannon and Weaver (1949). This means that Ward’s method produces more informative clusters than others because it is optimizing in function of the inertia and, thus, optimizing the pureness of information within the cluster. Similarly, the inertia is also proportional to the variability of the group. Hence why the goal is to cluster in a way to isolate specific information into classes. To group the two closest elements is equivalent to choose the pair of points which combination reduces at its lowest the inertia as all aggregation necessarily
reduces intra-cluster inertia by construction; just like Huygens-Steiner’s theorem in physics (Haas, 1928). In short, this method tries to minimize the intra-cluster inertia and maximize the inter-cluster inertia in order to reach the most homogeneous clusters. According to Steiner’s theorem, the sum of intra- and inter-class variances is constant and equal to the total data variance, regardless of the number of clusters or their compositions. Thus minimizing the intra-inertia or maximizing inter-inertia is equivalent (Goutte et al., 1999). (Lebart, Morineau, and Piron, 1995)

From a statistical perspective, the Ward’s method seems better than other hierarchical clustering strategies. This is because it has an objective function to minimize the within-group sum of variability and therefore to maximize the among-group variability; thus, it gives a natural connection to the analysis of variance (Crossa, Bellon, and Franco, 2002). Furthermore, Ward’s method is appropriate for multi-normal data distribution. One problem with this sequential clustering strategy, however, is that while all variables, continuous and discrete, are used to form the initial groups using the Ward’s method, only the continuous variables can be used in the mixture models (Crossa, Bellon, and Franco, 2002). As well, Ward’s method has a tendency to produce balanced and spherical classes which is not always representative of the analyzed data structure (Faye et al., 2011).

Nonetheless, it is common to realize diverse dendrograms on the same data using different techniques in order to reach the most reasonable clustering because of the wide variety of distances and aggregation techniques that exist. Therefore, various experiments among the clustering options were done using the samples and, as expected, Ward’s technique was the one giving the most reasonable dendrograms in terms of structure, hierarchic trees and inertia (heights of the branches).

A dendrogram is a hierarchical graph where every leaf represents one object. The complete set of leaves linked to a given internal node determines a cluster. The dendrogram shows how objects have been sequentially grouped in the clustering process. The vertical axis is related with cluster homogeneity. The lower the level of an internal node, the more similar the elements of this cluster. Low level cuts of the dendrogram produce much more specific clusters. High cuts produce few general clusters. Gaps in clusters homogeneity are good points to cut the tree, related with Calinski-Harabatz index (Calinski and Harabasz, 1974).

Once the method and the technique were chosen, another battery of experiments was executed to identify the best combination of variables considered for the clustering. The experiments conducted for the variable selection, as mentioned previously, were also used to try out the best combination for the clustering process. More than 27 experiments were conducted on just the original variables coming from the BIXI database.

4.5.2 Gower Distance

"An appropriate dissimilarity measure is far more important in obtaining success with clustering than choice of clustering algorithm."

(Hastie, Tibshirani, and Friedman, 2009)

The standard version of the Ward’s method uses the Euclidean distance as a reference. In our application, this is not suitable since our dataset contains also qualitative variables. Gower distance will be used instead.

Gower’s general similarity coefficient is one of the most popular measures of proximity for mixed data types. The clustering options available using Gower are restricted to those applicable to similarity measures, and not to dissimilarities. Thus,
for example, you will not be able to optimize the Euclidean sum of squares without first transforming your proximities into distances.

To cluster, it is preferable to work with dissimilarities over similarities because the coefficient becomes then proportional to the inertia which has to be minimized in the process. Gower dissimilarity is just 1 minus Gower similarity. So, they are "the same", and limitations of one are the limitations of the other. Gower (1971) demonstrates in his paper that the square of the Gower’s dissimilarity coefficient is Euclidean which is necessary for Ward’s clustering method. The calculation of the Gower dissimilarities gives a matrix containing its coefficients. From this matrix can be applied the hierarchical clustering using various aggregation methods: Ward, single, complete, average, McQuitty, median or centroid.

In the present case, variables are distinguished as logical, qualitative and quantitative. Qualitative characters may have many levels (e.g. black, green, yellow, blue) but unlike the levels of quantitative characters they do not form an ordered set. Gower (1971) explains its calculations as follow:

“Two individuals i and j are compared on a character k and assigned a score $s_{ijk}$: zero when i and j are considered different and a positive fraction, or unity, when they have some degree of agreement or similarity. There are many ways of calculating $s_{ijk}$, some of which are described [in equation 4.13]. Sometimes, no comparison is possible because information is missing, or in the case of [logical] variables a character is non-existent in both i and j. The possibility of making comparisons can be represented by a quantity $δ_{ijk}$, equal to 1 when character k can be compared for i and j, and 0 otherwise [see equation 4.12]. When $δ_{ijk} = 0$, $s_{ijk}$ is unknown but is conventionally set to zero. The similarity between i and j is defined as the average score taken over all possible comparisons:”

$$S(i,j) = \frac{\sum_{k=1}^{v} δ_{ijk}s_{ijk}}{\sum_{k=1}^{v} δ_{ijk}}$$

(4.11)

$$δ_{ijk} = \begin{cases} 
0, & \text{if either or both of } i_k, j_k \text{ are missing} \\
0, & \text{if } k \text{ is logical and both } i_k, j_k \text{ are non-existent} \\
1, & \text{otherwise} 
\end{cases}$$

(4.12)

where v is the total number of variables in the database. $δ_{ijk}$ indicates only when comparisons are possible. The scores $s_{ijk}$ are assigned as follows:

$$s_{ijk} = \begin{cases} 
1 - \frac{|x_i - x_j|}{R_k}, & \text{if } k \text{ is numerical} \\
1, & \text{if } k \text{ is qualitative and } i = j \\
0, & \text{if } k \text{ is qualitative and } i \neq j 
\end{cases}$$

(4.13)

where $R_k$ is the range of character k and may be the total range in the population or the range in the sample. In our case, the logical variables are following the same values as the qualitative ones since there are no missing values, thus all of them are comparable. Otherwise, Gower (1971) explains the possible scores and validity of the logical character comparisons, which will not be discussed here.
Thus, $S_{ij}$ is defined by a continuous value ranging from 0 to 1. A value of 1 means that the two individuals differ in no character (they are exactly similar) whereas 0 means they differ maximally in all their characters.

### 4.5.3 Cutting the Dendrograms

When it comes to clustering, the correct number of classes ($k$) is often ambiguous, with interpretations depending on the shape and scale of the distribution of points in a data set and the desired clustering resolution of the user. In addition, increasing $k$ without penalty will always reduce the amount of error in the resulting clustering, to the extreme case of zero error if each data point is considered its own cluster (i.e., when $k$ equals the number of data points $n$). Intuitively then, the optimal choice of $k$ will strike a balance between maximum compression of the data using a single cluster, and maximum accuracy by assigning each data point to its own cluster. If an appropriate value of $k$ is not apparent from prior knowledge of the properties of the data set, it must be chosen somehow. There are several categories of methods for making this decision and there is no only one best choice since cluster analysis is essentially an exploratory approach; the interpretation of the resulting hierarchical structure is context-dependent and often several solutions are equally good from a theoretical point of view. Cluster Analysis seeks to isolate groups of statistical units (whether it be individuals or variables) for exploratory or descriptive purpose, essentially. Hence, the interpretation of the output of the clustering scheme is a critical phase and several cluster solutions may be equally interesting. (Omatu et al., 2015)

The methodology used to cut the dendrograms and decide the number of classes was the one proposed by Gibert, Rodríguez-Silva, and Rodríguez-Roda (2010). First, it begins with an analysis of the resulting dendrogram to obtain the best horizontal cut, using heuristic criteria or automatic tools. A very common criteria in hierarchical clustering is to cut the tree by optimizing the Calinski-Harabatz index, which measures the ratio between the intra-classes inertia and the inter-classes inertia. This corresponds to choosing the cut with the longer gap in the dendrogram and this index can also be optimized graphically. Performing the cut of the dendrogram identifies a partition of the data in a set of classes.

To strengthen the cutting decisions, an analysis of the intra-cluster inertia is processed in parallel. This analysis evaluates the marginal inertia, i.e., how much the inertia gets reduced by the addition of another cluster. It is similar to the Elbow method. Ketchen and Shook (1996) explains it as graphing the agglomeration coefficient (i.e., a numerical value at which various cases merge to form a cluster) on a y-axis and the number of clusters on an x-axis. The first clusters will add much information (explain a lot of variance), but at some point the marginal gain will drop. A marked flattening of the graph suggests that the clusters being combined are very dissimilar, thus the appropriate number of clusters is found at the "elbow" of the graph.

Then, among all the best possible cuts, must be chosen the one which allows the best interpretation. To do so, concepts have to be associated with each resulting class within the chosen set.

Finally, a class panel graph and basic statistics per group to associate conceptual labels to the classes are used to interpret every class. The class panel graph is a compact visualization of the conditional distributions of the variables versus the classes that permits easy identification of the variables with distinctive behaviour in certain classes (Gibert et al., 2005). Consequently, together with basic statistics
Chapter 4. Methodology

per groups, it helps to understand the main characteristics of every class and allows expert conceptualization of the classes and further labelling with a domain concept.

The three selected sets of clusters are named H1, H6 and H11.

4.6 Robustness, Sampling and Variable Selection

First, the descriptive statistics of the variables gave an insight about which one would be more relevant to work with or not. However, only variables coming directly from the operation of the BIXI system were considered at first. These data concern mostly the bike trips intrinsic information such as time, locations and user type. The point being that the clustering analysis is principally applied to the bike trips to identify similarities between them and group them according to these similarities. Then, we do believe that the context of a trip is crucial for its reason of being. Consequently, external data containing contextual information about a trip influences the trip information and can greatly enhance its understanding, and help the decision-making process managing the system.

Additionally, we will use Ward algorithm in this project and, since it is quadratic, it cannot be used with more than three million trips. To bypass the time complexity of the hierarchical clustering method which is hardly scalable (at least $O(n^2 \log(n))$, generally $O(n^3 d)$) (Manning, Raghavan, and Schütze, 2009), random samples of data were created from the main working database. There are several strategies to accelerate complexity of hierarchical clustering methods to make it scalable to big data sets. One is CURE (Clustering Using REpresentatives) (Guha, Rastogi, and Shim, 1998) that goes over a reduced sample and connects with k-nearest neighbour to extend the classification among the the remaining objects of the big data set. In this sense, we start building the clusters of a reduced sample as a first part of a CURE strategy.

Let’s recall that the size of the samples used were of 5,000 objects whereas the whole dataset has 3,503,355, which means that we worked with samples of 0.1% of the population. Since sampling can entail misrepresentation and bias, several experiments were replicated with several random samples to evaluate the robustness of clusters discovered in order to ensure stability of the kNN step of CURE strategy. In kNN step, each remaining object is compared to the clusters minimizing the distance between the object itself and the centroid of the class or the class representatives.

Different trials using the same parameters gave a good insight about the stability of the clustering. Stability and scalability were two essential characteristics evaluated during the processes. The scalability is referring mostly to time scales. It turned out that when evaluating the robustness of profiles obtained on several samples, some of the variables showed more unstable with the others. For this reason, a second generation of experiments was conducted to analyze which is the subset of relevant variables that produces stable results over random samples of data set. Once a consistent, stable and confident clustering was achieved with the BIXI variables, a final set of clusters was chosen to act as the basis onto which the further experiments with the contextual variables would be done. Details about the clustering process are given in the next section 4.5.

Then, a series of experiments was conducted to learn how the BIXI variables were correlated and interacted between each other during the clustering process. These experiments consisted into applying the clustering process with different sets of variables and evaluating the dendrograms and the profiling. The study of the profiling of an experiment was only done when the dendrogram was reasonable
and offered good cuts, then each cut was assessed. In total, the results for 27 sets of clusters have been recorded and way more have been done that did not give satisfying results after the profiling examination.

These experiments taught that the main issue regarded the variables expressing the timestamps of the trips. For instance, when included, the variables Weekdays, Month and Weekend prevailed over the other variables in the clusters profiles. The Month variable was irrelevant if the sample was taken from a specific week only. Also, the specific Weekdays variable was producing unstable results on the various samples. Consequently, since the aim of the project is to reach a model that would help the decision making process and, hopefully, the prediction of the demand, the Month and Weekdays were left out. The literature review highlighted the fact that the users follow two distinct patterns depending if it is during the weekend or not. On working days, the usage is dominated by the members that transit to and from work, hence generating peaks of use at the rush hours, but there is no distinction from a Tuesday to a Thursday for example. For these reasons, the time variables Year, Month, Weekdays and Hour were left out for the clustering. The logical variable Weekend was preferred over Weekdays, and the qualitative variable Period was more practical for the clustering than Hour that was composed of too many modalities (every hour being one).

For the contextual variable, the whole experimental process previously described was repeated apart from the evaluation of the clustering methods. In the end, 12 more sets of clustering were recorded up to their profiling for the external information variables, way more were done but did not prove themselves relevant enough.

