POLITECNICO DI TORINO

Master's Degree in Data Science and Engineering

Master's Degree Thesis

Data characterization by means of novel spatio-temporal patterns

Supervisors

Prof. Paolo GARZA

Dr. Luca COLOMBA

Candidate

Martina TOMA

October 2021

Abstract

Bike Sharing Systems are sustainable transportation strategies able to reduce the greenhouse gas emission. In recent years, they registered a strong growth thanks to the various benefits they bring, such as solving the "first and last mile" problem, that consists in travelling for short distances to reach work and get back.

This thesis analyzes a data-set containing data from Barcelona's stations to extract some meaningful patterns from a specif position and timestamp into a discretized space and time. The research consists in conducting different analyses with the Prefix Span algorithm to search for some correlations to predict future events that could happen.

The procedure used in this research can be applicable to other data-sets that contain events that happen over time, for which the spatial information is known.

Summary

Bike Sharing system is a widespread sustainable transportation strategy adopted to solve problems related to environmental pollution. It was introduced by policy makers and decision makers to enhance the life quality by reducing the greenhouse gas emission. It has been estimated that with a distance of 200000 Km, 37000 Kg of CO2 are saved. Bicycles can be considered a substitute mean of transportation with respect to cars or moped.

At the basis there is the concept of "as-needed": people have the possibility to pick a bike up and drop it off whenever they want. Users are encouraged to put bicycles into their apposite docks to avoid paying a penalty.

Bike Sharing systems are mostly used for short distances, rather than long journeys. They are useful to solve the "first and last mile" problem, that consists in reaching a public station from home or for going to the office work from a station.

The benefits Bike Sharing systems provide are not only connected to the reduction of the fuel emission and to the environmental benefits. In addition to these last advantages, there are other benefits concerning the introduction of a new mean of transport, an increase of people's health benefits and the reduction of traffic congestion.

During the evolution of Bike Sharing systems, four generations were identified. At the beginning, bicycles were introduced only to solve the greenhouse gas emission, were free and were not blocked. However, this starting point revealed a failure due to thefts and damaged bikes. Later, thanks to the introduction of IT-based systems, like GPS, it was possible to track bicycles, to identify the users and so to limit the vandalism.

In some countries, different types of Bike Sharing systems have been adopted. In addition to the previously analyzed station based system, two others developed. They are the dockless system and the hybrid one. In the dockless system users

have the chance to leave the bike whenever they want; on the other hand, users to pick bikes up have to search for them, since they are not placed in some pre-defined docks. An hybrid system, instead, encourages people to leave bicycles into the apposite places, by introducing a reward in terms of money, and at the same time allows users to choose between the best option for them.

The main goal is to extract some meaningful spatio-temporal patterns within the bike sharing system by means of a discretized time and space encoding and sequence mining algorithms. To carry out this procedure a data-set containing data from Barcelona's stations was considered. The correlation between events is analyzed to predict future events in the real world that could happen in the neighborhood and in the future time of the considered events.

Data about the stations has been collected at every 2 minutes from 2008-05-15 10:02 to 2008-09-30 21:58. Each record contains information about the timestamp, station id, the number of used slots, the number of free slots, the number of total slots and the percentage of used slots.

After applying the pre-processing phase, different analyses have been conducted, by considering only:

- full and empty stations
- full and almost full stations
- full and almost full stations that change state with respect to the previous one

Then, data transformation is applied by grouping records by time intervals with the chosen parameters and by transforming data in a discretized manner. This step is done in order to pass the transformed data to the Prefix Span, the applied algorithm, to extract frequent sequential patterns. Patterns were analyzed by considering different parameters, such as the support or the confidence. Once data was extracted, some filters have been applied to analyze the obtained outputs.

What emerged from these analyses is that by increasing the value of support to train the Prefix Span algorithm, the number of extracted patterns decreases, because only patterns with a higher frequency are allowed. At the same time, the higher the search space, in terms of distance, the higher the number of patterns. This behaviour can be explained by considering the fact that when the search radius increases more stations are gathered and, consequently, the number of patterns increases.

The computational time and the number of items that are present inside the

patterns have a trend very similar to the one described for the number of extracted patterns.

This study can be extended to other data-sets, not only to the one analyzed. This procedure is general and can be applicable to data-sets that contain events that happen over time, for which the spatial information is known.

Acknowledgements

Having reached the end of this university career, I would like to thank all the people who have been close to me and have encouraged me during my course of study.

To my supervisor Prof. Paolo Garza,
for his infinite availability, patience, for his useful advice during the entire period of writing this thesis. His contribution was indispensable from the choice of the topic to the end of this path.

To my co-supervisor Dr. Luca Colomba,
for his very useful advice and for suggesting the right changes to be made to the thesis. He was always ready to give me the right indications to follow.

To my family,
who has believed in me, for always being by my side with their advice and for the values they have always transmitted to me.

To Gabriel,
who patiently endured me when I was under stress from exams and encouraged me during this path.

To Aurora, Martina, Eleonora, Milena and Sonia,
my old friends, for supporting each other both during the moments of difficulty and in the moments of joy and satisfaction in achieving our goals. Thanks because I know that I can always count on them.

To my friends Giada and Francesca,
for always being close to me even in moments of difficulty, for always believing in me and for encouraging me to reach this important goal.

To my roommate Valentina,
companion of adventures, always present, who has supported and appreciated me

for how I am, with my strengths and weaknesses, a sincere and loyal friend to whom I have confided my thoughts and emotions.

To Sofia and Bianca,
my friends and my colleagues, who remained close to me in these years of university.

Table of Contents

List of Tables

List of Figures

Chapter 1

Introduction

1.1 Background

Bike Sharing system is a sustainable transportation strategy that has increased during the last years. It has been adopted by policy makers and decision makers to solve problems related to environmental pollution. It does not introduce any emission transportation and helps to reduce the greenhouse gas emission.

Bike Sharing is widespread in Europe, North America, South America and Asia and can substitute cars, not always, but very often. Its main drawbacks are linked to the short distance and to the availability of bikes and bike lanes, since there is no guarantee that people will find a bicycle when needed.

In the last years, thanks to the development of high technological systems, such as GPS, it is possible to track bicycles in real time and reserve a bike to deal with the problem of unavailability of this mean of transport.

Behind Bike Sharing systems there is the concept of "as-needed": people can pick a bike up when they need and drop them off to the apposite stations or, in some cases, also to dockless stations. People have not worry about future thefts once they leave the bike, since it is no more under their responsibility. Users have to drop bicycles off to the apposite parking places, otherwise a penalty in terms of money applies.

As said previously, bicycles are very useful for short distances and not for long journeys. They solve the "first and last mile" problem, in fact, bicycles can be considered as a substitute product to cars or mopeds, especially for reaching a public station from home or for going to the office work from a station.

The benefits that bicycles bring are very high: they introduce a new mean of transport, enhance environmental benefits, increase health benefits, reduce traffic congestion and reduce the fuel emission.

1.2 Research study and goals

The research study consisted in extracting some meaningful spatio-temporal patterns within the bike sharing system by means of a discretized time and space encoding and sequence mining algorithms. To carry out this procedure a data-set containing data from Barcelona's stations was considered.

These analyses can be extended to other complex problems, such as predicting future events in the real world. An example about future predictions might be the following: knowing that in a specific place rains (event 1) and there is congestion in another place (event 2), then in another place in a radius of 1 Km (this distance is discretized using blocks for example of 500 meters) from the place of events 1 and 2 will rain (event 3) in the next 10 minutes.

The sequence of procedures used during the analysis can be gathered into the following steps:

- analysis of the data-set
- pre-processing procedures, such as data cleaning
- data transformation, in such a way that the future algorithm is able to take input values and transform them into an efficient output
- application of the Prefix Span algorithm
- analysis of the output values from the previous algorithm
- extraction of statistics from the previous study
- control phase to provide an explanation of the obtained results

1.3 Case study

The analyzed data-sets contain data from Barcelona's stations. There are two data-sets: the first one has data collected at every 2 minutes from 2008-05-15 10:02 to 2008-09-30 21:58; while the second one contains for each Barcelona's station

some information about their latitude, their longitude and their name.

Data is gathered thanks to the RFID cards that people use. Users are incentivized to take a bike for no more than 2 hours: a penalty of 4.20 € is applied to all people that exceeds this amount of time. For the first 30 minutes bikes are free, while for the next half-hours it costs 0,70 €.

When users want to pick a bike up, they have to be identified and stations have to be traced with the GPS, that registers the positions of the bikes against thieves and also for statistical and predictive reasons. Data, in fact, has a great potential: it is not just for statistical reason, but it also improves the service of bike sharing systems thanks to real-time data.

This data can be used to get some general estimations about the status of a station. Furthermore, the status produces some useful information to:

- improve the service by analyzing the areas in which, due to a high request, there are some rejections by the users
- inform users with some estimates about the predicted states of the stations to give the chance to better manage their time
- better manage the service quality, especially during high-demand events

All the technological systems used to track the position of the bicycles and to identify the users are not limited to the Bike Sharing analysis, but can be extended to other fields, such as electric-car rentals and online transportation service like Uber. Data driven decision making and Big Data tools are at the basis for analyzing and predicting future events. [1]

With this work some meaningful patterns are extracted to understand what the future events will be, given that some events occurred in the past. The objective is the prediction of the future states of some stations that are in the radius of a fixed distance from those observed.

The state of a station can be full, almost full or empty. The state of the station is empty if all bikes are used and all the slots are free. Figure 1.1 shows an empty station from the city of Barcelona: there are 0 available bikes, 25 free slots/ spaces and 3 are not usable, for a total of 28 docks. [2]

The state is full if all slots are used and no slots are free. Figure 1.2 illustrates the state of a full station: there are 23 bikes that are available, ready to be used, and 0 free slots, so there is no space to drop bicycles off. [2]

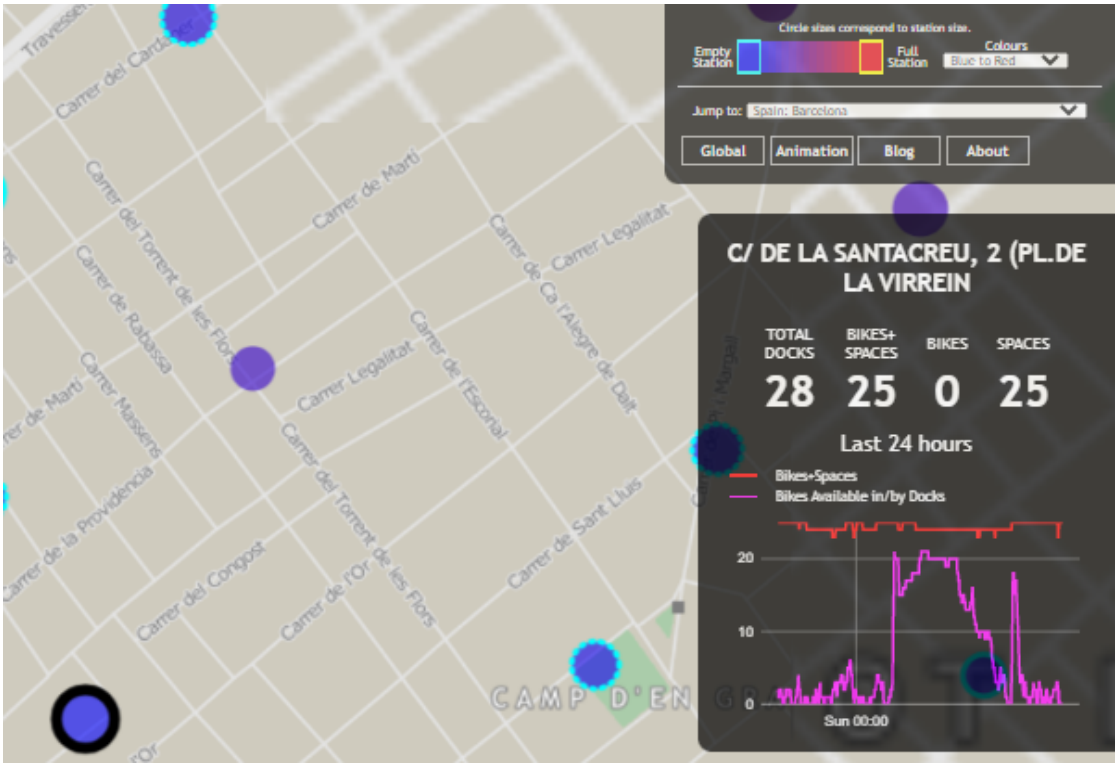


Figure 1.1: An empty station in Barcelona

In other cases the stations are neither full nor empty and are characterized by some available bikes. Figure 1.3 is an example of an almost full station: there is 1 available bike, 22 docks ready to be used and 5 are not usable, for a total of 27 docks. [2]

Figure 1.4 illustrates the arrival rate per hour in Barcelona during the 24 hours a day. The plot highlights three peaks in correspondence of the morning, afternoon and evening [3] .

1.4 Research organization

The thesis is organized in 4 chapters.

- Chapter 1 contains an overview of the thesis with a description of the general topic. A focus on the case study follows with the objectives to reach
- Chapter 2 illustrates the relative literature applied to the analyzed study. It illustrates the origin of the Bike Sharing systems and the impact the latter

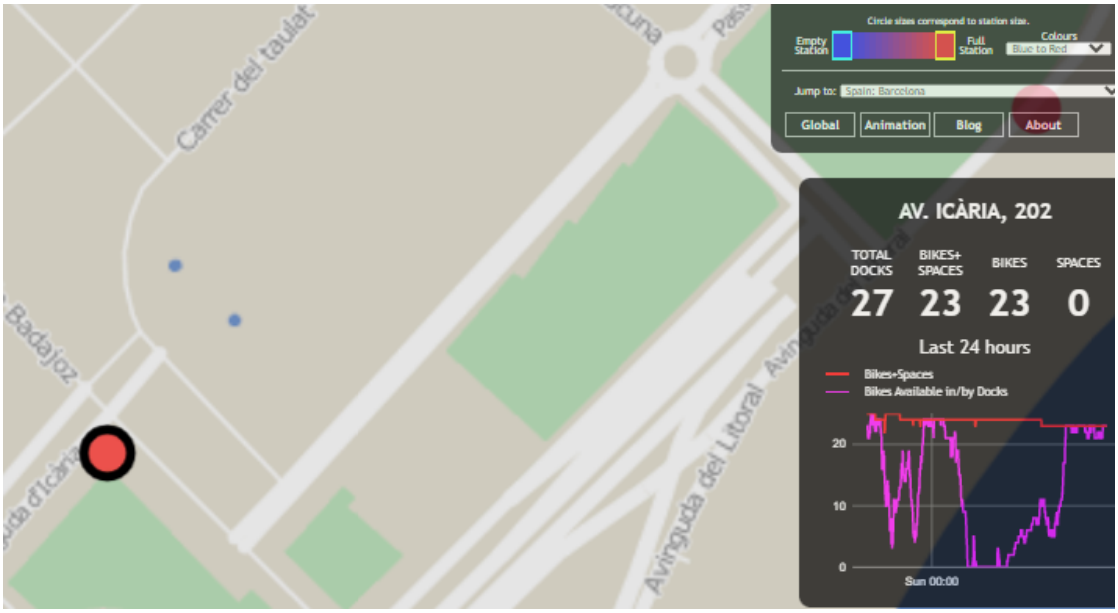


Figure 1.2: A full station in Barcelona

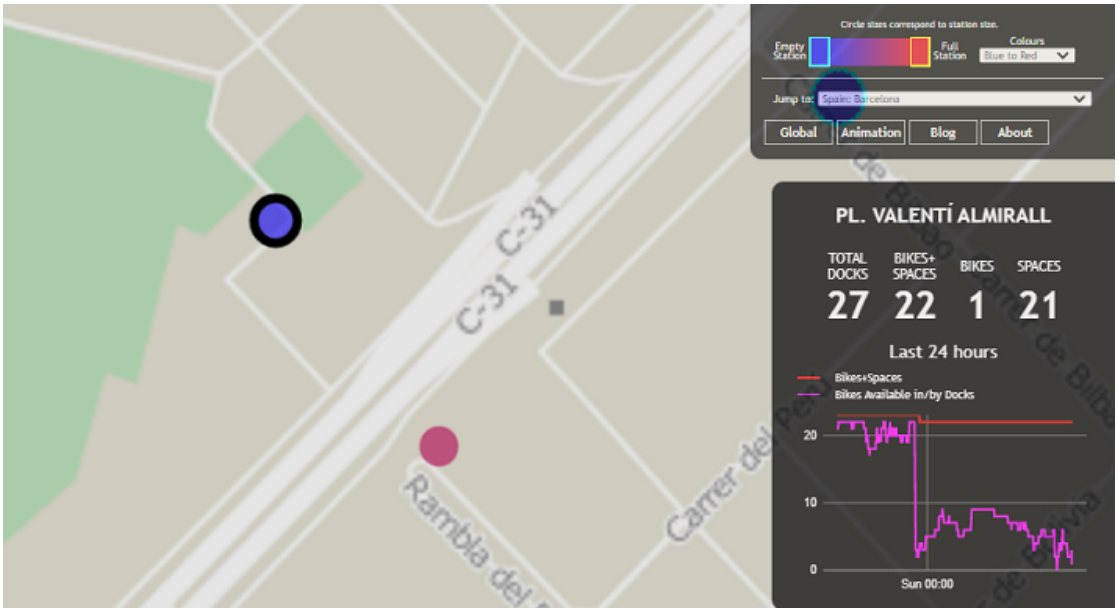


Figure 1.3: An almost full station in Barcelona

causes on the environment. Moreover, a description of the different bike sharing systems is provided with the relative advantages and disadvantages. Eventually, theory concerning the case study, such as the algorithm used, is

