POLITECNICO DI TORINO

Master Courses in
ICT for Smart Society

Master Thesis

Dimensioning Cellular Networks Photovoltaic Power Supply through Clustering Techniques

Supervisors:
Prof. Michela Meo
Greta Vallero

Candidate:
Edoardo Maria Sanna
ID Number: s245161

March 2020

Contents

List of Figures

List of Tables

Chapter 1

Introduction

With the huge growth of mobile technology in everyday life, the network efficiency of communication systems takes a very important role, in particular for the fruition of services that become more and more essential to life. Just think of all that range of applications that we use every day through the interface of a smartphone, and for which a solid, reliable and well-structured communication infrastructure is needed. To better understand, figure 1.1 shows the result of Ericsson Mobile Report [4], in which we can clearly see the enormous increase in terms of traffic consumption estimated by the year 2022, *[GB per month/per user]*.

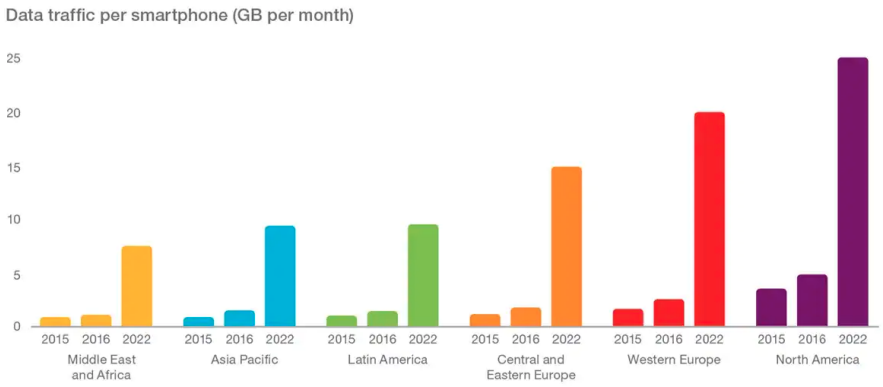


Figure 1.1: LTE mobile traffic consumption trend

Because of what we have previously said, the *Internet Server Providers* (ISP), have to face the problem of the continuous network improvement to make it efficient, reliable, and able to provide a high quality of service QoS. To achieve

these goals, it is necessary to increase the number of *Base Stations*, both for reasons of tradable traffic capacity and for coverage of new areas. With the increasing number of BSs, therefore, many tasks become more complex for those people who design and manage cellular networks.
Some typical examples of Machine Learning (ML) applications which can be helpful for ISP companies are: specific installation of renewable power supplies, energy saving algorithms thanks to the identification of similar patterns or particular behaviors, predictive maintenance thanks to the automatic identification of the risk of either outages or faults. All this seemed impossible until not so far ago. In fact, thanks to Machine Learning algorithms it is now possible to understand behaviors that the human eye is not sensitive to, similarities and correlations which are difficult to be detected when dealing with a large amount of data or understand the cause that leads to a certain event. This opens up a range of implementable applications: predictive maintenance, outages prediction, intelligent design, energy-saving algorithms, dimensioning of power supply plants, etc...
This thesis work follows this path: using different techniques of Machine Learning, in particular, Clustering Techniques, we aggregate traces of traffic by different features (Pattern, Traffic Consumption, Exceeding the consumption threshold, etc...) in order to apply different energy-saving algorithms to all the traces belonging to the same cluster, and therefore, hypothetically, with the same characteristics.
Once identified the characteristic properties of each cluster, a study upon them, in terms of their energy consumption considering both annual and seasonal consumption, has been conducted. Subsequently, the same analysis has been carried out in the hypothesis of traces coming exclusively either entirely from macrocells or entirely from microcells.
The final goal of this work is to be able to dimension a photovoltaic power supply system based on the results of the cluster analysis. By being able to identify the best size of a photovoltaic power supply, starting from the characteristics of the individual clusters, it would be possible save time since the design operator is provided with an initial indication of the size of the optimal plant. In order for this procedure to be applied, it is necessary to know only few characteristics of the traffic traces related to the area where the installation is intendeed to be made, that is the pattern of the traces and the behavior of the traces with respect to the threshold.
This is just one of the possible applications deriving from this type of analysis, as a matter of fact there are many interesting other perspectives which can be followed. As an example, one could focus the analysis on the identification of which areas need intensification of the towers fleet, or again which is the best area

whether to install a photovoltaic system or not. However, the applications are not only limited to the technological aspects but they also embrace the sociological ones. In fact, through this analysis the population behavior can be studied for example to understand the people flows in different areas and at different times (both of the day and of the year). All this to say that the potential applications of the custer analysis regarding cellular networks are numerous.
In the next chapter, a brief introduction to the technologies covered and the used tools is provided.

Chapter 2

What we are talking about

Before getting in to the hearth of my work, it is a must to introduce the themes and the technologies, tools and techniques utilized to reach my goals.

2.1 Cellular Network

A cellular network is a communication network that allows the communication point to point through a mobile phone and a wireless medium. In fact, all the communication in object are radio ones. Trying to cover all the spaces, the cellular network divide the territory roughly into small areas called "cells", each served by a Base Transceive Station, from now on BS. The name "Cellular" refers in fact, the technique used to cover the territory.
Due to the everyday higher request of bandwidth, it has become more and more difficult to well fit well people needs. So ISP have to introduce new cells, reducing the dimension of the previous ones, increasing the number of BS in that area.

2.1.1 Cell Types

Principally, the cell types are two: Macro Cells and Micro Cells.There also exist other types of cells, smaller and bigger, but we are interested in only the first two.

- **Macro Cells** : is a cell that provides coverage by an high power cell site for a large amount of territory. Usually the antennas for these kind of cells are ground-based mounted or roof-base mounted, due to their large dimension. Macro cells are large in output power.
- **Micro Cells** : The main difference the micro and macro cells, is the purpose. If the macro cell has the task of cover a very large area, the

micro cell covers very little areas such as mall or hotel. The success of the micro cell is reached with the advent of Long-Term Evolution (LTE/4G) technology. In figure 2.1 the cellular network coverage system architecture is shown. [6].

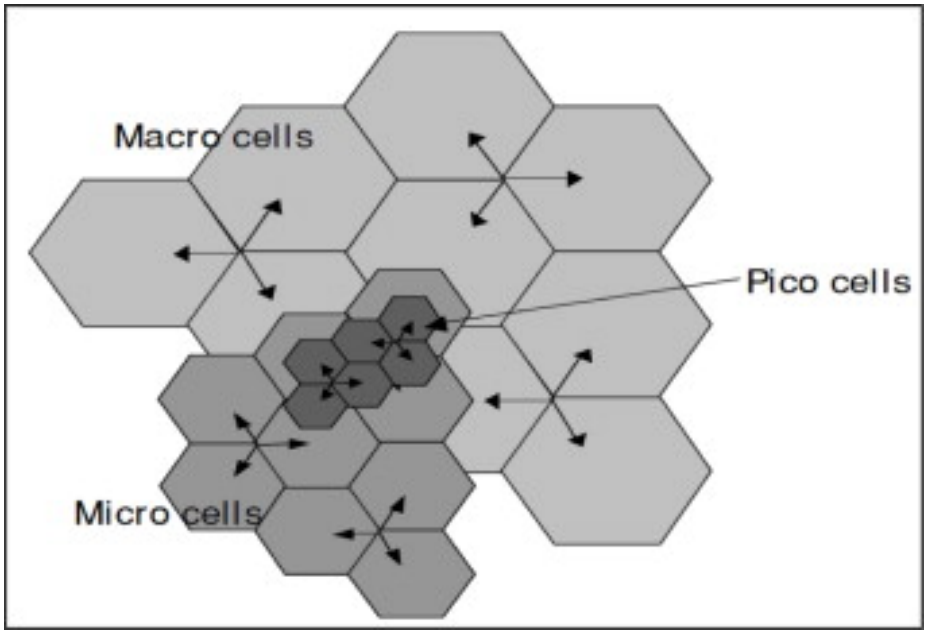


Figure 2.1: Cellular Network coverage system

2.1.2 Standards

- **2G** : It is the first full digital wireless communication standard. Thanks to this innovation, more data service were introduced, the famous one is the Short Message Service, better known as SMS, that is a technology for exchanging information through plain-text messages. The 2G were commercially launched on the GSM standard in Finland in 1991.

- **3G** : This standard increases principally the ratio of transmission that used to reach 114 *kbit/s*. With the later releases, throughput reaches several *Mbit/s*. This amount, makes possible a stable internet access from mobile, video calls and mobile TV technologies. The first release of the 3G were introduced in 1998.

- **4G** : 4G provides new features with respect to the services introduced by 3G, for example mobile broadband access for laptops with wireless modem.

A wide range of new possibilities were reached with the advent of 4G: IP telephony, gaming service, high-definition mobile TV, cloud computing, etc...
It reaches a ratio of 100 *Mbit/s*, and it was introduced in 2008, but in Italy comes up only in 2012.

2.1.3 Power Consumption

In [ref-articolo vallero-meo], a LTE tower power consumption model were presented. That model was used in my work to understand the consumption of each "aggregation of towers" in an area, considering each area in both cases, as Macro and Micro cell.

$$P_{in} = N_{trx} * (P_0 + \Delta_p P_{max} \rho), \qquad 0 \leq \rho \leq 1 \tag{2.1}$$

The equation 2.1 taken by [14], shows the relation between the percentage of traffic exchanged with respect to the maximum capacity of the tower, to understand the input power needed. In the table, are explained the meaning of the variables that compose the equation.
P_{in} is the input power required to power the BS, P_0 is the power required to power the tower in Idle, i.e. when no antenna is transmitting or receiving a signal, N_{trx} is the number of transmitters in the Base Station, Δ_p is the load slope dependent on power consumption, P_{max} is the power required by the single transmitter when maximum capacity is reached, eventually ρ is the traffic load.
In the table 2.1, the parameter values are summarized both in case of a Micro cell and in case of a Macro cell.

BS Type	N_{trx}	P_{max} **(W)**	P_0 **(W)**	Δ_p
Macro	6	20	84	2.8
Micro	2	6.3	56	3.6

Table 2.1: Row data parameters

2.2 Solar Irradiance

Solar radiation is the energy emitted by the sun, generated by thermonuclear fusion reactions that take place in the Star and produce electromagnetic radiation at different wavelengths, which propagates into space carrying solar energy.
In order to take advantage of the opportunites offered by this energy, that we

can find in nature and in great quantity, man has been at work since its origins, maturing more and more a deep knowledge. For more than half a century now, researchers have found a way to convert solar energy into electrical energy, and to exploit it for his needs. Although the invention of the solar panel dates back to the early 1990s, it is only in recent years that the real potential of this and other low environmental impact technologies has been understood, and efforts have been made to heavily invest in these technologies.
In the figure 2.2, extract from [5] we can in fact appreciate the increasing trend of solar energy produced in recent years, and a rosy perspective for the coming years. The scenarios presented in figure are two, a bad one and a good one, respectively coloured with a dark and a light yellow. The evolution towards one or the other scenario depends mainly on the amount of *GW* of power installed in the coming years.
This graph comes from a study that is annually conducted by *Solar Power Europe*, which elaborates an account of the last seven years and produces a prediction of the future five ones.

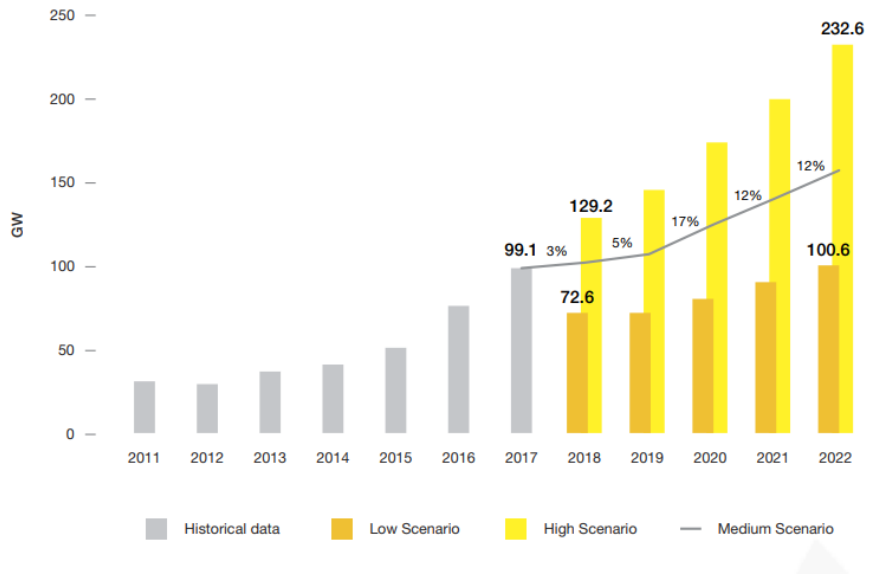


Figure 2.2: Solar energy production trend

2.2.1 Photovoltaic Solar Panel

The solar panel is an optoelectric device capable of converting solar energy into electrical energy through the photovoltaic effect.
The atomic element of which a solar panel is composed is the *photovoltaic cell*

or *solar cell*. The panels commonly available on the market consist of 46 to 96 cells each. To give an idea of what the photovoltaic effect is, we can say that it is a physical phenomenon that occurs when an electron passes from the valence band to the conduction band due to a sufficiently energetic photon incident on the material.
In order to calculate the amount of power generated by a solar panel, in first approximation we can use the formula expressed by the equation 2.2.

$$P = \eta I_0 sin(\alpha) S \tag{2.2}$$

- I_0 : it's the solar irradiance perpendicular to the the direction of the solar rays expressed in $Watt/m^2$
- α : it is the angle of inclination with respect to the incidence solar irradiance
- S : is the surface of the PV panel expressed in m^2
- η : is the efficiency that is a performance indicator

This type of technology is used today for different aims and purposes, in the context of my work, it will be considered the case in which this technology is used for the power supply of the Base Stations described above. In fact it is increasingly common practice to apply this element to the towers that make up the cellular network, trying to understand with different technologies to which repeaters it is more convenient to apply them.

2.3 Cluster Analysis

To identify similar patterns and behaviors among the different BSs under study, it was decided to approach the problem with cluster analysis in order to emerge the above-mentioned characteristics. Once you have decided to face a problem with cluster analysis, you have to ask yourself some questions and make some choices, among the most important ones, the choice of the features and the clustering algorithm.
Basically there are four processes to be carried out [12]:

- Identify and choose the features that are most representative, or that you want to highlight, to be taken into account to group the various elements.
- Once the choice of the characteristics has been made, it is necessary to proceed with the choice of the metric that will indicate the similarity or dissimilarity of one element with respect to another. One of the most common and used is the Euclidean distance.

- Proceeding to the next step, the third phase is the choice of the grouping algorithm.
- Finally, to conclude our study, we must validate the obtained results.

In the next paragraphs, we will analyze point by point the various phases of a cluster analysis, referring to the algorithms and metrics used during this thesis work.

2.3.1 Identification and choice of features

Although in recent years the piers of data generated and exchanged around the world have grown dramatically and although the technologies for accessing such data have changed, the structures with which they are identified and structured have remained almost the same. These structures are the matrices. In the matrices each row corresponds to an observation, an element or a sample, while each column defines a characteristic of that row, the so-called features.
Obviously the choice of features will be made in relation to the objective of the analysis. In the specific case of this work, the features defining the trend pattern of each single track were used in the first analysis, i.e. the sampling at constant time intervals (15 minutes) of the traffic volume exchanged in a particular area. In second analysis, it was chosen to introduce, among the clustering features, parameters indicating the time when the traffic of a particular area was above a certain consumption threshold and the volume of traffic exchanged above that threshold.

2.3.2 Identification and choice of similarity criterion

The general objective of clustering is to group elements together, in a certain number of groups, or rather clusters, according to a certain criterion of similarity or dissimilarity.
This type of calculation can become very heavy in the case of a large number of samples and a large number of features, and it can take some time for the machine to complete this task. For this reason it is convenient to perform it only once, going to build a symmetrical matrix that contains all the combinations of similarity between each elements. In this case it will not be necessary to recalculate these values but it will be enough to extract them from the matrix. Below a clarifying example 2.3.

$$P = \begin{bmatrix} p_{11} & \dots & p_{1n} \\ p_{21} & \dots & p_{2_n} \\ \dots & \dots & \dots \\ p_{n1} & \dots & p_{nn} \end{bmatrix} \tag{2.3}$$

Once of the mostly used indicator of similarity, already mentioned before, is the Euclidean distance. This distance indicates how close two elements are on an N-dimensional plane. The smaller this distance is, the more similar to each other these samples are. The distance will be equal to zero when we calculate this parameter between two equal elements.
However, since it is not a normalized distance, i.e. with values included in a range, using this parameter, we have to think about what are the ranges of values we will have to deal with. Depending on the plane and the type of data we are using, a distance that at first glance may seem very large, and therefore give us the idea that we are dealing with two very different elements, could be small in reality. For this reason using such metrics it is good to pay attention to the order of measurement. The equation 2.4 shows how to calculate this parameter between two elements.

$$d(i,j) = \sqrt{\sum_{k=1}^{n}(x_{ik} - y_{jk})^2} \tag{2.4}$$

Where:

- $d(i,j)$: is the Euclidean distance between element i and element j
- x_{ik} : is the k feature of the element i
- y_{jk} : is the k feature of the element j
- n : is the length of the elements i,j

2.3.3 Choice of aggregation algorithm

Generally speaking, a Clustering algorithm comes out from the combination of the choice of the proximity metrics of the elements and the choice of a basic criterion to be used to group these elements. By defining these factors, we will implicitly decide the "type of groups" we want to form.
Our goal will be to choose the metrics and the algorithm that will create Clusters that better respect the structure and meaning of the data, and in particular that can enhance all those features that were identified previously in the choice of the features.