4.7 Interpretation of the Clusters

One of the crucial issues in clustering is to understand the underlying meaning of clusters. The interpretation of the classes can be difficult and time consuming. It also holds a big part of subjectivity through its interpretability criteria. First, for all the variables, the relevance of differences between classes is usually assessed using the corresponding statistical test (ANOVA, Kruskall-Wallis test or chi-2 independence test) according to Gibert, García-Alonso, and Salvador-Carulla (2010). However, the amount of observations comprised in the working dataset (over 3,500,000) bias the \( p \)-value that can recognize significance where there isn’t really. In this case, graphical tools oriented to support the interpretation of classes is better used. These visual interpretation are done using a class panel graph of significant variables where conditional distributions of the variables through the classes, displayed through multiple histograms or bar-charts, are shown. (Gibert, Rodríguez-Silva, and Rodríguez-Roda, 2010; Gibert, Conti, and Vrecko, 2012; Gibert, García-Rudolph, and Rodríguez-Silva, 2008)

This marked class panel graph allows to easily observe why the variables are significant and relevant for the description of the cluster profiles; particularities of the variables in one specific class regarding the others can be seen and evaluated and, as coherences are found, this helps to develop a conceptualization process which leads to a class-labelling proposal regarding the semantic entity represented by each class (Gibert, García-Alonso, and Salvador-Carulla, 2010).

From the class panel graphs of each set of clusters, a visual comparison assessment is done. The comparison is mostly between the common variables shared among certain pairs of cluster to see how each particularities are shared between the clusters. For instance, a variable can be strictly confined to one class and the same
variable be more spread between the classes of another set of clusters. Therefore, in such case, the former clustering set would be said to be more clearly clustered with less skewing between the variables.

After the comparisons between the sets of clusters, a multi-view cross-clustering was done between two selected sets of clusters that were made independently using different variables. This was done by considering the classification of both sets within the sample, thus labelling them with a binomial nomenclature. From there, deeper analysis were carried out. First, the distribution of the later set within each class of the former set was calculated to see if some of the classes were more likely to happen during specific contexts. Then, the two distinctive most frequent cases were studied as a local analysis of the most common patterns to provide insights about important decision making.

### 4.8 Integrated Multi-view Clustering

The clusters’ profiles of H6 were neither quite appreciable nor totally relevant. Clustering of the trips might be skewed because the big amount of variables include a wide range of aspects and not all variables play the same role. Indeed, some are contextual and others are intrinsic to the trips. Thus, the multi-view clustering methodology is proposed to get a profiling of the trips properly conceptualized that can distinctively take into account both BIXI variables on the one hand and contextual variables on the other hand. (Sevilla-Villanueva, Gibert, and Sánchez-Marrè, 2017)

The multi-view clustering deals with the high dimensionality of data by splitting the variables in several subgroups. In the present case, there will only be a distinction between two groups, the BIXI variables and the external variables, similarly to what Bickel and Scheffer (2004) did. Then, the objects are clustered under each group of variables independently. In this project, two views are used: one coming from the BIXI variables and the other one from the external variables only.

There is no consensual technique to combine the various sets of clusters obtained independently and Abdullin and Nasraoui (2012) particularly stress the need for it. Ensemble methods and semi-supervised techniques are proposed among others. Authors claim that most of the effort must concentrate on the agreement between the different views or resulting partitions (Sevilla-Villanueva, Gibert, and Sánchez-Marrè, 2017).

In this work, multi-view clustering is proposed to find conceptual patterns on a sample of trips from the working data frame described earlier in section 3.2.5. The variables in the dataset describe different aspects of the user’s trip like the intrinsic space-time characteristics or contextual factors like the weather information and the geographical particularities. The idea is to split the original BIXI variables from the added contextual information. These subsets can be analyzed locally to address the high dimensionality, as the original multi-view clustering approach proposes. Then, trips are clustered using a either one of the two sets of variables each time, generating independent views.

However, giving several disconnected views of the data does not help decision-making. The views need to be integrated in some ways. Thus, (Sevilla-Villanueva, Gibert, and Sánchez-Marrè, 2015) propose to extend the multi-view clustering by an integrative step where the results of each view are properly combined to give a single partition combining the structure found under all views. To do so, the views are simply combined through a binomial labelling that consider both clustering on each objects.
Once the integrated multi-view clustering is done, an overview of the resulting classes distributions help determining if the results make sense and are relevant.

4.8.1 Local Analysis of Recurrent Patterns

Two cases stood out after the multi-view clustering. Afterwards, a local analysis is done on these two specific cases to elaborate recommendations to optimize the management. The local analysis is resumed by a repetition of the dynamic analysis explained here (4.3.3) and the complex network analysis explained here (4.4). The results are then compared with the ones from the whole working data set and discussed in section ???. Finally, two other local analysis are conducted using only the dynamic approach and applied on classes from the uncrossed sets of clusters used for the multi-view clustering. The goal was to strengthen the power of the model developed to extract decisional knowledge.

4.9 Description of the Tools Used

Two software have been principally used for the development of this project: R and QGIS. The first one for everything related with data manipulation and mathematics and the second one for the visualization, rendering and conversion of geographical data into the working database in R. This section aims to briefly present these two software and their contributions.

4.9.1 R

Presentation of the Software

R is a language and environment for statistical computing and graphics. The term “environment” is intended to characterize it as a fully planned and coherent system, rather than an incremental accretion of very specific and inflexible tools, as is frequently the case with other data analysis software. It is a GNU project which is similar to the S language and environment which was developed at Bell Laboratories (formerly AT&T, now Lucent Technologies) by John Chambers and colleagues. R can be considered as a different implementation of S. There are some important differences, but much code written for S runs unaltered under R. R, like S, is designed around a true computer language, and it allows users to add additional functionality by defining new functions. Much of the system is itself written in the R dialect of S, which makes it easy for users to follow the algorithmic choices made. Advanced users can write C code to manipulate R objects directly.

R provides a wide variety of statistical (linear and nonlinear modelling, classical statistical tests, time-series analysis, classification, clustering, …) and graphical techniques, and is highly extensible. The R language is the Open Source vehicle of choice for participation for research in statistical methodology. R can be extended (easily) via packages. There are about eight packages supplied with the R distribution and many more are available through the CRAN family of Internet sites covering a very wide range of modern statistics.
It is important to understand that R was built by statisticians, not by data miners. Thus, its focus is on statistical expressiveness, not on scalability. As well, implementations in R aren’t the best, except for core R which usually has a competitive numerical precision with other software.

One of R’s strengths is the ease with which well-designed publication-quality plots can be produced, including mathematical symbols and formulae where needed. R has its own LaTeX-like documentation format, which is used to supply comprehensive documentation, both on-line in a number of formats and in hardcopy.

R is available as Free Software under the terms of the Free Software Foundation’s GNU General Public License in source code form. It compiles and runs on a wide variety of UNIX platforms and similar systems (including FreeBSD and Linux), Windows and MacOS. (Hornik, 2016)

Contributions and Packages

Within R, the `daisy()` function in the `cluster` package (Maechler et al., 2017) was used to calculate the dissimilarities or distances between the entries. `daisy()` computes all the pairwise dissimilarities (distances) between observations in the data set. The main feature of `daisy()` is its ability to handle variables of mixed types (e.g. nominal, ordinal, (a)symmetric binary, etc), even when different types occur in the same data set. In that case, or whenever metric = "gower" is set, a generalization of Gower’s formula is used. The handling of nominal, ordinal, and (a)symmetric binary data is achieved by using the general dissimilarity coefficient of Gower (1971). If $x$ contains any columns of these data-types, both arguments metric and stand will be ignored and Gower’s coefficient will be used as the metric. This can also be activated for purely numeric data by metric = "gower". With that, each variable (column) is first standardized by dividing each entry by the range of the corresponding variable, after subtracting the minimum value; consequently the re-scaled variable has range $[0,1]$, exactly (Maechler et al., 2017).

After applying the `daisy()` function, the result is a matrix containing the dissimilarity Gower’s coefficient. From this matrix can be applied the hierarchical clustering function `hclust()` (from the `stats` package (R Core Team, 2017)) using various aggregation methods: Ward, single, complete, average, McQuitty, median or centroid. `hclust()` applies an aggregative hierarchical cluster analysis on a set of dissimilarities using one of the specified method for analyzing it.

To form the three samples of 5,000 objects the base R function `sample()` was applied on the whole working database.

The network study has been done using R (R Core Team, 2017) and three different packages: `statnet` (Handcock et al., 2003), `igraph` (Csardi and Nepusz, 2006) and `tnet` (Opsahl, 2009). The three different tools have been used since different methodologies can be applied to obtain the general network metrics, hence giving different results. There is not a better method, but there might be a more appropriate one.

`statnet`: `statnet` is a suite of packages that have been developed by the `statnet` development team based out of University of Washington. It includes the packages `network` (Butts, 2015; Butts, 2008), `ergm` (Handcock et al., 2017; Hunter et al., 2008), `latentnet` (Krivitsky and Handcock, 2015; Krivitsky and Handcock, 2008), `sna` (Butts, 2016), `dynamicnetwork` (Bender-deMoll, Morris, and Moody, 2008), rSoNIA (Bender-deMoll, Morris, and Moody, 2008), `networksis` (Admiraal and Handcock, 2015) and `netperm` (Butts, 2006; Handcock, 2003). These packages allow researchers to create additional functions on top of existing ones. This ability greatly reduces
the time spent on programming, and let researchers focus on the contribution to
the literature instead. For example, if someone has already written a function for
identifying the shortest paths in a network, a researcher that would like to extend
this measure can simply work on this code without programming the function from
scratch. \texttt{statnet} allows to statistically test models based on an Exponential-Family
Random Graph Model framework. In fact, statistical modelling is the explicit pur-
pose for the development of this package. The biggest advantage of \texttt{statnet} is that it
is expansive and probably allows to do most things ones might want to do in social
network analysis. The trade-off, of course, is that because it is so comprehensive, it
can be daunting and complicated. (Shizuka Lab, 2017)

\textit{igraph}: \texttt{igraph} is a somewhat more basic approach than \texttt{statnet} to network anal-
ysis according to Shizuka Lab (2017). However, it still contains a lot of functional-
ity, including calculating network properties, generating random graphs for simula-
tions, etc. and will probably fit most needs. \texttt{igraph} seems to be more efficient than
\texttt{statnet}, and the many of the functions (particularly manipulating data and dealing
with vertex attributes) can seem more intuitive (Shizuka Lab, 2017). The list of
available measures is slightly more complete than for \texttt{statnet}. Indeed, \texttt{igraph} allows
calculations for the clustering coefficient or betweenness centrality using the edges
and not only the nodes like \texttt{statnet}. If the strength of \texttt{statnet} is statistical modelisa-
tion, \texttt{igraph}'s one is its capacity to generate classical graph models such as random,
small-world and scale-free (Beauguite, 2012).

\textit{tnet}: This third package has been developed by Tore Opsahl, an active researcher
in the field of weighted networks. He believed that there were no programs that
could both analyze the richer forms of network data and allow users to create their
own functions. He stated that, programs like UCINET and Pajek have a small set of
functions for weighted and two-mode networks, but they do not allow users to code
additional functions. Therefore, researchers proposing new measures must create
stand-alone programs to deal with a single aspect of weighted networks (Opsahl,
Agneessens, and Skvoretz, 2010). He added that a number of packages for ana-
lyzing networks has been created within R, notably the \texttt{sna} and \texttt{statnet} packages
mentioned previously. However, these packages rely on the basic \texttt{network} package
for data structures to represent networks (Butts, 2008). This basic package does not
have data classes for weighted, two-mode, and longitudinal networks. Therefore, to
ease the development of new functions for these networks, Tore Opsahl developed
such a platform with \texttt{tnet}. Although it is a user-written package in R, it does not rely on the \texttt{network} package (Opsahl, Agneessens, and Skvoretz, 2010). Basically,
this tool allows to modulate the consideration of the edges’ weight by varying an \( \alpha \)
parameter. When \( \alpha \) is 0, the weight of the edges is not taken into account at all. As it
gets closer to 1, the weight is more considered. Hence, one must set the \( \alpha \) according
to the extent of the weight of the edges. The calculations will be explained more in
details later on.

The \texttt{tnet} package is the one used for the complex network analysis given that
it has been developed specifically for the type of network that is the BIXI one. The
other packages have also been used as experimentation and to compare the results
from \texttt{tnet}, but since the complex network analysis is not the core of this project, no
comparisons between the packages are discussed here.

4.9.2 QGIS

Presentation of the Software
QGIS (previously known as Quantum GIS) is a cross-platform free and open-source desktop geographic information system (GIS) application that provides data viewing, editing, and analysis. It allows users to analyze and edit spatial information, in addition to exporting graphical maps. QGIS supports both raster and vector layers; vector data is stored as either point, line, or polygon features. Multiple formats of raster images are supported, and the software can georeference images. QGIS supports shapefiles, coverages, personal geodatabases, DXF, MapInfo, PostGIS, and other formats. Web services are also supported to allow use of data from external sources. QGIS integrates with other open-source GIS packages, including PostGIS, GRASS GIS, and MapServer. Plugins written in Python or C++ extend QGIS’s capabilities. Plugins can geocode using the Google Geocoding API, perform geoprocessing using fTools, which are similar to the standard tools found in ArcGIS (the most popular proprietary GIS software), and interface with PostgreSQL/PostGIS, SpatiaLite and MySQL databases.