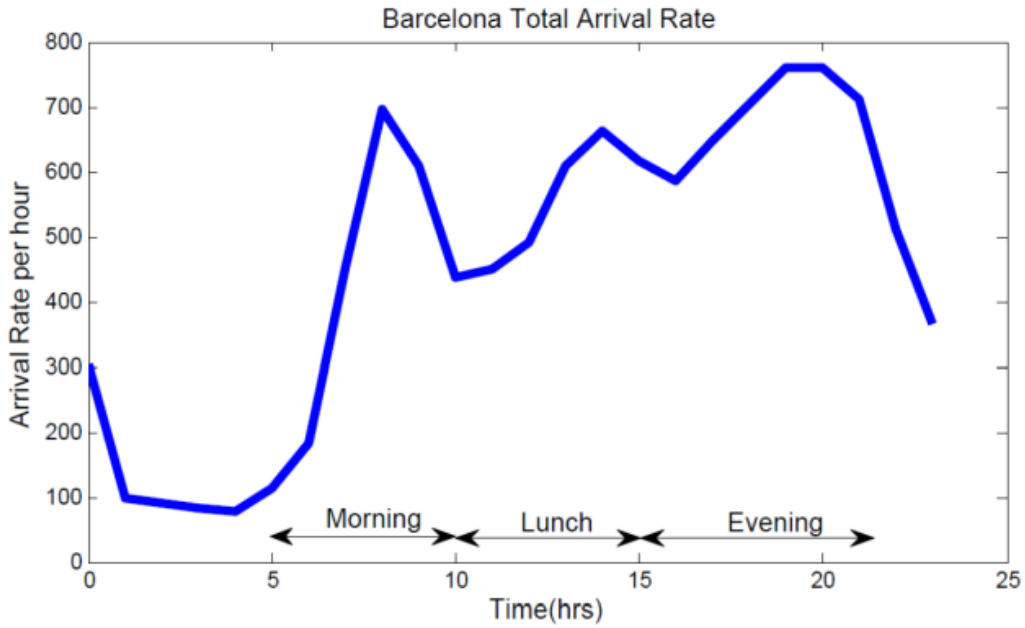


Figure 1.4: Arrival rate per hour in Barcelona

presented

- Chapter 3 presents a more accurate description of the data-sets used to perform the study and an analysis of the discussed problem and of the proposed solution. It discusses how starting data have been transformed, how the patterns are extracted, the applied filters on the results
- Chapter 4, the last one, contains a description of the obtained experimental results during the analysis
- Chapter 5 contains the conclusions about the thesis with the proposed future works

Chapter 2

Related work

2.1 Background

Bike Sharing is a low-cost system developed around the world to provide a way to substitute cars and other means of transportation. It has already existed for decades, but in the last years evolved, thanks to the diffusion of IT, it was possible to track them and enhance communication. During the evolution of Bike Sharing systems, four generations were identified.

In the first generation bikes were not blocked and were free. The first phase started with the "Provos' White Bike" plan and takes the name by the fact that bikes were coloured of white. This project has the origin in Amsterdam and took place in July 1965. During this initial phase fifty bicycles were open to the public for free. Unfortunately, this plan failed because of thefts and damaged bikes. In this phase there were other plans that were adopted, like "Vélos Jaunes", the one adopted in La Rochelle, France, in 1974 and the "Green Bike Scheme" in Cambridge, United Kingdom, in 1993. After the failure of the first phase, because the only emphasis was the environmental issue, the second phase took place.

The second phase, instead, is characterized by coin-deposit systems. To solve the problem of thefts, a new system was introduced to prevent it. Users had to deposit some money, about $4, before using bicycles. Furthermore, docking stations were adopted to lock bikes ones returned. With this new system, the only method to take a bike was to introduce some coins. During this phase an improvement was visible with respect to the previous system; however, it didn't prevent vandalism, since users were anonymous, not registered and could not be traced.

The third system is distinguished by IT-based systems. These systems make

use of electronic tools to track, pick up and drop-off bicycles. Users are no more anonymous, but are identified with advanced technologies, thanks to magnetic cards and smart cards. The investments were paid off thanks to great benefits that offer. IT systems allowed to reduce drastically thefts, due to the users' identification.

The fourth generation introduced some improvements with respect to the previous one, such as demand-responsive and multimodal systems. One of the technological improvements was about the introduction of GPS for real time tracking. Stations were flexible and this means that the presence of docking stations was not a requirement, in fact, "dockless" bicycles were introduced. [4]

2.2 Bike Sharing and its impact on the environment

Bike Sharing is a system that does not provide any emission transportation. Differently from other means of transportation, bikes have the potential to reduce greenhouse gas emission. It has been calculated that on average 50,000 bikes cover a distance of 200,000 Km per day. That distance would produce 37,000 Kg of CO2 (carbon dioxide) if cars were used.

Bike Sharing does not include only one country or one continent, but it is widespread in four continents: Europe, North America, South America and Asia. Despite being very helpful to encourage people to reduce pollution, there are still some obstacles. Among them there is the limited amount of infrastructures: there is no guarantee that people find at least an available bike and a bike lane. Other drawbacks are linked to the high technology cost and safety issues. For sure this system is an environmental benefit, that leads to an improvement of the quality of life.

Because of global climate change, policy experts had to analyse the necessity for more sustainable strategies. The strategies concerned clean fuel, new vehicle strategies and transportation strategies. Bike sharing is one of the adopted strategies to address this topic. In the last years there was a growth in the use of bicycles. Bike sharing is a useful tool to integrate not only cycling, but also to provide a valid transportation alternative.

The concept behind bike sharing system can be enclosed into the concept "as-needed". It means that people that want to take a bike do not have high responsibilities like checking that nobody thieves the bike after that the bike is left into a docking station. The advantages that bikes arise concern the possibility to reserve a bike

through the app, to pick a bike up if available and to drop it off whenever users want. They provide a short-term access and they are very attractive to users that want to move for a short distance. Users do not have to bear the burden of buying a bike, but they pay depending on the amount of time they spend using it.

In addition to this, people do not have to care about the maintenance: it is a cost that only the Bike Sharing system suffers. Users have the possibility to leave the bike wherever they want if there is a bike station with an empty slot to put it inside. Bike sharing systems induce many benefits such as:

1. increasing mobility options
2. reducing traffic congestion
3. reducing fuel use
4. increasing health benefits
5. enhancing environmental benefits

A Bike Sharing system can be seen as a substitute product to cars, mopeds and other means of transport. Bike Sharing can be a good method to solve the "first and last mile" problem: bikes are an important mean to move for short distances, such as moving from home to a public station or from a public station to the work office. [5] However, bikes can't substitute cars for long distances, but only for some short routes.

2.3 Bike Sharing systems

Different types of bike sharing services have been adopted: station based system, dockless system and an hybrid one. Both the station based system and the dockless one present some advantages and disadvantages.

Station based systems, having fixed docks, provide some predefined places where to take and leave bikes. This is a benefit for users that want to pick a bike up because it is located at a specific station. On the other hand, it will be a drawback when users want to drop a bicycle off, especially if there are no available slots to insert it or if the station is far from the destination place. By placing bicycles into some predefined locations, cities are more ordered because users have not the possibility to leave them wherever they want.

With dockless stations, instead, people have to search for bikes, that are not collocated in a prefixed space. One of the main advantage is that users can drop

Figure 2.1: Barcelona Bike Sharing

the bicycle off wherever they want. With stations based systems users have to reach a station and if there are no available space to insert the bike they have to search for another station. Dockless stations can benefit from a reduced cost of the initial investment: less money is spent because less fixed costs are present. For instance, there is no more the cost of the docking stations, but only of the bikes.

Hybridization is a sort of mix between these two different methods. Users have the possibility to choose among the best option to use. Thanks to this service, it is possible to benefit from advantages of both station based systems and dockless ones.

An example of the hybridization system has been implemented in Cracow, Poland. Users can pick bikes up from both a station based system and a dockless one and can drop them off by using both two systems. This hybrid system encourages users to leave bicycles into the dock stations. The Polish system introduced a penalty in terms of coins, 3 PLN, for users that do not leave bicycles in a station based system. Users that take a bike, that is around, and leave it into a docking station

are rewarded with a credit on their account. [4]

Figure 2.2 shows an hybrid system adopted in the USA, in which both a station based system and a dockless one are present.

Figure 2.2: Hybrid Bike Sharing system in the USA

2.4 Knowledge discovery

Data Mining discipline provides some useful tools to extract some meaningful information, patterns or rules from a large amount of data. The objective is to obtain some patterns that have some frequent occurrences in the sequential data: the support gives a hint about the importance of the set of items. The extraction of the rule consists in extracting some strong association rules such as X $\rightarrow$ Y: if something happens then another rule follows. The patterns are characterized by a set of items and their importance depends on the number of times in which a pattern appears. The database is a sequence of patterns, each pattern is a transaction and it is a set of items (itemset). Each pattern has some time windows ordered by time. Sequential patterns can obtain a significant information from the time factor. The discovery of patterns is useful for different purposes, such as the

description and the prediction of events. The frequent patterns can be used to obtain rules to describe connections between events.

The process to discover knowledge is iterative and consists of the following steps:

1. Data cleaning: to remove noisy data
2. Data integration: to combine different data sources
3. Data selection: to filter relevant data for the analysis
4. Data transformation: to transform data in a more appropriate way to perform some operations
5. Data mining: a process to extract data patterns
6. Pattern evaluation: to identify some patterns based on some statistics
7. Knowledge presentation: to present knowledge to people

Figure 2.3 shows a clear image of the previously described steps. [6]

2.5 Sequential patterns

A sequential pattern is a frequent subsequence: a subsequence whose support is greater or equal to minsup. The support of a subsequence is obtained by dividing the number of sequences containing that subsequence by the total data sequences. A pattern is frequent if its frequency is $\geq$ to the minimum frequency. The obtained frequencies can be used to obtain rules. To obtain the frequent patterns it is useful to start from one episode with one event, to do a level search in the lattice and on each level to get the candidates and check the relative frequencies. The procedure consists in projecting sequences into a set of smaller projected databases and in growing each fragment in each projected database. For example with min support = 0.5 are taken the sub-sequences that can be extracted having at least 50 % of support.

2.6 Prefix Span

The algorithm used to perform the analysis is the Prefix Span. Prefix Span is a sequential pattern algorithm, that is described in “Pei et al., Mining Sequential Patterns by Pattern-Growth: The Prefix Span Approach” [7]. The used algorithm considers the following parameters:

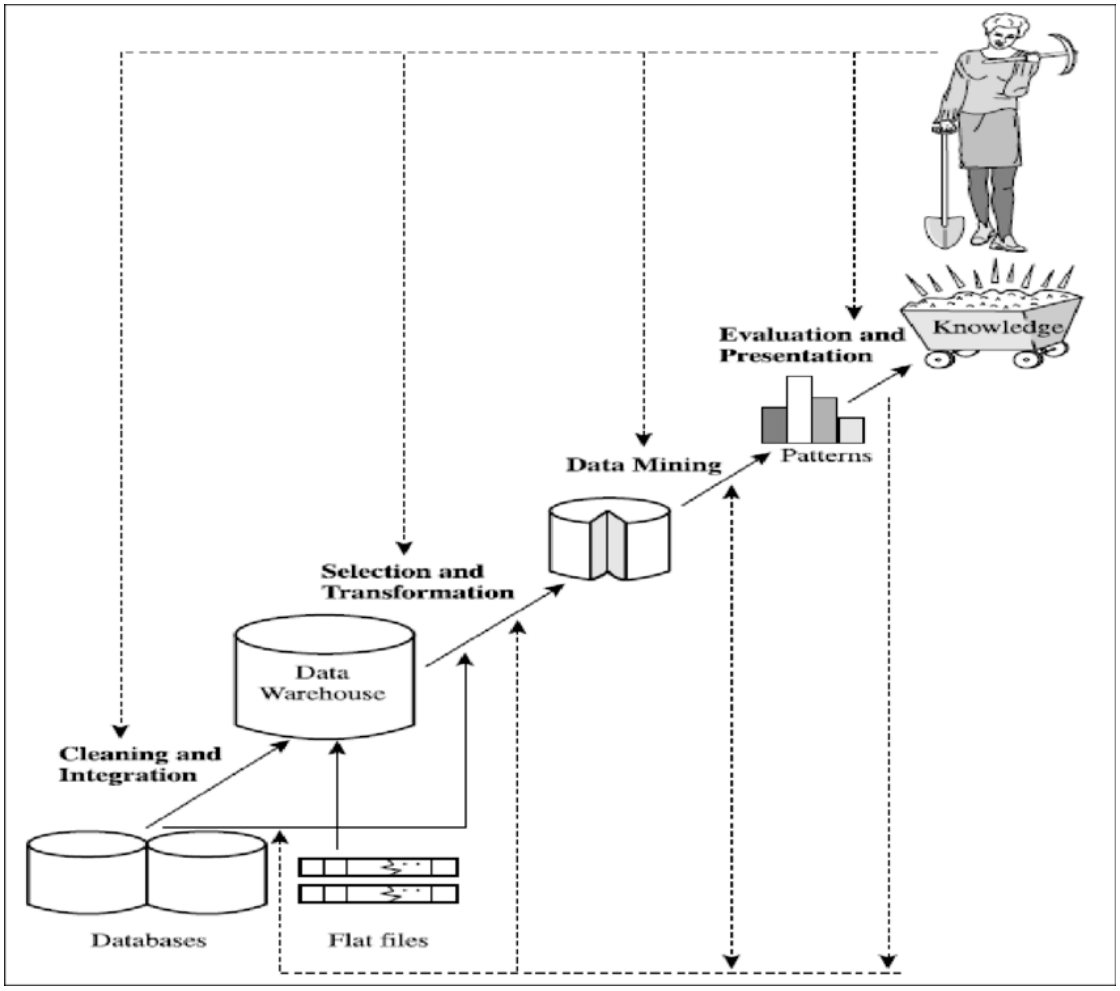


Figure 2.3: Knowledge discovery process

1. *minSupport*: the minimum support to be considered a frequent sequential pattern;

2. *maxPatternLength*: the maximum length for a frequent sequential pattern;

3. *maxLocalProjDBSize*: the maximum number of items allowed in a prefix-projected database before local iterative processing of the projected database starts;

4. *sequenceCol*: rows with null values in this column are ignored; by default it was set to "sequence".

2.7 Conclusion

This study provides a general method to analyze a data-set containing spatial and temporal information and extract patterns. This procedure is not only applicable to this data-set, but also to other data-sets containing events that happen over time, for which the spatial information is known.

Chapter 3

Work description

3.1 Introduction

The bicycle-sharing system in Barcelona allows citizens to move toward small and medium routes eliminating pollution, roadway noise and traffic congestion. Barcelona makes use of different public transports from metro and buses to the Bicing service, that eliminates the pollution, noise and traffic congestion. It can be considered as the best means of transport in the city.

In order to rent a bicycle each user has to swipe the RFID card at a service station to be identified by the system, so that a bike is unlocked from the support frame. Bikes are free for the first 30 minutes, while the subsequent half-hours cost 0.70 €. A penalty of 4.20 € per hour is applied when the use of the bicycle exceeds 2 hour. In addition to this, there might be the possibility to have the membership cancelled after three times in which the use is in excess for 2 hours. When the users want to return a bicycle, they just place it inside a free slot, then it is automatically recognized and is locked.