Obviously, these combinations between the type of grouping and choice of metric have been extensively studied. A wide range of algorithms has been developed and presented in the literature.
Unfortunately, there is no specific algorithm for each type of problem able to meet all the needed requirements. Furthermore, it is impossible to derive a mathematical equation that, according to the type of data and the fixed target, allows to derive the optimal clustering algorithm. At this point, it is important to reduce the possible amount of choices to a small number and proceed with experimental tests. Only after the accurate analysis of the experimental results, the algorithm that suits the problem the most can be chosen. In this work two types of algorithms have been tested. Which are **K-Means**, adopting a few variants in terms of comparison metrics, and an algorithm of **gerarchical ascending** aggregation.
The figure 2.3 shows the result of applying different clustering algorithms to the same datasets. Depending on the algorithm used, the result is different and the clusters are not always the same but are subject to different logics.

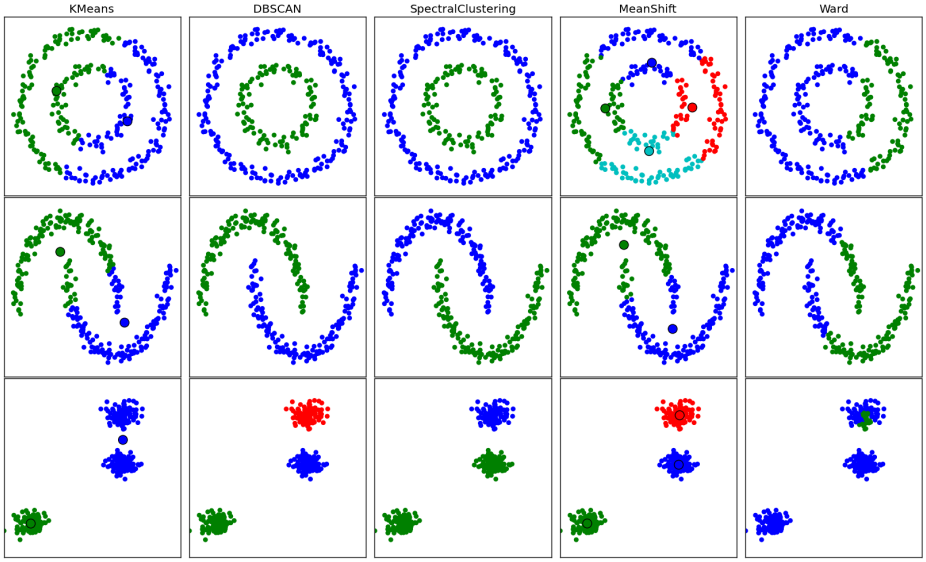


Figure 2.3: Clustering Technique Comparison [10]

2.3.4 Validation of results

Once the clusters are generated, it is important to find a method that can define how well these clusters are actually formed.
Cluster validation is a very important procedure and becomes fundamental in two cases. The first is when you cannot visually find an association for the elements

that make up a cluster. The second case, which is the one presented in this thesis work, is when we need a value to help us determine the optimal number of clusters. In our case, in fact, having used two algorithms in which the choice, of the number of clusters, must be made a priori, this type of evaluation is essential. The two methodologies utilized are **Davies Bouldin Index** [3] and the **Elbow Method** [8].

2.4 Clustering Algorithms

In this section, I will present the two clustering algorithms used in analysis and testing in my work. The first one is a hierarchical clustering algorithm and it is presented in [15] which aims to aggregate the two most similar paths at each iteration. The second is a classic K-Means algorithm that has proved to be very useful in this kind of work.

The starting point for both algorithms is a list of features vectors, which in our case will be the aggregate consumption traces of the various areas. The data we will have to deal with will then be widely discussed in chapter 4.1.

The primary objective of both these algorithms, in this context, will be to be able to put together all those tracks that have a similar pattern, thus being able to identify a more or less small number of key patterns. Patterns to which each type of track can be related to.

2.4.1 Hierarchical Ascendant Clustering Algorithm

This algorithm has as its starting point, as we have just said, in the set of tracks we have in our possession.
When the algorithm starts its life cycle, it considers each input element as a single cluster composed of the element itself. Subsequently, in an iterative way, the two clusters that have the smallest distance between them are merged into a single cluster. This iterative process will continue until the stop condition is reached, i.e. until the minimum distance at each step continues to be below a certain threshold.
It was decided to use the Euclidean Distance as distance metrics, equation 2.4, and to consider the centroids of the various clusters to calculate the distance between them.
The system, discussed before is presented in the algorithm in figure 2.4.

Algorithm 1: Traffic Patterns Identifier

Input: Cell towers number M, Threshold value T,
Traffic vector X_i, for $i = 1, 2, 3...M$
Output: Cluster labels of tower i, L_i, for $i = 1, 2, 3...M$
Initialize:
  Clusters: $c_k \leftarrow [X_k]$, for $k = 1, 2, 3...M$,
  Cluster set: $C \leftarrow [c_1, c_2...c_M]$
  Cluster number: $N \leftarrow M$
  Distance matrix: $D \leftarrow Inf$
  Stop index: $stop \leftarrow false$
while $stop == false$ **do**
  $D \leftarrow Inf$
  for $\forall c_i, c_j \in C, i \neq j$ **do**
    $D_{i,j} \leftarrow compute\_distance(c_i, c_j)$
  $[Mindistance, index1, index2] \leftarrow find\_min(D)$
  if $Mindistance > T$ **then**
    $stop \leftarrow true$
    $break$;
  $merge(c_{index1}, c_{index2})$
  $N \leftarrow N - 1$
for $i = 1$ to N **do**
  for $\forall X_k \in c_i$ **do**
    $L_k \leftarrow i$
Return L

Figure 2.4: Hierarchical Algorithm Pseudocode [15]

2.4.2 K-Means Algorithm

The second algorithm I am going to present is the K-Means. To be able to use this tool you need to be sure that the objects you want to group can be represented as vectors, and therefore form a space vector.
The main objective of this algorithm is to minimize the intra-cluster distance, i.e. the distance between all the elements within a single cluster. The one who will use this tool will have to choose at the beginning the value of K, which is the number of clusters into which the elements of our whole dataset will be divided. At the first iteration, these clusters will be initialized either with random vectors or with elements randomly extracted from the total population. Each cluster is identified by its centroid, which is the barycenter of all the elements that compose it. In this case, at the beginning, the centroid will coincide with the element itself. It should be noted that K-Means is also an iterative algorithm. At each iteration, the distance between each element of the population and each centroid is calculated. This sample is then assigned, obviously, to the cluster that

has the minimum distance. Also in this case, for uniformity, it was decided to use the Euclidean distance as metric, equation 2.4.
At the end of each iteration, the new centroids of the newly formed clusters are calculated, and all objects are reassigned until the stop condition is met. In this case, the Mean Square Error (MSE) equation 2.5, had to stand under a certain threshold.

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(Y_i - \widehat{Y}_i)^2 \tag{2.5}$$

In figure 2.5 is presented the utilized algorithm.

K-MEANS(P, k)

Input: a dataset of points $P = \{p_1, \ldots, p_n\}$, a number of clusters k
Output: centers $\{c_1, \ldots, c_k\}$ implicitly dividing P into k clusters

```
1  choose k initial centers C = {c_1, ..., c_k}
2  while stopping criterion has not been met
3      do ▷ assignment step:
4          for i = 1, ..., N
5              do find closest center c_k ∈ C to instance p_i
6                  assign instance p_i to set C_k
7          ▷ update step:
8          for i = 1, ..., k
9              do set c_i to be the center of mass of all points in C_i
```

Figure 2.5: Hierarchical Algorithm Pseudocode [1]

2.4.3 Davies-Bouldin Index

Unfortunately, we are not a priori aware of the optimal number of key patterns we can identify within our set of traffic volume traces. In this regard, we need a parameter that tells us which is the optimal number of clusters (K-Means) or when to stop the algorithm (Hierarchical Clustering).
The Davies-Bouldin Index is a very powerful indicator as it takes into account both the internal distance to the elements of the single cluster and the distance existing between each pair of clusters. The smaller the internal distance within a cluster is and the distance between the different clusters will be large, the better our index will be. Our goal will then be to minimize that index.
The mathematic formulation is the follows,

Minimize:

$$\frac{1}{R}\sum_{i=1}^{R}\max_{j=1,j\neq i}^{R}\frac{S_i+S_j}{M_{i,j}}, \tag{2.6}$$

subject to:

$$M_{i,j} = ||A_i - A_j||_2, \tag{2.7}$$

$$S_i = \frac{1}{T_i}\sum_{k=1}^{T_i}||X_k - A_i||_2 \tag{2.8}$$

Where:

- R: is the number of clusters;
- S_i: is the intra cluster of the cluster i
- $M_{i,j}$: is the intercluster distance between clusters i and j
- A_i: is the centroid of the cluster i
- T_i: is the number of traces in cluster i
- X_k: is the vectorized data of the tower k

2.4.4 Elbow Method

The second method presented to understand the best number of cluster, is the Elbow Method, presented in [13].
Usually, it is calculated with respect to **Distortion** or **Inertia**:

- **Distortion**:It is calculated as the average of the square distances from the cluster centers of the respective clusters. Typically, the Euclidean distance metric is used.
- **Inertia**: It is the sum of squared distances of samples to their closest cluster center.

Going to calculate one or both of these values for each value of K, where K is the number of clusters, and going to plot all the obtained values, it will be possible to recognize in the graph a curve that can be traced back to an arm shape. The "elbow" of the curve, will indicate the optimal number of clusters.
The typical trend of the curve is due to the fact that, for a smaller number of clusters the elements that compose it will be many and therefore more heterogeneous. By increasing the number of clusters, you will get increasingly more homogeneous clusters that will necessarily reduce these two parameters.
In figure 2.6, an example of a generic curve.

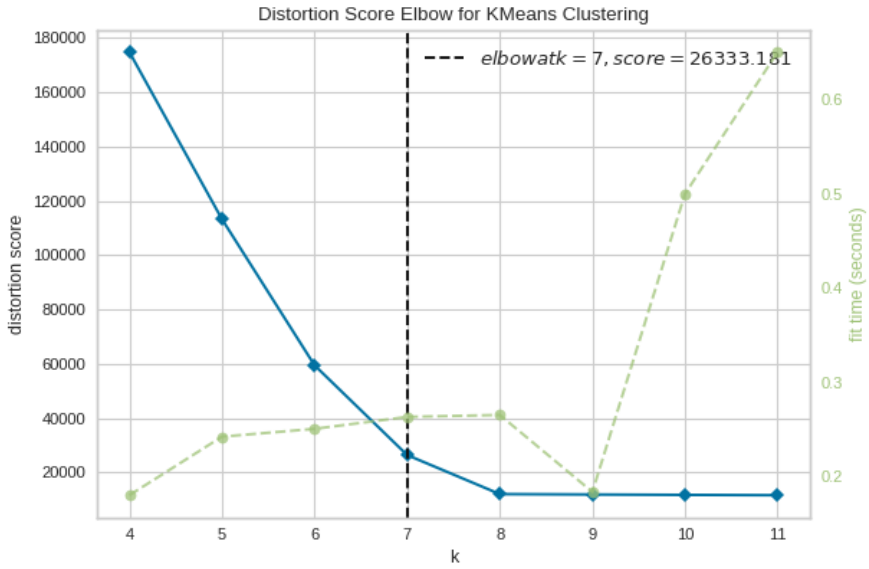


Figure 2.6: Generic Elbow Curve - *SciKit Documentation*

Chapter 3

The Dataset

3.1 The Dataset

The dataset used is related to the mobile Internet consumption in the city of Milan, form the 28th of February 2015 h.23.00 to the 29th of April 2015 h. 22.15. Roughly two months of continuous gathering. The sampling frequency is equal to 15 minutes: every 900 seconds, a picture of some parameter of the BS is snapshotted. This data was divided into 59 folders, and each of them contains all the measurements of the day of each area considered. The before mentioned areas come from an ad-hoc created grid, composed by 1419 elements, as shown in figure 3.1
As previously said, the amount of traffic of a single user is referred to an area, and not to a single BS. In the picture, we can see the grid used for dividing the city of Milan. Each area has an ID that allows us to match the geographic position on the map with the amount of traffic consumed in that area.
Using the software named QGis, and matching the geographical position of the known areas with the coordinates of the same Base Stations available online [9] [7] [2], we can have an idea of how many towers in a single area are present. The results are shown in figure 3.2.

Parameter	**Unit of Measure**
Timestamp	$[s]$
Area ID	/
Duration of connection	$[ms]$
Country Code	$+^{**}$
Traffic in Download	$[bit/s]$
Traffic in Upload	$[bit/s]$

Table 3.1: Row data parameters

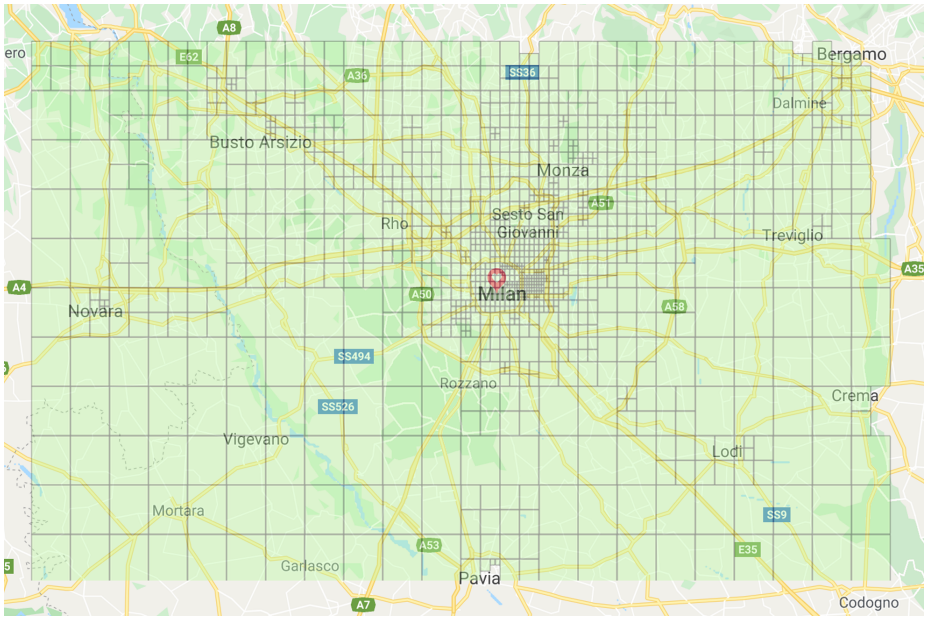


Figure 3.1: Milan grid over map

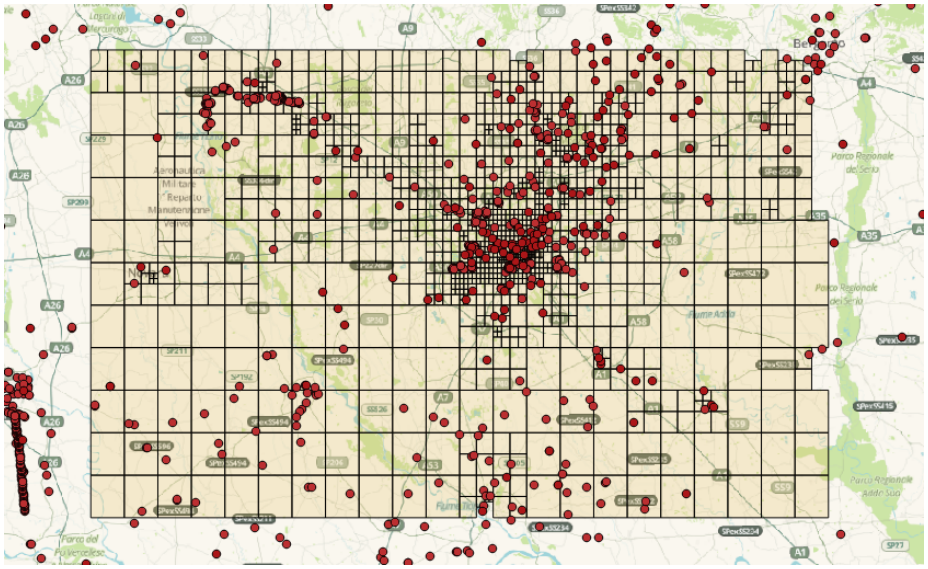


Figure 3.2: Milan grid and towers positions over map

3.2 Data Aggregation

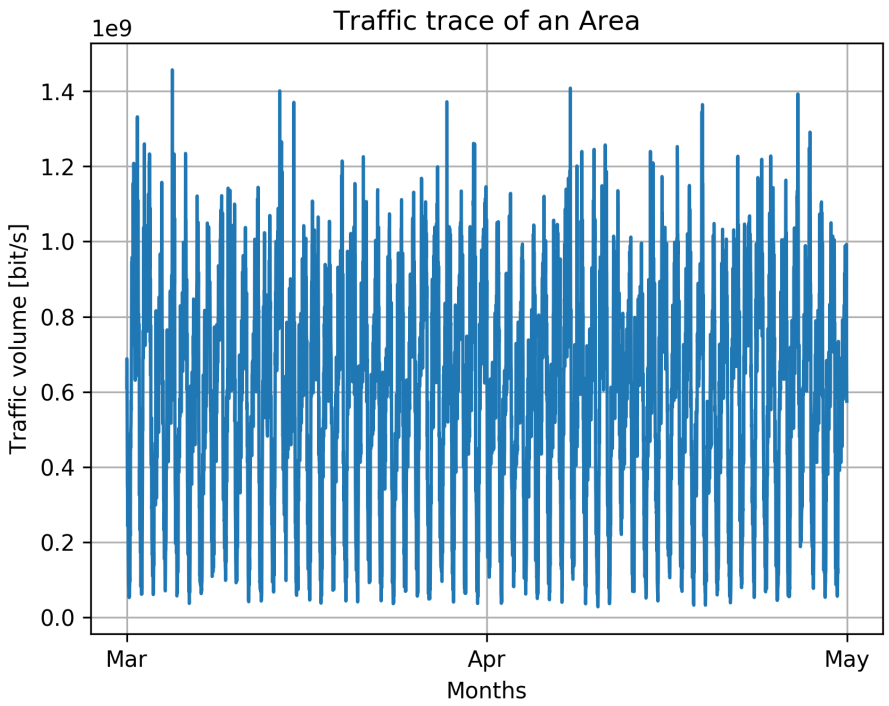
To make the data suitable to our purpose, we have to transform the daily disaggregate by user structure, to a unique trace that lasts for all the gathering

period. This trace has to take into account all the users that generate traffic at the same instant of time.
The procedure of data aggregation was made with a Python script that finds all the connections in the same area in the same instant of time and sum all the traffic generated. In this way, we generate a single value for each instant of time that represents the whole amount of traffic (downloaded and uploaded) in that instant.
By putting these values in time order, and performing this procedure for each Id, the dataset that we would use for subsequent analysis was built.