Gary Sherman is at the origin of QGIS as he began to develop it in early 2002. Then, it became an incubator project of the Open Source Geospatial Foundation in 2007 and the first version was released in January 2009. QGIS is written in C++. QGIS has a small file size compared to commercial GIS’s and requires less RAM and processing power; hence it can be used on older hardware or running simultaneously with other applications where CPU power may be limited. QGIS is maintained by volunteer developers who regularly release updates and bug fixes. As a free software application under the GNU GPL, QGIS can be freely modified to perform different or more specialized tasks. As of 2012, developers have translated QGIS into 48 languages and the application is used internationally in academic and professional environments. As of 2017, QGIS is available for multiple operating systems including Mac OS X, Linux, Unix, and Microsoft Windows. A mobile version of QGIS was under development for Android as of 2014. (QGIS Project, 2014)
Chapter 5

Applications: Presentation of the Results

5.1 Data Descriptive and Visualization

In this section are presented the data in a more understandable and visual way through descriptive statistics. The descriptive of each variable is carried out with some basic statistical analysis to give the reader a good idea about the profile of the dataset. Then, the data are presented in a geographic representation using QGIS to better understand the context of the dynamic. Further on is presented a short complex network analysis approach towards the BIXI network’s topology and structure. In the end are presented the results from the clustering manipulations.

5.1.1 Descriptive Statistics

The figure 5.1 mostly presents the use of BIXI in time discriminated between members and casual users. Figure (a) presents the total number of trips during the three years where data are available and an increase in use of the service every year is noticeable. For all the other figures, data from the 2015 season are used and all bar plots are divided between members and casual users except if specified otherwise.

Figures (b), (c), (d) and (e) respectively reveal the proportions of trips taken by members versus casual users, coming back to their starting stations (looping trips), taken during the weekend, and taken on Friday or Saturday nights between 20:00 and 5:00. Let’s note that all figures (b), (c), (d) and (e) are all on the same y-scale where 100% is 3,500,000. From these figures, one can observe that the trips are predominantly done by members (83.6%) (figure b)). Very few trips end at the same station where they started (figure (c)) but, in proportion, casual users do more loop trips, probably due to tourism. In fact, the stations with most of the loop trips are located in the Old Montreal and on Sainte-Hélène Island where there is the famous Parc Jean-Drapeau. Casual users are more active on weekends, where the members are the least actives (figure (d)).

Figure (f) shows the distribution of the duration of the trips. However, the bar plot is skewed by outlying values. Thus, the figure (j) offers a zoom on the single bar of figure (f), hence showing only the trips with duration less than one hour. The duration of the trips seems to follow a power law. It has many outliers which are important to know about in order to expect some possibly skewed results because of them.

Figures (g), (h) and (i) expose the amount of trips for each month, weekday and period of the day respectively. The most active months for both members and casual users are July and August, when the city is animated by several festivals and outside activities (figure (g)). Let’s recall that a BIXI season goes from mid-April until
Figure 5.1: The caption being too big, please report to the footnote where the figure is mentioned for its detailed description.
mid-November which explains the very low use shown for both these half-months. Nevertheless, the harsh weather reduces greatly the use of the service during the months of April, October and November. Actually, the activity follows the weather: the warmer it is, the more cyclists there are. Another instance of this is when looking at the daily use throughout the whole year (figures (l) and (m)). Interestingly, members grow a motivation to use BIXI along the week and reach a peak on Thursdays. In descending order, the busiest days of the week are: Thursday, Friday, Wednesday, Tuesday, Monday, Saturday and Sunday (figure (h)). The period of the day that knows the highest activity is the afternoon (figure (i)), but this is better detailed in figures (l) and (m).

Figure (k) gives an overview of the amount of trips taken each day for the whole season. Note that all the specified dates on the x-axis are Wednesdays. It is also possible to identify the weekly pattern of use by the members with a decrease in use during the weekends whereas the casual users show peaks on weekends. Some days seem to have very few trips (from May 9th to 12th, June 28th, from September 12th to 14th, etc). This might be caused by bad weather as it will be discussed later on.

Figures (l) and (m) both exhibit the sum of the amount of trips taken during the whole year at each specific hour, but discriminated in two ways: by weekend and user-type respectively. As expected from the literature review, the graphs exhibit two daily peaks of use on labour days; one in the morning from 8:00 to 9:00 and one in the afternoon at the end of the working hours from 16:00 to 19:00. These correspond to the rush hours when people transit from home to work and back. On weekends, there are no peaks - the variation of usage from hour to hour is more mellow - but the activity is higher in the afternoon from 14:00 to 17:00 (figure (l)). For casual users, there are no peaks and the activity is also higher during the afternoon from 14:00 to 18:00 (figure (m)) at all time during the week. Very few trips are taken between 4:00 and 6:00 in the morning.

Figure (n) exhibits the elevations of the trips (difference in altitude between the destination and the origin stations). The trips go slightly down on average, but the distribution of the elevation of a trip seems to follow a normal distribution (figure (n)).

The descriptive of the variables \textit{Day}, \textit{Time}, \textit{TenS}, \textit{Start.Station.number}, \textit{End.Station.number}, \textit{Altitude.S} and \textit{Altitude.E} were kept out of the report because they were either not showing any relevant or interesting information or they were not translatable into a proper visual format.

Then, the figure 5.2 presents the graphical statistical descriptive of the stations related information. Figures 5.2(a) and 5.2(b) respectively indicate the distribution of the altitude and the capacity (in date of the 2017 season) of the 541 stations recorded in the BIXI live data base. The capacity distribution seems to point out some clear steps. The stations being modular, the modules might come in defined amount of docks. Figures 5.2(c) and 5.2(f) give information about the location of the stations as if they are located in the city centre or within 250 metres from a metro station respectively. A relatively low proportion of stations are located in the downtown area in comparison with other North American cities where the BSS is mostly in the centres. Then, the remaining figures show proportions of trips related with the downtown area and the metro stations. The figures 5.2(d) and 5.2(e) indicate the percentage of trips respectively beginning and ending in the city centre and the results are also differentiate between members and casual users. Remembering that 16.4\% of the trips are taken by casual users, these two graphs suggest that the casual users hang out more in the city centre. The four remaining figures (5.2(g), 5.2(h), 5.2(i), 5.2(j))
TABLE 5.1: Descriptive Statistics of the trips divided in three sections.

<table>
<thead>
<tr>
<th>Variable</th>
<th>Mean</th>
<th>Standard deviation</th>
<th>Min.</th>
<th>1st Q</th>
<th>Median</th>
<th>3rd Q</th>
<th>Max.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Number of trips per Month (ave.)</td>
<td>437,889</td>
<td>203,218</td>
<td>128,825</td>
<td>292,790</td>
<td>543,538</td>
<td>568,974</td>
<td>630,873</td>
</tr>
<tr>
<td>Number of trips per Day (ave.)</td>
<td>16,294</td>
<td>5,529</td>
<td>4,091</td>
<td>11,841</td>
<td>17,672</td>
<td>21,200</td>
<td>25,541</td>
</tr>
<tr>
<td>Number of trips per Hour (ave.)</td>
<td>954</td>
<td>648</td>
<td>61</td>
<td>371</td>
<td>968</td>
<td>1,271</td>
<td>2,439</td>
</tr>
<tr>
<td>Trip duration (mm:ss)</td>
<td>15:49</td>
<td>248.51</td>
<td>00:00</td>
<td>06:03</td>
<td>10:46</td>
<td>18:17</td>
<td>(174 days)</td>
</tr>
<tr>
<td>Elevation (m)</td>
<td>-2.73</td>
<td>20.40</td>
<td>-112.32</td>
<td>-14.62</td>
<td>-1.40</td>
<td>9.35</td>
<td>106.11</td>
</tr>
</tbody>
</table>

give all information about the presence of a metro station nearby the starting or ending station or both. It is striking that the proportions of of members versus casual users is exactly the same for the four cases and that it is also exactly the same as the proportion of member and casual users trips in general. There is apparently no tendency from one or the other to cycle more from/to a metro station.

Finally, the figure 5.3 presents the statistical descriptive of the variables describing the weather during the trips. Figure 5.3(a) is a bar plot of the temperature during the trips and it exhibits the propensity of users to cycle more when the temperature is higher. Figure 5.3(b), 5.3(c) and 5.3(d) are also bar plots, but showing respectively the relative humidity, the speed of the wind and the atmospheric pressure during trips. The figure 5.3(e) show the proportions of the cardinal origins of the wind. It is consistent with the general values as the BIXI users don’t show a tendency to cycle more with a specific wind. Finally, the two last figures 5.3(f) and 5.3(g) are about the weather conditions during trips. Figure 5.3(f) show that the trips are majorly taken during cloudy or clear weather conditions. Figure 5.3(g) supports this fact by confirming that about 93% of the trips are executed during proper weather conditions, namely during "clear", "cloudy" and "foggy" weather.

The tables 5.1, 5.2 and 5.3 complement the graphs by presenting the statistical summaries of the numerical variables. First, table 5.1 describes the trips in space and time. It gives the summaries for the average number of trips per month. Let’s remember that two of the eight months are half months (April and November), hence skewing a bit the values (minimum, mean and standard deviation). It also gives the average number of trips per day and per discrete hour.

Table 5.2 shows the comparison between the statistics of the weather during the trips with the weather in general. Again, it indicates the propensity of users to cycle more during warmer temperature, less humid and slightly lower atmospheric pressure. However, there is no difference when looking at the wind speed.

Table 5.3 describes the stations capacities and altitudes. The stations with a capacity of zero are probably temporary stations that were not in use at the moment of the collect of data. Indeed, BIXI uses temporary stations to cope with over-demand of bikes and docks. For example, on a normal working day, some delimited zones are established by BIXI employees in the downtown area were the bike traffic is intense during the morning peak where BIXI users can just drop their bikes. Then, when the peak has passed, the employees go on with the redistribution. Technically, the highest (or lowest) elevation a user can ride between two stations is 112.32 meters.
Chapter 5. Applications: Presentation of the Results

(a) Stations’ Altitude
(b) Stations’ Capacity
(c) Stations Located Downtown
(d) Trips beginning downtown
(e) Trips ending downtown

(f) BIXI stations’ close to metro sta-
(g) Trips beginning from metro sta-
tion
(h) Trips ending at metro station
(i) Trips beginning OR ending at
metro station
(j) Trips beginning AND ending at
metro station

FIGURE 5.2: The caption being to big, please report to the footnote where the figure is mentioned for its detailed description.
Chapter 5. Applications: Presentation of the Results

(a) Temperature during trips
(b) Relative humidity during trips
(c) Wind speed during trips
(d) Atmospheric pressure during trips
(e) Wind direction during trips
(f) Weather conditions during trips
(g) Trips during proper weather conditions

Figure 5.3: The caption being too big, please report to the footnote where the figure is mentioned for its detailed description.
Table 5.2: Descriptive Statistics of the trips divided in three sections.

<table>
<thead>
<tr>
<th>Variable</th>
<th>Mean</th>
<th>Standard deviation</th>
<th>Min.</th>
<th>1st Q</th>
<th>Median</th>
<th>3rd Q</th>
<th>Max.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Trip Temperature (°C)</td>
<td>19.2</td>
<td>6.3</td>
<td>-4.7</td>
<td>15.5</td>
<td>20.4</td>
<td>23.8</td>
<td>31.7</td>
</tr>
<tr>
<td>Temperature (°C)</td>
<td>14.5</td>
<td>8.1</td>
<td>-9.8</td>
<td>8.4</td>
<td>15.8</td>
<td>20.9</td>
<td>31.7</td>
</tr>
<tr>
<td>Trip Rel. Humidity (%)</td>
<td>61</td>
<td>16</td>
<td>11</td>
<td>49</td>
<td>60</td>
<td>73</td>
<td>99</td>
</tr>
<tr>
<td>Rel. Humidity (%)</td>
<td>68</td>
<td>18</td>
<td>11</td>
<td>55</td>
<td>69</td>
<td>82</td>
<td>99</td>
</tr>
<tr>
<td>Trip Wind speed (km/h)</td>
<td>16</td>
<td>8</td>
<td>0</td>
<td>10</td>
<td>15</td>
<td>21</td>
<td>53</td>
</tr>
<tr>
<td>Wind speed (km/h)</td>
<td>16</td>
<td>8</td>
<td>0</td>
<td>9</td>
<td>14.0</td>
<td>21</td>
<td>53</td>
</tr>
<tr>
<td>Trip Atm. Pressure (kPa)</td>
<td>101.08</td>
<td>0.69</td>
<td>98.52</td>
<td>100.67</td>
<td>101.04</td>
<td>101.48</td>
<td>103.24</td>
</tr>
<tr>
<td>Atm. Pressure (kPa)</td>
<td>101.15</td>
<td>0.81</td>
<td>98.52</td>
<td>100.63</td>
<td>101.10</td>
<td>101.64</td>
<td>103.68</td>
</tr>
</tbody>
</table>

Table 5.3: Descriptive Statistics of the trips divided in three sections.