Bicycles are flexible means of transports, in fact people can go wherever they want, provided that a station is present in the neighbourhood. The service provides assistance 24 hours a day in case of failure and a personal insurance. It is an ecologic and cheap solution and allows people to take advantage of a very pleasant bike ride. Moreover, thanks to the Mediterranean climate and to the various cycle paths, getting around by bike is facilitated.

This study provides a general method to analyze data-sets and extract patterns from them. It is applicable to data-sets that contain records about events that occur in a specific time and space. The future predictions, that can be obtained,

depend on the correlation between spatio-temporal events.

3.2 Data sets

Two data sets are used to do all the analyzes. The first one contains registered data about 283 stations of Barcelona. Each record contains information about:

1. timestamp
2. station Id
3. used slots
4. free slots
5. total slots
6. used slots %

This data-set contains data from 2008-05-15 10:02 to 2008-09-30 21:58. The other one, instead, contains information about each station. In fact, for each station the analyzed variables are:

1. station Id
2. longitude
3. latitude
4. name

3.3 Data analysis

The first data-set contains for a fixed station and timestamp the number of used slots, the number of free slots, the total number of slots and the percentage of used slots, given by the ratio between the difference of total slots and free slots and the total slots available. The second data set contains records about 3301 stations, but those used are only 283: the ones that are in the first data set as foreign key. For each station data were collected at each 2 minutes: from 2008-05-15 10:02 to 2008-09-30 21:58. Since data has been filtered in order to extract some useful patterns, it results that there are some missing timestamps. The goal was to analyze the entire data set to extract some spatial and time patterns.

3.4 Data cleaning

The data-set used to perform the entire analysis was filtered by null values because they do not provide any useful information. In addition to this, records containing information about no free available slots and at the same time no used slots were deleted, since they do not give any hint about the state of the stations: empty, full and almost full.

3.5 Statistics

To better explore the data-set some statistics have been computed. It resulted that:

1. The average number of stations that are at a distance smaller or equal to 0.5 Km is 3.51;
2. The average number of stations that are at a distance smaller or equal to 1 Km is 14.32;
3. The average number of stations that are at a distance smaller or equal to 2 Km is 49.84;
4. The average number of stations that are at a distance smaller or equal to 3 Km is 94.91;
5. The average number of stations that are at a distance smaller or equal to 4 Km is 140.07;
6. The average number of stations that are at a distance smaller or equal to 5 Km is 179.53.

As can be supposed, the higher the radius from the location of one station, the higher the number of stations around it.

3.6 Implementation details

The analyses have been carried out by grouping the timestamps by different time intervals: 10, 20 and 30 minutes. Data have been grouped by time intervals: in each group there is the list of registered stations with the corresponding state. The defined states are:

- Full: there are no free slots;
- Empty: all the slots are empty;

- Almost full: there are 1 or 2 available slots.

The applied parameters are:

- spatial delta: the maximum number of spatial deltas;
- time delta: the maximum number of time windows;
- support: the support used to train the Prefix Span algorithm;
- spatial threshold: the distance in meters for one spatial delta

The experiments have been carried out by applying the following values:

- spatial deltas: 3 and 4;
- time deltas: 3 and 4;
- supports: 0, 0.05, 0.1, 0.5, 0.6, 0.7, 0.8, 0.9;
- spatial thresholds: 100, 500, 1000 meters.

Different filters have been used to analyze the data-set:

1. sequences with at least 2 windows, at least a T0 (event that happens at time t=0) and at least a spatial delta of 0 (reference event)

Once this filter has been applied, other 2 filters have been applied on it:

1. sequences with at least one event "Almost Full" and one event "Full"
2. sequences with at least one event that happens at time T0 and ΔS=0 and at least one event with $\Delta S \neq 0$ and $\Delta T \neq 0$. This filter extracts patterns with events that happen at a given station at a certain time and events that happen in its neighbourhood in the next time intervals.

3.7 Data transformation

3.7.1 First step

Starting from the data in the data set different transformations have been used. Data has been grouped by time intervals: for each time interval there is a list containing the stations with their relative state. The following example shows two interval times of 10 minutes, i.e. from 15/5/2015 10:00 to 15/5/2015 10:09 and from 15/5/2015 10:10 to 15/5/2015 10:19. In the first time interval station id 132 is almost full, while station id 61 is full; in the following time interval station id 52 and 61 become almost full.

```
[((2008, 5, 15, 10, 0), '132_AlmostFull, 61_Full),
 ((2008, 5, 15, 10, 1), '52_ AlmostFull, 61_ AlmostFull)]
```

3.7.2 Second step

The second step consists in combining different windows according to the chosen parameter. If the parameter is 2, then for each time interval it is considered that line and the following one obtaining a time window equal to 2 timestamps, apart from the last one that contains only one.

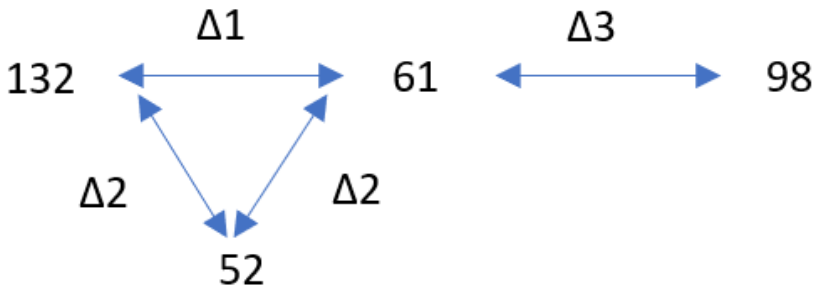
```
[((2008, 5, 15, 10, 0), '132_ AlmostFull, 61_ AlmostFull),
 ((2008, 5, 15, 10, 1), '52_ AlmostFull, 61_ AlmostFull),
 ((2008, 5, 15, 10, 2), '52_ AlmostFull, 69_ AlmostFull),
 ((2008, 5, 15, 11, 0), '17_ AlmostFull, 54_ AlmostFull)]

Window 1- [[‘132_T0_ AlmostFull, ‘61_T0_ AlmostFull],
           [‘52_T1_ AlmostFull, ‘61_T1_Full, ‘98_T1_Full]]
Window 2- [[‘52_T0_ AlmostFull, ‘61_T0_ AlmostFull],
           [‘52_T1_ AlmostFull, ‘69_T1_ AlmostFull]]
Window 3- [[‘52_T0_ AlmostFull, ‘69_T0_ AlmostFull],
           [‘17_T1_ AlmostFull, ‘54_T1_ AlmostFull]]
Window 4- [[‘17_T0_ AlmostFull, ’54_T0_ AlmostFull]]
```

3.7.3 Third step

The third step consists in computing the spatial step application. For each window the list containing the first timestamp, that is the one at time t=0, and the projection of the considered window with respect to the different stations are considered. To make an example consider as spatial threshold 500 meters, as time delta 3 and the following distances between the stations:

1. station 132 and station 61 -> 400 meters

2. station 132 and station 52 -> 800 meters
3. station 132 and station 98 -> 1800 meters
4. station 61 and station 52 -> 700 meters
5. station 61 and station 98 -> 1400 meters



This implies that station 132 is 0 km away from itself, so 0 spatial deltas; 400 meters away from station 61, within 1 delta; 800 meters from station 52, within 2 deltas; 1800 meters from station 98, so 4 deltas. Since in this example only stations that are at most 1.5 Km are considered, the last one is not taken into consideration. In addition to this, station 61 is 700 meters away from station 52, that is 2 deltas, station 61 is 150 meters away from station 98, that is 1 delta.

Window 1|132 – [[AlmostFull _T0_ Δ0, AlmostFull _T0_ Δ1],
[AlmostFull _T1_ Δ2, Full_T1_ Δ1]]
Window 1|61 – [[AlmostFull _T0_ Δ1, AlmostFull _T0_ Δ0],
[AlmostFull _T1_ Δ2, Full_T1_ Δ0, Full_T1_ Δ1]]

3.8 Haversine formula

To calculate the distance between 2 stations, given their latitude and longitude, the Haversine formula has been applied. The Haversine formula calculates the great-circle distance between two points: the shortest distance over the earth's surface, 'as-the-crow-flies' distance between the points, ignoring any hills. The Haversine formula is used to calculate the distance according to these following formulas:

$a = \sin^2 \frac{\Delta\varphi}{2} + \cos\varphi_1 \cdot \cos\varphi_2 \cdot \sin^2 \frac{\Delta\lambda}{2}$
$c = 2 \cdot a \tan 2(\sqrt{(a)}, \sqrt{(1-a)})$
$d = R \cdot c$

where $\Delta\varphi$ is given by the difference of the two latitudes and φ_1 is the latitude of the first station, φ_2 is the latitude of the second station, $\Delta\lambda$ is the difference of the longitude and R is the radius of the hearth, that is approximately 6373 Km.

3.9 Confidence

One of the metrics used to analyze data is the confidence. It is a probability, which is given by the number of times in which a pattern appears divided by the number of times in which that pattern appears without considering the last time window. Due to the fact that it it is a probability, it assumes values from 0 to 1. The confidence is defined as a conditional probability: P(Y|X). It is the probability of registering the Y event given that X is satisfied.

3.10 Support

The support is defined as P(X ∪ Y): the rule has a support s over the data set if s % of the transactions of the data contain both X and Y. A potential sub-sequence is frequent if the support is greater or equal to a fixed threshold.

3.11 Experiments

To analyze and explore the data-set some parameters have been fixed and others are left free. The number of extracted patterns are plotted with respect to:

1. the confidence, by fixing the support and the value of spatial delta
2. the different values of spatial and time deltas and fixing the value of spatial threshold (for example 100, 500, 1000 meters)
3. the support, by fixing the values of spatial and time deltas

3.12 Conclusion

This chapter provides details about the pre-processing phase, the algorithm used, some definitions and how the experiments have been carried out. In the next chapter the results about the experiments will be analyzed.

Chapter 4

Experimental results

4.1 Introduction

This chapter is used to analyze the obtained results through several experiments. The following results concern full and empty stations; full and almost full stations; state changes with full and almost full stations.

4.2 Full and empty stations

One of the analyses consists in analyzing only full and empty stations. To do that, after the data cleaning phase, only full and empty events are considered. Data obtained during this experiment related to these 2 events (empty and full) is about 22,5 % of the data-set, after data cleaning is performed. When the Prefix Span algorithm is applied on this filtered data, events that contain only full and empty events, most of the extracted patterns is about empty stations, stations with 0 used slots. The total distinct stations, after data cleaning and the filtering of empty and full stations are applied, are 272. Most of the data is about the empty stations: the full stations are only 4.31 %. This experiment regards both full and empty stations with an interval time of 30 minutes, 3 spatial deltas and 3 time deltas. By analyzing the obtained patterns, the following results are visible:

- when the spatial threshold is 100 meters and the support is 0, patterns with both full and empty stations coexist. Starting from a slightly higher support, like 0.05, only empty stations are present

- when the spatial threshold is 500 meters, instead, patterns with both full and empty stations are present only if the support is under 0.1. With a support of 0.5, the only extracted patterns are those with empty stations

- when the spatial threshold is 1000 meters, patterns with only empty stations exist starting from at least 0.7 support

By increasing the spatial threshold, there are still patterns concerning full stations even with a high support.

Figure 4.1 shows the number of extracted patterns in function of the supports (0, 0.05, 0.1, 0.5, 0.6, 0.7, 0.8, 0.9), depending on the spatial threshold (100, 500, 1000 meters) with an interval time of 30 minutes, 3 spatial deltas and 3 time deltas. The higher the support, the lower the number of extracted patterns and the higher the spatial threshold, the higher the number of patterns.

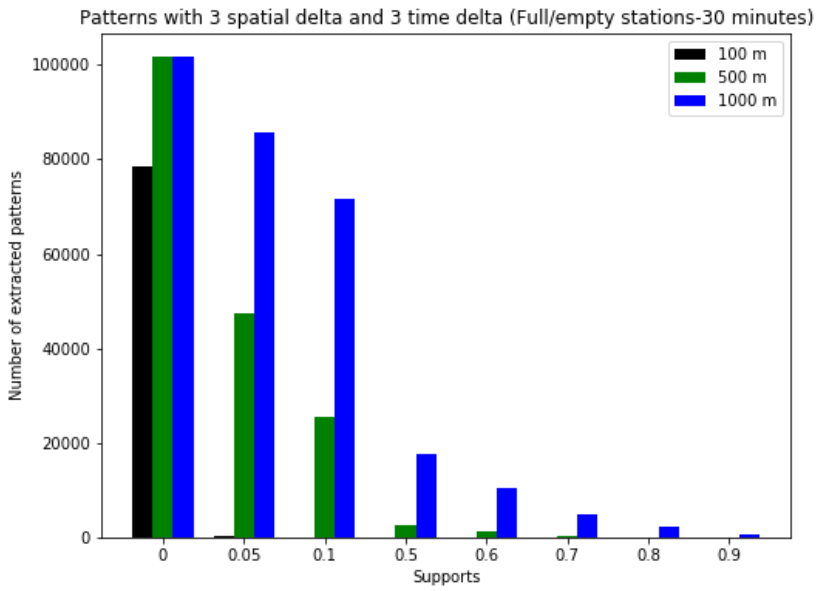


Figure 4.1: Extracted patterns with full/empty stations every 30 minutes

Table 4.1 shows how the spatial threshold and support affect the number of extracted patterns just after the Prefix Span algorithm is applied and after some chosen filters. This table shows for a fixed spatial threshold (meters) and support the following cases:

- pre-filter: the number of patterns just after the Prefix Span algorithm has been applied, without considering any filter
- 2 windows T0Δ0: the number of patterns that have at least 2 time windows, in

which T0Δ0 is present, that is at least an event that happens at the reference station Δ0, whose distance is 0 Km from itself, and at time t=0

- Full & empty events: the number of patterns in which there is at least a full and an empty event

- S=0 T=0, S≠0 T≠0: the number of patterns that contain an event that happens at a specific position at time t=0 and then, after some time, another event occurs at another station in its neighbourhood

spatial threshold [m]	**support**	**pre-filter**	**2 windows T0Δ0**	**Full & empty events**	**S=0 T=0, S≠0 T≠0**
100	0	78533	25672	23542	18870
500	0	101579	29408	27058	21872
1000	0	101579	29408	27058	21872
100	0.05	179	84	0	39
500	0.05	47377	10311	8085	7570
1000	0.05	85570	19891	17541	14799
100	0.1	72	32	0	13
500	0.1	25455	5072	3754	3621
1000	0.1	71521	17242	15125	12675
100	0.5	7	3	0	0
500	0.5	2537	730	0	502
1000	0.5	17645	2704	1543	1929
100	0.6	3	1	0	0
500	0.6	1295	301	0	175
1000	0.6	10577	1580	714	1095
100	0.7	1	0	0	0
500	0.7	415	58	0	24
1000	0.7	26742	3749	3092	2596
100	0.8	1	0	0	0
500	0.8	71	0	0	0
1000	0.8	2336	6	0	0
100	0.9	0	0	0	0
500	0.9	6	0	0	0
1000	0.9	566	0	0	0

Table 4.1: Comparison between different filters applied at different spatial thresholds and supports with full and empty stations

Figure 4.2 shows the number of patterns depending on the values of confidence from 0 to 1 with an interval time of 30 minutes, 3 spatial deltas and 3 time deltas. With small values of confidence there is a pick of the number of extracted patterns, then the trend decreases and only after 0.5 support starts to increase.

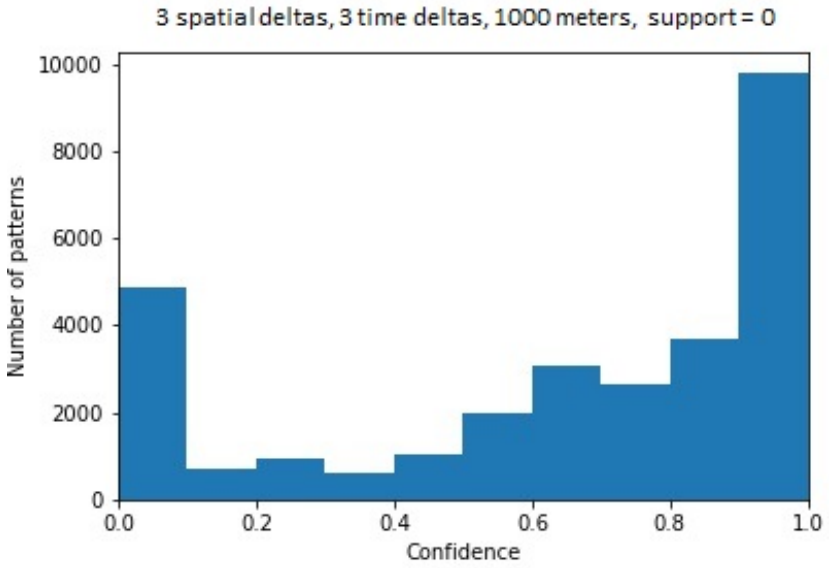


Figure 4.2: Extracted patterns with full/empty stations and different values of confidence every 30 minutes

Figure 4.3 shows how the execution time changes in function of different values of support and of different spatial thresholds. The execution time is very sensitive to the different applied parameters. It increases when:

- the spatial threshold increases
- the support decreases

A reasonable explanation is that when the spatial threshold increases, the search space is enlarged, there are many patterns and much more time is needed to process data. At the same time, when the support increases, the number of patterns decreases, since less patterns are filtered through the Prefix Span algorithm.