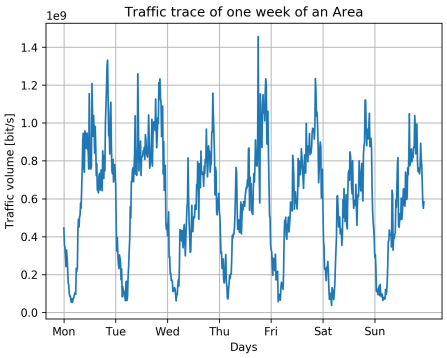
3.3 Data Visualization

Before getting into the details of the analysis, it is good to get familiar with the data you are dealing with, trying to visualize them from different points of view. In this case, we decided to paint them in two ways. First in the temporal aspect, aiming to visualize the behavior with respect to the time of our curves. Secondly under the spatial aspect, going to create a heatmap that shows us the distribution of the BSs use, and then the amount of traffic exchanged, above the map.
Figure 3.3 shows the new structure of the data. The trace in the plot is one out of 1419, randomly chosen. Taking into account the whole trace, figure 3.3a, which lasts two months, is not that easy to identify some repetitive pattern. But, looking deeply, it's easy to see that the data has a pretty evident pattern during the day, figure 3.3b. Finally, going even deeper, in figure 3.3c, a typical daily behavior is shown. This is characterized, as it was easy to predict, by a very low amount of traffic during the night hours and the residual part, the bigger, during the rest of the day. The traffic grows in the first hours of the morning and reaches two main peaks. One during lunch hour and the other during dinner hours, respectively 13.00 and 20.00. We have to pay attention, because the daily behaviour is largely affected by these trend, but the position and the time of the peaks mainly depends on the position of the tower. This curve, probably belongs to a tower situated in a residential neighborhood. In fact, the second peak, the one that falls during dinner hours, is the highest.

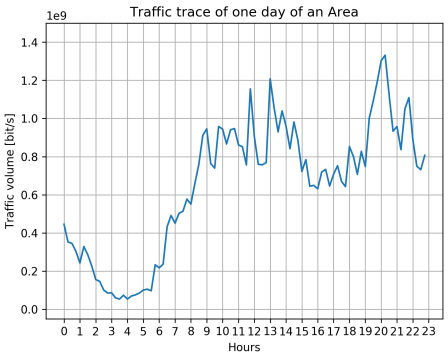
The figure 3.5 shows us different points of view of the distribution of use of the BSs related to the city of Milan, expressing the volume of traffic with a color that goes from red to blue, where red represents a high quantity and blue a low quantity of traffic exchanged. Let's discuss more detail the individual images. Figure 3.4 represents the daily distributions of internet traffic exchange. We can notice how the biggest volume of traffic is exchanged in the city center, where there are offices, shops, i.e., where the city lives more. Figure 3.5a represents the



(a) Monthly point of view



(b) Weekly point of view



(c) Daily point of view

Figure 3.3: Trace with three different point of view

distribution of traffic at 4 A.M.. The scenario that is shown is very different from the daily situation, the traffic areas are no longer displaced in the city center, but in the places that have a life also at night. In this case, we are referring to "Milano Centrale" station. We can also see at the top-left handside an area that pops out and that cannot be seen in any of the other time zones: the "Milano RHO"

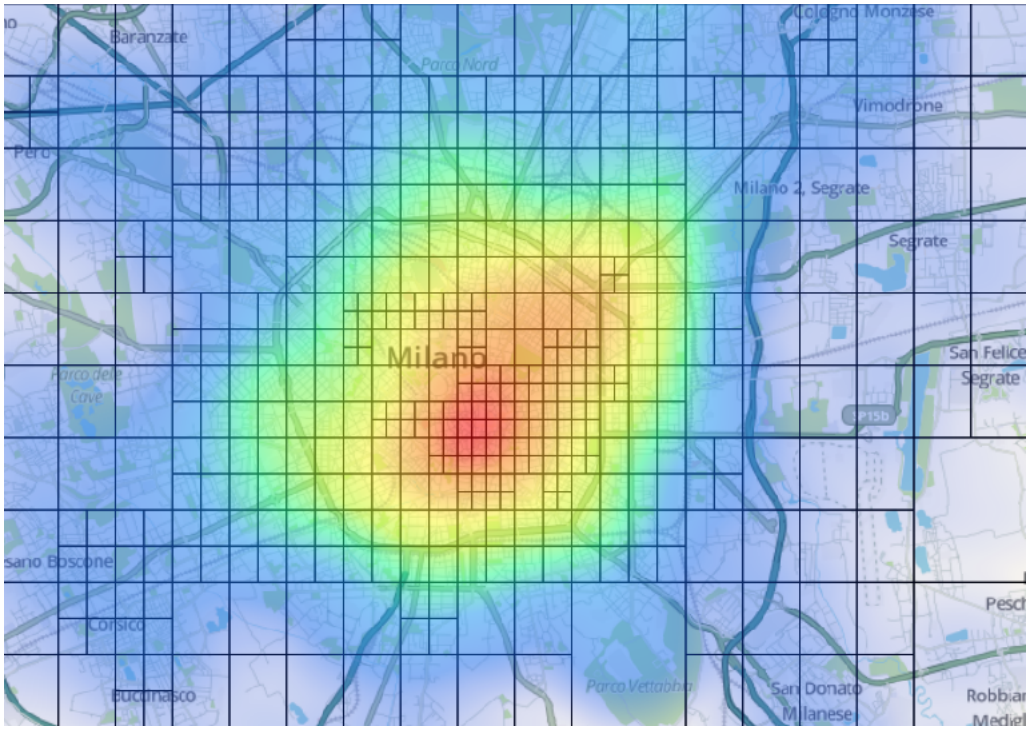
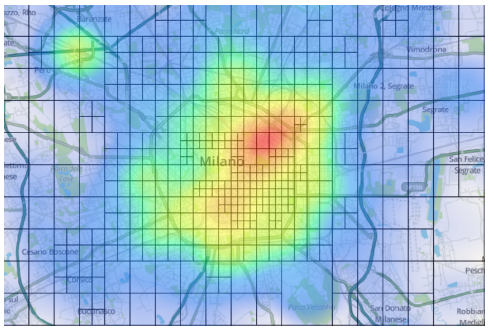
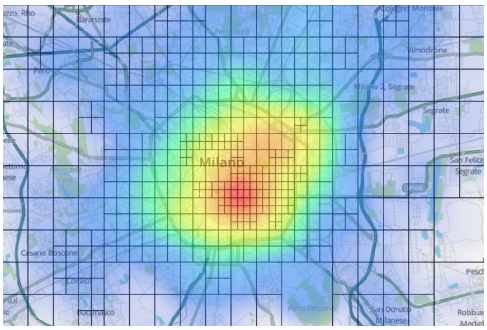


Figure 3.4: Daily Traffic Consumption Map Distribution

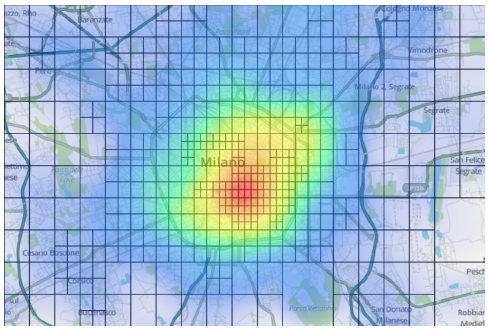
station. Figures 3.5b and 3.5c, respectively 8 A.M. and 12 P.M., are similar to each other, they tend to be closer to the daily version. The traffic moves towards the city center, as workers and non-workers tend to reach this area. Another thing you can notice is how the radius of the amplitude of the traffic tends to collapse and then expand. To confirm this, figure 3.5d shows the distribution at 8 P.M., where you don't have a real fulcrum. As a matter of fact, it is the time when are supposed people go back home and use the Internet connection, this is the reason why the area expands.



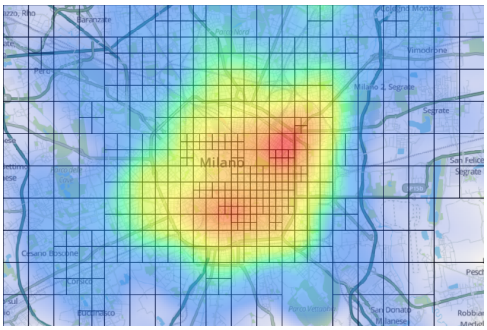
(a) 4 A.M.



(b) 8 A.M.



(c) 12 P.M.



(d) 8 P.M.

Figure 3.5: Total and hourly heatmaps

Chapter 4

Data Analysis

4.1 Data Pre-Processing

Data pre-processing is a fundamental procedure in the context of data mining. Data collection methodologies are often flawed in many ways and often unsupervised. For example: values over thresholds, impossible data combinations or missing values. Proceeding with a data analysis without taking this into account would lead to misleading results. Often the pre-processing phase is the most important in a Machine Learning algorithm.
In fact, when we approach a task in which is involved the data, we have to begin doing ourself these questions: "The data is ready to use for my goal? Has the form that best suits my requirements? All the data is gathered correctly or I need to do some adjustments?"
The answer, most of the time, is not. Data in the real world is "*dirty*". So, we have to apply different procedures to "*clean*" the data.

4.2 Data Normalization

In statistics, data normalization is a procedure that essentially, limits the domain of a certain set of values into a default range. To compare, in terms of pattern and behavior, traces that belong to a different area, with a different number of towers inside, so a different amount of traffic exchanged, we need to perform some normalization technique.
In this section I will describe the utilized technique, describing also, pros and cons of that technique.

4.2.1 Relative Maximum Normalization

The first normalization presented, and also the most utilized in my work is the so-called *"Relative Maximum Normalization"*. It consists in relating its own partial set of data to the maximum of itself.
In this way, all the data that comes from this normalization are in the range of 0 - 1. This procedure allows the comparison in terms of shape, pattern or behavior of curves that, originally, were very different in terms of absolute value.

$$v_n = \frac{v}{max_r} \tag{4.1}$$

4.2.2 Absolute Maximum Normalization

The second normalization technique proposed, is very similar to the previous one. The main difference is that all the sets of the dataset are normalized by a single value. This is the absolute maximum of the entire dataset. This kind of normalization allows us to maintain the proportion of the volume between the different traces but using elements significantly smaller. This can be useful in terms of computational complexity. In essence, it transforms all the elements of our dataset in percentages with respect to the absolute maximum values.

$$v_n = \frac{v}{max_a} \tag{4.2}$$

4.2.3 Z-Score Normalization

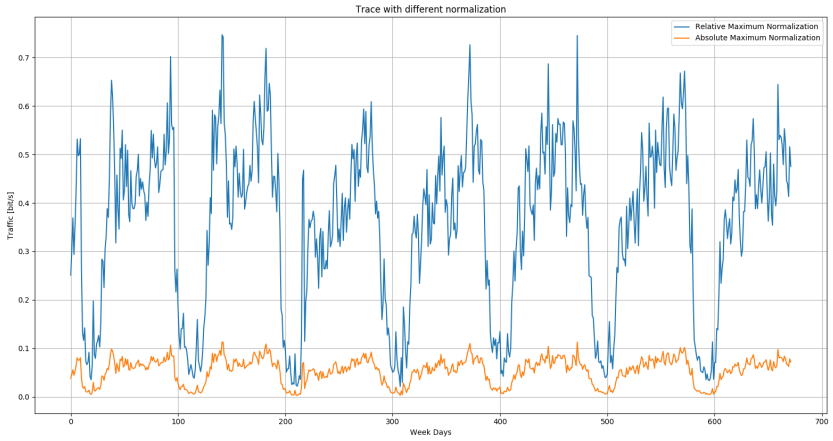
This normalization is useful when we want to avoid outlier issues. By the way, it changes the scale that is present between the different samples, losing this information that can be important. The formula of this normalization is expressed below:

$$v_n = \frac{v - \mu}{\sigma} \tag{4.3}$$

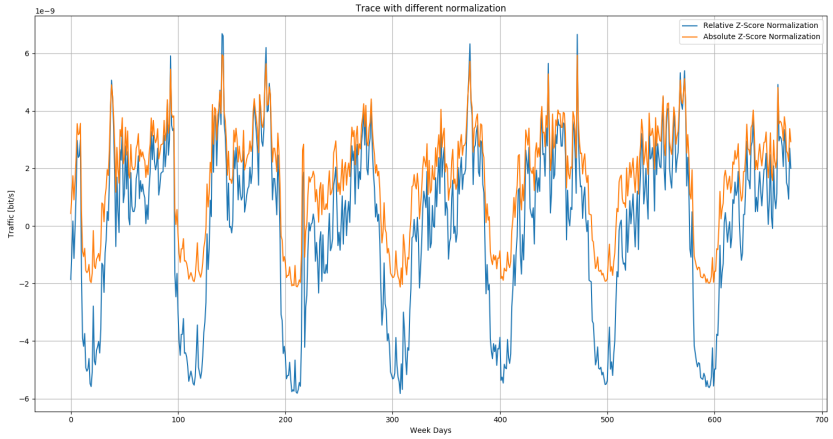
Where μ is the mean value and the σ is the standard deviation of the dataset.

4.3 Data Cleaning

As we have previously said, the data we use, in all of the situations, are not perfect. It can be affected by different kind noises, which can be due to the nature of the measurement or due to the mistake of the human or the measurement tool. For this reason, we have to perform some technique to identify and remove the outliers.
Caused by the presence of N BSs inside each area, and not knowing that number,



(a) Maximum normalization



(b) Z-Score normalizations

Figure 4.1: Different normalization technique

it is possible that the traffic exchanged in different areas can be very different. So, to avoid the presence of too much different volume exchanged, we removed all the traces that belong to the same cluster that is very distant from others.
Regarding the main data cleaning procedures, they were already implemented during the data aggregation phase. In fact, in that phase, we paid attention to not generate arrays that were of different sizes, or that within them had impossible values for our domain. For this reason, the procedures to be implemented to complete the data cleaning remain those of identification and elimination of

outliers.

4.3.1 Outlier Remotion

In order to eliminate outliers and proceed with a better structured and correct analysis, the first step is to understand how to identify these elements.
But what are these outliers in our context? They are all those traces that will be characterized by impossible patterns, or traces that belong to areas whose exchanged volume is too big or too small to be compared to others. This huge difference would introduce errors and flawed results in the analysis.
In order to simplify the work, and not to go into too complex cleaning techniques, since we already have data 'of quality', it was decided to proceed with the use of Percentiles. *"The percentile is a statistical measure that tells us under what value a certain percentage of the other elements under observation falls"*. This is a very simple technique that involves the use of percentiles to identify the cut-off point of the dataset, eliminating a certain portion of data.
In general, the 2^o and the 98^o percentile are used as reference values. All traces that had a certain value below the 2^o percentile and all those that had it above 98^o have been removed from the population. This turned out to be a good choice even during clustering.
As an indicator of the track, it was decided to opt for the average traffic volume. Once this value was calculated for each BS, they were sorted in ascending order and represented by a bar chart. The position of the $2^o r$ and the 98^o percentile was also highlighted in red. The procedure descripted above is shown in figure 4.2.