<table>
<thead>
<tr>
<th>Variable</th>
<th>Mean</th>
<th>Standard deviation</th>
<th>Min.</th>
<th>1st Q</th>
<th>Median</th>
<th>3rd Q</th>
<th>Max.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Capacity (docks)</td>
<td>32</td>
<td>18</td>
<td>0</td>
<td>19</td>
<td>31.0</td>
<td>44</td>
<td>66</td>
</tr>
<tr>
<td>Altitude (m)</td>
<td>59.19</td>
<td>19.71</td>
<td>14.83</td>
<td>45.14</td>
<td>57.80</td>
<td>71.85</td>
<td>127.15</td>
</tr>
</tbody>
</table>

Interesting Remarks

The trips duration statistical summary seems to show outliers since both its extreme values are way off reasonable values (trips of 0 seconds and 174 days). From the previous figures 5.1(f) and 5.1(j), it seems that trips lasting more than one hour are exceptional. About the lower outliers, a quick observation of the shortest trip demonstrated that many trips in the database present a duration of less than 10 seconds. To visually express it, two histograms were made zooming on the trips with duration of less than two minutes. These histograms are presented in figure 5.4. Our first hypothesis was that the lower outliers were either due to system bugs or a user that checked out a bike that was defective (flat, broken pedals, etc.) and put it back right away. Therefore, these trips should appear as looping trips. Figure 5.4(a) highlights the presence of an outlying class of trips below the 45 seconds mark. According to this figure, the shortest real trips should be around 40 seconds; point where the fat tail distribution begins. Then, figure 5.4(b) confirm the common feature of the lower outliers. As expected, they are all loops. All the trips below 30 seconds are loops which confirm the supposition that they must be bikes that users put directly back in the dock after they took notice of some defects.

An inconsistency in the figure 5.1(k) might catch the attention. Indeed, some days encounter a significantly low amount of trips. One of the reason can be because of bad weather, consequently leaving the users reluctant to cycle. To test this hypothesis, another bar plot was made while highlighting the number of trips made under proper weather conditions as show figure 5.5. As a matter of fact, a correlation is observed between fewer trips and days marked by bad weather (rain or storms).

5.1.2 Visualization of BIXI within the city

To better understand the environment and the geographical context into which BIXI operates, a presentation of Montreal City, its neighbourhoods and its topography is essential. The presence of a big hill (the Royal Mount) in the middle of the city influences the cycling behaviour of the BIXI users (see the topographical map at figure 5.6(a)). From figure 5.6(b), the most important neighbourhoods to point out...
are the Plateau Mont-Royal which hold the highest density of BIXI stations and of cyclists in the city, and Ville-Marie which contains the downtown area. Sherbrooke street which is mentioned a couple of times delimits the border along the North-South axis between these two neighbourhoods and goes along the edge of a plateau about 10 to 20 meters high.

The figure 5.7 shows the location of all the active stations and their capacity as of 2017. As well, on the map are identified the natural areas such as parks, wooded zones and water elements and the existing cycle paths. Therefore, this map presents the relations between the stations with the built cycling infrastructure and the natural environment.

The figure 5.8 shows the location of all the active stations and their capacity as of 2017 in regard with the population and the metro stations. The population density was simply calculated using the total population of each district divided by its total area. Hence, the natural areas are also there to nuance the fact that some districts might present a low population density due to the presence of major green or blue areas which increase the area for the calculation. Nonetheless, the BIXI stations are still mostly located in the most densely populated area (Plateau Mont-Royal) and
Chapter 5. Applications: Presentation of the Results

Figure 5.6: Topographic and urban features of Montreal.

(a) Topographic map of Montreal

(b) Map of Montreal’s neighbourhoods with the main and secondary streets.
Chapter 5. Applications: Presentation of the Results

Figure 5.7: Map representing the locations of the stations in relation with the natural environment.
in the city centre (Ville-Marie). Moreover, this map shows how almost all metro stations have a BIXI station within a 250 meters buffer zone and that these BIXI stations have a greater capacity.

The figure 5.9 shows the location of all the active BIXI stations and their capacity as of 2017 in regard with the urban usage and infrastructure. It is important to notice how densely served is the commercial zone (magenta) and the skyscrapers (light green) in the downtown area. As well, most of the schools and universities (light blue) are well deserved.

5.1.3 Dynamic study of Usage and Users’ Behaviour

Both figures 5.10(a) and 5.10(b) present the annual frequency of use of all stations as starting points or destination points. This way, we can see the zones where the service is more used and maybe understand the motivations for using it. On the other hand, it might be more interesting to compare both graphs and focus the attention on the stations that vary, i.e. that are either more popular as an origin than a destination or vice versa, because these stations are those that create a greater need for redistribution of bikes. To do so in a general aspect, the "balance" throughout a year between the number of incoming and outgoing bikes for each station has been calculated simply by deducting figure 5.10(a) to figure 5.10(b) to make figure 5.11. In other words, to each station were added the total number of bikes arriving and subtracted the number of bikes leaving during the year. Consequently, figure 5.11 illustrates well how the bottom of the plateau delimited by Sherbrooke street (both neighbourhoods Ville-Marie and The Sud-Ouest) acts as an attracting zone and the neighbourhood Plateau Mont-Royal as a repulsing one.

Based on the information gathered from the two previous figures, a GIF image was created to dynamically illustrate the average working days imbalance of each station every hour. For obvious reasons, the GIF is not shown in this report. Nonetheless, four characteristic hours are presented at figure 5.12. A station that has as many arrivals than departures during a specific hour would appear as beige. An attracting station (green or blue) is a station that receives more bikes than it sends during a specific hour, hence becoming an attractor of bikes. An extreme attractor (in blue) would be a station suffering of an over-demand of docks and an over-offer of bikes (a full station). A repulsing station (orange or red) is the opposite. An extreme repulsor (in red) would be a station suffering an over-demand of bikes and an over-offer of docks (an empty station).

Accordingly, at 8:00 in the morning (see top left map in figure 5.12), the stations located in the residential areas are all orange and red since everybody is going to work or to school. Hence, a concentration of blue stations in the city centre at the same time since a lot of people work there and four universities are also in this area. At 12:00, during lunch (see top right map in figure 5.12), the city is more stable but there is still some activity going on around the city centre. During the afternoon peak, at 17:00 (see bottom left map in figure 5.12), the opposite phenomenon from the morning is observed. Everybody leaves the commercial and job zones to return home. Then, at 3:00 in the morning (see bottom right map in figure 5.12), the city is asleep except for 9 stations (2 green and 7 orange). From the 7 orange stations (higher demand of bikes), 5 are located where the famous bars are (2 on Mont-Royal street and 3 on lower Saint-Denis street) and, because the bars are closing at 3:00, this higher demand is probably coming from users that go out on Thursday nights.

Interestingly, the four stations located close to the Montreal University campus are repulsors from 11:00 until 0:00. Considering that these stations are among the
FIGURE 5.8: Locations of the stations in relation with the population density and the proximity with metro stations.
Figure 5.9: Locations of the stations in relation with the infrastructure and the urban use of the environment.
Chapter 5. Applications: Presentation of the Results

(a) As origin station  
(b) As destination station

**Figure 5.10:** Popularity of each station during the year 2015.

**Figure 5.11:** Difference between incoming and outgoing trips for each station along 2015.
highest ones in altitude, it is probably due to the mountain effect where the users
are willing to go down the hill, but reluctant to go up it, hence creating repulsing
stations. It is easier to observe this effect on the actual GIF than on figure 5.12.

The figure 5.13 displays a rendering of the most popular trips taken between
stations. The darker is a line connecting two stations, the more often the trip occurs.
Note that the connections are directed, but the lines on the map are not. Most of the
most common trips are being part of multimodal transportation means and solving
the last mile problem since many darker lines are connected to metro stations. These
trips are meant to connect homes to the closest metro station that was too far to walk
as it is the case with all the area east of Mont-Royal metro station. The utility of this
map is to provide knowledge about the location of the origins and destinations of
the users as well as the potential routes they might follow, thus giving decisional
information about road maintenance.

5.2 A Complex Network Approach

The degree distribution calculated with tnet is represented under various forms in
figure 5.14. The distribution does not seem to follow a power law like most networks
usually are. Indeed, it almost looks like a normal distribution, which is rarely seen
in real networks but more common in theoretical ones. A deeper analysis is needed
to confirm this hypothesis.

The figures 5.15(a) and 5.15(b) illustrate how the tnet algorithm varies according
to the tuning parameter \( \alpha \). For the figure 5.15(a), a value of 0 was used, whereas a
value of 1 was used for figure 5.15(b). Note that the degree is differentiated between
in and out degrees. The 10 stations with the highest values for the in-degree, out-
degree, closeness centrality and betweenness centrality are shown on the following
maps in figure 5.15.

What struck at first is how the results don’t vary much with the \( \alpha \) and that an \( \alpha \)
of 0.5 is not literally between the results of \( \alpha = 0 \) and \( \alpha = 1 \). However, it is worth
noting that when the weight is considered (\( \alpha = 1 \)), the stations with the highest
closeness centrality form a line along the Mont-Royal street. It makes sense accord-
ing to the popularity and the diversity of the famous neighbourhood Plateau Mont-
Royal of which Mont-Royal street is the spinal core. It is common knowledge that
people from all over the city go to this street for whatever reasons: work, shopping,
restaurants, bars, sociocultural activities and else. The reasons related with leisure
and sociocultural features are probably what increase the centrality of these stations
over the downtown ones which is a more business related area only. They are also
more geographically central in the city, again compared to the downtown area, thus
more accessible from anywhere on the island.

Then, as proposed by Opsahl, Agneessens, and Skvoretz (2010), when \( \alpha = 0.5 \),
the results are relatively different from the results with values of 0 and 1; in the sense
that there seems to be more similarities between the results of figure 5.15(a) and
figure 5.15(b) than any of those with figure 5.15(c). Nonetheless, the differences are
negligible if we consider that the stations represented are still in the same zones of
the city (the neighbourhoods of Plateau Mont-Royal and Ville-Marie). If the network
was represented as geographical patches instead of individual stations, probably
that the results would be very more similar.

Indeed, the relative consistency between the results for different \( \alpha \) is partly due
to the amount of trips used. In other words, there were so many trips between
many stations that the weights are very large. Let’s note that the directed network
Figure 5.12: The stations’ balance is presented here for four specific hours. The top left image is the morning peak at 8:00. The top right is a calmer and balanced moment during lunch time at 12:00. The bottom left image is the afternoon peak, when workers and students are transiting back home at 17:00. The bottom right image is when the city is asleep and some people are leaving the bars at 3:00.
Figure 5.13: Map of the most frequent trips between 2 stations. This map was generated using R, hence the inferior quality of the image. Nonetheless, the scale is more or less the same as for the previous maps.

Figure 5.14: The degree distribution for an $\alpha = 0.5$ using tnet.
Chapter 5. Applications: Presentation of the Results

Figure 5.15: The top 10 stations for the four principal metrics of the Opsahl’s method using tnet.

(a) $\alpha = 0$

(b) $\alpha = 1$

(c) $\alpha = 0.5$
is 68% complete which means that 68% of the possible directed trips between all the stations have been realized. Also, as the bottom-left graph shows in figure 5.14, there is about 96% of probability of having a degree of at least 400 in this network, which is quite high for a 460 nodes network. Therefore, since the measures are calculated dividing by the weights, the values get very small and insignificant at some point, hence small differences between the popular and central stations.

As future lines of study, it would be interesting to pursue this complex network approach. The principal challenge of weighted networks that include flux of some kind is to take the dynamics into account with capacities and loads involved. Therefore, based on the works of Motter and Lai (2002), one could carry out the dynamic analysis and study, for example, the resilience of the system and its reaction when some nodes are removed. As well, it would be interesting and more relevant for decision making to include the spatial location of the station into the network study.

### 5.3 Clustering

After many trials and errors, the most stable and resilient combination of original BIXI variables was achieved. This set of clusters will be referred to as H1. Then were added the contextual information to study their impact. The resulting set of clusters using "all" variables is referred to as H6. Finally, for the multi-view clustering, a third set of clusters using only the contextual variables was made and will be referred to as H11 and the set of clusters emerging from the combination of H1 and H11 is referred to as H12. The results of the clustering process for H1, H6 and H11 are presented below. Beforehand, the statistical summary of their respective dissimilarity matrices is presented in table 5.4.

<table>
<thead>
<tr>
<th>Clustering</th>
<th>Mean</th>
<th>Standard deviation</th>
<th>Min.</th>
<th>1st Q</th>
<th>Median</th>
<th>3rd Q</th>
<th>Max.</th>
</tr>
</thead>
<tbody>
<tr>
<td>H1</td>
<td>0.3167</td>
<td>0.0935</td>
<td>0.0005</td>
<td>0.2465</td>
<td>0.3105</td>
<td>0.3824</td>
<td>0.7289</td>
</tr>
<tr>
<td>H6</td>
<td>0.3507</td>
<td>0.0641</td>
<td>0.0267</td>
<td>0.3078</td>
<td>0.3520</td>
<td>0.3950</td>
<td>0.6342</td>
</tr>
<tr>
<td>H11</td>
<td>0.3704</td>
<td>0.0845</td>
<td>0.0000</td>
<td>0.3141</td>
<td>0.3743</td>
<td>0.4308</td>
<td>0.6913</td>
</tr>
</tbody>
</table>

#### 5.3.1 Clustering of the Original BIXI Data Variables: H1

Let’s recall the variables used to calculate the dissimilarity matrix: *Day, Weekend, Nightout, Period*, the four variables of the geolocation, *Total.Duration* (trip duration), *Loop* and *Account.Type* (user type). From there, the Ward method produced the dendrogram of figures 5.16.