Figure 4.4 shows the average number of events inside the extracted patterns. On average, the higher the spatial threshold, the higher the average number of events inside the patterns. At the same time, on average, the higher the support, the lower the number of patterns. With the 0.7 support, there are no events for a

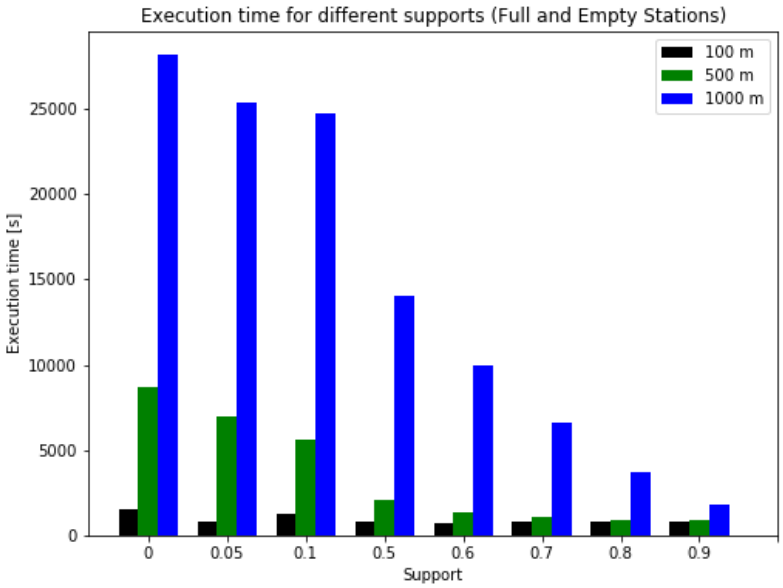


Figure 4.3: Execution time with full/empty stations every 30 minutes

spatial threshold of 100 meters. With the 0.8 support there are only events starting from a spatial threshold of 1000 meters. Using a support of 0.9, instead, there aren't events for any spatial thresholds.

Figure 4.5 shows the maximum number of events in the patterns. Up to a support of 0.1 the number of events is 5 and it is the same for the three different spatial thresholds used. Furthermore, the number of maximum patterns with a spatial threshold of 500 and 1000 meters remains the same up to a support of 0.7. The number of patterns for supports 0.5 and 0.6 and for a spatial threshold of 100 meters shows a decreasing trend, also with respect to lower values of supports. With a spatial threshold of 100 meters, there are no patterns from a support of 0.7.

Figure 4.6 shows the minimum number of patterns with 3 spatial deltas, 3 time deltas every 30 minutes. The plot illustrates that the only present values are 2 and 0. From a support of 0.7 onward, there no events with a spatial threshold of 100 meters. With a spatial threshold of 500 meters, there are no events starting from 0.8, a higher support. Eventually, with a support of 0.9 there aren't events with any spatial thresholds.

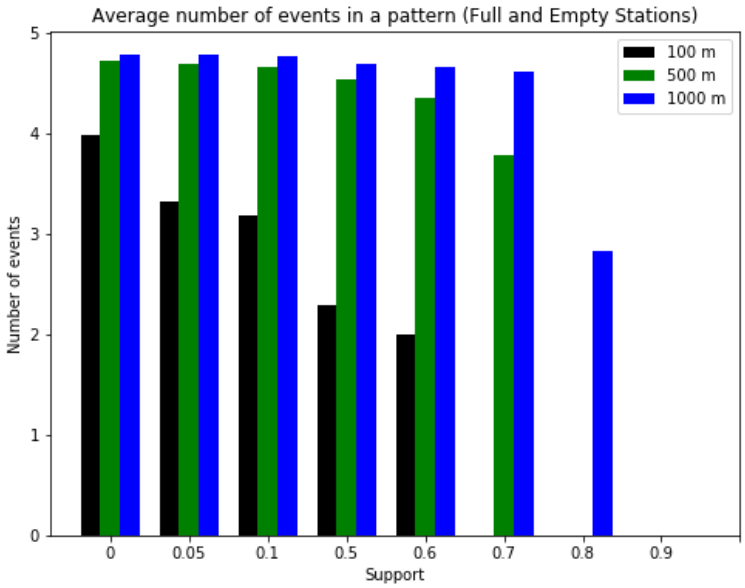


Figure 4.4: Average number of events with full/empty stations every 30 minutes

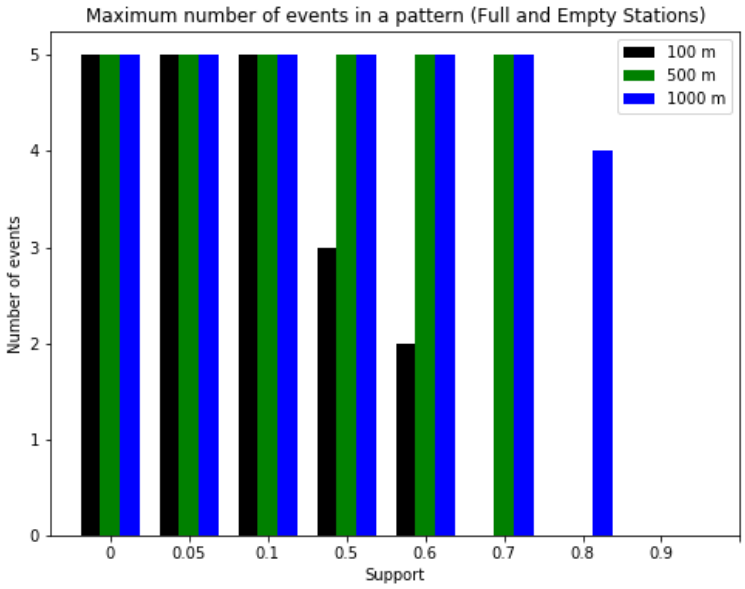


Figure 4.5: Maximum number of events with full/empty stations every 30 minutes

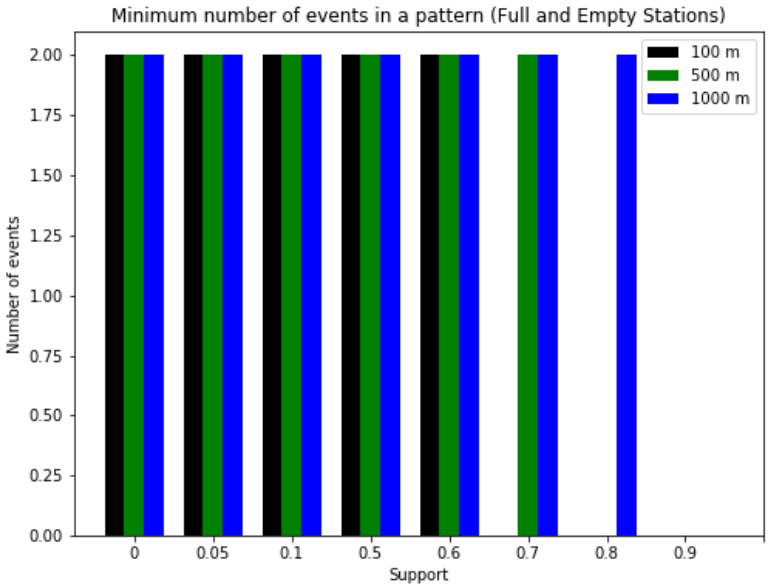


Figure 4.6: Minimum number of events with full/empty stations every 30 minutes

4.2.1 Example of an extracted pattern

The extracted patterns are ordered according to decreasing confidence and with equal confidence to decreasing frequency. The following pattern is characterized by two time windows, one at time t=0 and the other at time t=2, using a spatial threshold of 100 meters, time delta=3 and spatial delta=3. The pattern presents in the first time window 2 empty events: the first one is the reference station $\Delta 0$, while the second one is referred to a station that is 2 spatial deltas distant from the first one, that is between 100 and 200 meters. The second window presents 3 events with 3 full stations: one at a distance of 2 deltas, while the others at a distance of 1 spatial delta, within 100 meters of distance. The pattern illustrates that if there are 2 empty stations at time t=0, of which the second one is at a distance between 100 and 200 meters with respect to the first one, then, there will be 3 full stations in the neighborhood of the first station, used as reference, of which two are within 100 meters and the other between 100 and 200 meters.
$[['Empty\_T0\_\Delta 0, Empty\_T0\_\Delta 2'],$
$['Full\_T2\_\Delta 2, Full\_T2\_\Delta 1, Full\_T2\_\Delta 1']]$
This pattern can be used to predict the future behaviour, given that some observations are registered at the reference time t=0. Furthermore, this pattern is registered in all the applied filters because there are 2 windows with at least an event that presents $T0\_\Delta 0$; both full and empty events; and there is at least an

event that happens at a specific reference station at time t=0 (S=0 and T=0) and at least an event that happens at time different from 0 and at a distance different from 0.

Visualization of the pattern

Figure 4.7 is a visualization of the pattern previously analyzed. The red circles represent events at time t=0, while green circles concern time t=2. At the center of the circular sector there is the station with $\Delta 0$, the reference station; the circles that are positioned near 100 are the stations that are within 100 meters with respect to the reference station; the circles near 200 are stations that are at a distance between 100 and 200 meters from the reference station.

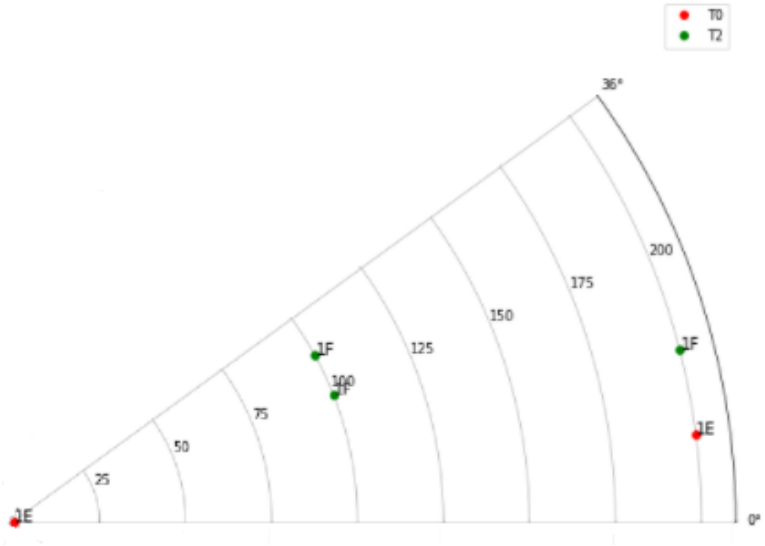


Figure 4.7: Visualization of a pattern with full and empty stations

4.3 Full and almost full stations

Other analyses have been carried out by considering not only full, but also almost full stations. Starting from the entire data-set, data cleaning phase is performed and then only full and almost full stations are considered. If 500 and 1000 meters as spatial thresholds are considered, the number of extracted patterns with a 0 support is the same and also the types of patterns are identical in both 20 and 30 minutes as shown in Figure 4.8 and Figure 4.9.

Figure 4.10 and Figure 4.11 show the number of patterns in function of the

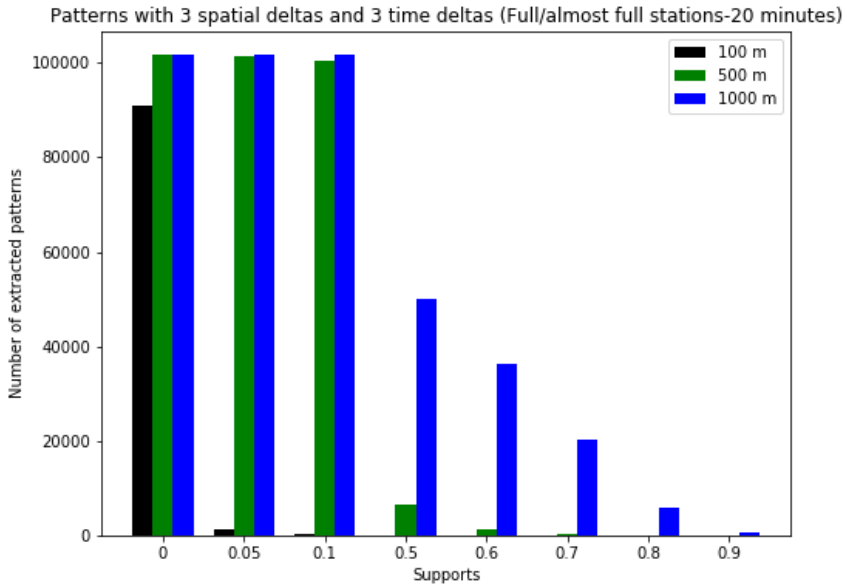


Figure 4.8: Extracted patterns with full/almost full stations every 20 minutes

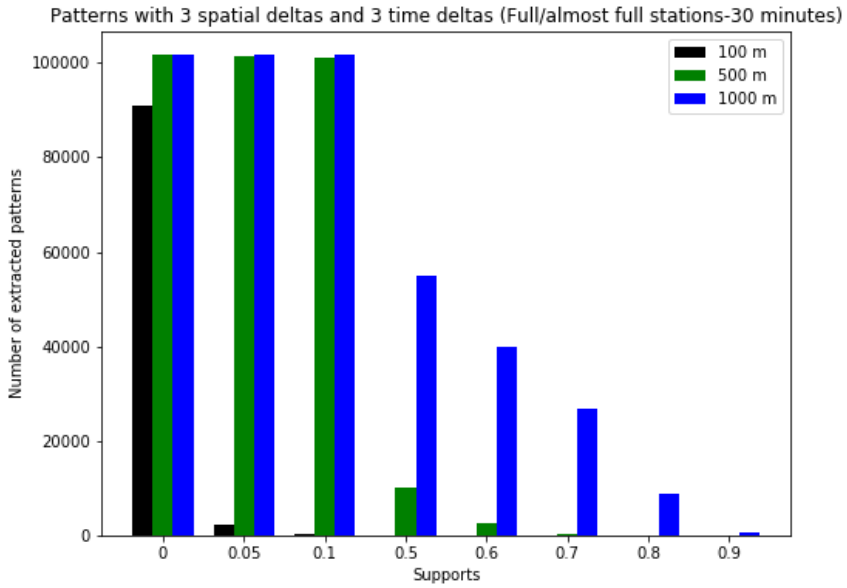


Figure 4.9: Extracted patterns with full/almost full stations every 30 minutes

values of confidence, from 0 to 1. In both graphs there is on average an increasing behaviour: the higher the values of confidence, the higher the number of patterns.

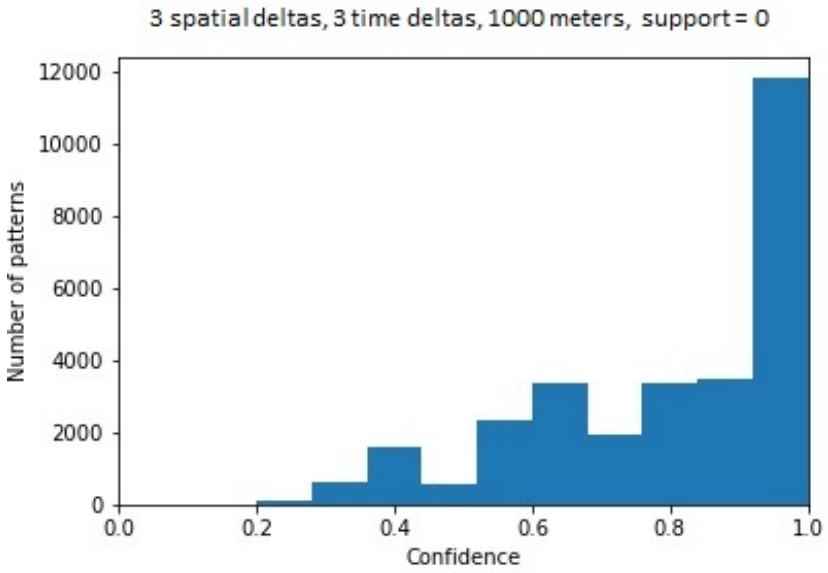


Figure 4.10: Extracted patterns with full and almost full stations and different values of confidence every 20 minutes

The last two columns of Table 4.2 represent filters that have been applied at the "2 windows T0Δ0" filter. The third column has the highest value of extracted patterns because no filters are applied. The fourth one, instead, has less patterns than the previous one, since patterns with only one window were removed. The fifth and sixth column present less patterns than the fourth one and, consequently, even less than the "pre-filter" column.
By increasing the length of the radius from a fixed station, the number of reachable stations increases, as a consequence. At the same way, the number of extracted patterns increases as the spatial threshold increases. On the other hand, the number of extracted patterns decreases as the number of the support increases.