As you can see, the elements excluded from the analysis can, in fact, be considered outliers. In particular, having a look at the last elements on the right, i.e. the elements that have a larger average traffic volume. We can see how they are actually incomparable in comparison to the others. The last 3 elements have an average volume that is more than three times the 98^o percentile.

4.4 Peaks Analysis

In order to better understand the type of tracks we were dealing with, it was decided to proceed with a peaks analysis, from three points of view:

- **Hourly**: see the distribution of peaks consumption by time of day;
- **Amplitude**: display the distribution of peaks by their amplitude in absolute value;

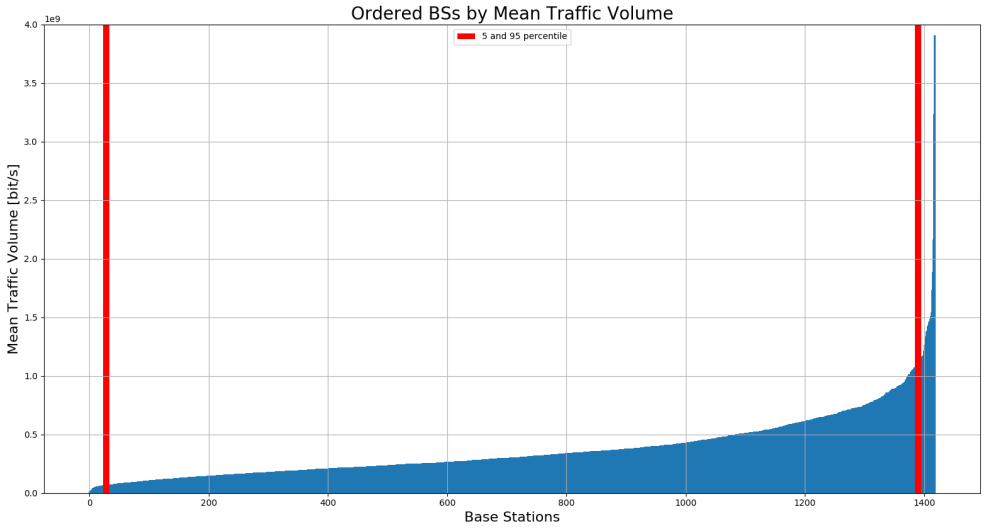


Figure 4.2: The sorted BSs, ready for the application of Percentile Cleaning

- **Normalizated Amplitude**: display the distribution of peaks by their normalized amplitude.

It has been decided to choose the peak as the identifying element because it is an excellent index to understand the position where a traffic trace reaches a certain value for a certain period. Let's assume that the peaks that we take into consideration are not impulses, but values that have in their around, elements of comparable size.

4.4.1 Hourly Distribution

Through a histogram it was decided to represent the distribution of the peaks with respect to the time of day, to identify which were the times of the day when, in the city of Milan, the greatest moments of stress were verified. All of this is shown in figure 4.3.

As you can see, the total number of spikes is spread between 6:00 a.m. and midnight. The particular thing, however, is that in the vast majority of cases, the maximum daily peaks are found in the evening time slot, particularly between 8 and 9 p.m.

4.4.2 Amplitude Distribution

In this case, however, we decided to show the distributions in terms of peak amplitude. In the first case, figure 4.4a, taking into account the peak amplitude

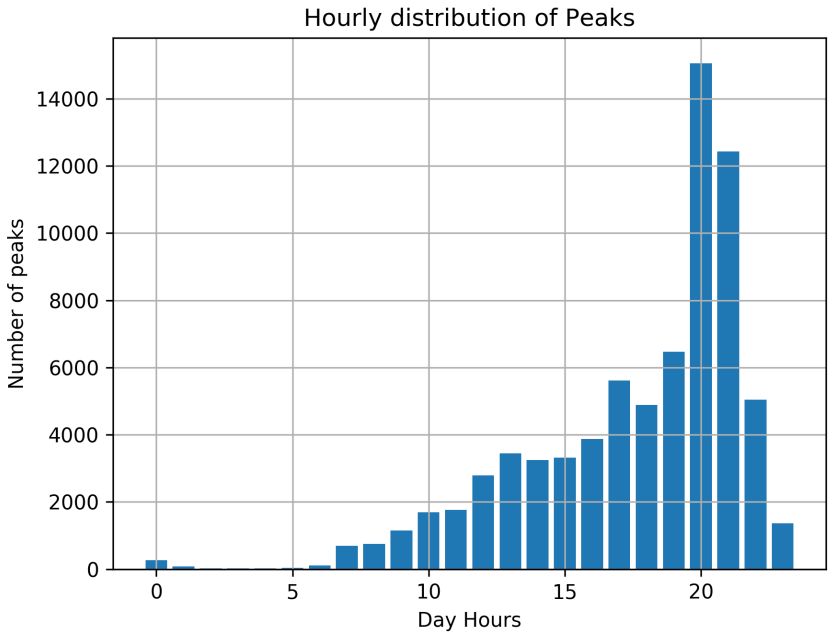
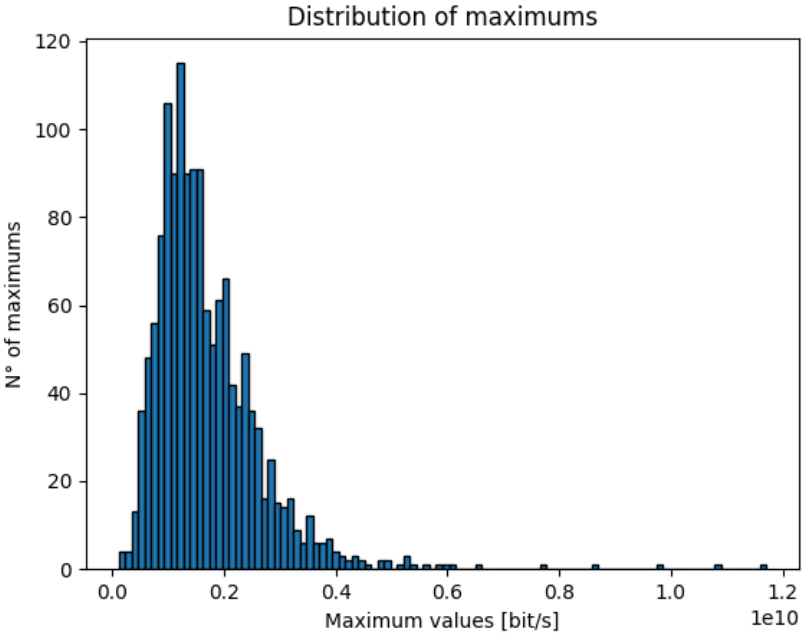


Figure 4.3: Houly distributed peaks

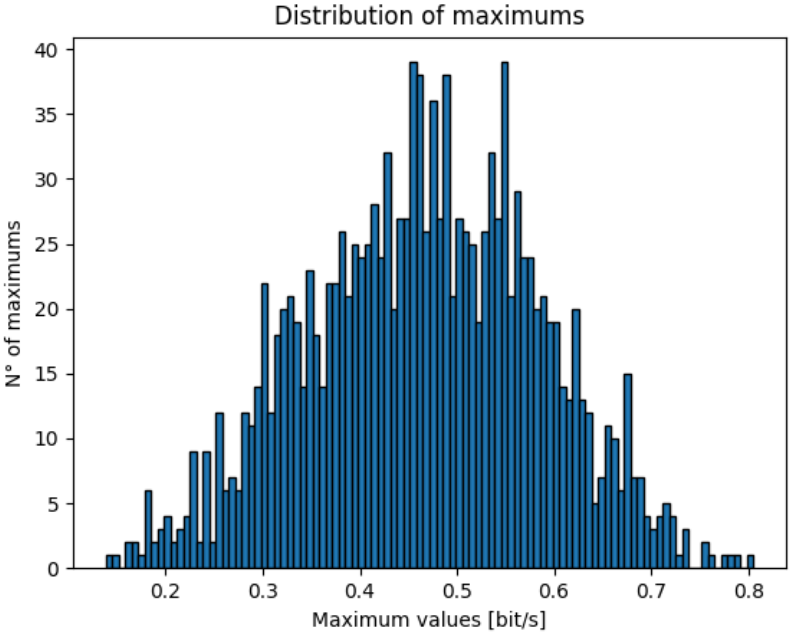
in absolute value. In the second case, represented in figure 4.4b, taking into account the amplitude of the peaks, but in the case of normalization concerning its maximum.
So, in the case of hourly distribution, we tried to answer the question "how are the peaks distributed over the course of a day?". In this case, we are trying to answer the question "Now that I know how they are distributed during the day, which is the order of magnitude? How are they distributed in terms of this value?".

4.5 Weekly Pattern

As a matter of efficiency, we thought to try to transform the two months of surveys in our possession, which consisted of 5757 elements each for a total of 1419 tracks, for a total of 8169183 elements of the order of magnitude of $10^8 - 10^9$, into tracks that considered only the typical weekly pattern identifying that specific track. In this way, they contained a small number of elements, only 672. In fact, performing a large number of calculations, on such a large amount of data, it can lead to large computational times and long waits for simple operations.



(a) Distribution of Peaks by aplitude



(b) Distribution of Peaks normalized by amplitude

Figure 4.4: Different normalization technique

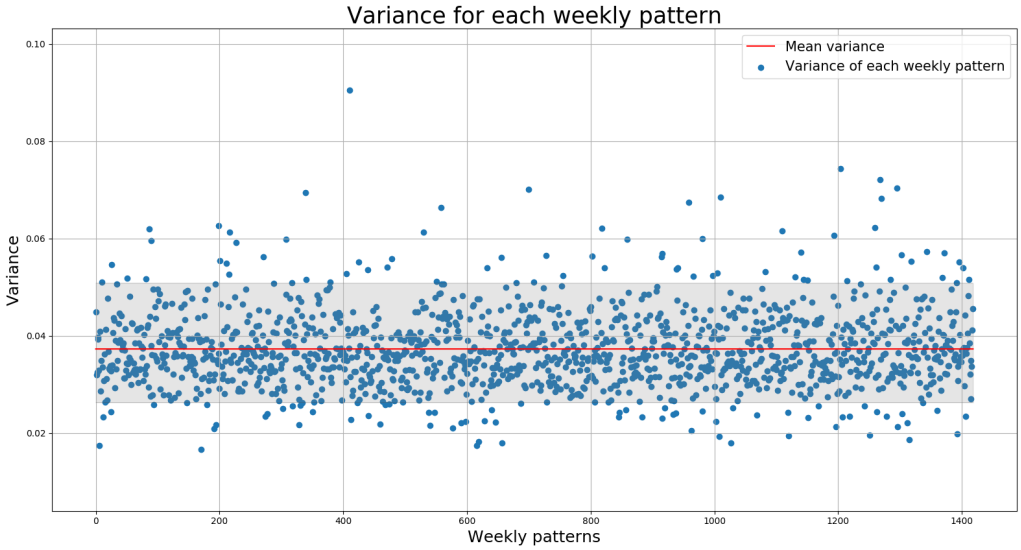


Figure 4.5: Weekly Pattern Variance

The first step was to divide the data-trace into days, putting together every Monday, Tuesday and so on. Once we obtained the 7 groups, one for each day, we averaged them, thus obtaining a curve representing the standard day. Concatenating among them the vectors that composed the days, it was finally possible to obtain the typical week for that track.

In order to validate the use of the weekly paths instead of the entire data carrier, a reasoning had to be made regarding the variance of the weekly data. For this reason, the variance was calculated between all the same days belonging to a path, and therefore every week. It is on average equivalent to 0.0373039, or 3.7% of the normalized data. A value that allows us to consider the weekly paths valid and confirms that the consumption over time in a given area remains more or less the same during the time and in particular during the weeks. Figure 4.5 shows us the variance values for each weekly pattern. We can also see that 90% of the variance values are within the range 0.02633 - 0.05083.

Figure 4.6 highlights its variance for a weekly type trace, obviously we can notice how this is greater in the daytime hours and minimum in the night hours. To investigate which were the factors that influenced the variance in a typical weekly pattern, we visualized, in the figures 4.7a and 4.7b, respectively the track with the maximum and the minimum variance. Then we went on to identify which area these tracks belonged to. The first, the trace with the highest variance belongs to a district of "RHO Fiere", so it's easy to understand the reason of such

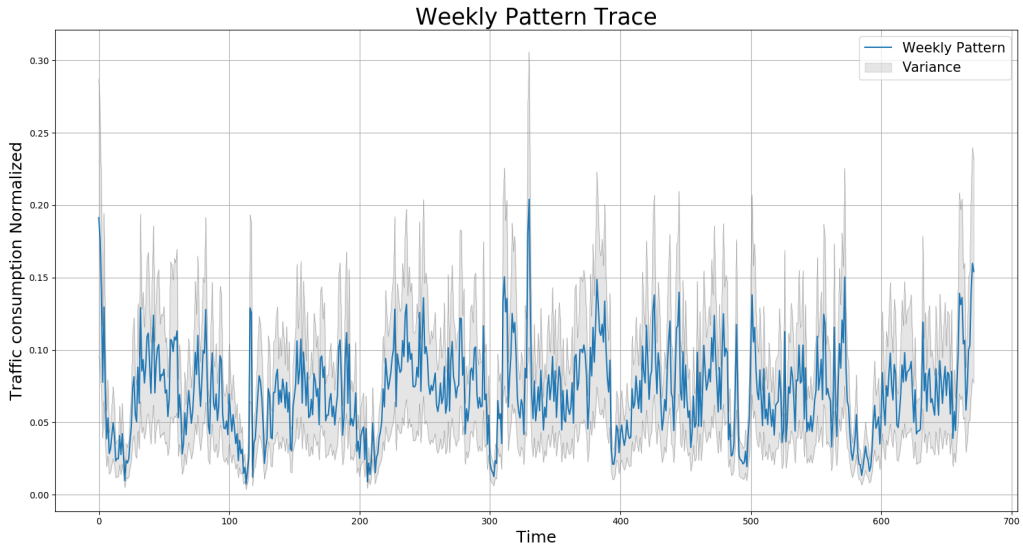


Figure 4.6: Weekly Pattern Variance

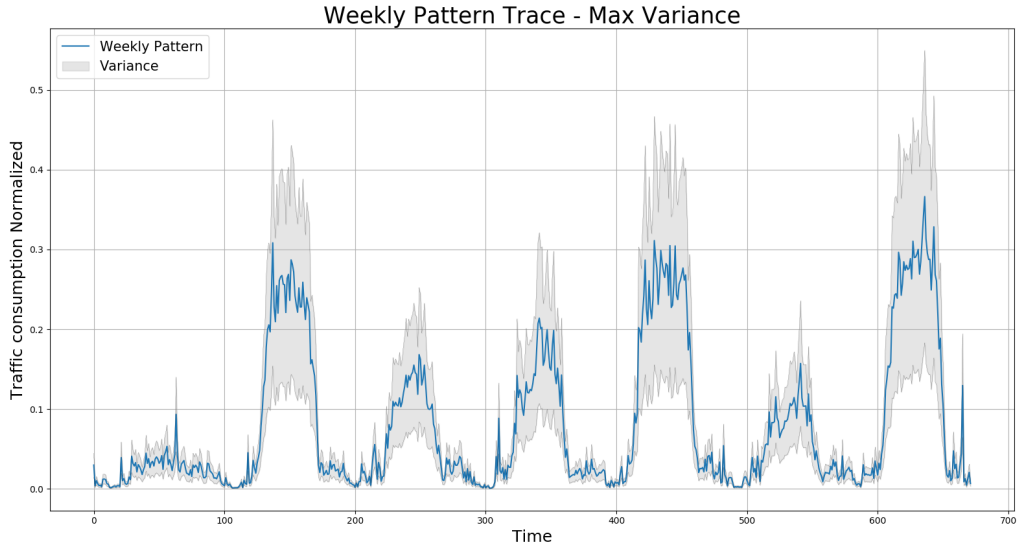
a great variance. There will be a very high consumption on exhibition days and a very low consumption on other days. The second track, the one with the minimum variance, belongs to an area that contains a small grove, the Wood "La Goccia", behind the detached headquarters of the Politecnico di Milano "La Masa". Since this area is almost completely occupied by this wood, consumption will be low and constant over time. At first glance, it can be understood that the main cause that affects the variance is the area, and therefore the habits of the people who populate it.