**Dendrogram of H1**

The structure of the tree was studied and the Calinski-Harabatz index optimized to find the most suitable number of classes. Instinctively, the dendrogram 5.16(a) allows a fair cut into six clusters. The Calinski-Harabatz index is maximized upon the dendrogram to determine the number of classes on the basis of what is suggested by the structure of analyzed data itself, rather than a priori supposing a blind number of classes as occurs in other clustering methods. In our case, the objective is to find
the number of cluster closer to eight that has the highest reduction of inertia. Eight because the goal is to segregate the data into classes with limited but clear similar information and eight variables are considered for this clustering. Therefore, from the inertia plot of figure 5.16(b) the blue circle presents a good option. Even though it does not show the biggest drop of inertia, relatively to its number of classes it still is a fair reduction and it confirms the cut into six classes. The other coloured circles indicate other potential good cuts. Finally, figure 5.16(c) exhibits the result of the cutting into six classes.

Profiling and interpretation of the classes

Once the cut is made and the objects are grouped into their classes, a study of the profile of the clusters is conducted to interpret each class. The graphs are gathered into figure 5.17 highlight the profile of each cluster in function of the variables used to do the clustering.

According to figure 5.17(a), classes 2, 3 and 5 have significantly longer trips and class 4 has slightly shorter trips. Figure 5.17(b) indicates that class 4 is only composed of weekend trips and classes 5 and 6 have a majority of weekend trips. From figure 5.17(f), one can conclude that class 6 has only and all the Nightout trips. Figure 5.17(c) shows that almost all looping trips are in class 2 and some in class 6. Figure 5.17(d) demonstrates that casual users are split between classes 3 and 5, but also that some are in classes 2 and 6. Finally, from figure 5.17(e) class 1 has a predominance of morning trips and class 6, as expected, has only evening and night trips. The other classes respect the general tendencies with a peak in the afternoon. Since the variable Day was not affecting significantly the profiles of the classes, it was left out for reasons of concision. Interestingly, the geolocation of the origin and destination stations do not have any impact in the clustering. Therefore, the clustering is not geographically constrained. This is probably due to the data structure that indicates trips with an origin and a destination, hence unable to properly cluster the stations geographically. Table 5.5 shows the characteristic profiles of each cluster and the classes are given a name for further reference.
Chapter 5. Applications: Presentation of the Results

Figure 5.17: Interpretation of H1.
Chapter 5. Applications: Presentation of the Results

### Table 5.5: Characterization of H1

<table>
<thead>
<tr>
<th>Cluster</th>
<th>Number of objects</th>
<th>%</th>
<th>Name</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>3056</td>
<td>61.1</td>
<td>Mundane</td>
<td>Contains the majority of the trips. Contains most of the morning trips. Typical trips by members on working days.</td>
</tr>
<tr>
<td>2</td>
<td>156</td>
<td>3.1</td>
<td>Loop</td>
<td>Contains all the looping trips.</td>
</tr>
<tr>
<td>3</td>
<td>397</td>
<td>7.9</td>
<td>Week Casual</td>
<td>Contains all the trips by casual users done during working days.</td>
</tr>
<tr>
<td>4</td>
<td>721</td>
<td>14.4</td>
<td>WE Member</td>
<td>Contains all the trips taken by members on weekend.</td>
</tr>
<tr>
<td>5</td>
<td>318</td>
<td>6.4</td>
<td>WE Casual</td>
<td>Contains all the trips taken by casual users on weekend.</td>
</tr>
<tr>
<td>6</td>
<td>352</td>
<td>7.0</td>
<td>Nightout</td>
<td>Contains all the trips on Friday and Saturday nights between 20:00 and 5:00.</td>
</tr>
</tbody>
</table>

### 5.3.2 Clustering of All the Variables Including the External Information: H6

Once the H1 classes are defined, let's study how they are affected by the addition of external variables describing the context of the trips. The variables added to those of H1 to calculate the dissimilarity matrix are: $\text{Capacity.E}^1$, $\text{Capacity.S}$, $\text{Metro.station.S}$, $\text{Metro.station.E}$, $\text{Downtown.S}$, $\text{Downtown.E}$, $\text{Altitude.S}$, $\text{Altitude.E}$, $\text{Elevation}$, $\text{Temperature}$, $\text{Rel.Humidity}$, $\text{Wind.speed}$, $\text{Wind.cardinal}$, $\text{Atm.pressure}$ and $\text{Weather.main}$. From there, the Ward method produced the dendrogram of figures 5.18.

#### Dendrogram

This dendrogram (figure 5.18(a)) is less instinctive to cut. However, a visual optimization of the Calinski-Harabatz index leads to a cut for seven clusters. From the plot of inertia, it can appear more clearer. Again, the blue circle in the figure 5.16(b) indicates the best option to have the highest number of cluster that has the highest reduction of inertia while having the longest branches in the dendrogram. Finally, figure 5.16(c) exhibits the result of the cutting into seven classes.

#### Profiling and interpretation of the classes

Once the cut is made and the objects are grouped into their classes, a study of the profile of the clusters is conducted to interpret each class. The graphs are gathered into figure 5.19 highlight the profile of each cluster in function of the variables used to do the clustering.

According to figure 5.19(a), class 2 has longer trips. Figure 5.19(b) indicates that class 6 is only composed of weekend trips, class 7 has a majority of them, classes 3 and 4 have some, and 1 and 5 have very few. From figure 5.19(c), class 7 is primarily composed of Nightout trips, but class 2 also has some. From figure 5.19(d) class 1 has a predominance of morning trips and class 7, as expected, has only evening and night trips. Class 2 has also a significant proportion of evening and night trips.

---

$^1$Note: the suffixes "S" and "E" to the variables mean "of starting station" and "of ending station" respectively.
Class 5 have a noteworthy amount of afternoon trips. The other classes respect the general tendencies of period distribution. As oppose to H1, figure 5.19(e) shows that none of the classes are defined by looping trips. Figure 5.19(f) demonstrates that casual users are mostly in class 3, but there are also few of them in classes 4, 6 and 7. Figure 5.19(g) does not say much, but it serves as a comparison point with the next clustering: H11. Similarly, figure 5.19(h) is also for comparison purpose, but it does not say much apart that the “no wind” trips are in class 7, that class 2 has more North wind and that class 6 has more wind for the East. Note that this figure shows the proportion from all the modality X contained in the class A. The rest is dominated by class 5 because it has more trips in its group. From figures 5.19(i) and 5.19(j), class 4 is strongly characterized by trips in the downtown area and class 1 by trips mostly ending in the downtown area but starting from other places in the city. Figure 5.19(k) indicates that class 4 is only one with a positive elevation, class 1 has trips strongly descending and class 2 has trips more moderately descending. Finally, figure 5.19(l) shows that trips under clear weather are mostly in class 2, those under cloudy weather mostly in class 5 and those under bad weather (rain, storm, snow, etc) are in class 1. The other classes have both mostly clear and cloudy weather. Some variables used for the clustering were left out because they were not relevant for the interpretation or the comparison with H1 or H11. Table 5.6 shows the characteristic profiles of each cluster and the classes are given a name for further reference.

### 5.3.3 Clustering of the External Variables Only: H11

Finally, this last trial was motivated by an intuition to make a multi-view cluster combining a clustering made only using original BIXI data (H1) and a clustering applied only over the contextual variables because the H6 was not enlightening and conclusive enough to carry on. Therefore, the H11 dissimilarity matrix was calculated using only these variables: Capacity.E, Capacity.S, Metro.station.S, Metro.station.E, Downtown.S, Downtown.E, Altitude.S, Altitude.E, Elevation, Temperature, Rel.Humidity, Wind.speed, Wind.cardinal, Atm.pressure and Weather.main. From there, the Ward method produced the dendrogram of figures 5.20.
Chapter 5. Applications: Presentation of the Results

Figure 5.19: Interpretation of H6.
TABLE 5.6: Characterization of H6

<table>
<thead>
<tr>
<th>Cluster</th>
<th>Number of objects</th>
<th>%</th>
<th>Name</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>743</td>
<td>14.9</td>
<td>Bad Morning</td>
<td>Contains the morning trips of members. Contains trips that end in the downtown area. Contains trips that go down (negative elevation). Contains trips during bad weather.</td>
</tr>
<tr>
<td>2</td>
<td>883</td>
<td>17.7</td>
<td>Clear Nights</td>
<td>Contains trips of members during clear weather. Contains night and evening trips, mainly on working days.</td>
</tr>
<tr>
<td>3</td>
<td>529</td>
<td>10.6</td>
<td>Casual</td>
<td>Contains trips by casual users. Contains long duration trips. Contains trips during cloudy weather.</td>
</tr>
<tr>
<td>4</td>
<td>614</td>
<td>12.3</td>
<td>Clear Up</td>
<td>Contains trips of members that go up. Contains trips during clear weather.</td>
</tr>
<tr>
<td>5</td>
<td>1313</td>
<td>26.3</td>
<td>Loop</td>
<td>Contains looping trips of members. Contains trips during cloudy weather.</td>
</tr>
<tr>
<td>6</td>
<td>650</td>
<td>13.0</td>
<td>WE</td>
<td>Contains trips of members during weekend.</td>
</tr>
<tr>
<td>7</td>
<td>268</td>
<td>5.4</td>
<td>Nightout</td>
<td>Contains the Nightout trips of both members and casual users.</td>
</tr>
</tbody>
</table>

**Dendrogram**

Again, the dendrogram 5.20(a) is not so obvious to cut just by looking at it. However, it seems way more clear from the plot of inertia in figure 5.16(b). There are two good options (red and blue circles). Again, the blue circle is preferred for its more numerous classes. Finally, figure 5.16(c) exhibits the result of the cutting into four classes.

![Dendrogram](image1)

![Inertia](image2)

![Cut into 4 classes](image3)

**Figure 5.20**: Dendrogram analysis and cut for clustering H11.
Table 5.7: Characterization of H11

<table>
<thead>
<tr>
<th>Cluster</th>
<th>Number of objects</th>
<th>%</th>
<th>Name</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1476</td>
<td>29.5</td>
<td>Clear</td>
<td>Contains all the trips on clear weather.</td>
</tr>
<tr>
<td>2</td>
<td>709</td>
<td>14.2</td>
<td>DTe Cloudy</td>
<td>Contains trips that go down and end in the downtown area. Trips during cloudy weather.</td>
</tr>
<tr>
<td>3</td>
<td>1669</td>
<td>33.4</td>
<td>Flat Cloudy</td>
<td>Contains trips with a null elevation. Trips during cloudy weather. Trips during East and Northwest winds.</td>
</tr>
<tr>
<td>4</td>
<td>1146</td>
<td>22.9</td>
<td>Bad Weather</td>
<td>Contains all trips during bad weather. Contains a lot of trips ending close to a metro station.</td>
</tr>
</tbody>
</table>

Profiling and interpretation of the classes

Once the cut is made and the objects are grouped into their classes, a study of the profile of the clusters is conducted to interpret each class. The graphs are gathered into figure 5.21 to highlight the profile of each cluster in function of the variables used to do the clustering.

Like for the two other cases, some variables used for the clustering were left out because they were not relevant for the interpretation or the comparison with H1 or H11. Nonetheless, some profiles stand out. Figure 5.21(b) presents the class 4 with a lot of trips ending close to a metro station and some for class 2. Class 2 is strongly characterized by trips ending in the downtown area. Actually, this is also linked to trips ending at high capacities and at lower altitudes, but these were not significant enough to be shown; it is only pointed out to show how some variables are correlated. As well, this include in class 2 trips that are going down (that have a relatively negative elevation), as exposed by figure 5.21(e), since the downtown area is lower in altitude. According to the same figure, class 3 has “flat” trips, i.e. trips with zero elevation. Figure 5.21(f) presents strong profiles as class one has almost all and only trips during clear weather; classes 2 and 3 are mostly composed of trips during cloudy weather; and class 4 has almost all the trips during bad weather (rain, showers, storms, snow, etc.). Finally, figure 5.21(c) shows that all the trips when there is no wind are in class 1 as well as most of the dominant winds from West and Southwest. The class 3 has most of the less common winds (East, Northeast, North and Northwest) while the two other classes are not characterized by anything special. As recalled before, this figure presents the proportion from all the modality within a class. Therefore, class 3 contains almost 80% of all the trips with East wind. Table 5.7 shows the characteristic profiles of each cluster and the classes are given a name for further reference.

5.3.4 Multi-view Clustering

The results of the composition of the classes are numerically expressed as a bivariate distribution in table 5.8 where the rows are the H1 classes and the columns are the H11 classes. This matrix indicates the number of objects within each cross class.