Figure 4.12 illustrates the execution time in function of different values of support and spatial thresholds. The trend, in general, confirms what stated before with full and empty stations: the smaller the support, the higher the execution time and the higher the spatial threshold, the higher the time needed.

Figure 4.13 shows the average number of events inside the extracted patterns, considering full and almost full stations. The higher the spatial threshold, the

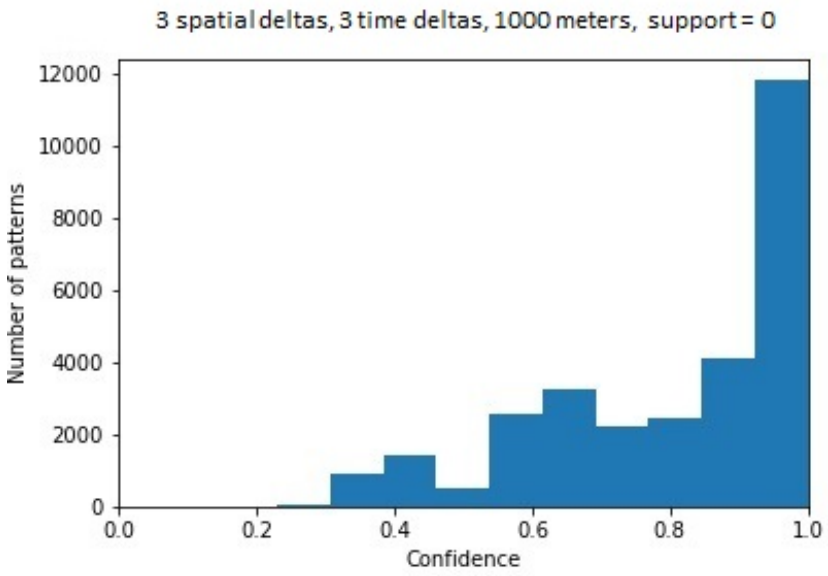


Figure 4.11: Extracted patterns with full and almost full stations and different values of confidence every 30 minutes

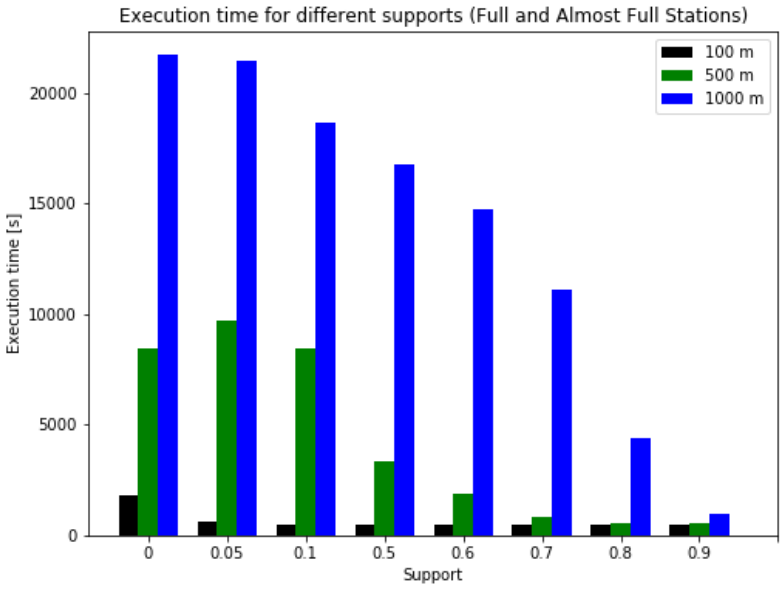


Figure 4.12: Execution time with full/almost full stations every 30 minutes

spatial threshold [m]	support	pre-filter	2 windows T0Δ0	Full & almost full events	S=0 T=0, S≠0 T≠0
100	0	91067	28340	26210	21012
500	0	101579	29408	27058	21872
1000	0	101579	29408	27058	21872
100	0.05	2332	1403	1266	780
500	0.05	101530	29404	27056	21868
1000	0.05	101579	29408	27058	21872
100	0.1	254	130	75	27
500	0.1	101103	29305	26987	21793
1000	0.1	101579	29408	27058	21872
100	0.5	9	3	0	0
500	0.5	10098	2276	1635	1525
1000	0.5	54935	11039	9850	7988
100	0.6	5	2	0	0
500	0.6	2693	619	364	368
1000	0.6	39952	7050	6157	5057
100	0.7	3	1	0	0
500	0.7	381	71	2	32
1000	0.7	26742	3749	3092	2596
100	0.8	1	0	0	0
500	0.8	55	8	0	0
1000	0.8	8849	1017	504	662
100	0.9	1	0	0	0
500	0.9	4	0	0	0
1000	0.9	518	5	0	0

Table 4.2: Comparison between different filters applied at different spatial thresholds and supports with full and almost full stations

higher the average number of events inside the patterns. Moreover, the higher the support, the lower the number of patterns. Starting from a support of 0.8, there are no events for a spatial threshold of 100 meters. Using a support of 0.9, instead, there are no events for both 500 and 1000 meters as spatial thresholds.

The maximum number of events in the patterns is shown in Figure 4.14. Up to a support of 0.1 the number of events is 5 and it is the same for the three different used spatial thresholds. In addition to this, the number of maximum patterns with a spatial threshold of 500 and 1000 meters remains the same up

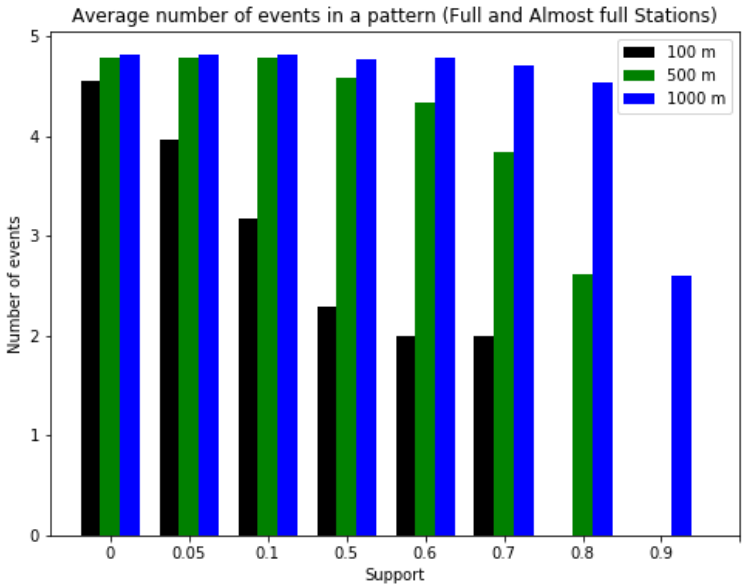


Figure 4.13: Average number of events with full/almost full stations every 30 minutes

to a support of 0,7. The numbers of patterns for supports 0.5, 0.6 and 0.7 and for a spatial threshold of 100 meters decreases also with respect to lower values of supports. With a spatial threshold of 100 meters, there are no patterns from a support of 0.8 and with a support of 0.9 there are only patterns for a spatial threshold of 1000 meters.

Figure 4.15 shows the minimum number of patterns with 3 spatial deltas, 3 time deltas every 30 minutes. The plot illustrates that the only present values are 2 and 0. From a support of 0.8 onward, there no events with a spatial threshold of 100 meters. With a spatial threshold of 500 meters, there are no events starting from 0.9, a higher support. Eventually, with a support of 0.9 there are only events with a spatial threshold of 1000 meters.

4.3.1 Example of an extracted pattern

The following pattern is an example of the extracted patterns with a spatial threshold of 100 meters, 0 support, 3 spatial deltas and 3 time deltas. It is composed of three time windows: one at time t=0, one at time t=1 and the last one at time t=2. The first window presents an almost full station at the reference

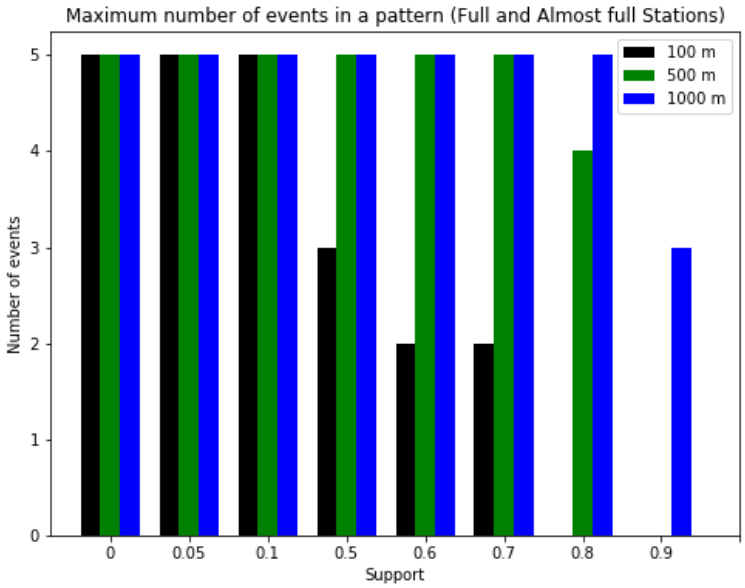


Figure 4.14: Maximum number of events with full/almost full stations every 30 minutes

station; the second one, instead, a full station at time t=1 at a distance between 200 and 300 meters; the last one presents an almost full station at a distance between 200 and 300 meters with respect to the reference station. The described pattern demonstrates that if an almost full event at time t=0 with $\Delta 0$ and another event at time t=1 of a station that is between 200 and 300 meters occur, then another event with almost full station between 200 and 300 meters at time t=2 occurs. $[['AlmostFull\_T0\_\Delta 0'], ['Full\_T1\_\Delta 3'], ['AlmostFull\_T2\_\Delta 3']]$

Visualization of the pattern

Figure 4.16 is a visualization of the pattern previously analyzed. The red circle represents events at time t=0, while the blue circle is about time t=1 and the green circle concerns time t=2. At the center of the circular sector there is the station with $\Delta 0$, the reference station; the circles that are positioned near 300 meters are the stations that are between 200 and 300 meters with respect to the reference station.

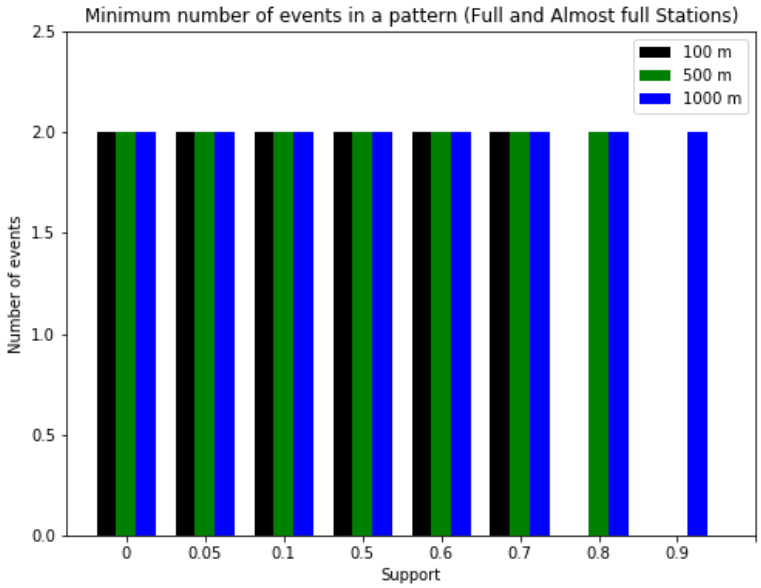


Figure 4.15: Minimum number of events with full/almost full stations every 30 minutes

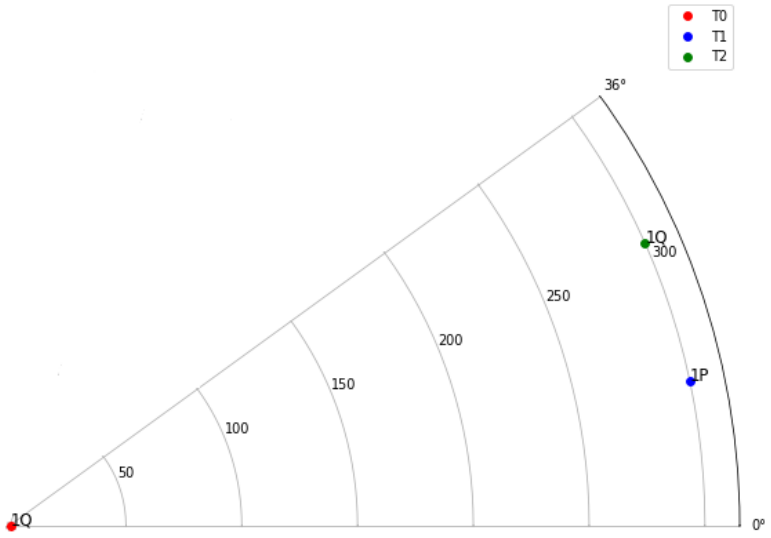


Figure 4.16: Visualization of a pattern with full and almost full stations

4.4 State changes with almost full and full stations

The last analysis consisted in analyzing the data-set with almost full and full stations and in extracting patterns that change state when going from a timestamp to another. As happened previously, after the data cleaning step, only data concerning almost full and full stations have been extracted.

When there is no state change, as shown in the first two rows of Figure 4.17, that means that the state remains the same from one timestamp to another, the only registered pattern is the first one and if in the future there aren't any changes for that station there won't be any extracted patterns.

The third row is characterized by full and almost full stations in the first two timestamps and both events remain as extracted patterns because two states coexist for the same timestamp; the third timestamp has a full state, that is different from the previous state.

In the fourth row, instead, there is a situation in which all the states are different from the previous timestamps: for this reason, all the states continue to be registered.

In the fifth case for a specific station there is an event in the first timestamp, in the first half an hour, but there are no other events in the following timestamp and there is an event in the third timestamp. In this case, the events in the first and third timestamp remain because in the middle there is not an event.

In the last case of the table, in the second timestamp there are no extracted events, since the event at the first timestamp is the same in the second one, but in the third timestamp there are two events and this implies that they are registered.

Table 4.3 shows for a fixed spatial threshold (meters) and support the extracted patterns before and after having applied some filters. The same considerations made in the previous tables, can be done on this table. From 0.5 support on, there are at most three observable patterns and for this reason, not so many meaningful patterns are present.
Figures 4.18, 4.19, 4.20 show the number of patterns depending on the values of confidence from 0 to 1. The plots show that the trend is not so much different when the patterns are analysed at each 10, 20 and 30 minutes. With some exceptions, the trends show a decreasing behaviour.