4.6 Cluster Analisys

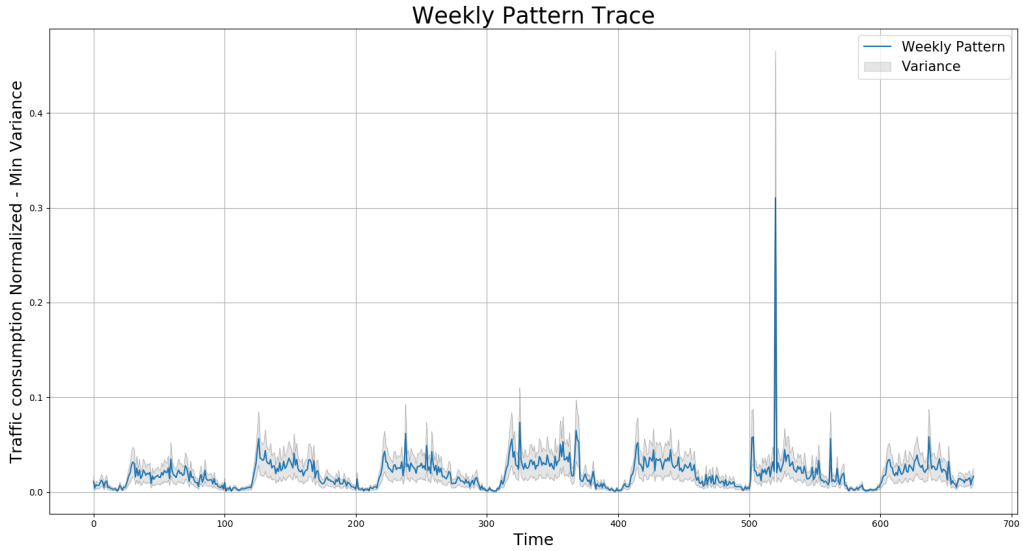
This section will describe the procedures and results of the application of the algorithms presented in 2.4 to the Dataset, resulting from the procedures described above, and to the raw data. The logical thread that will be followed is to tell first how the final clusters have been obtained and then how they will be used for the application described in the next chapter, that is the energy-saving techniques through solar panels.

4.6.1 Choice of the Best K

Before proceeding with the proper analysis, it will be necessary to define which is the best number of clusters able to contain our data. Then we evaluate the



(a) Weekly Pattern with maximum variance



(b) Weekly Pattern with minimum variance

Figure 4.7: Minimum and Maximum Variance Weekly Pattern

Davies-Bouldin index for all K ranging from 2 to 20, and the dispersion index for the same range of K in order to build the elbow curve. Applying these two indices to the weekly patterns, we obtained an indication on which was the best number of clusters to proceed with the analysis. As previously said, there is not a number that is the best in any case but, you must also take into account the type of analysis you want to perform.

Davies-Bouldin Index - DBI You can see that, excluding the K value equal to 2 and 3, which is a number of clusters that would be too reductive for the type of analysis we are going to perform, the best number of clusters according to the Davies-Bouldin Index is 6. In table 4.1 all the values for the DBI are shown.
In the next paragraph, through the elbow method, we will consolidate this choice.

K	**DB Index**	K	**DB Index**
2	0.89713	11	1.42947
3	1.04911	12	1.84396
4	1.13020	13	1.79177
5	1.14702	14	2.04450
6	1.05948	15	2.01010
7	1.34991	16	1.90426
8	1.59414	17	1.88280
9	1.61731	18	2.19020
10	1.39810	19	2.06749

Table 4.1: Davies-Bouldin Indexes from 2 to 19

The Elbow Method As mentioned earlier, this empirical and visual method has been used to give other validity to the number of clusters obtained through the Davies-Bouldin Index. The "Elbow" extracted in the calculation of the Dispertion for each K considered is shown in figure 4.8.

4.6.2 Application to the Raw Data

For the first approach with the clustering algorithms, it was decided to apply the "Raw" dataset, to demonstrate how the implementation of the cleaning procedures is fundamental for this type of work.

Hierarchical Algorithm Once defined the number of K, we proceeded with the application of the algorithms for the value of K defined, which is 6 according to the aforementioned example.

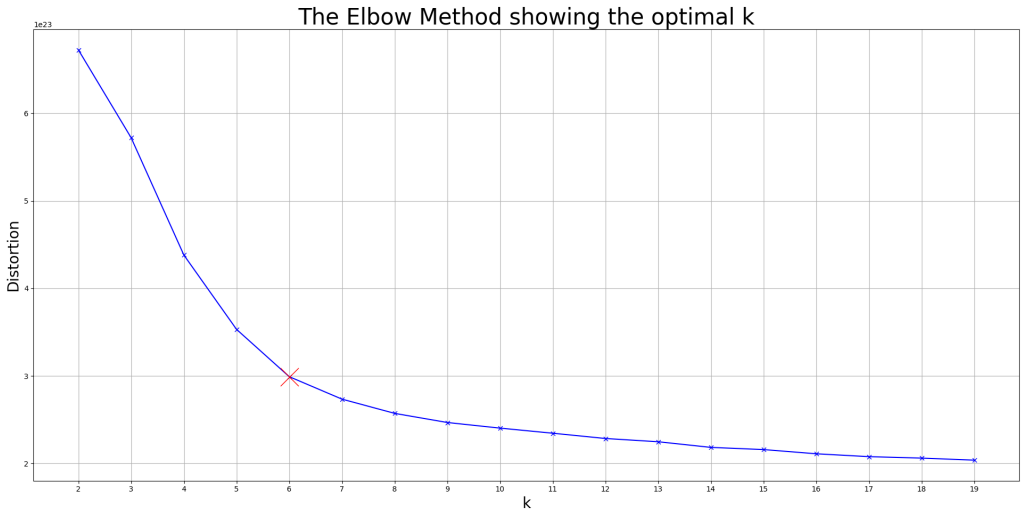


Figure 4.8: The Elbow for K in the range 2:19

In figure 4.9 the centroids of the six clusters obtained are shown. For the sake of clarity, only a slice of the whole length of the centroids are shown, precisely one week. As we can see, the various curves are confused among them, and specific patterns cannot be easily identified. This is because it has been noticed that through this algorithm we obtain very unbalanced clusters and, almost all the tracks are merged into a single one. In fact, one cluster contains 1394 tracks, while the remaining 25 are divided into the other 5 clusters.
In addition, the hierarchical algorithm has proved to be extremely inefficient in terms of execution time. Having to necessarily recalculate at each iteration, all the combinations between the Euclidean distances of the new centroids and all the tracks. For this reason, the executions took more than 5 hours, using a standard processor.
Taking all these factors into account, it was decided not to continue using this algorithm in subsequent analyses.

K-Means Clusterization The other clustering algorithm used in this work is the K-Means. This algorithm responded very well to the first executions on the raw dataset compared to the hierarchical algorithm.
We can notice, in figure 4.10, how the subdivision of the clusters is more evident in this case and, the six clusters are well distinguishable from each other. In this case what is highlighted is a division based on the volume of the tracks, not having yet applied any kind of normalization or cleaning. In fact, you notice that the centroid of the "*Cluster 1*" is significantly larger than all the others. The

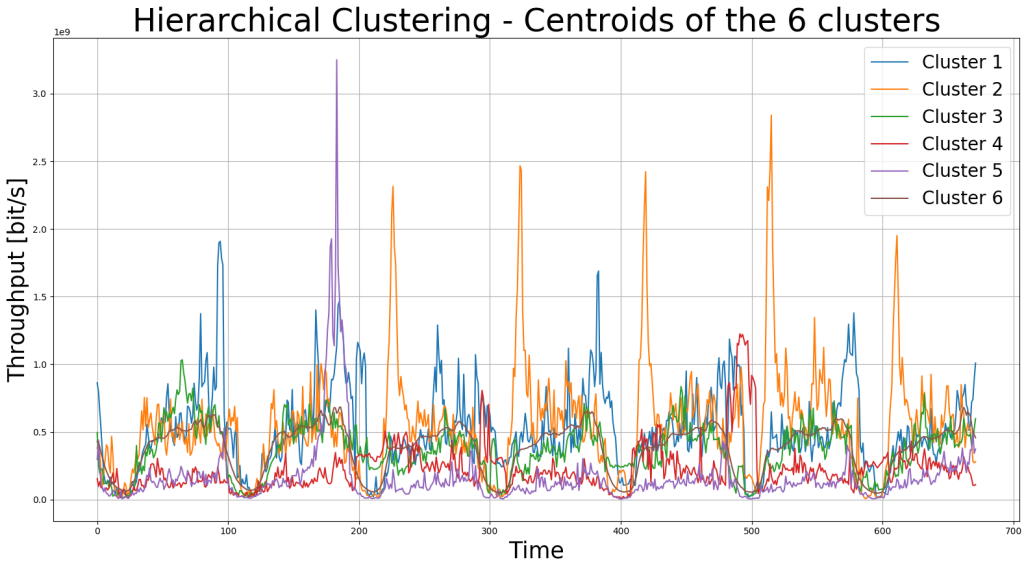


Figure 4.9: Hierarchical Clusterization for K = 6 and Raw Data

reason is because it is composed by few elements, and those elements are the ones that will be eliminated with the cleaning procedure described in 4.3.1. Also in this case, for a matter of clarity, in the figure is represented only a slice of the centroid trace, as previoulsy one week.
Due to the aforemntioned reasones, it has been decided not to proceed any further with this algorithm. On the other hand it has been decided to continue with the following analyses, exclusively with the K-Means algorithm.

4.6.3 Application to the Elaborated Data

Once we have chosen the algorithm that best fits our dataset, we proceed with the analysis. In this section we will analyze the clustering procedure by applying to the training set the cleaning procedure, 2^o and 98^o percentile, and different types of normalization.
The figure 4.11 shows the same clustering with $K = 6$, considering the outlier removal procedure. In fact, you can see that all clusters are closer together and, you can notice the absence of the cluster that was much higher with respect to the others in the previous example. Also, in this case, you can see the separation between the various centroids. They are much more balanced than the previous case. In fact, even if with some difference, the number of elements for each cluster is comparable.
Having not yet applied any kind of normalization, even in this case the subdivision that we can see is mainly characterized by the volume of traffic exchanged, even

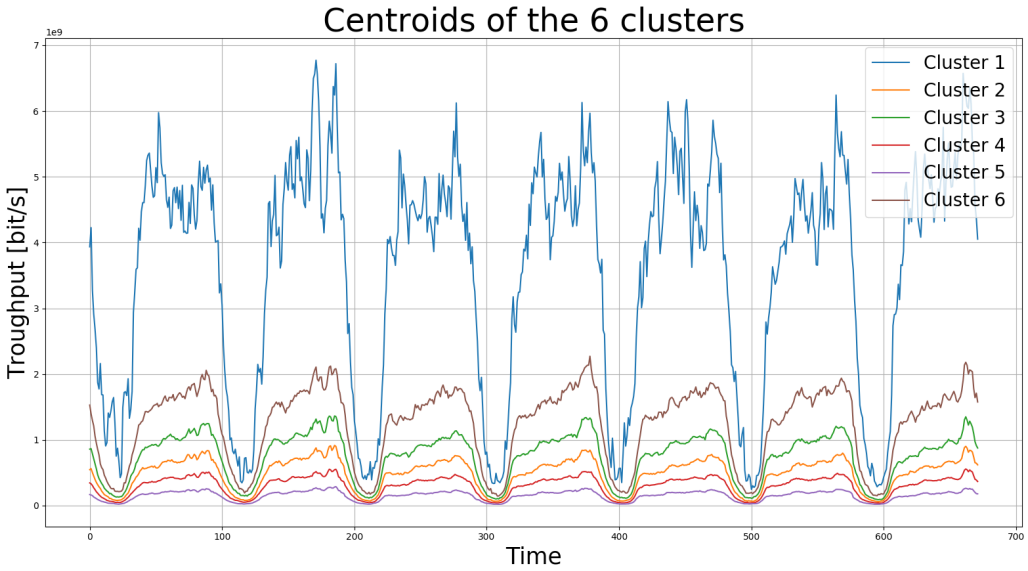


Figure 4.10: K-Means Clusterization for K = 6 and Raw Data

if we look at the "*Cluster 6*", we can appreciate same differences in the pattern, with peaks that are not in the evening but in the morning.

In the next paragraphs, we will discuss the effects of the application of various normalization techniques on data, in order to try to identify the best one for our objective. We will look for the normalization that will highlight the features that we are more interested in.

Relative Maximum Normalization This type of normalization, which exploits the maximum of each track, allows us to emphasize the curve for what is its pattern, going to decrease the influence of the volume exchanged in a given area. This is because with this type of normalization we will put all tracks on the same level since they are normalized with respect to themselves.
To avoid the use of peaks that can be too large, or not real, for this normalization it has been performed another practise for the outlier remotion. All the values over the 98^o percentile were associated with the values of that percentile.
Figure 4.12 shows the threshold for which the elements are assigned to the 98^o percentile value. The figure 4.13 shows us the result of the K-Means method as a result of the normalization, described above, on the data. We can see how our goal to reveal the characteristics of the curves has been successful. In fact the clusters are no longer clearly distinct for their volume, but also for their trend.
In particular, we can see the *Cluster 5* and the *Cluster 2*, differ from each other in their countercurrent trend. That is, for the fact that the majority of the volume

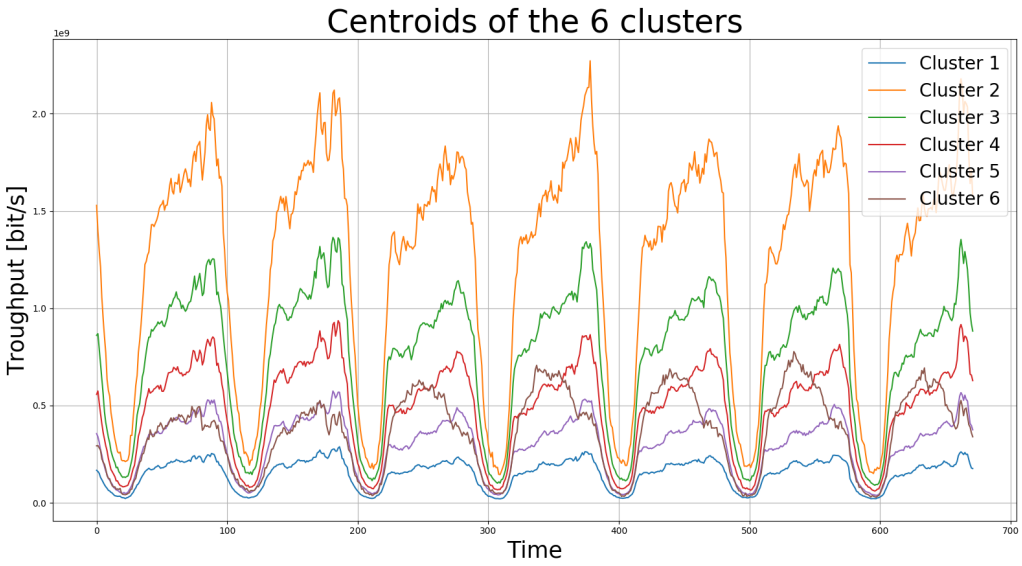


Figure 4.11: K-Means Clusterization for K = 6 and Cleaned Data

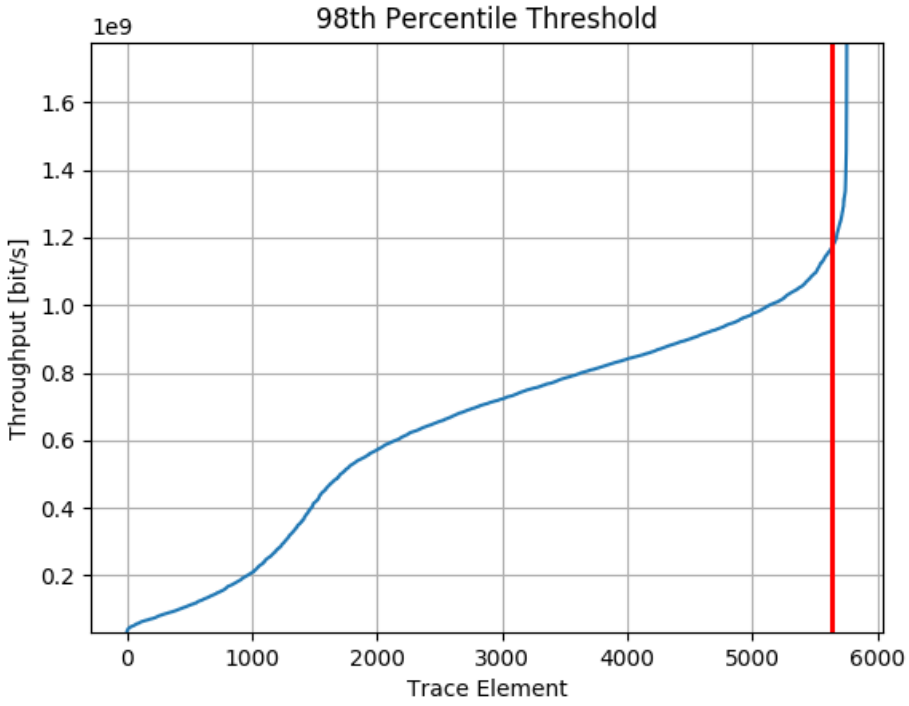


Figure 4.12: 98^o percentile threshold

of traffic exchanged during the day is in the first half of the day and not in the evening.