At first sight, the proportions of the H11 classes seem to be preserved within each H1 class. To verify this statement, a bar plot was made to show the proportions of each H11 classes into the H1 classes as presented in figure 5.22. Even though the
Chapter 5. Applications: Presentation of the Results

(a) H11: Origin is downtown

(b) H11: Metro at destination

(c) H11: Wind direction

(d) H11: Destination is downtown

(e) H11: Elevation

(f) H11: Weather conditions

FIGURE 5.21: Interpretation of H11.
### Table 5.8: Number of objects from H11 clusters into the H1 clusters.

<table>
<thead>
<tr>
<th>Class</th>
<th>Clear</th>
<th>DTe Cloudy</th>
<th>Flat Cloudy</th>
<th>Bad Weather</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mundane</td>
<td>894</td>
<td>462</td>
<td>979</td>
<td>721</td>
</tr>
<tr>
<td>Loop</td>
<td>44</td>
<td>14</td>
<td>65</td>
<td>33</td>
</tr>
<tr>
<td>Week Casual</td>
<td>127</td>
<td>63</td>
<td>130</td>
<td>77</td>
</tr>
<tr>
<td>WE Member</td>
<td>217</td>
<td>82</td>
<td>268</td>
<td>154</td>
</tr>
<tr>
<td>WE Casual</td>
<td>81</td>
<td>57</td>
<td>119</td>
<td>61</td>
</tr>
<tr>
<td>Nightout</td>
<td>113</td>
<td>31</td>
<td>108</td>
<td>100</td>
</tr>
</tbody>
</table>

Proportions are close to be kept, there are some standouts. Both the Mundane and the WE Casual classes have the same proportions.

**Figure 5.22:** Multi-view clustering using H1 and H11. The labels of clusters’ names were shortened to their initials for graph convenience.

Experts with good knowledge of the city of Montreal helped to explain the resulting profiles. The Loop class has more Flat Cloudy and less DTe Cloudy trips. This is because most looping trips are done by tourists in the touristic area that are the Old Montreal and the Old Port, and the Sainte-Hélène Island which are all out of the downtown area. Both are quite low (almost sea level) and don’t have any elevation around. More obviously, according to how the elevation is calculated, all looping trips have a null elevation since they end exactly where they started. More trips during clear weather would have been expected for the looping trips, but, on the other hand, tourists that decide to come to visit Montreal must not really care about cloudy weather. Anyways, looping trips are so not frequent that these percentages can change quickly if very few more observations are added. This whole Loop class is negligible.

The class WE Member is quite close to original proportions, but still show slightly less DTe Cloudy and more Flat Cloudy trips. This is in line with the habits of the inhabitants (members) to be more tempted to go spend a weekend day in a park or neighbourhood nearby than go downtown (usually where work is), hence the trips with no elevations.

The WE Casual class contains more Flat Cloudy and DTe Cloudy trips. One would suppose that these are tourists going to visit the downtown area or the
Old Montreal and lodging from either the city centre (hence the flat trips) or in the Plateau Mont-Royal neighbourhood which is famous for renting touristic apartments. The Flat Cloudy trips can also be generated by tourists in the Plateau Mont-Royal visiting the surroundings. At last, it is expected that there are less trips when the weather is bad as the casual users would prefer other means of transportation or other activities during their weekend.

The Nightout class shows quite more Bad Weather and less DTe Cloudy trips. These results are harder to interpret. Normally, people would not cycle under bad weather if they are on their way to go out. Unless it is on their way back and users do not care anymore about the weather and they prefer a quick way home than waiting for night buses or pay a taxi. Then, it is strange that users are less tempted to go out downtown since there are a bunch of famous bars as well as the renowned Sainte-Catherine street. However, maybe the people that do use BIXI when they go out are those that prefer their local bars nearby their home and those that go downtown favour the means of public transportation, hence reducing the share of DTe Cloudy. Note that experts with

5.4 Local Analysis

To follow through with the local analysis, the resulting sets of clusters H12 from the integrated multi-view clustering is preferred over the H6 set of clusters. First, the H1 showed to be more stable since the same results were obtain over and over with many different samples with always a very consistent dendrogram. As well, it is way more scalable than H6 since the same clusters were observed again when looking only at one month or one week. The H6 set of clusters had to adapt to the varying scales. In conclusion, H1 represents more the interactions between the variables instead of the weight of every single variable, like in H6, which disperses the values within the clusters.

All these characteristics from H1 are transferred to H12 because the multi-view clustering does not affect the intrinsic properties of the independent clusters, but combine them using their classes. This is why the crossing of H1 and H11 gives a final set of clusters that are more homogeneous. The classes are clearly define and avoid faint tendencies.

The local analysis consists in looking at every cluster independently to analyze local interactions between variables inside each specific profile. Once isolated, a case of study can provide useful information and guidelines for decision-making and the management in function of the patterns brought to light. In this thesis, the local analysis are limited some specific clusters coming from the multi-view clustering method; the goal being to show the kind of knowledge that can be withdrawn from such an analysis. Then, visualizations of the dynamical balancing factor inside the clusters are made.

In this section, two cases of study are analyzed. These two cases are the two most frequent classes (containing more objects) that are not sharing any other class from the cross-clustering. The first case is the Mundane trips (from H1) during Flat Cloudy (from H11). The second one is the case crossing the Weekend Member trips (from H1) with Clear (H11).
5.4.1 Local Analysis of the Mundane & Flat Cloudy Trips

Let’s first recall what composes this class. On one hand are the most common trips namely trips on working days by members. Therefore, it also includes most of the morning trips. On the other hand, the contextual class indicates flat trips (no elevation) during cloudy weather with East and Northwest winds. Interestingly, winds from the East and the Northwest are not the most common, which clashes with the rest of the information. However, the class Flat Cloudy is characterized by the fact that it contains most of the East and Northwest winds, but it does not mean that it only deals with those winds. Indeed, it still has about 20% of the trips during Southwest and West wind which, in absolute, might be more than all the East and Northwest winds trips it has.

We are dealing here with a class that is quite classical. One could expect from the dynamic analysis to see the patterns where people move from the residential area to the commercial one in the morning and back in the afternoon. Also, there might be a lot of use as multimodal transportation where users go to or from a metro station.

First, the results from the sheer annual balance of the stations are quite interesting. On figure 5.23, four maps are presented. The top ones indicate with a colour scale the total number of starting and ending trips per station throughout the whole season. The bottom left map combines the two top ones and provides the annual balance of each station. The bottom right map displays the results from the complex network analysis. Note that the colour scale for the starting trips figure is inverted to be consistent with the two other related figures.

The results here are surprising because it does not support the hypothesis. The movement observed is not from the residential area to the commercial one. In fact, the movement is mostly concentrated into the neighbourhood Plateau Mont-Royal. It seems to have the busiest stations for departures and arrivals. This is consistent with the flat factor (no elevation). It might also suggest the development of the job sector in the neighbourhood; people do not have to travel until the skyscrapers in the city centre for work anymore.

The bottom left map of the annual balance highlights a different information. It points out the translation of the heart of the city from Ville-Marie towards the Plateau Mont-Royal. Throughout the year, more users seem to go from Ville-Marie to the border between the two neighbourhoods and from the outskirts of Plateau Mont-Royal into it.

Finally, the complex network analysis provides another proof to support the previous statement by locating the most central stations the heart of the Plateau Mont-Royal, along the streets Saint-Denis and Mont-Royal. Because of the few amount of trips, many stations have the same centrality measures. For that reason, the number of dots for each metric is not necessarily the same. The top eleven stations are represented for the in-degree and the out-degree, nine for the closeness and ten for the betweenness.

To complete the spatial analysis, a map of the all the trips was drawn and colour-scaled to highlight the frequency of the trips. Unfortunately, the small size of the sample does not allow to observe a meaningful gradient of colour as the most frequent trips have been realized twice. Still, they are mostly located in the Plateau Mont-Royal neighbourhood and along its border with Ville-Marie.

Now that the spatial context of the trips is better understood, let’s have a look at when the trips occur more precisely. The total number of trips of the class being too low for a dynamic analysis showing every hour, the balances were calculated over the periods of the day instead. Figure 5.26 stresses which stations are attractors and
Figure 5.23: Balance Analysis of the Mundane & Flat Cloudy Trips
Figure 5.24: Most Frequent Mundane & Flat Cloudy Trips

Figure 5.25: Degree Distribution of the Mundane & Flat Cloudy Network
which ones are repulsors during four periods of the day: morning, noon, afternoon and evening. The night period was left out for irrelevancy due to a lack of data.

Note that for the local analysis, the balance of trips for each station has not been divided by the amount of working days to avoid having to deal with very small fraction, thus making it quite hard to differentiate the colour-scale. It is also due to the fact that we are still only dealing with a small sample of the whole database. As well, the trips are not limited only to working days since the classes already do some “filtering” and some do not have any working days trips at all or night trips. These adjustments are valid for all the subsequent analyzed cases.

There is an interesting dynamic going on along the four images. First, the strong repulsors (in red) move from the Parc Lafontaine (the green area surrounded by red dots in the morning map) in the morning slowly towards the downtown area in the afternoon. There is also a high activity going on along the Saint-Laurent and Saint-Denis streets in the afternoon and the evening. Secondly, the strong attractors (in blue) are more dispersed in general, but there number increase in the residential area in the afternoon and evening. It is also interesting to remark that there are well-balanced stations in the morning and the afternoon and that they are mostly located on the opposite side of the Plateau Mont-Royal, along the border with Rosemont-La Petite Patrie neighbourhood.

These images are quite powerful in terms of operational knowledge. They give clear hints about from/to where distribute the bikes in the network. Moreover, with a clustering well extended within the whole working data set, the results could be even more meaningful and defined by hour or even by every interval of ten minutes.

In conclusion, this local analysis proves that the most mundane and classical trips are mostly done in one neighbourhood: the Plateau Mont-Royal and that the repulsing stations are condensed in a very specific area. This gives important operational knowledge about where to bring extra bikes in specific periods of the day. Similarly, the same kind of analysis could be done for a specific periods like the week of festivities between Saint-Jean-Baptiste and the Canada Day or the weeks when school begins in September using the same methodology.

5.4.2 Local Analysis of the Weekend Member & Clear Trips

Let’s first recall what composes this class. On one hand, there is the second most common type of trips which are those from members during the weekend. These trips are usually done later in the day, like afternoon or evening, and they are shorter in duration than the average. On the other hand, the contextual class indicates clear weather. This class is also dominated by the most common winds from West and Southwest and some of the trips end up downtown.

Therefore, this case is made of trips done by members going out on weekend when the weather is nice. Another classical and normal case from which we could expect to see small and short trips around the residential area or towards parks and centres of activities during the afternoon during this dynamic analysis.

On figure 5.27, the same four maps of annual balance as previously explained are presented. The main difference being that instead of scaling on the quantiles, the Jenks natural break optimization (Jenks, 1967) implemented in QGIS was used to form the four ideal groups because the values are less spread. It is worth noting that all the stations with values of zeros have been removed from the map because they were too dominant and hiding the other stations.

The two top maps show again greater activity around the same area corresponding to the junction between Ville-Marie and Plateau Mont-Royal, but slightly more
Figure 5.26: Dynamic Analysis of the Mundane Trips During Flat Cloudy Context
towards downtown. Actually, the stations with highest activity are located along
the cycle path on Maisonneuve Boulevard, the only cycle path crossing Ville-Marie
parallel to the Saint-Lawrence River (on the Northeast-Southwest axis).

Nevertheless, in this case, it is the complex network analysis figure that tells more
significant information. It clearly lays out the main difference with the previous case
as the most central stations are, indeed, those along the Maisonneuve cycle path in
Ville-Marie. Fewer dots are shown because of the low amount of trips. The stations
with the highest degree have only 5 connections (2 stations) as indicates figure 5.29.
Therefore, the top five stations are mapped for the in-degree, four for the out-degree
and the closeness, and six for the betweenness. There is no distinction on the
\( \alpha \) since
it made no difference here. The results were exactly the same for \( \alpha = 0 \) and \( \alpha = 0.5 \).
This due to the fact that all edges have a weight of one or all stations are connected
only once. Consequently, the map of the trips is simpler than usual since no pairs
of stations have a weight greater than one. Thus, only single trips are drawn in
figure 5.28.

Once the spatial information about the behaviour of this class is set out, the dy-
namic of the trips can complete the analysis with more relevance. Since this class
is made of fewer objects, the rendering of the dynamic is less striking since fewer
trips are spread between as many stations. Generally, the network seems to be quite
balanced. Hence, these results shade the ones from figure 5.27 that were stressing a
high centrality of the downtown stations. However, even though it reaches its peak
of use in the afternoon as expected, there is no drastic hubs of attractors or repul-
sors that appear at any time. This also confirms some conclusions from the literature
review about the lower and more consistent use of the system during weekends.
Therefore, the operator can focus more on adapting its system to any special events
that could be going on during weekends as well as assuring an impeccable service
for the casual users who are more present on weekends.

In conclusion, on weekends, members use BIXI to go downtown on clear days,
but in a disperse manner in time which avoids over-demand situations. As a future
analysis, it would be interesting to see if these trips are related in anyway with the
several festivals and activities going on in Montreal from May until October. The
famous Quartier des Spectacles is located where most of the activity is taking place,
thus maybe a high proportions of the trips happen to be during the FrancoFolies, the
Montreal International Jazz Festival or the Just For Laugh Festival.