Data			Extracted patterns		
0-30min	**31-60min**	**61-90min**	**0-30min**	**31-60min**	**61-90min**
Full	Full	Full	Full	---	---
Almost full	Almost full	Almost full	Almost full	---	---
Full/ Almost full	Full/ Almost full	Full	Full/ Almost full	Full/ Almost full	Full
Full	Almost full	Full	Full	Almost full	Full
Full	---	Full	Full	---	Full
Full	Full	Full/ Almost full	Full	---	Full/ Almost full

Figure 4.17: State change from one timestamp to one other

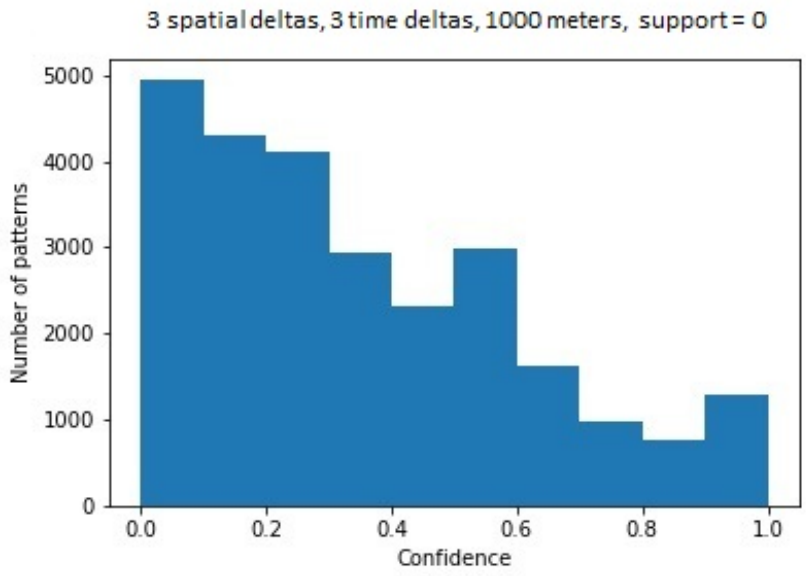


Figure 4.18: Extracted patterns with full and almost full stations and different values of confidence that change state every 10 minutes

Figure 4.21 shows the execution time depending on different values of supports and spatial thresholds. Apart from the 0 support, that has high values of execution

spatial threshold [m]	**support**	**pre-filter**	**2 windows T0Δ0**	**Full & almost full events**	**S=0 T=0, S≠0 T≠0**
100	0	141	70	58	18
500	0	18970	7320	6630	4725
1000	0	47493	16455	15244	11625
100	0.05	7	2	1	0
500	0.05	44	10	4	1
1000	0.05	229	84	31	15
100	0.1	3	0	0	0
500	0.1	8	0	0	0
1000	0.1	65	21	3	1
100	0.5	3	0	0	0
500	0.5	3	0	0	0
1000	0.5	3	0	0	0
100	0.6	2	0	0	0
500	0.6	2	0	0	0
1000	0.6	2	0	0	0
100	0.7	1	0	0	0
500	0.7	1	0	0	0
1000	0.7	1	0	0	0
100	0.8	1	0	0	0
500	0.8	1	0	0	0
1000	0.8	1	0	0	0
100	0.9	1	0	0	0
500	0.9	1	0	0	0
1000	0.9	1	0	0	0

Table 4.3: Comparison between different filters applied at different spatial thresholds and supports with full and almost full stations that change state

time, the others do not follow a particular trend. A reasonable explanation is that with 0 support much more patterns are registered; while by increasing the support, not so many patterns are extracted and have a similar behaviour.
Other analyses have been carried out with full and almost stations that change state every 30 minutes using different values of spatial and time deltas. The modifications concern the following cases:

- 3 spatial deltas and 4 time deltas
- 4 spatial deltas and 3 time deltas

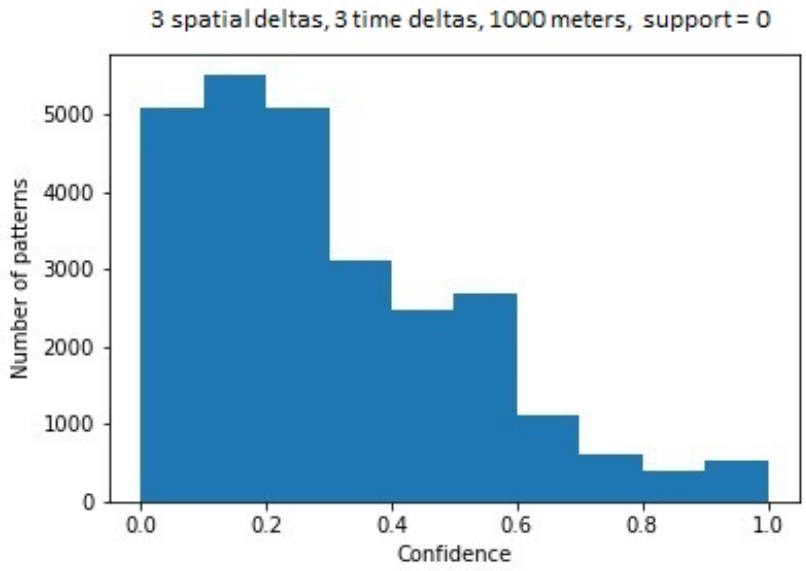


Figure 4.19: Extracted patterns with full and almost full stations and different values of confidence that change state every 20 minutes

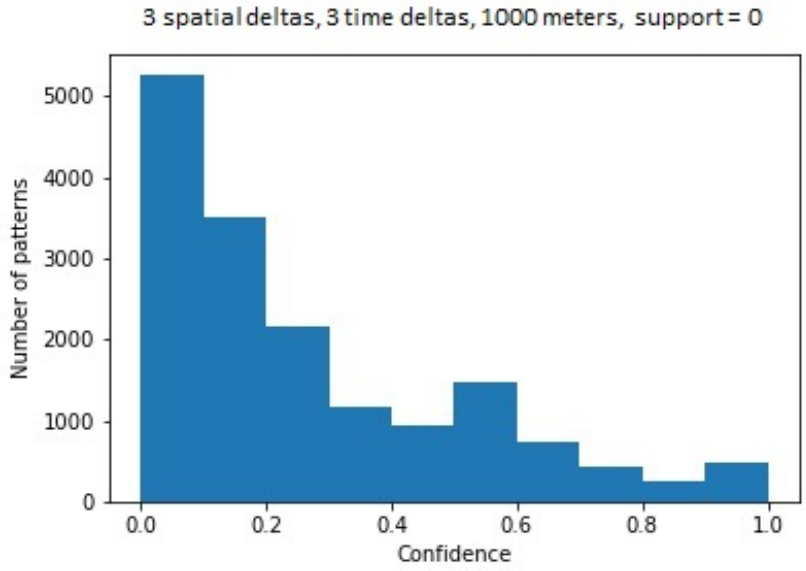


Figure 4.20: Extracted patterns with full and almost full stations and different values of confidence that change state every 30 minutes

In both these last cases, it is visible an increasing number of extracted patterns with respect to the case with 3 spatial deltas and 3 time deltas. By increasing the

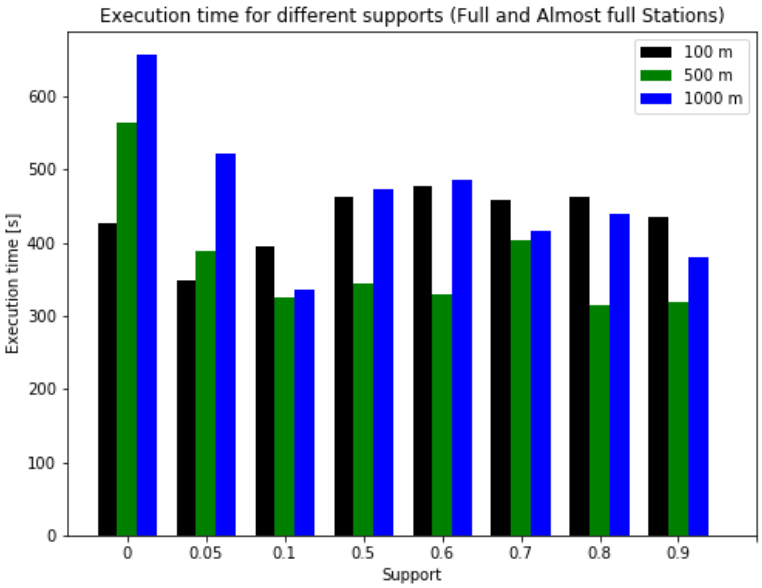


Figure 4.21: Execution time with full/almost full stations that change state every 30 minutes

number of maximum time windows and the spatial delta, it is reasonable that more patterns are present.

Figure 4.22 shows the average number of events inside the extracted patterns, considering full and almost full stations that change state. On average, the higher the spatial threshold, the higher the average number of events inside the patterns. Furthermore, the higher the supports, the lower the number of events inside the patterns. Starting from a support of 0.1, there are no events for spatial thresholds of 100 and 500 meters: there are only events for 1000 meters. Starting from a support of 0.5, instead, there aren't events for any spatial thresholds.

Figure 4.23 shows the maximum number of events in the patterns. Up to a support of 0.05 there are events for the three different spatial thresholds used. With 0 support, there are 5 events for all the three spatial thresholds considered. With a support of 0.05 it is visible a decreasing trend with respect to the previous analyzed support and at the same time the number of maximum events increases when the spatial threshold increases. With a support of 0.1 the only registered events are those with a spatial threshold of 1000 meters. Starting from a support of 0.5 no events are present.

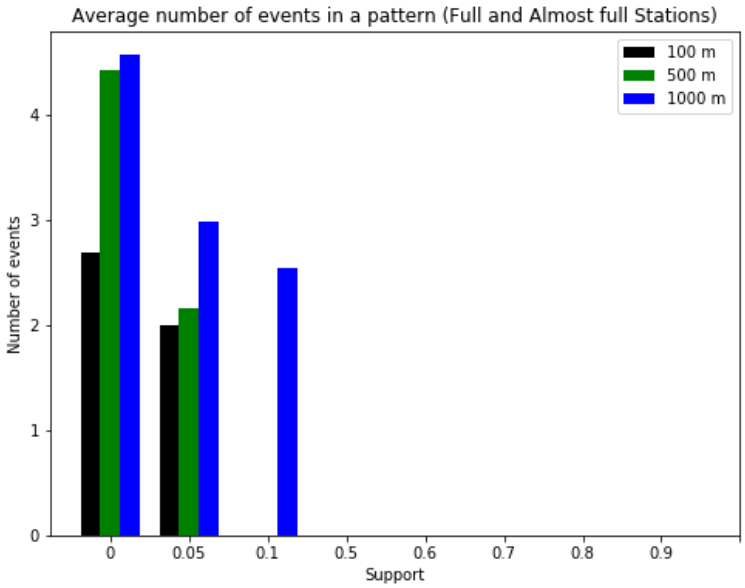


Figure 4.22: Average number of events with full/almost full stations that change state every 30 minutes

Figure 4.24 shows the minimum number of patterns with 3 spatial deltas, 3 time deltas every 30 minutes. The plot illustrates that the only present values are 2 and 0. Both supports 0 and 0.05 present the same behaviour: in all the three spatial thresholds, 100, 500 and 1000 meters, the minimum number of events is 2. With a support of 0.1 there are only events with a spatial threshold of 1000 meters. Eventually, starting from a support of 0.5 no patterns are extracted.

4.4.1 Example of an extracted pattern

The following pattern refers to a spatial threshold of 1000 meters, 0 support, 3 spatial delta, 3 time deltas and has 3 time windows: one at time t=0, one at time t=1 and the last one at time t=2. The first window presents 2 full events, of which the first one is the reference station and the second one is a station that is at a distance between 2000 and 3000 meters, The second window presents 2 full stations both at a distance between 1000 and 2000 meters; while the last window has only one almost full event at time t=2 at a distance between 1000 and 2000 meters.
$[['Full\_T0\_\Delta0, Full\_T0\_\Delta3'], ['Full\_T1\_\Delta2, Full\_T1\_\Delta2'],$
$['AlmostFull\_T2\_\Delta2']]$

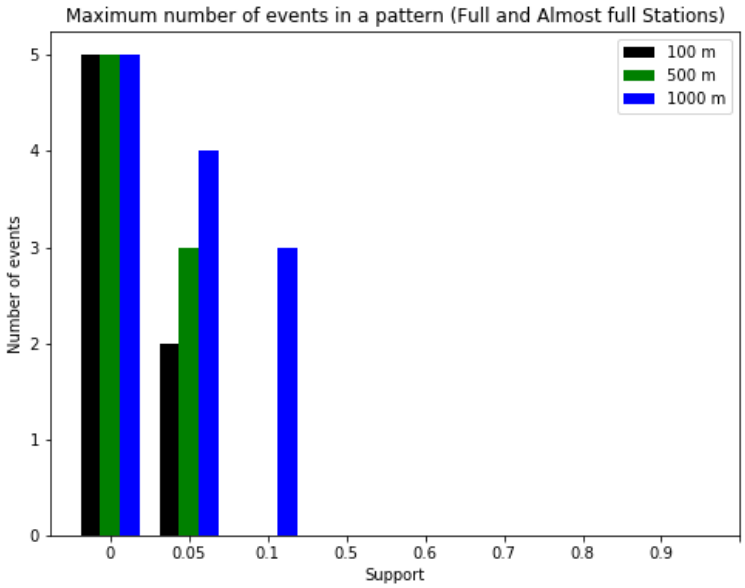


Figure 4.23: Maximum number of events with full/almost full stations that change state every 30 minutes

Visualization of the pattern

Figure 4.25 is a visualization of the pattern previously analyzed. The red circles represent events at time t=0, while blue circles concern time t=1 and the green circle time t=2. At the center of the circular sector there is the station with $\Delta 0$, the reference station; the circles that are positioned near 1000 are the stations that are within 1000 meters with respect to the reference station; the circles near 2000 are stations that are at a distance between 1000 and 2000 meters from the reference station, that is at the center; the circle near 3000 meters is the station that is between 2000 and 3000 meters.

4.4.2 3 spatial deltas and 4 time deltas

Other experiments concerned the use of different parameters for the spatial delta and time delta. Table 4.4 shows the results by using 3 spatial deltas and 4 time deltas. As seen in the previous cases, when the support increases, the number of patterns decreases or remains the same. This behaviour is due to a greater restriction of the number of filtered patterns. By using small values of supports, such as 0, 0.05 or 0.1, the number of patterns follows an increasing trend by increasing the spatial threshold. With a 0 support and 1000 meters of spatial threshold more

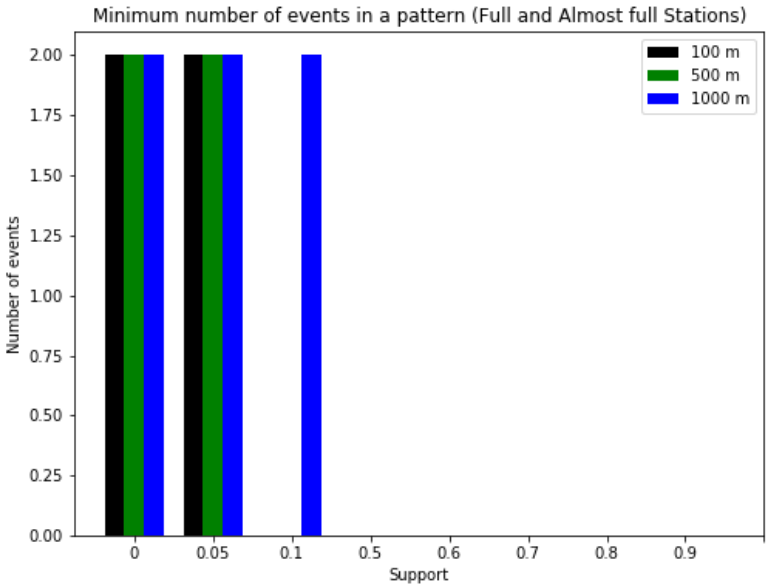


Figure 4.24: Minimum number of events with full/almost full stations that change state every 30 minutes

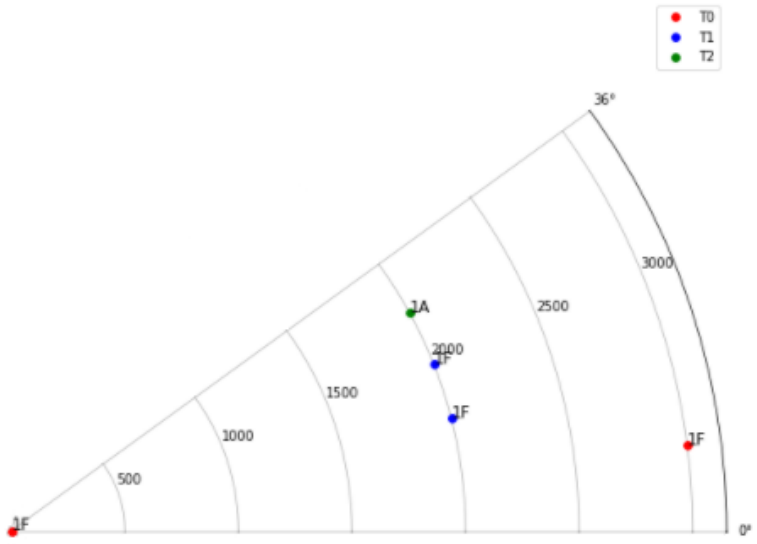


Figure 4.25: Visualization of a pattern with full and almost full stations that change state (3 spatial deltas-3 time deltas)

than 100 thousand patterns are present; while by using 0.9 as support and 1000 meters as spatial threshold, only one pattern is extracted.

spatial threshold [m]	**support**	**pre-filter**	**2 windows T0Δ0**	**Full & almost full events**	**S=0 T=0, S≠0 T≠0**
100	0	174	87	68	23
500	0	30391	11887	10494	7483
1000	0	132141	39721	36765	26904
100	0.05	6	2	1	0
500	0.05	27	6	1	0
1000	0.05	231	91	27	10
100	0.1	3	0	0	0
500	0.1	5	0	0	0
1000	0.1	52	21	1	1
100	0.5	3	0	0	0
500	0.5	3	0	0	0
1000	0.5	3	0	0	0
100	0.6	2	0	0	0
500	0.6	2	0	0	0
1000	0.6	2	0	0	0
100	0.7	1	0	0	0
500	0.7	1	0	0	0
1000	0.7	1	0	0	0
100	0.8	1	0	0	0
500	0.8	1	0	0	0
1000	0.8	1	0	0	0
100	0.9	1	0	0	0
500	0.9	1	0	0	0
1000	0.9	1	0	0	0

Table 4.4: Comparison between different filters applied at different spatial thresholds and supports with full and almost full stations that change state with 3 spatial deltas and 4 time deltas

Table 4.4 with a support of 0 presents many patterns with respect to the 0 support of Table 4.3. In fact, with a support of 0 and a spatial threshold of 1000 meters, in Table 4.4 the number of patterns is more than twice patterns in Table 4.3, in both the pre-filter and the post-filter cases.