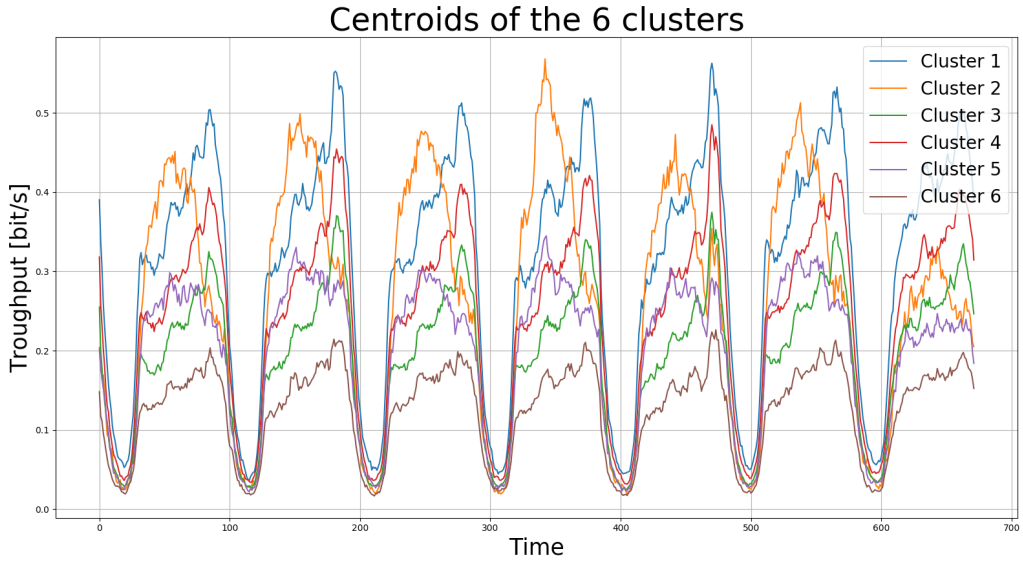


Figure 4.13: K-Means Clusterization for K = 6 and Relative Maximum Normalization

Absolute Maximum Normalization This type of normalization uses the absolute maximum of the entire dataset. The hypothesis that is made is that all the BSs considered, have the same maximum capacity, which coincides with the absolute maximum.
This normalization will scale the traces between 0 and 1, but without losing the ratio, in terms of volume exchanged, between them. In figure 4.14 we can see the clustering done with the dataset after the application of such normalization. The subdivision in clusters obtained is very similar to the one shown by figure 4.11. This result is easily explained by the fact that the traces have only been scaled all with respect to a single value.

Z-Score Normalization This normalization was carried out on the assumption that the data had a Gaussian distribution within the week. The normalization, in fact, was done by subtracting and dividing the single values by a variance and a dynamic average. By dynamic we mean that these two parameters are calculated not for the whole trace but for the single days and hours. For example, the value of Monday at 10 a.m. has been normalized with the average and variance calculated on all Mondays present on the track at the same time.

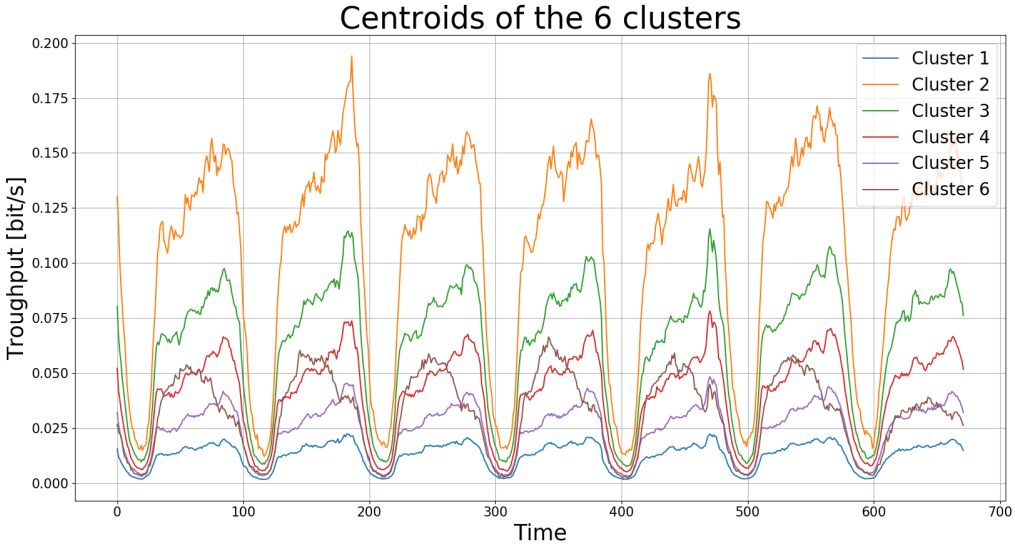


Figure 4.14: K-Means Clusterization for K = 6 and Absolute Maximum Normalization

Figure 4.15 shows the result of clustering with this kind of normalization.

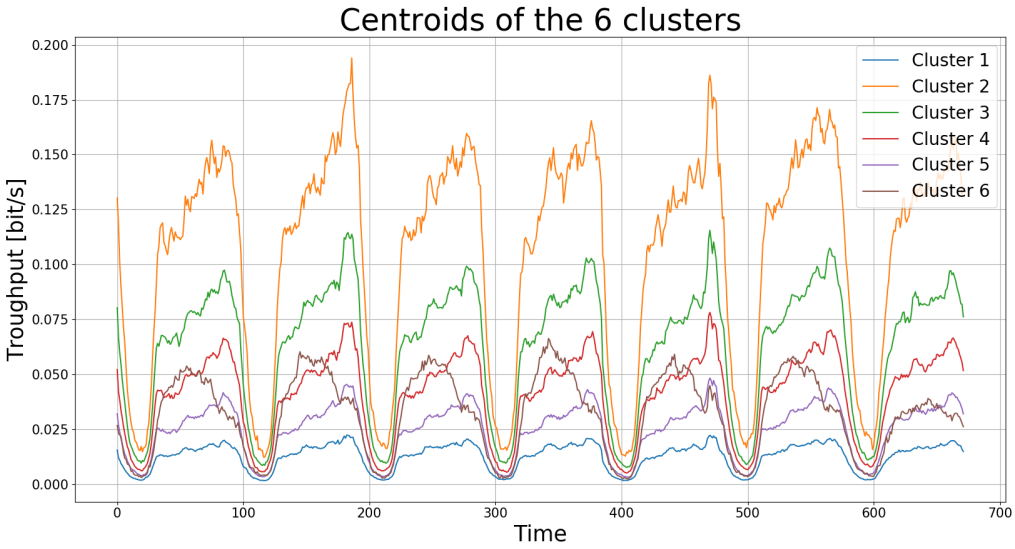


Figure 4.15: K-Means Clusterization for K = 6 and Z-Score Normalization

4.6.4 Application to the Weekly Pattern

In this section, the results obtained by the k-means clusterization, starting from the average weekly traces, that are built according to 4.5, will be discussed.
The use of the weekly traces is based not only on the computational time reduction due to the data dimension decrease, but it also represents the minimum time interval in which a trace trend can be correctly compressed and expressed.
It was decided to apply to the weekly traces the normalization to the relative maximum, which is the one that has shown to be the most influential in the subdivision of clusters according to the pattern. In this case, the maximum taken into account is the one averaged over the weeks, i.e. the maximum of the average week. Therefore it was not necessary to apply the procedure of assigning the 98^o percentile to all values exceeding this threshold since the maximum is already an average value and thus not too large concerning the other trace values.

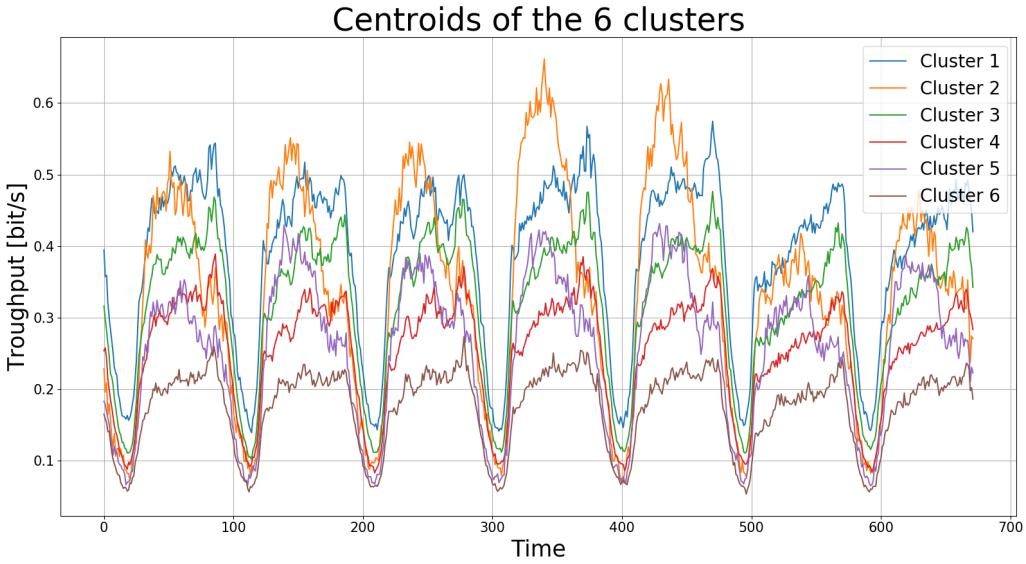


Figure 4.16: K-Means Clusterization for K = 6 with Weekly Traces and Relative Maximum Normalization

From a first observation of the figure 4.16 it could be noticed how by using the weekly average traces as a starting point for K-Means clustering, a more confusing and less linear result was obtained compared to the others. This is partially true, in fact, by condensing two months of data in a week, the curves will be more jagged, but the "type" pattern will emerge. For these reasons, the next analyses will therefore be carried out taking into account the average weekly tracks.

4.6.5 Analysis of Consumption Threshold Violation

The first step in order to proceed with this type of analysis is to define what the consumption threshold is and how much its value is. The consumption threshold is a value, expressed as a percentage that in various energy-saving algorithms, in particular green networking, represents the percentage of traffic management capacity used by a microcell under which it is more convenient not to keep the BS active and turn the traffic load on the reference macrocell. This procedure can be applied only in case the remaining capacity of the macrocell allows this transfer.
The objective to be achieved with this analysis is to group the traces that have a similar behavior towards the consumption threshold. It has been decided to apply the same algorithm used previously, the K-Means, also in this type of analysis, to maintain uniformity in the method.
Once identified the goal and the methodologies, it was necessary to understand what were the useful features that would give us the right information to achieve the goal. Two features have been identified:

- **Time above threshold**: the first feature considered is the time for which the percentage of capacity used by BS is above the threshold;
- **Volume above the threshold**: the second feature considered is the volume of traffic exchanged by BS above the threshold.

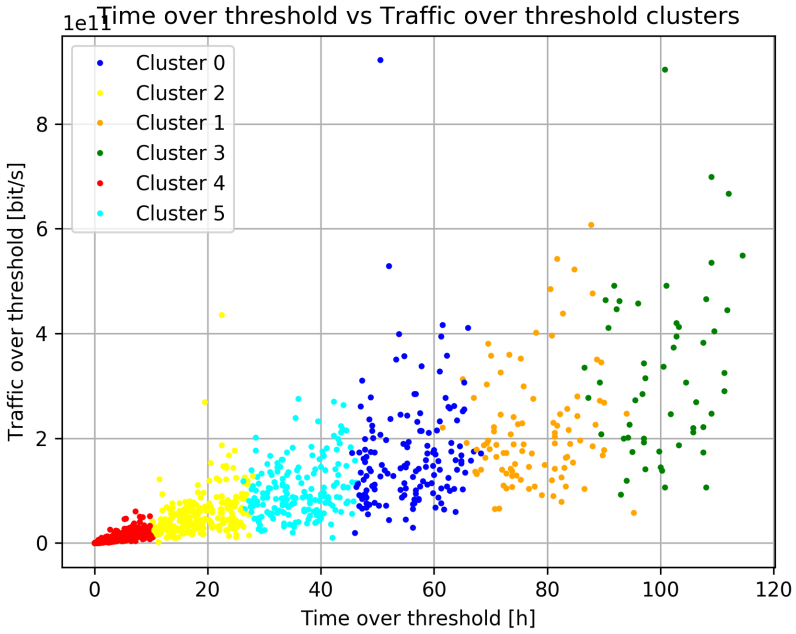
Only the time above the threshold is inadequate to allow us to identify the behavior of the trace with respect to the threshold. The reason lies in the fact that we do not know the volume of traffic exchanged above that threshold in that time interval. Two different paths could spend the same time above the threshold but exchange very different amounts of traffic. For this reason, the volume of traffic exchanged above the consumption threshold has been included in the features taken into account for the analysis.
Also in this case, the K number of clusters was chosen to be 6. Again, the best number of clusters was identified using the Davies Bouldin Index (DBI), which, although in this case the value 4 was better than 6, the 6 continued to be a better number with respect to 5. For this reason, it was preferred to keep the same of the previous analysis the number of clusters. It should be noted that in this situation the difference between the indices for the choice of the best number of clusters is in the order of 10^{-3}. The DBI value for $K = 6$ was calculated at 0.499214.
Figure 4.17 shows us the result of the clustering described above and, what differs between the two graphs is the normalization. Figure 4.17a shows the data not

affected by normalization. On the other hand, the chart 4.17b shows the data clustered after the normalization procedure.

As we can see, in both cases the clusters are well defined and distinct from each other. Moreover, we can notice a certain linear dependence between time and volume, which in the case of normalized data becomes much stronger. The first thing we can notice looking at the figure 4.17a is how, as the time spent above the threshold increases, the traffic exchanged is getting bigger and bigger ,and more and more affected by variance. So as the time spent above the threshold increases, so will the variance of the exchanged volume in that time interval. We can see that some traces that remain for a long time above the threshold, actually exchange the same amount of traffic that is exchanged by some traces that spend much less time above the threshold. This will turn into clusters that will be more and more spread as time goes by.
As for figure 4.17b, it presents a different scenario. The linear dependency is almost perfect, even if you still notice an increase in variance as time increases. What does not happen in this case, however, is that two elements belonging to two different clusters, which spend a very different time above the threshold, exchange the same amount of "normalized traffic". It should be reminded that the first figure described is, in any case, the result of a clustering carried out through normalized data. However, the volume of traffic returned to the original value has been shown.

4.6.6 2nd Level Clusterization

The next step would be to put together the results of the clustering carried out using the weekly patterns of the traces and the clustering carried out using the features that describe the behavior of BS in relation to the consumption threshold. The problem could easily have been solved by creating a single element containing both the threshold and the pattern features. Unfortunately, these two features are dimensionally very different, the first ones are 1×1 elements, the second ones 1×5757 vectors. This makes difficult to use them together without changing their meaning. In particular, the features used for the threshold exceeding analysis were composed of elements that identified a different characteristic (time and volume). In the case of the pattern analysis, the BS trace had to be considered as a single element itself to preserve its meaning. To solve this intricate enigma it was decided to proceed with a clusterization on two levels.
Clustering on two levels means the application of the K-Means algorithm two times: for the first layer using only the features characterizing the behavior towards the threshold; for the second layer reapply the same algorithm considering, this time, the weekly average trace, that represents the pattern. This allowed to



(a) Time vs Traffic clustering not Normalized

(b) Time vs Traffic clustering Normalized

Figure 4.17: K-Means clusterization for threshold behavior

have a primary subdivision based on the behavior of the single BS with respect to the consumption threshold and then a subdivision based on the typical behavior, within each primary cluster, giving rise to 36 clusters.
This clustering on two levels allows us to take into account again the behavior of a trace not only concerning its pattern but also its volume. In fact, traces that spend a similar amount of time and exchange a similar amount of traffic volume in that time interval will be labeled under the same cluster. This aspect will be important to take into account in the next chapter when it will be discuss the calculation of energy consumption of BSs.

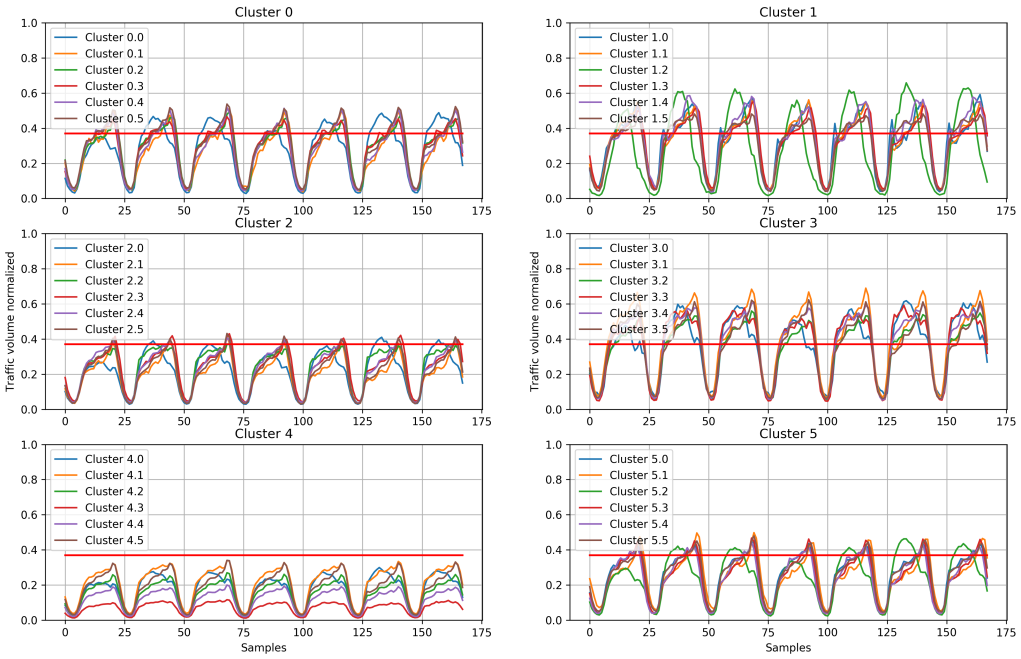


Figure 4.18: Representation of the 36 clusters

Figure 4.18 shows the thirty-six clusters obtained following the procedures described above. The figure is divided into six sub-figures, each of which represents a cluster obtained from the first application, the one through the use of the features of volume and time with respect to the threshold. We can agree with what has just been said by comparing the figure in object with figure 4.17b. The shining example is the one offered by cluster 4 which is composed by those traces that exchange less traffic in less time above the threshold, and it is in fact, the cluster that holds the closest position to the origin in the previous chart. Cluster 4 also has another feature that makes it different from the others. The average of the tracks that compose it never exceeds the threshold. Going to investigate the individual tracks that make it up, they exceed it for less than 10 hours each

in a whole week. Ideally the best choice for these BSs is to hold them in a sleep state.
Generally, we can notice in the first analysis the differences in volume between the various clusters. Entering more into detail, with this double level analysis we can better appreciate the trend differences of the various BSs. The most meaningful examples are: *Cluster* 0.0, *Cluster* 1.2, *Cluster* 2.0, *Cluster* 3.0, *Cluster* 4.0 and *Cluster* 5.2. All of these are composed of traces that have, on average, the highest consumption during daytime. This aspect may become interesting in the next chapter where we will address the problem of dimensioning a photovoltaic power supply that will be done through this clusters analysis.