5.4.3 Local Analysis of the Nightout and the DTe Cloudy trips

To complete and conclude this local analysis section, two more cases were looked at.
The particularity of these cases is that they are not from the integrated multi-view
clustering H12, but from each independent set of cluster composing it, H1 and H11.
The interest was to see the convergence of the trips in the dynamic analysis for the
Nightout and the DTe Cloudy classes specifically. Therefore, these analysis are more
superficial and only look at the dynamic.

Since these local analysis are additional ones aiming to confirm the usefulness of
the method, the figures discussed are available in the appendix A.

First, the DTe Cloudy is totally in accordance with its description as all the trips
shown seem to be connected with the downtown area (see figure A.3(a)) . As well,
the map shows two levels of frequency meaning that some of the pairs of stations
have a weight of 2. On the other hand, the nightout map is more chaotic, but there
seems to be a hub appearing on the border between the Plateau Mont-Royal and
Figure 5.27: Balance analysis of the Weekend Member & Clear Trips
Figure 5.28: Most frequent Weekend Member & Clear trips

Figure 5.29: Degree Distribution of the Weekend Member & Clear Trips
Figure 5.30: Dynamic Analysis of the Weekend Member Trips During Clear Context
Ville-Marie about at the intersection with Saint-Denis street. This would make total sense since this street is highly frequented at night for its numerous bars.

Some interesting and curious features stand out from the dynamic analysis over the Nightout class, but do not lead to any conclusion (see figure A.3(c)). For example, it is quite curious to see a hub of four attractors in the evening surrounded by several orange dots (mellow repulsors). Then, on the night time frame, the repulsor located in Longueil (the red dot on the other side of the river) is also surprising since one would expect it to be blue as people are coming back from Montreal to Longueil at night, not the opposite. Nevertheless, there is also some operational knowledge as most of the Nightout repulsor are located along one street in Ville-Marie and the attractors are in Plateau Mont-Royal not too far.

Finally, the dynamic analysis of the DTe Cloudy class clearly exhibits what was expected: there is a densely populated hub of attractors located downtown (see figure A.4). However, what is worth noting is from where do the users come from to go downtown. Well, in the morning they principally come from the lower Plateau Mont-Royal (close to Ville-Marie); at noon the activity is way more calm; in the afternoon they all come from the Parc Avenue (the street going along the Mont-Royal along the Northwest-Southeast axis); and in the evening they also come from the Plateau Mont-Royal, but from a different area. In the morning, there is also a fair amount of users coming from the Côtes-des-Neiges - Notre-Dame-de-Grâce and Westmount neighbourhoods along the cycle path of Maisonneuve Ouest Boulevard.

This last analysis is less powerful in terms of decision-making since it does not provide any temporal information. It is limited to only geographical information, hence lacking an element for a truthful space-time knowledge.

5.5 Sources of errors

Notable sources of errors are the possible pre-processing of the data by BIXI before publishing it on the web page. For example, the "trips" generated by the redistribution of bicycles between stations are not included into the open database, but must have been present in the original one as Imani et al. (2014) indicated.

Interpretability criteria: Another source of errors is the subjectivity behind some analysis, for instance, for some preprocessing manipulations, for variable selection, for the cut of the dendrograms and the multi-view clustering analysis. Moreover, all these decisions are related and lead to one another. Therefore, if subjectivity generates some bias, the bias is increased at every step where another subjective decision is made.

Why we choose H12 over H6. Why is it better. H6 has dirty profiles while H12 has purer profiles. It allows a better combination between the primary variables and the external information. Some variables can standout more as influential. Discuss the dimensions taken into account (6 of H1 and 4 of H11). Develop a local analysis of a profile for better management.
Chapter 6

Conclusion

This final master thesis has presented a transdisciplinary approach to the modelisation of the public bike-sharing system of BIXI using mainly data science techniques along complex network theory and geographical visualization in order to extract decisional knowledge and help to manage the operations. This thesis hopes to contribute to the scientific state of the art by using a data mining approach methodology with the open data of the Montreal BIXI system. As well, it hopes to contribute to the characterization of the spatio-temporal patterns of BIXI as well as increasing knowledge about its users’ behaviour.

Analyzing BIXI data falls within the concept of sustainability by using a data mining approach in order to understand the BIXI network system and improves the urban mobility in Montreal.

The proposed model uses a data mining approach to process the open data from the BIXI bike-sharing system in Montreal in order to evaluate how data science techniques can provide added value to the company oriented to enhance the management of the service, in particular to avoid over-demand situations of available bikes or free docks. To do so, open data available about trips were exploited to identify patterns and discover relevant decisional knowledge about user’s behaviour.

Also, the thesis tackle the issue of using additional available information about the city or the context of the trips that can be found in open data sources and to evaluate the impact of enriching the company data with external information into the modelling.

Such an approach presents the advantage of make emerge behavioural patterns in relation with contextual information. Clustering allows the flexibility of not knowing beforehand the number or the types of classes. Using hierarchical clustering method, it also allows easier and more efficient results of pattern identification in comparison to other classical clustering methods. However, it is an expansive method to compute, hence the need to work with small samples from the whole database. Also, working with samples might bring its share of subjective bias. To compensate this effect, extensive experimentation using several samples have been carried out to ensure stability in the results.

Researches in the field of bike-sharing systems are recently booming and have a bright future considering the constant progress in the field of artificial intelligence and data science. In the past decades, there has been at first plenty of surveys and researches about the implementation process, the early adopters, the impacts of the BSS and the characteristics of its users. Then came more and more studies concerning the user’s behaviour until now where most projects are regarding predicting the demand. This sector is evolving at the speed of the technologies which makes it a promising and trendy subject. Within the next few years, the prediction models will get better and better until a certain threshold where it cannot bring anymore advantage and it faces the natural chaos like the weather. Then, the BSS will seek a
way to become less financially dependent from the government and maybe become self-sufficient if they succeed reducing sufficiently their operational costs with their models. Then, it would be a new era.

The shallow information emerging from the statistical descriptive already supported some statements from the literature review. There is a greater share of members than casual users using the service. Members use it more during working days and casual more during weekends. The members generate two peaks of intense use during the day. These peaks correspond to the transit hours when people go to and back from work namely at 8:00 and around 17:00. Thus, members use BIXI to transit to and from work. The weather is a factor having an impact on the use of the BIXI service. When the weather is not proper to cycle (e.g. when it rains), the usage is affected and reduced. Then, discoveries that emerged later in the project also confirmed that there are attracting and repulsing hubs distributed around the city according to the hours of the day. Most users go in the city centre in the morning for work and come back at the end of the labour day.

A simple complex network approach revealed underlying features of the network like the most central stations in terms of topological network. It identified which stations are the most popular (with the most arrivals of bikes), which ones are the most gregarious (with the most departures of bikes), which ones are the most topologically close to the other stations (with the lowest average shortest path), and the most topologically in between stations (which is the most transited by shortest path). This approach aims at simply give an overview of the potential of complex network theory as well as complement the information obtained from the data science approach. Important decisional knowledge could rise from pursuing the study, for example through a percolation experiment to study the resilience of the network and how users react to specific changes. In my knowledge, there has been no study of bike-sharing network or any public transportation network using the complex network theory, but it definitely hides very groundbreaking discoveries.

At first, the goal was to use scalable clustering methods such as the CURE methodology which allows to work with massive data bases of high dimensionality. To do so, the first step was to form clusters using samples. For complexity reasons, this project is limited to only working with samples and left the next phase of extending clustering to the remaining data for future lines of study. We focused on interpreting the clusters obtained from the sample. The stability of these results were evaluated using extensive experimentation using numerous samples.

Throughout the project, three analysis have been realized to study how are the patterns emerging were generating useful profiles of user’s behaviour. The first one was using only the BIXI variables. The second one only the contextual information. Then, the third analysis was about evaluating the impact of adding the contextual variables to the original BIXI variables in two ways: one by simply applying the same hierarchical clustering process using all the variables, and the other on using a multi-view clustering method.

The final conclusion was that the multi-view clustering gave better results from the analysis. It combines in an additive way the characteristics of the original BIXI variables as well as those from the external variables containing contextual information. This means that the characteristics from each cluster does not get lost after being mixed up. From there, profiles were defined taking into account all this information and, thus, users’ behaviour was characterized under various classes.

Once the bike trips are segmented in such ways, it becomes easy to make a local analysis to each one of the class in order to establish management strategies for the system in a local way for each profile studied. The last part of the project was
about incorporating the complex network approach and a dynamic analysis of the balancing factor locally to every cluster in order to extract operational knowledge to manage the network. Through the conclusions drawn from the analysis of a pair of concrete cases with specific profiles using the same methodology as for the whole data base, the project illustrated the type of knowledge that can be extracted applying this kind analytic methodology. For example, the results from the analysis of the Mundane & Flat Cloudy class taught us in which zones there is an overflow of bikes, at what time precisely and from which kind of users. This kind of results give clear guidelines about the operational management to optimize the redistribution process and enhance the service.

**Future Perspectives**

The first thing to do to bring this project forward would be to extend the clustering classification to the whole database. Then, the local analysis would be more meaningful and more relevant information could be reaped from it. Also, the local analysis should be extended to all the classes to identify all possible patterns and a marginal analysis should be done over the exceptional patterns (like long looping trips during a rainy night).

Alternatively, the database of external contextual information should be extended with, for instance, a calendar of sociocultural activities. For example, the closest station to the Bell Centre might know different patterns on game nights of the Montreal Canadians. Or, when the city is pulsating with all the numerous festivals going on, the BIXI network must also be following the same rhythm. Moreover, if one could get access to the personal data about the BIXI members, then really precise patterns could emerge for specific user profiles. Finally, applying different clustering methods could broaden the perspectives on the resulting patterns and diversify the knowledge extracted from it.

Accordingly, as Blain (2017) explained during its interview, BIXI simplifies its whole network into geographically constrained clusters (or zones), just like Chen et al. (2016) also propose in their paper. Not only it lightens the computing requirements, but it also simplifies the results since neighbour stations usually follow the same patterns of usage.

Then the decision-making guidelines should be automatized in a way that there is a live feed of information based on the evolution of the network. The next step would be to further develop the models to achieve a proper predictive model of the over-demand in the system.

On the other hand, the complex network study should be deeply extended and not only to the BIXI system, but also globally to public transportation in general. There is a definitive lack of research in this specific field, but there is definitely a lot of knowledge hidden there.

The optimization is the ultimate challenge for the BSS operators. With the advent of the machine learning and the boom in artificial intelligence, we are at a bright dawn for the demand prediction. Redistribution will always be needed. Nevertheless, no matter the prediction model, the breadth of an algorithm or the fancy technologies used to cope with the various challenges facing the operations of a bike-sharing system, because in the end, it is the field reality combined with the incredible human adaptability that dictate the path to follow.
Appendix A

A.1 Metadata Table

A.2 Extra Results from the Complex Network Analysis

The results shown on figure A.2 were the first calculated and not quite those expected for the betweenness and closeness, which led to some suspicions about the methodologies used and try different algorithms. The effect of considering the weight have a major impact in the present case since the trips forming specific edges occur way more than once (up to 4952 in 2015). Let’s recall that when the weight is not considered, the measures are calculated using only the number of ties and their frequency does not matter. This is why the results obtain at figure A.2 are interesting since they shed light upon some unexpected stations’ characteristics. The red dots might be funnels between neighbourhoods and the core of the city as theoretically explained earlier. Thus, these stations identify some hubs in the city such as the student neighbourhood close to Université de Montréal behind the Royal Mount (lower left corner of the map). As for the green dots, even though they are not in the geographic center of the network, they are “closer” to all the stations in the network in terms of connections when considering only the sheer amount of ties. Without being funnels between the city and further neighbourhoods, they probably experienced incoming and outgoing trips from and to almost all the other stations of the city.