However, by increasing the support, even if of small values, like 0.05, and by using 100 and 500 meters as spatial thresholds, there are less patterns in Table 4.4, with 3 spatial deltas and 4 time deltas, than Table 4.3, with 3 spatial deltas and 3 time deltas. This behaviour is due to the fact that, even if there are more patterns, with a higher support, are taken only those patterns that are more frequent. In other words, even if the algorithm extracts more patterns, there are less frequent patterns that repeat themselves.

In other cases, instead, with 1000 meters and 0.05 support or with 500 meters and 0.1 support, there are more patterns in Table 4.4 with respect to Table 4.3: there are more frequent patterns with 3 spatial deltas and 4 time deltas.

Figure 4.26 shows the number of patterns in function of different values of confidence. The trend is not so much different from the 3 cases analyzed in Figure 4.18, 4.19 and 4.20 with 3 spatial deltas and 3 time deltas, 1000 meters, support=0: in all cases it is clear a decreasing behaviour. This behaviour shows that when probability of the future events given that other events occurred is near 0 there are much more patterns; while, when the probability is near 1 there are not so many patterns. This implies that the future events are not so much dependant on the old events.

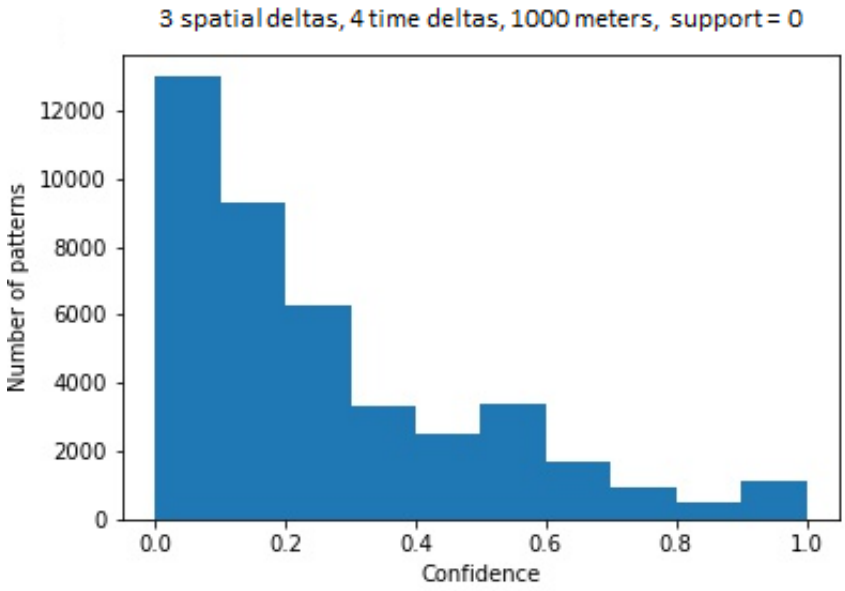


Figure 4.26: Extracted patterns with full and almost full stations that change state every 30 minutes for different values of confidence (3 spatial deltas-4 time deltas, 1000 meters and support=0)

The execution time can be visible from Figure 4.27. By using small values of support it is clear that the higher the spatial thresholds, the higher the number of supports. While, starting from 0.6 support, there is not a specific trend. In fact, when the support is 0.7 the number of extracted patterns with 100, 500 and 1000 meters is the same and the execution time does not follow a particular trend.

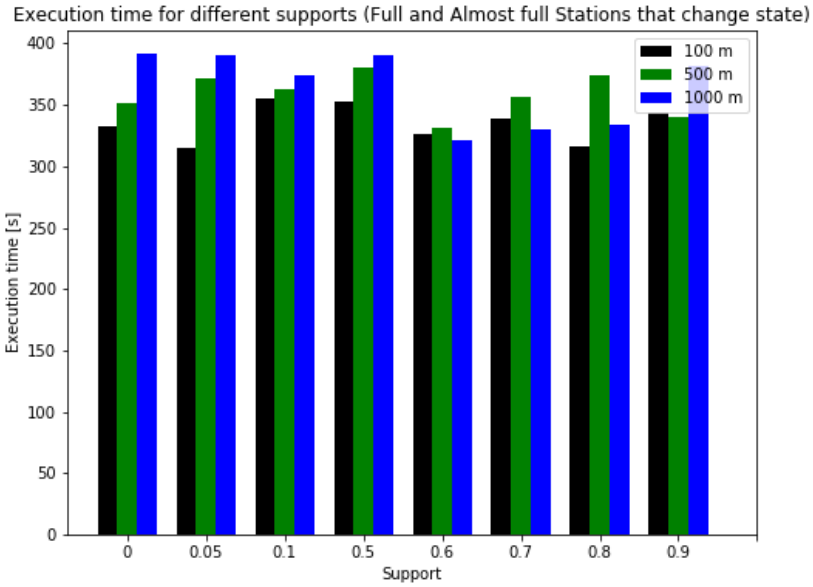


Figure 4.27: Execution time with full/almost full stations that change state every 30 minutes (3 spatial deltas, 4 time deltas)

The average number of events with full and almost full stations that change state every 30 minutes with 3 spatial deltas and 4 time deltas is shown in Figure 4.28. When the spatial thresholds increase, the number of patterns increases or is at least the same. When the support is 0.1 the only visible patterns are with 1000 meters. Starting from 0.5 supports, there aren't any visible patterns and, as a consequence, there are no events.
With a support of 0, the number of maximum events in a pattern is 5 with 100, 500 and 1000 meters as spatial thresholds, as shown in Figure 4.29. When the support is 0.05, instead, the maximum number of events in a pattern is 2; while, with 1000 meters, there are 5 maximum number of patterns. With a 0.1 support, there are at most 3 events in 1000 meters as spatial thresholds and 0 events with 100 and 500 meters. Starting from 0.5 supports there aren't any visible events in

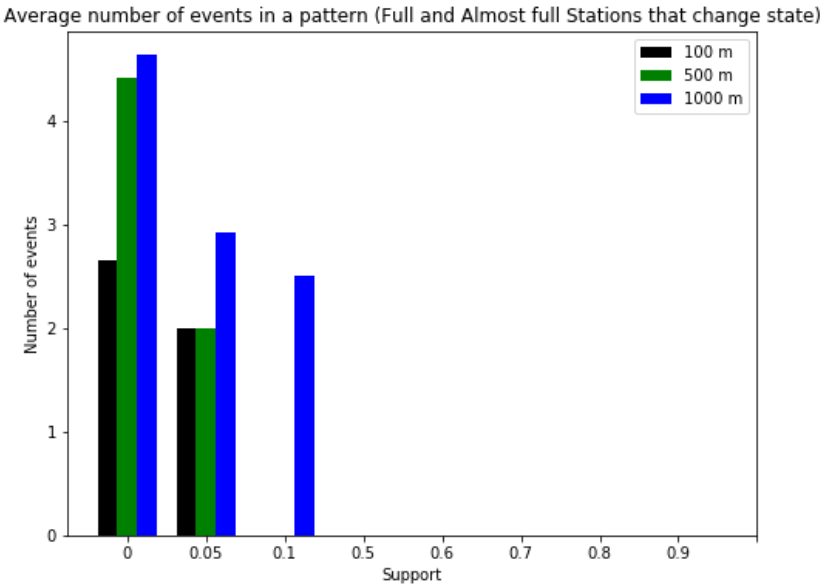


Figure 4.28: Average number of events with full/almost full stations that change state every 30 minutes (3 spatial deltas-4 time delta)

any considered spatial thresholds.
About the minimum number of events, instead, for every applied spatial threshold it is 2 for 0 and 0.05 supports. With 0.1 support, there are patterns only if the spatial threshold is 1000 meters. The absence of minimum number of events for 100 and 500 meters of spatial thresholds is not surprising: in fact, in Figure 4.30 for these spatial thresholds there are no maximum number of events, so there isn't neither a minimum of events.

4.4.3 Example of an extracted pattern

The following pattern refers to a spatial threshold of 1000 meters, 0 support, 3 spatial delta, 4 time deltas and has 3 time windows: one at time t=0, one at time t=1 and the last one at time t=3. The first window presents 2 full events, of which the first one is the reference station and the second one is a station that is at a distance between 1000 and 2000 meters. The second window presents one full station at a distance between 1000 and 2000 meters; while the last window has only one almost full event at time t=3 at a distance between 2000 and 3000 meters.
$[['Full\_T0\_\Delta0, Full\_T0\_\Delta1'], ['Full\_T1\_\Delta2']$
$['AlmostFull\_T3\_\Delta3']]$

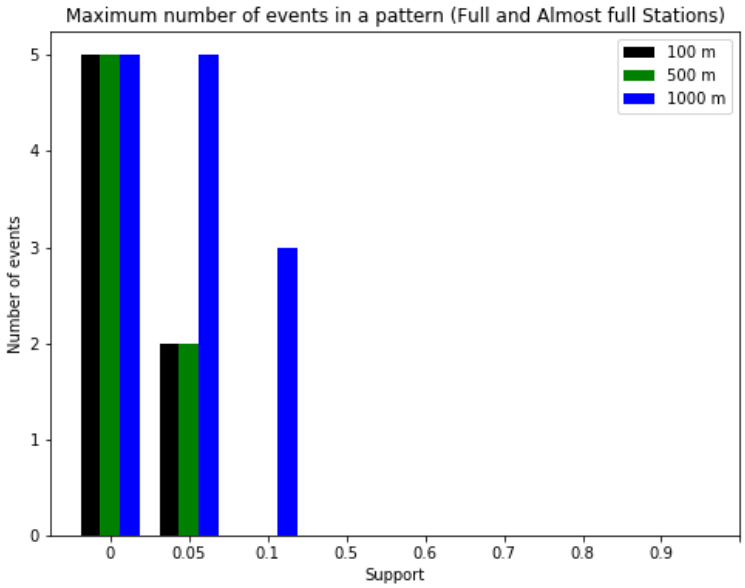


Figure 4.29: Maximum number of events with full/almost full stations that change state every 30 minutes (3 spatial deltas-4 time deltas)

Visualization of the pattern

Figure 4.31 is a visualization of the pattern previously analyzed. The red circles represent events at time t=0, while the blue circles concern time t=1 and the black one represents events at time t=3. At the center of the circular sector there is the station with $\Delta 0$, the reference station; the circle that is positioned near 1000 is the station that is within 1000 meters with respect to the reference station; the circle near 2000 is the station that is at a distance between 1000 and 2000 meters from the reference station, that is at the center; the circle that is positioned at 3000 meters is the station that is at a distance between 2000 and 3000 meters.

4.4.4 4 spatial deltas and 3 time deltas

Another experiment has been conducted with 4 spatial deltas and 3 time deltas. Table 4.5 shows the results by using 3 spatial deltas and 4 time deltas: in all these analyzed experiments, with different supports and spatial thresholds, there are more or at least the same number of patterns than the experiments with 3 spatial deltas and 3 time deltas and with 3 spatial deltas and 4 time deltas.
A reasonable explanation is that by increasing the quantity of spatial deltas, the search space is enlarged and there are more patterns that have been extracted.

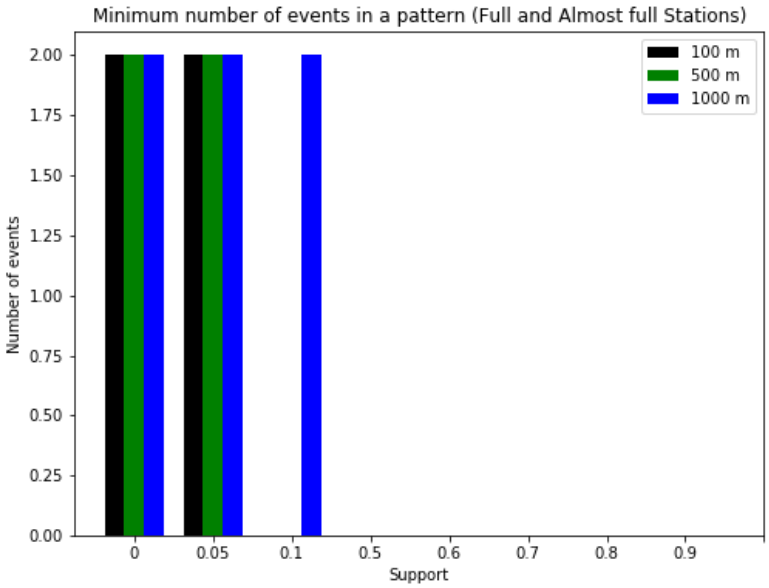


Figure 4.30: Minimum number of events with full/almost full stations that change state every 30 minutes

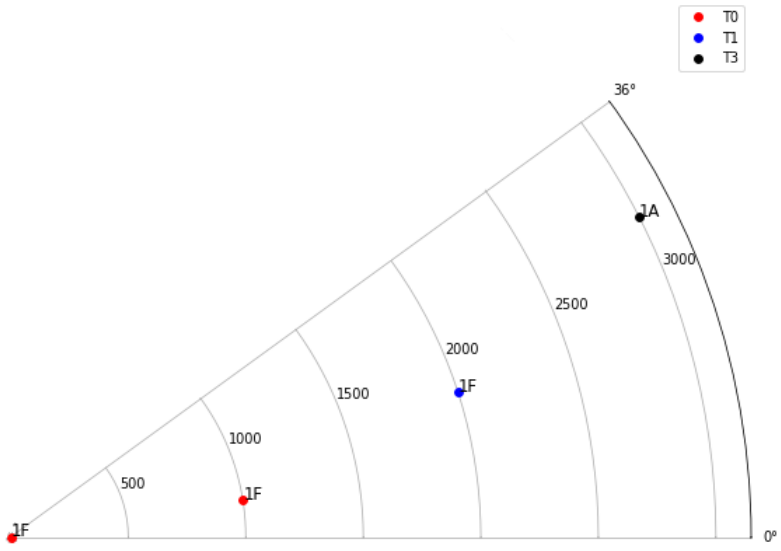


Figure 4.31: Visualization of a pattern with full and almost full events that change state (3 spatial deltas-4 time deltas)

Furthermore, the number of frequent patterns that have been extracted with 4 spatial deltas and 3 time deltas is more or at least the same number of frequent patterns than the experiments with 3 spatial deltas and 3 time deltas and with 3 spatial deltas and 4 time deltas.

spatial threshold [m]	**support**	**pre-filter**	**2 windows T0Δ0**	**Full & almost full events**	**S=0 T=0, S≠0 T≠0**
100	0	299	166	137	76
500	0	58573	19567	17762	12756
1000	0	144443	42011	39112	29899
100	0.05	7	2	1	0
500	0.05	61	13	5	1
1000	0.05	368	139	44	28
100	0.1	3	0	0	0
500	0.1	10	0	0	0
1000	0.1	86	28	3	2
100	0.5	3	0	0	0
500	0.5	3	0	0	0
1000	0.5	3	0	0	0
100	0.6	2	0	0	0
500	0.6	2	0	0	0
1000	0.6	2	0	0	0
100	0.7	1	0	0	0
500	0.7	1	0	0	0
1000	0.7	1	0	0	0
100	0.8	1	0	0	0
500	0.8	1	0	0	0
1000	0.8	1	0	0	0
100	0.9	1	0	0	0
500	0.9	1	0	0	0
1000	0.9	1	0	0	0

Table 4.5: Comparison between different filters applied at different spatial thresholds and supports with full and almost full stations that change state with 4 spatial deltas and 3 time deltas

Figure 4.32 shows the number of patterns in function of different values of confidence. It is shown a decreasing trend, as happened previously with 3 spatial deltas and 3 time deltas (Figure 4.18, Figure 4.19 and Figure 4.20) and with 3 spatial deltas and 4 time deltas (Figure 4.26).