Chapter 5

Dimensioning of Photovoltaic Power Supply

The dimensioning of a photovoltaic system is a very important procedure for achieving the optimum profit from this kind of installation. However, this procedure is not always simple and intuitive and it must take into account a large number of parameters that may confuse the operator who will be in charge of the sizing. This chapter is addressing the problem of dimensioning a photovoltaic power supply thanks to the cluster analysis conducted in the previous chapter. The objective is to provide an immediate indication of the optimal size of the photovoltaic system by taking into account the installation area. From this information you can easily obtain an indication of both the quantity and the type of the traffic exchanged. Once this information in known it will be easier to associate the BS, for which you want to design the system, to an already studied cluster being able to have an immediate indication of the recommended size.

First of all, it was decided to carry out the analysis over the entire solar year in order to take into account the seasonality factor. This is a fundamental characteristic in order the choose the right photovoltaic system. Secondly, it was also decided to carry out this analysis for three different climate zones: a temperate, warm and cold one.

A starting hypothesis had to be made to achieve this objective: the traffic patterns of the city of Milan are the same of the other considered cities. This idea is based on the fact that, regardless of latitude, the people behavior is the same: low traffic at night, high traffic during the day and variable one in particular city areas. The considered traces, however, cover only two months of a year. This is the reason why the traffic pattern is considered uniform throughout the whole

year.
Starting from this hypothesis, a new trace dataset has been created by replicating the trends of the aforementioned two months considered, being aware of keeping the correct position of the real months (March and April). Eventually, for this analysis the traces are separately considered either as Macro or Micro cells.

5.1 Irradiation Data

In order to perform this step of analysis, the download of the solar radiation data is necessary.
The used data has been downloaded from [11], which is a tool that provides solar radiation data for the years between 2005 and 2016. It is important to notice that datasets can be generated with different methods of solar radiation models. Maoreover, the tool allows the customization of the considered photovoltaic system and the position, expressed in latitude and longitude, of the place whose analysis is intended to be referred to.

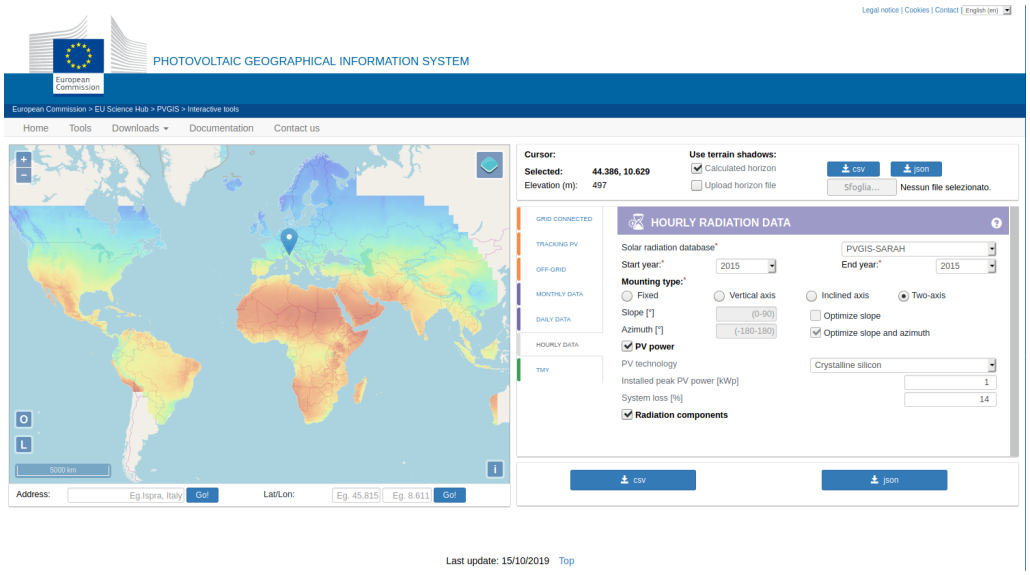


Figure 5.1: European Commission Tool for Solar Radiation Data Download

Figure 5.1 shows the entire interface used to download the data. It is basically composed of two main parts: the map from which you can select the position, and the panel from where you can choose the settings.
As far as the locations are concerned, data from three cities, belonging to three different climate zones, has been downloaded. More precisely:

- **Milan**: for the Tempered belt

- **Oslo**: for the Arctic belt
- **Cairo**: for the Tropical belt

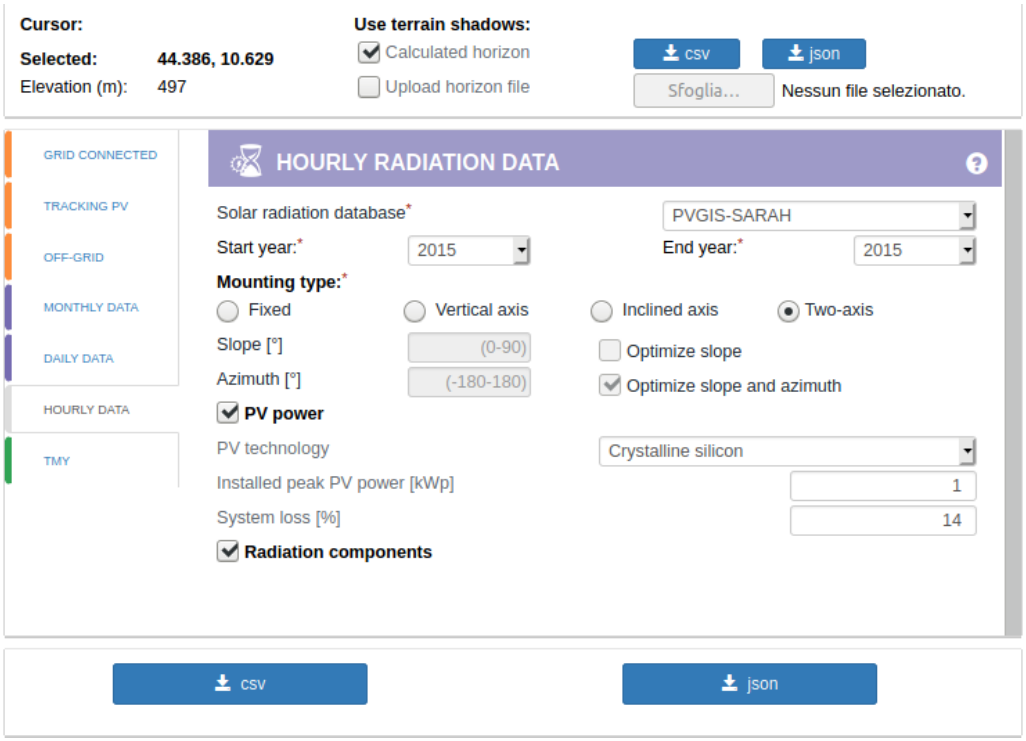


Figure 5.2: Detail of European Commission Tool for Solar Radiation Data Download

Figure 5.2 shows the detail of the configuration panel. In detail, the meaning of every panel feature and the motivations of the taken choices will be explained. It is important to notice that the parameters shown in the figure are exactly the ones selected at download time and which will then be used for the analysis.

- Solar Radiation Database: these are the solar radiation databases calculated by different satellites;
- Start and End Yeat: these are the start and end year you want to download.
- Mounting Type: the type of freedom of the solar panel to react to the position of the sun;
- PV Technology: the technology of the photovoltaic system taken into consideration;

- Installed peak PV power [kWp]: is the maximum peak power that the system can achieve;
- System loss [%]: the percentage of loss in the conversion phase.

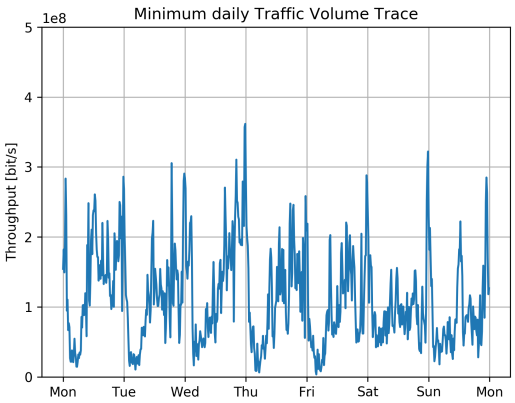
Among the several solar radiation databases, the one called "PVGIS-SARAH" has been chosen. This is the one with the best measurement accuracy for Europe and Africa.
The period of time chosen has been decided not to be limited to just one year, since it could be particularly influenced by the atmospheric factors of that particular year. Due to this, the solar radiation data of ten years has been considered and it has been then averaged in order to obtain one year average radiation.
The Mounting Type has instead been chosen as the one able to guarantee freedom on both $x\,y$ axis. The feature which specify the optimum slope/azimuth has been selected. This results in an adaptive solar panel which automatically tracks the position in which the generated power is maximized.
The photovoltaic panel technology keeps on being the default one, which is based on Crystalline Silicon, as in the majority of cases.
In order to investigate the trend of the power generated with respect to the solar panel dimension [kWp], data of different dimension panels have been downloaded. This has allowed us to find a linear relationship between the aforementioned parameters. Thanks to this, the power generated by a $1\,kWp$ panel is considered as the unit value, and all the power values related to the other panels are directly calculated by means of a Python script during the dimensioning phase.
Finally, the System Loss has been left at its typical default value equals to 14%.

5.2 Building Consumption Traces

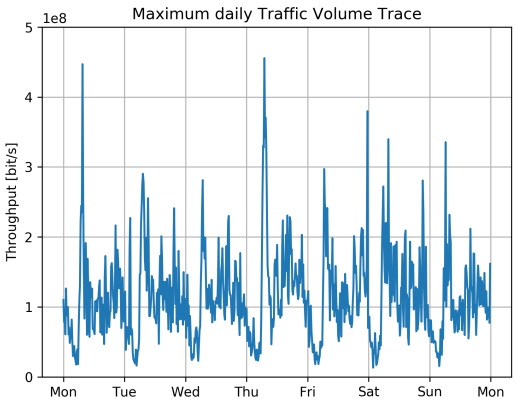
The second component, in order to evaluate the energy savings and to proceed with the dimensioning of the photovoltaic system, is the set of curves for that indicate the energy consumption for each trace. In practice, how many kWh the single BS needs to fulfill all the required tasks. In a system with no other type of power supply, the entire amount of power required by the tower is retrieved from the grid.
We proceeded with the generation of two datasets: the first one assuming that the traces came from Macro Cells, and the second one assuming that the traces came from Micro Cells. For the generation of these two datasets it was enough to apply the equation 2.1 to each single element of each trace in its annual form.
In order to make the generated data consistent with the solar radiation one, it was necessary to proceed with a transformation of it. In fact, the solar data

are sampled with an hourly frequency (for each hour we know the average solar radiation). It is good to remember that the data on the traces had a sampling frequency of 15 min, meaning that 1/4 of the frequency is used to sample the solar radiation. For this reason we averaged the measurements present in every single hour, to create a "reduced" vector that reduces itself from 35040 to 8760 elements, which is exactly the number of hours present in a year.

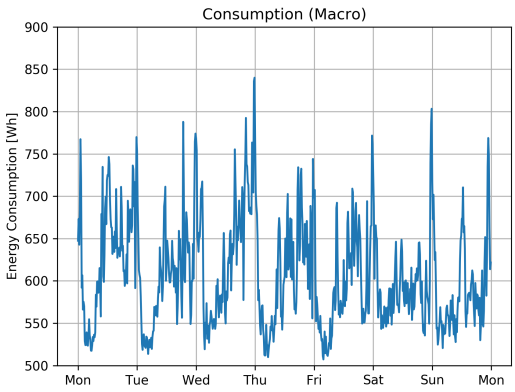
Figure 5.3 shows the transformation from traffic volume, expressed in $[bit/s]$, to energy consumption expressed in $[Wh]$. The results of the equation is actually in W but since we are considering the mean power related to one hour, it follows that the absolute value of the energy is the same. In figure, two different scenarios are investigated: the first one, whose 3 graphs of the left column belong to, takes into account the weekly average traces of the BS that has the lowest traffic exchange during the sunny hours. The second case, whose graphs in the right column belong to, instead, takes into account the average weekly trace of the BS that has the highest volume of traffic exchanged during sunny hours. The hours of sunshine of the whole year have been calculated considering the vector of the solar radiation data. Daylight hours were considered to be all the ones during which there was solar radiation different from 0.
Figures 5.3a and 5.3b show the two average weekly traces taken into account. It is important to remember that in this phase, for the calculation of consumption, the non-averaged traces replicated throughout the whole year, have been used. For the sake of clarity, it was decided to show the transformation of the weekly average traces.
Figures 5.3c and 5.3d represent the traces converted into energy consumption in the hypothesis that the originary BSs were Macrocells. On the other hand, figures 5.3e and 5.3f, represent the traces of energy consumption in the hypothesis that the originary BSs were Microcells.
At a first glance, the patterns remain the same as the traffic traces. This is because the used model is linear and it is composed of a fixed part, which is the power needed to sustain the tower as it is, and of a variable part caused by the traffic exchanged by the tower. The second thing that immediately pops up is that the transformation tends to flatten the trace because, for any element of the curve, the transformation returns a value between P_0 and $P_0 + (P_{max} * \delta P)$. This is due to the fact that the fixed part is much larger than the variable part.
Generally, consumption remains in a range between 500 and 900 Wh in one hour for Macrocells, while, in a range between 110 and 150 Wh for Microcells.



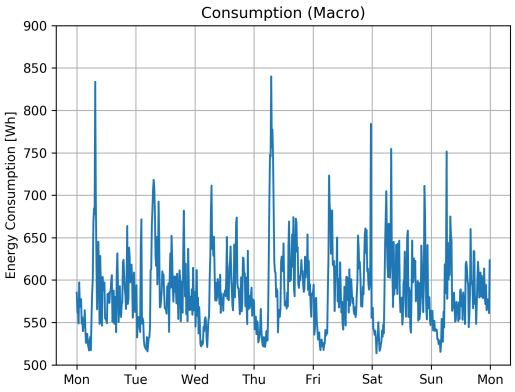
(a) Minimum daytime traffic volume trace



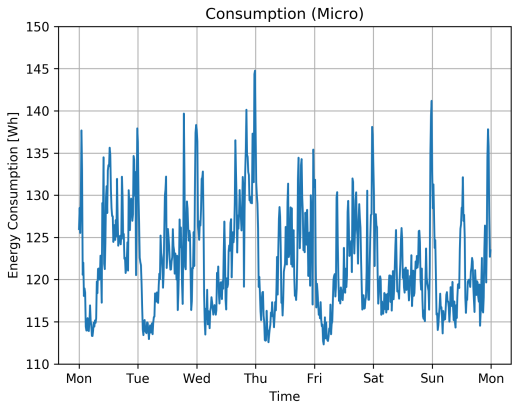
(b) Maximum daytime traffic volume trace



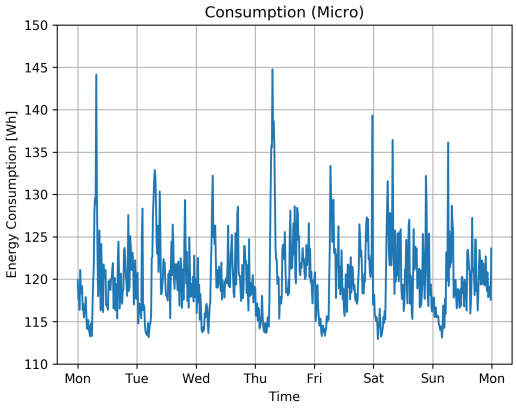
(c) Consumption considering the minimum daytime traffic volume BS as Macro cell



(d) Consumption considering the maximum daytime traffic volume BS as Macro cell



(e) Consumption considering the minimum daytime traffic volume BS as Micro cell



(f) Consumption considering the maximum daytime traffic volume BS as Micro cell

Figure 5.3: Conversion from traffic volume to power consumption for the Lowest and Highest daytime traffic volume BSs

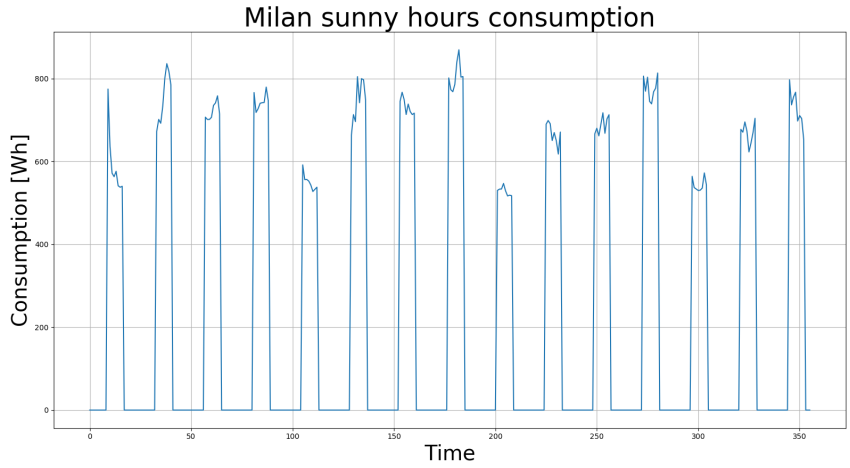
5.3 Energy Saving

Once the annual consumption traces have been generated and the solar radiation traces have been downloaded, it is possible to calculate the energy savings for the different traces. To simplify the work and consider only the portion of data of our interest, a vector of ones and zeros has been used in order to indicate the daylight hours for the three different cities. In particular, for the elements whose position corrisponds to a one in the daylight hours vector, solar radiation is present. For the elements whose positions corresponds to zero no solar radiation in present. To generate the consumption vector related to the sunshine hours, it was enough to multiply the two vectors element by element. Then, by summing all the resulting elements it was possible to calculate the amount of energy needed to keep the relative BS active during the sunny hours.