A.3 Figures accompanying the complementary local analysis

The figures A.3(a) and A.3(b) present the trips as edges over the map of the city for the DTe Cloudy and the Nightout classes respectively. Figure A.3(c) present the dynamic of the Nightout trips. Only two time frame are shown since this class is only composed of evening and night trips.
<table>
<thead>
<tr>
<th>Variables</th>
<th>Modalities</th>
<th>Meaning</th>
<th>Type</th>
<th>R Class</th>
<th>Measuring unit</th>
<th>Range</th>
</tr>
</thead>
<tbody>
<tr>
<td>Start.date</td>
<td>Date of the beginning of the trip</td>
<td>Date</td>
<td>Date</td>
<td>Date POSIXlt</td>
<td>Date</td>
<td>[2015-04-15 00:00:00, 2015-11-15 23:58:00]</td>
</tr>
<tr>
<td>End.date</td>
<td>Date of the arrival</td>
<td>Date</td>
<td>Date</td>
<td>Date POSIXlt</td>
<td>Date</td>
<td>[2015-04-15 00:00:00, 2015-11-17 13:27:00]</td>
</tr>
<tr>
<td>Year</td>
<td>2015</td>
<td>Year of the trip</td>
<td>Qualitative</td>
<td>Factor</td>
<td>year</td>
<td>[2015]</td>
</tr>
<tr>
<td>Month</td>
<td>4, 5, 6, 7, 8, 9, 10, 11</td>
<td>Month of departure of the trip</td>
<td>Qualitative</td>
<td>Factor</td>
<td>Month identifier number</td>
<td>[4, 11]</td>
</tr>
<tr>
<td>Day</td>
<td>1 : 31</td>
<td>Day of departure of the trip</td>
<td>Qualitative</td>
<td>Factor</td>
<td>Day of the month</td>
<td>[1, 31]</td>
</tr>
<tr>
<td>Start.weekday</td>
<td>Monday, Tuesday, Wednesday, Thursday, Friday, Saturday, Sunday</td>
<td>Day of the week when the trip begins</td>
<td>Ordinal</td>
<td>Ordered factor</td>
<td>Day of the week</td>
<td>[Monday, Sunday]</td>
</tr>
<tr>
<td>Weekend</td>
<td>TRUE, FALSE</td>
<td>If the trip begins during the weekend (Saturday or Sunday)</td>
<td>Boolean</td>
<td>Logical</td>
<td>Logical</td>
<td></td>
</tr>
<tr>
<td>Nightout</td>
<td>TRUE, FALSE</td>
<td>If the trip begins Friday or Saturday night between 20:00 and 5:00.</td>
<td>Boolean</td>
<td>Logical</td>
<td>Logical</td>
<td></td>
</tr>
<tr>
<td>Period</td>
<td>Morning &lt;Noon &lt;Afternoon &lt;Evening &lt;Night</td>
<td>The period of the day when the trip begins</td>
<td>Ordinal</td>
<td>Ordered factor</td>
<td>[Morning, Night]</td>
<td></td>
</tr>
<tr>
<td>Start.time</td>
<td>Time when the trip starts</td>
<td>Character</td>
<td>[0:00, 23:59]</td>
<td>Character</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Hour</td>
<td>Time when the trip starts rounded down to the hour</td>
<td>Character</td>
<td>[0:00, 23:00]</td>
<td>Character</td>
<td></td>
<td></td>
</tr>
<tr>
<td>TenS</td>
<td>Time when the trip starts rounded down to the ten of minutes</td>
<td>Character</td>
<td>[0:00, 23:50]</td>
<td>Character</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Start.station.number</td>
<td>460 modalities corresponding to the ID of the stations</td>
<td>Starting station’s terminal identifier</td>
<td>Qualitative</td>
<td>Factor</td>
<td>[5002, 10002]</td>
<td></td>
</tr>
<tr>
<td>End.station.number</td>
<td>460 modalities corresponding to the ID of the stations</td>
<td>Ending station’s terminal identifier</td>
<td>Qualitative</td>
<td>Factor</td>
<td>[5002, 10002]</td>
<td></td>
</tr>
</tbody>
</table>
### Table A.2: Metadata Table of the Working Data Frame (Part 2)

<table>
<thead>
<tr>
<th>Variables</th>
<th>Modalities</th>
<th>Meaning</th>
<th>Type</th>
<th>R Class</th>
<th>Measuring unit</th>
<th>Range</th>
</tr>
</thead>
<tbody>
<tr>
<td>Capacity.S</td>
<td></td>
<td>Bike capacity of the origin station</td>
<td>Num</td>
<td>Integer</td>
<td>Number of docks</td>
<td>[0, 66]</td>
</tr>
<tr>
<td>Capacity.E</td>
<td></td>
<td>Bike capacity of the destination station</td>
<td>Num</td>
<td>Integer</td>
<td>Number of docks</td>
<td>[0, 66]</td>
</tr>
<tr>
<td>Metro.station.S</td>
<td>TRUE, FALSE</td>
<td>If the station of origin lies within 250 of radius from a Metro station</td>
<td>Boolean</td>
<td>Logical</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Metro.station.E</td>
<td>TRUE, FALSE</td>
<td>If the station of arrival lies within 250 of radius from a Metro station</td>
<td>Boolean</td>
<td>Logical</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Metro.sORe</td>
<td>TRUE, FALSE</td>
<td>If the station of origin OR arrival lies within 250 of radius from a Metro station</td>
<td>Boolean</td>
<td>Logical</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Metro.sANDe</td>
<td>TRUE, FALSE</td>
<td>If the station of origin AND arrival lies within 250 of radius from a Metro station</td>
<td>Boolean</td>
<td>Logical</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Downtown.S</td>
<td>TRUE, FALSE</td>
<td>If the station of origin is within the determined downtown area</td>
<td>Boolean</td>
<td>Logical</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Downtown.E</td>
<td>TRUE, FALSE</td>
<td>If the station of arrival is within the determined downtown area</td>
<td>Boolean</td>
<td>Logical</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Altitude.S</td>
<td></td>
<td>Altitude of the station of origin</td>
<td>Num</td>
<td>Num</td>
<td>meters</td>
<td>[14.83, 127.15]</td>
</tr>
<tr>
<td>Altitude.E</td>
<td></td>
<td>Altitude of the station of arrival</td>
<td>Num</td>
<td>Num</td>
<td>meters</td>
<td>[14.83, 127.15]</td>
</tr>
<tr>
<td>Elevation</td>
<td></td>
<td>Difference in altitude between the stations of arrival and origin</td>
<td>Num</td>
<td>Num</td>
<td>meters</td>
<td>[-112.323, 106.108]</td>
</tr>
<tr>
<td>Start.lat</td>
<td></td>
<td>Starting station’s latitude according to the CSR WGS84</td>
<td>Spatial</td>
<td>Num</td>
<td>Latitude</td>
<td>[45.43, 45.58]</td>
</tr>
<tr>
<td>End.lat</td>
<td></td>
<td>Ending station’s latitude according to the CSR WGS84</td>
<td>Spatial</td>
<td>Num</td>
<td>Latitude</td>
<td>[45.43, 45.58]</td>
</tr>
<tr>
<td>Start.long</td>
<td></td>
<td>Starting station’s longitude according to the CSR WGS84</td>
<td>Spatial</td>
<td>Num</td>
<td>Longitude</td>
<td>[-73.67, -73.50]</td>
</tr>
<tr>
<td>End.long</td>
<td></td>
<td>Ending station’s longitude according to the CSR WGS84</td>
<td>Spatial</td>
<td>Num</td>
<td>Longitude</td>
<td>[-73.67, -73.50]</td>
</tr>
<tr>
<td>Variables</td>
<td>Modalities</td>
<td>Meaning</td>
<td>Type</td>
<td>R Class</td>
<td>Measuring unit</td>
<td>Range</td>
</tr>
<tr>
<td>----------------------</td>
<td>--------------</td>
<td>--------------------------------------</td>
<td>----------</td>
<td>---------</td>
<td>----------------</td>
<td>------------------------</td>
</tr>
<tr>
<td>Total.duration</td>
<td></td>
<td>Transit’s total duration</td>
<td>Period</td>
<td>Period</td>
<td>seconds</td>
<td>[0.079sec, 174d 4H 21M 11sec]</td>
</tr>
<tr>
<td>Loop</td>
<td>TRUE, FALSE</td>
<td>Is true if the trip ends at the same station where it started</td>
<td>Boolean</td>
<td>Logical</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Account.type</td>
<td>Member</td>
<td>Type of user</td>
<td>Quali</td>
<td>Factor</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Temperature</td>
<td></td>
<td>Air temperature</td>
<td>Num</td>
<td>num</td>
<td>°C</td>
<td>[-4.70, 31.70]</td>
</tr>
<tr>
<td>Rel.humidity</td>
<td></td>
<td>Air relative humidity</td>
<td>Num</td>
<td>Integer</td>
<td>%</td>
<td>[11, 99]</td>
</tr>
<tr>
<td>Wind.speed</td>
<td></td>
<td>Wind speed</td>
<td>Num</td>
<td>Integer</td>
<td>km/h</td>
<td>[0, 53]</td>
</tr>
<tr>
<td>Wind.cardinal</td>
<td>N, NE, E, SE, S, SW, W, NW, No wind</td>
<td>Wind direction</td>
<td>Qualitative</td>
<td>Factor</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Atm.pressure</td>
<td></td>
<td>Atmospheric pressure</td>
<td>Num</td>
<td>Factor</td>
<td>kPa</td>
<td>[98.52, 103.24]</td>
</tr>
<tr>
<td>Weather.main</td>
<td>report to table ??</td>
<td>Simplified weather description</td>
<td>Qual</td>
<td>Factor</td>
<td>Weather</td>
<td></td>
</tr>
<tr>
<td>Proper.conditions</td>
<td>TRUE, FALSE</td>
<td>If the weather conditions are proper to cycle (Clear, Cloudy or Fog)</td>
<td>Boolean</td>
<td>Logical</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
### Table A.4: Metadata of the Stations Data Frame

<table>
<thead>
<tr>
<th>variable</th>
<th>Modalities</th>
<th>Meaning</th>
<th>Type</th>
<th>R Class</th>
<th>Measuring unit</th>
<th>Range</th>
</tr>
</thead>
<tbody>
<tr>
<td>terminalName</td>
<td>544 modalities corresponding to the ID of the stations</td>
<td>Station’s terminal identifier</td>
<td>Qualitative</td>
<td>Factor</td>
<td></td>
<td>[5002, 10002]</td>
</tr>
<tr>
<td>name</td>
<td></td>
<td>Name of the station taken from the name of the streets crossing nearby</td>
<td>String</td>
<td>Character</td>
<td></td>
<td></td>
</tr>
<tr>
<td>long</td>
<td></td>
<td>Station’s longitude according to the CSR WGS84</td>
<td>Spatial</td>
<td>Num</td>
<td>Longitude</td>
<td>[-73.67, -73.50]</td>
</tr>
<tr>
<td>lat</td>
<td></td>
<td>Station’s latitude according to the CSR WGS84</td>
<td>Spatial</td>
<td>Num</td>
<td>Latitude</td>
<td>[45.43, 45.58]</td>
</tr>
<tr>
<td>Capacity</td>
<td></td>
<td>Capacity of the station</td>
<td>Num</td>
<td>Integer</td>
<td>Number of docks</td>
<td>[0, 66]</td>
</tr>
<tr>
<td>MetroST</td>
<td>TRUE, FALSE</td>
<td>If the station lies within 250 of radius from a Metro station</td>
<td>Boolean</td>
<td>Logical</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Downtown</td>
<td>TRUE, FALSE</td>
<td>If the station is within the determined downtown area</td>
<td>Boolean</td>
<td>Logical</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Altitude</td>
<td></td>
<td>Altitude of the station</td>
<td>Num</td>
<td>Num</td>
<td>meter</td>
<td>[14.83, 127, 15]</td>
</tr>
</tbody>
</table>
Figure A.1: Centralities calculated using the Barrat et al. (2004) algorithm

Figure A.2: Centralities calculated using the Freeman (1978) algorithm
Appendix A

109

(a) Most Frequent DTe Cloudy Trips
(b) Most Frequent Nightout Trips

(c) Dynamic Analysis of the Nightout Trips

Figure A.3: Local Analysis of the Nightout Class and the DTe Cloudy Class independently
Figure A.4: A dynamic analysis of the DTe Cloudy context class.
Bibliography


BIBLIOGRAPHY


Côme, Étienne and Oukhellou Latifa (2014). “Model-Based Count Series Clustering for Bike Sharing System Usage Mining : A Case Study with the Vélib System of...


Décarie, Jean-Philippe (2016). Faire des affaires pour aller ailleurs. Montreal. URL: http://plus.lapresse.ca/screens/deb013d2-44cc-42b4-8d3d-6db400155cb7\%25257C\%2525252025252025252025252025252025252025252025252025252025252025252025252025252025252025252025252025252025252025252025252025252025252025252025252025252025252025252025252025252025252025252025252025252025252025252025252025252025252025252025252025252025252025252025252025252025252025252025252025252025252025252025252025252025252025252025252025252025252025252025252025252025252025252025252025252025252025252025252025252025252025252025252025252025252025252025252025252025252025252025252025252025252025252025252025252025252025252025252025252025252025252025252025252025252025252025252025252025252025252025252025252025252025252025252025252025252025252025252025252025252025252025252025252025252025252025252025252025252025252025252025252025252025252025252025252025252025252025252025252025252025252025252025252025252025252025252025252025252025252


Fuller, Daniel et al. (2012). “The impact of implementing a public bicycle share program on the likelihood of collisions and near misses in Montreal, Canada”. In: Preventive Medecine 57.6, pp. 920–924.


BIBLIOGRAPHY


Haas, A.E. (1928). Introduction to Theoretical Physics, Volume 1. Introduction to Theoretical Physics. Constable & company ltd. URL: https://books.google.es/books?id=VavvAAAAMAAJ.

Han, Yufei, Etienne Come, and Latifa Oukhellou (2014). “Towards bicycle demand prediction of large-scale bicycle sharing system”. In: Transportation Research Board, p. 17.


Singhvi, Divya et al. (2015). “Predicting Bike Usage for New York City’s Bike Sharing System”. In: *Computational Sustainability*.


Wang, Xize et al. (2013). “Modeling Bike Share Station Activity: the Effects of Nearby Businesses and Jobs on Trips to and from Stations”. In: *Transportation Research Record*.