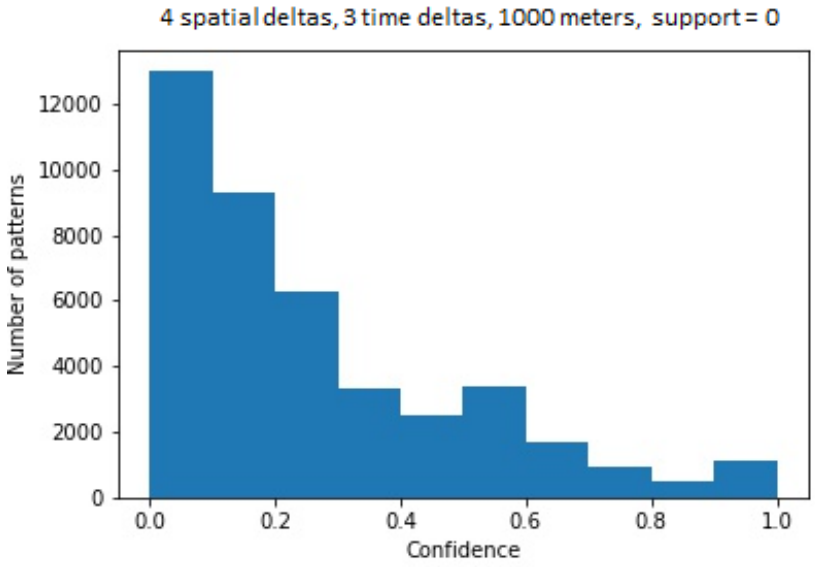


Figure 4.32: Extracted patterns with full and almost full stations that change state every 30 minutes for different values of confidence (4 spatial deltas-3 time deltas, 1000 meters and support=0)

The execution time can be visible from Figure 4.33. By using a support of 0 there are much more patterns than the other supports, especially for 500 and 1000 meters as spatial thresholds, and this consideration reflects the execution time. Up to 0.1 support the trend shows an increasing behaviour when the spatial threshold increases.
The average number of events with full and almost full stations that change state every 30 minutes with 4 spatial deltas and 3 time deltas is shown in Figure 4.34. When the spatial thresholds increase, the number of patterns increases or is at least the same. When the support is 0.1 the only visible patterns are with 1000 meters. Starting from 0.5 supports, there aren't any visible patterns and, as a consequence, there are no events.
With a support of 0, the number of maximum events in a pattern is 5 with 100, 500 and 1000 meters as spatial thresholds, as shown in Figure 4.35. When the support is 0.05, instead, the maximum number of events in a pattern is 2 with 100 meters as spatial thresholds; while, with 500 meters, there are 3 maximum number of events with 500 meters; with 1000 meters there are 4 maximum number of events. With a 0.1 support, there are at most 3 events in 1000 meters as spatial thresholds and 0 events with 100 and 500 meters. Starting from 0.5 supports there aren't any visible events in any considered spatial thresholds.
About the minimum number of events, instead, for each applied spatial threshold

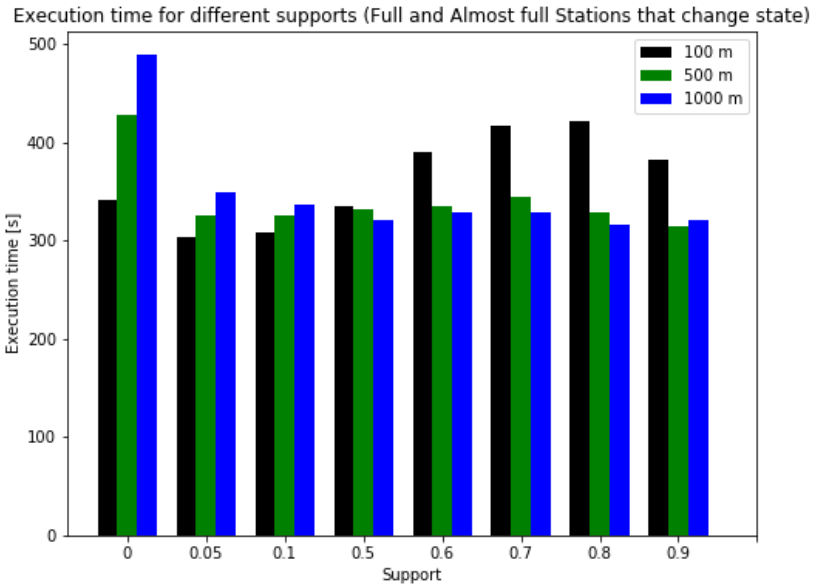


Figure 4.33: Execution time with full/almost full stations that change state every 30 minutes (4 spatial deltas, 3 time deltas)

it is 2 for 0 and 0.05 supports. With 0.1 support, there are 2 patterns only when the spatial threshold is 1000 meters.

4.4.5 Example of an extracted pattern

The following pattern refers to a spatial threshold of 1000 meters, 0 support, 4 spatial delta, 3 time deltas and has 2 time windows: one at time t=0 and one at time t=4. The first window presents 2 events: one almost full station at a distance delta 0 and one full station at a distance within 1000 meters. The second window presents one almost full station at a distance between 3000 and 4000 meters at time t=2.
$[['AlmostFull\_T0\_\Delta0, Full\_T0\_\Delta1'], ['AlmostFull\_T2\_\Delta4']]$

Visualization of the pattern

Figure 4.37 is a visualization of the pattern previously analyzed. The red circles represent events at time t=0, while the green circle concerns time t=2.

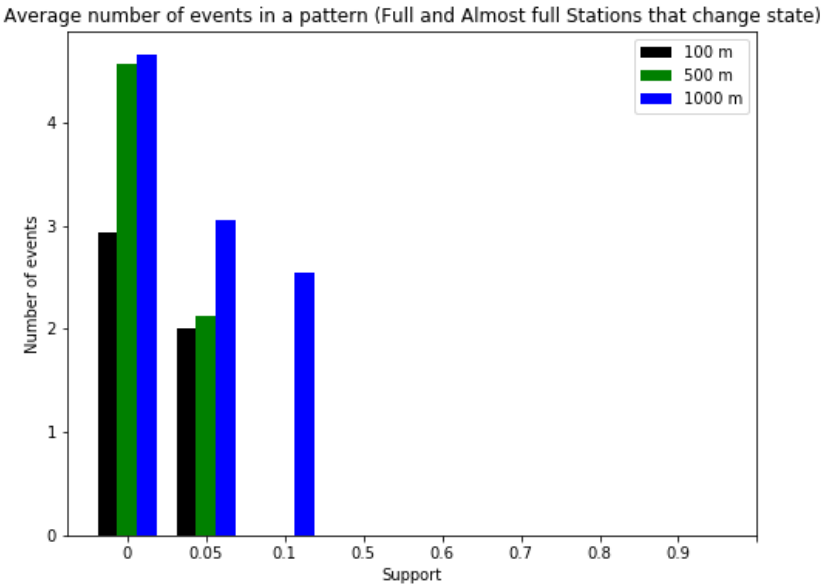


Figure 4.34: Average number of events with full/almost full stations that change state every 30 minutes (4 spatial deltas-3 time delta)

4.5 Comparison between the two strategies of full and almost full stations

Figure 4.10 and Figure 4.11 have a completely different trend from Figures 4.18, 4.19, 4.20: the first two figures exhibit an increasing trend, while the last three a decreasing one. The increasing trend shows that there is a pick toward 0.95-1 of confidence. The conditional probability is given by this formula: $P(A|B) = \frac{P(A \cap B)}{P(B)}$. When $P(A \cap B) = P(B)$ the conditional probability $P(A|B)$ is 1: this means that the whole is conditioned by the part $P(B)$, a subset of the set A. In this specific case, when the probability of a future event, given that other events occurred, is 1 or near 1, these last events are a subset of the future events. On the other hand, when the conditional probability is 0, the two events A and B are disjoint, there are no events in common. In the analysis a zero conditional probability means that the future events and the previous events have no elements in common.

In the last analysis, in which full and almost full stations that change state are considered, with a slightly higher value of support like 0.05 the number of extracted patterns decreases a lot, especially if 100 meters are used as spatial threshold and with a support of 0.10, there are even less patterns. A reasonable

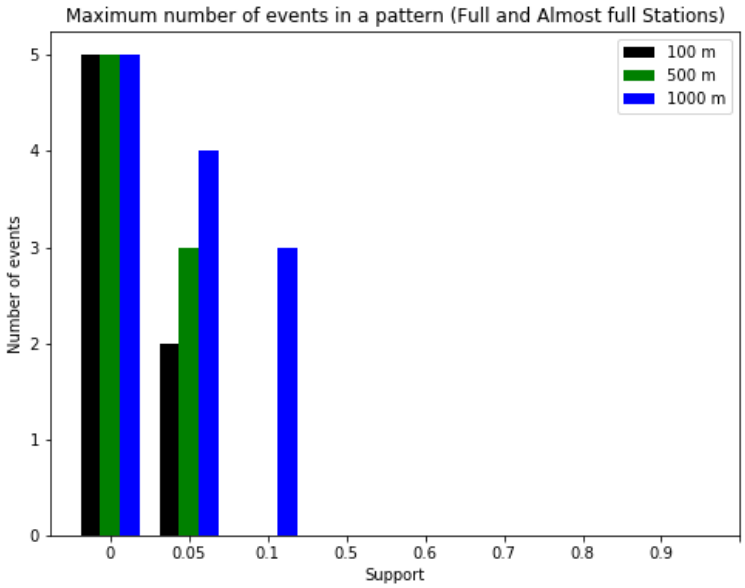


Figure 4.35: Maximum number of events with full/almost full stations that change state every 30 minutes (4 spatial deltas-3 time deltas)

explanation is that the extracted patterns are present only when there is a change of state and since the state of a station very often remains the same, the number of patterns decreases. In the previous analyses with all full and almost full stations, instead, all patterns are considered, not only those that change state when going from a timestamp to one other: the number of extracted patterns is much higher, as shown in Table 4.2.

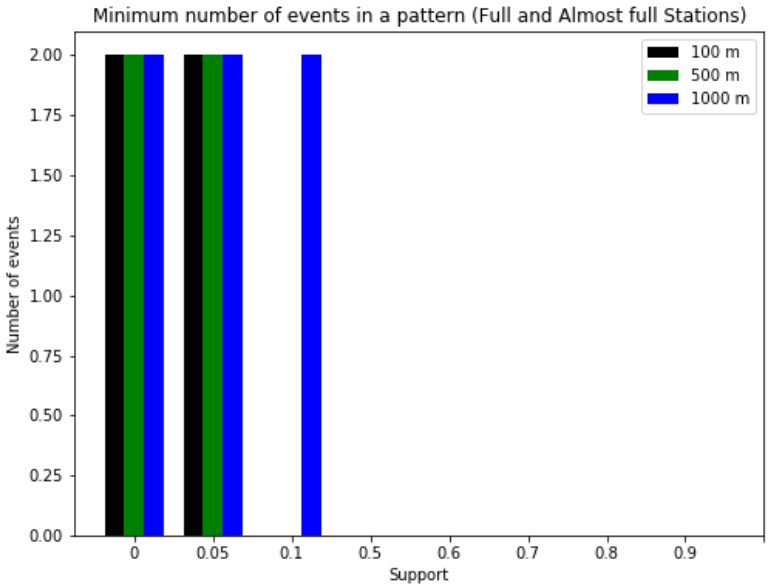


Figure 4.36: Minimum number of events with full/almost full stations that change state every 30 minutes (4 spatial deltas-3 time deltas)

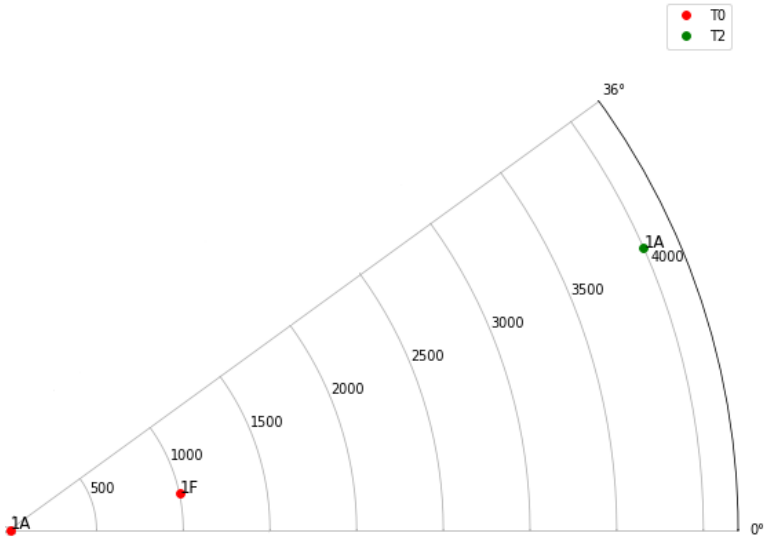


Figure 4.37: Visualization of a pattern with full and almost full stations that change state (4 spatial deltas-3 time deltas)

Chapter 5

Conclusions

5.1 Results

This thesis allowed me to carry out what I had intended to do.
Some experiments have been performed by using different spatial thresholds, spatial deltas, time deltas and time intervals. The extracted patterns have been saved into some files, in order to store them into external documents. Once these patterns have been extracted, some filters have been applied on them and their results were saved into some external files.

The number of extracted patterns just after the algorithm has been applied in all the experiments is higher compared to the other results when the filters have been considered. This behaviour reflects what expected: in fact, the filters reduce the number of extracted patterns, by taking only those of interest.

Immediately after the patterns have been extracted from the Prefix Span algorithm, patterns have been filtered and the number of patterns is reduced, as expected.

By increasing the spatial delta, there are much more patterns that have been extracted, especially if the support is 0: more stations are considered when the spatial delta increases.

By increasing the time delta, instead, more subsequent time intervals are considered and with a 0 support there are more patterns. When the support increases, there is not always an increasing trend because, even though there are much more patterns, they might be not so much frequent.

After having conducted different analyses, on average, the results show that the number of patterns increases as:

1. the spatial delta, time delta and time threshold increase because the search space is enlarged
2. the support decreases because there are less restrictions that allow to accept some patterns even though they are not so much frequent

5.2 Future works

In future some extensions could be applied on this thesis work. The proposed method it is a general one and can be applicable to other data-sets containing spatial and temporal information. An implementation could be the analysis of other data-sets of different cities to compare the extracted patterns and their relative statistics.

Bibliography

[1] Jose L. Walteros. *Improving the Service Quality of Bike Sharing Systems via the Analysis of Real-Time User Data*. 2018 (cit. on p. 3).

[2] Bike Share Map. URL: `https://bikesharemap.com/` (cit. on pp. 3, 4).

[3] Ahmadreza Faghih-Imani†, Robert Hampshire, Lavanya Marla, and Naveen Eluru†. *An Empirical Analysis of Bike Sharing Usage and Rebalancing: Evidence from Barcelona and Seville*. 2015 (cit. on p. 4).

[4] Susan A. Shaheen, Elliot W. Martin, and Adam P. Cohen Rachel S. Finson. *Public Bikesharing in North America: Early Operator and User Understanding*. 2012 (cit. on pp. 8, 11).

[5] Susan A. Shaheen, Stacey Guzman, and Hua Zhang. *Bikesharing in Europe, the Americas, and Asia: Past, Present, and Future*. 2010 (cit. on p. 9).

[6] Jiawei Han, Micheline Kamber, and Jian Pei. *Data Mining Concepts and Techniques*. 2012 (cit. on p. 12).

[7] Jian Pei, Jiawei Han, Behzad Mortazavi-Asl, and Helen Pint. *PrefixSpan: Mining Sequential Patterns Efficiently by Prefix-Projected Pattern Growth*. 2011 (cit. on p. 12).