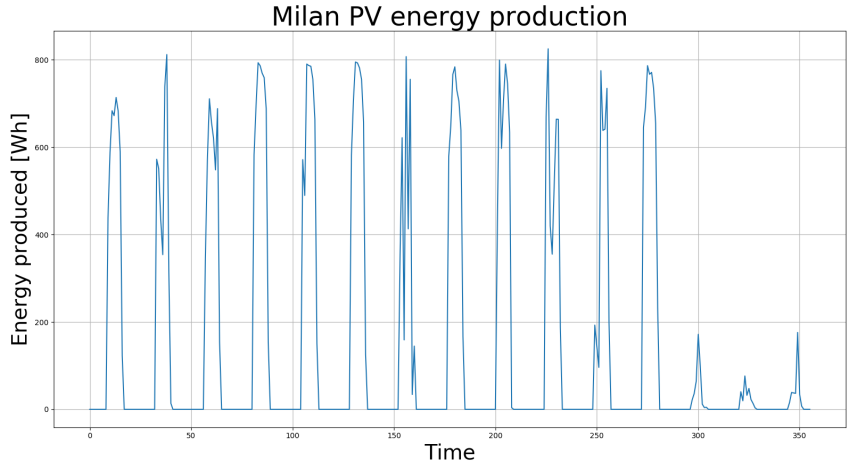
Figure 5.4 shows a portion of the vector of zeros and ones indicating the hours of sunshine in the city of Milan (figure 5.4a), and, with the application of the solar hours' vector, the corresponding consumption (figure 5.4b). Finally, figure 5.4c shows the energy produced by a solar panel of 1 kWp for the shown portion of time.
To calculate the energy saving produced by the installation of a photovoltaic system on a given BS, it will be enough to make the difference between the traces shown in the figures 5.4b and 5.4c. The result will exclusively depend on the consumption trace, since the energy produced is the same inside each specific city.

Once all the necessary elements have been calculated, it is possible to determine the energy saving for the various cases presented above.
As previously mentioned, the calculation was carried out by subtracting the solar production vector of the three cities from each consumption trace. The result of this operation returns a vector that shows the actual consumption once given the application of that particular photovoltaic power supply. At this point, a small trick has to be made on this vector. Since in some points the produced energy exceeds the consumpted one, a negative and non-sense value would came out and therfore it is necessary to zero them. It is important to notiche that, without applying what has just been said, the energy saved would be the same irrespective of the trace.
Finally, to obtain the energy saved it will be enough to subtract the consumption obtained with the application of photovoltaic supply from the total consumption. By adding together all the elements of the resulting trace, we will obtain a unique value that will identify the energy saved throughout the year.

(a) Daylight hours vector



(b) Daylight consumption (partial) for a random trace



(c) Daylight consumption (partial) for a random trace

Figure 5.4: Daylight hours vector and daylight consumption in Milan

Macro Cell Regarding the case in which all the traces come from macrocells, the used parameters, necessary for the power calculation, are the ones present in

table 2.1.
Keeping for the moment the solar panel size to 1 kWp, the table 5.1 shows the average energy saved for the three cities considered.

	Mean Saved Energy [kWh/year]	**PV System Dimension** [kWp]
Milan	1741.614	1
Oslo	1188.703	1
Cairo	2219.263	1

Table 5.1: Mean Saved Energy for the considered cities - Macrocells

As you can see, the average savings reflect what one may imagine that is, the highest saving in Cairo, the intermediate one in Milan, and the lowest in Oslo.

Micro Cell The calculations were also made on the assumption that all the traces came from microcells. The average energy savings for the three cities are also shown in table 5.2.

	Mean Saved Energy [kWh/year]	**PV System Dimension** [kWp]
Milan	435.119	1
Oslo	341.093	1
Cairo	481.427	1

Table 5.2: Mean Saved Energy for the considered cities - Microcells

Firstly, the main difference we can appreciate between the first case and the second one is that the amount of energy saved has considerably decreased. Secondly, the ratios of difference between the various cities have also slightly decreased. As a matter of fact, the saving caused by the city position is less appreciated.

5.4 Dimensioning

In this section, we will address the decision-making process of sizing photovoltaic power supplies for different BS. Then, by cross-referencing with the clusters results deriving from the previous steps, a table describing the best sizing for each cluster has been generated.

In order to size the photovoltaic system, the first topic of discussion is the choice of metric. Ideally, we would like the solar panel, in the absence of a battery,

to be able to generate enough power to cover 100% of the consumption during the solar hours of the day. This is unfortunately impossible, primarily because of the unpredictability of solar panels which depend on the weather (in the absence of sun or cloudy days, production will be almost zero). Secondly, going to size a system that is able to provide almost 100% of the required power, would lead to a big waste of money, since you would be installing a very oversized system.

Cell type	**Spring**	**Summer**	**Autumn**	**Winter**
Macrocell	548.205 [kWh]	564.030 [kWh]	327.809 [kWh]	330.343 [kWh]
Microcell	130.402 [kWh]	131.238 [kWh]	84.707 [kWh]	83.898 [kWh]

Table 5.3: Seasonal Energy Saved

The table 5.3 shows the average seasonal energy savings if the traffic comes from macrocells and microcells, with the installation of a 1 kWp photovoltaic power system. As can be seen, the difference between the energy saved during the spring and summer seasons is more than 200 kWh. Therefore, trying to reach a production close to 100% during all the solar hours of the year, you would fall into overproduction during the spring and summer. Being the case under analysis, a case without battery, we are interested in finding the right compromise between the size of the photovoltaic system and savings. In order not to fall into oversizing, it was decided to size the system in such a way that it is able to cover on average 70% of the daily requirements of a given BS.

The procedure for finding the best sizing was simple. In an iterative way, the average energy saving percentages given by the installation of a certain size of the photovoltaic system were calculated. The size was increased at each step of the process by 0.1 kWp. The process was interrupted when an average energy saving of more than 70% of the total was achieved. Once the dimensioning for one BS was calculated, the process was able to move to the next one, generating a dataset that contained for each BS the best dimension for the photovoltaic power supply. Besides this, the energy savings for the various seasons was also calculated.

5.4.1 PV dimensioning from Clusters

Once obtained the dataset with the dimensions of the photovoltaic systems for the single traces, it is time to cross-reference this dataset with the clusters obtained, as described in 4.6.6. Therfore, it is necessary to create a data structure that contains the optimal dimensions of the photovoltaic system, for each cluster and

sub-cluster. In order to estimate the best dimension for every cluster, all the elements of the corresponding vector are averaged.
The results of this process will also be a useful guide in the planning phase for a new plant. In fact, knowing the area of installation, it is possible to trace back the traffic pattern so that the information about the best size of the plant is roughly available in a few seconds.

Macrocells Tables 5.4, 5.5 and 5.6 show the best dimension for a photovoltaic power supply, in the hypothesis of macrocells traces, for the cities of Milan, Oslo, and Cairo, respectively. It means that they show the minimum panel size that guarantees an average coverage of 70% of the daily energy requirements.
Going to consider separately the three cities, it is possible to see the big difference, in terms of size of photovoltaic system, able to guarantee the same percentage of energy saved. This leads to very different expenses for the installation and consequently different strategies in terms of investment and expenditure amortization.

Starting from table 5.4, which concerns the city of Milan, we notice a sizing range from 4.6 to 5.2 kWp. What is expected and confirmed, is that the clusters that need the bigger photovoltaic systems, correspond to the ones which have higher consumption during the daytime hours. Contrarly, the clusters which have lower consumption during daytime, need small photovoltaic systems.
Having a look at the city of Oslo (table 5.5), the overall situation is different from the previous one. In fact, although we expected a larger size due to the much greater latitude, we did not expect to have to install such large systems to reach the 70 % average coverage. The reason lies in the fact that the days are really short in autumn and winter and, in order to guarantee such a large production during the daytime, it is necessary to have a very large plant. For the same reason, we can see that the energy production in spring and summer is larger than the other cities. In Oslo, the sizing range is wider than in the city of Milan and it is in the range of 17.1 to 19.6 kWp.
In conclusion, table 5.6, which refers to the city of Cairo, points out that for very low latitudes, the ones close to the equator, the traces pattern type present in a cluster is not so relevant. In fact, the size range varies only between 1.2 and 1.4 kWp. This is because the seasonal variations are much more attenuated compared to the cities analyzed previously and the photovoltaic systems are able to produce the energy demanded by a tower even with very small plant size.

Microcells Tables 5.7, 5.8, and 5.9 show the best dimension for a photovoltaic power supply in the hypothesis of microcells traces, for the cities of Milan, Oslo,

Name	Best [kWp]	Spring [kWh]	Summer [kWh]	Autumn [kWh]	Winter [kWh]
Cluster 0.0	4.9	763.315	789.978	460.578	465.670
Cluster 0.1	4.9	743.394	764.463	445.014	447.378
Cluster 0.2	5.1	778.451	800.922	465.489	469.087
Cluster 0.3	4.9	736.558	755.569	438.986	442.154
Cluster 0.4	4.9	771.711	799.023	469.439	477.813
Cluster 0.5	4.8	721.490	738.213	425.739	425.233
Cluster 1.0	4.9	741.843	760.723	443.519	447.139
Cluster 1.1	4.8	740.411	764.508	446.570	449.372
Cluster 1.2	5.1	771.505	790.674	460.736	463.200
Cluster 1.3	4.8	756.241	783.927	458.182	462.461
Cluster 1.4	5.0	772.282	795.699	463.016	466.579
Cluster 1.5	4.7	716.437	735.106	426.729	432.144
Cluster 2.0	5.0	751.995	771.702	449.061	451.975
Cluster 2.1	4.7	720.628	740.054	430.375	432.771
Cluster 2.2	5.1	777.109	800.775	464.748	468.020
Cluster 2.3	4.9	745.318	764.566	444.196	446.876
Cluster 2.4	5.0	762.845	782.788	454.531	457.788
Cluster 2.4	4.9	746.003	765.807	445.265	448.066
Cluster 3.0	5.0	767.607	789.441	458.461	461.310
Cluster 3.1	4.6	689.369	709.103	411.232	414.365
Cluster 3.2	5.1	792.821	817.882	477.360	480.743
Cluster 3.3	4.9	741.084	761.999	443.542	446.533
Cluster 3.4	5.1	767.616	788.724	457.622	459.979
Cluster 3.5	4.8	725.688	743.905	432.111	434.604
Cluster 4.0	4.9	750.071	771.129	448.417	451.128
Cluster 4.1	4.8	724.604	743.411	431.965	434.547
Cluster 4.2	5.2	788.156	809.214	469.930	473.465
Cluster 4.3	4.8	728.877	746.479	433.334	435.696
Cluster 4.4	5.0	755.987	776.384	451.796	454.735
Cluster 4.5	4.8	736.854	757.224	440.650	443.299
Cluster 5.0	4.9	742.072	760.054	441.550	444.072
Cluster 5.1	4.7	707.771	726.920	423.034	425.139
Cluster 5.2	5.2	775.146	797.159	461.658	464.420
Cluster 5.3	4.9	753.960	774.644	450.565	453.544
Cluster 5.4	4.9	746.349	770.507	446.040	449.607
Cluster 5.5	4.9	754.120	779.459	454.476	457.900

Table 5.4: Dimensioning PV power supply through the cluster - Macrocells Milan

Name	Best [kWp]	Spring [kWh]	Summer [kWh]	Autumn [kWh]	Winter [kWh]
Cluster 0.0	18.1	828.305	829.497	309.880	304.699
Cluster 0.1	18.3	809.798	806.682	299.416	291.138
Cluster 0.2	18.9	847.590	844.466	312.833	304.536
Cluster 0.3	18.3	803.292	798.568	295.315	286.651
Cluster 0.4	18.0	835.504	837.348	316.233	311.320
Cluster 0.5	18.4	786.126	780.355	285.512	276.234
Cluster 1.0	18.5	808.803	804.026	297.405	289.150
Cluster 1.1	17.9	804.693	804.645	300.620	294.195
Cluster 1.2	19.1	841.034	835.596	310.013	300.454
Cluster 1.3	17.9	820.466	823.143	308.538	303.579
Cluster 1.4	18.8	840.445	838.550	311.796	304.200
Cluster 1.5	17.8	784.615	777.738	285.552	279.928
Cluster 2.0	18.7	819.785	815.644	302.154	293.171
Cluster 2.1	17.9	785.552	781.778	289.579	281.267
Cluster 2.2	19.0	845.724	843.472	312.496	304.859
Cluster 2.3	18.5	812.899	808.156	298.975	289.966
Cluster 2.4	18.9	831.774	826.971	305.659	296.917
Cluster 2.5	18.5	812.974	808.681	299.676	290.846
Cluster 3.0	18.8	835.740	832.264	308.351	299.873
Cluster 3.1	17.1	751.383	747.899	276.386	270.392
Cluster 3.2	19.0	861.458	860.498	321.544	313.892
Cluster 3.3	18.2	807.906	803.964	298.550	290.494
Cluster 3.4	19.1	837.016	832.944	308.438	299.311
Cluster 3.5	18.1	791.963	786.834	290.652	281.683
Cluster 4.0	18.5	818.235	814.366	301.726	293.057
Cluster 4.1	18.0	790.475	785.788	290.691	282.074
Cluster 4.2	19.6	859.390	854.457	316.058	306.837
Cluster 4.3	18.3	795.725	789.718	291.700	282.317
Cluster 4.4	18.6	824.390	819.559	303.713	295.154
Cluster 4.5	18.1	803.503	799.509	296.697	288.385
Cluster 5.0	18.5	809.703	803.972	297.076	287.913
Cluster 5.1	17.5	771.106	767.564	284.680	276.724
Cluster 5.2	19.5	846.145	841.536	310.133	301.770
Cluster 5.3	18.6	821.592	817.640	303.126	294.362
Cluster 5.4	18.9	809.883	815.033	301.209	291.375
Cluster 5.5	18.1	819.239	819.501	305.532	299.728

Table 5.5: Dimensioning PV power supply through the cluster - Macrocells Oslo

Name	Best [kWp]	Spring [kWh]	Summer [kWh]	Autumn [kWh]	Winter [kWh]
Cluster 0.0	1.3	697.061	731.635	525.500	501.745
Cluster 0.1	1.3	680.342	705.435	512.637	487.344
Cluster 0.2	1.3	708.277	735.406	533.355	507.256
Cluster 0.3	1.3	671.488	693.244	505.003	479.938
Cluster 0.4	1.3	701.376	738.316	527.099	505.814
Cluster 0.5	1.3	660.694	681.435	496.800	470.193
Cluster 1.0	1.3	675.737	698.763	508.683	483.388
Cluster 1.1	1.3	680.848	711.169	514.270	489.637
Cluster 1.2	1.3	703.689	725.769	529.087	502.121
Cluster 1.3	1.3	694.736	730.240	525.092	501.298
Cluster 1.4	1.3	704.329	732.741	530.675	504.645
Cluster 1.5	1.3	657.549	679.119	492.217	472.011
Cluster 2.0	1.3	684.757	707.680	515.006	489.157
Cluster 2.1	1.3	660.299	683.064	497.066	472.508
Cluster 2.2	1.3	707.904	735.989	533.562	507.604
Cluster 2.3	1.3	679.578	701.982	511.216	485.508
Cluster 2.4	1.3	694.528	717.386	522.080	495.832
Cluster 2.5	1.3	680.720	703.635	512.135	486.630
Cluster 3.0	1.3	699.962	725.937	527.245	500.878
Cluster 3.1	1.2	634.180	657.337	477.400	454.698
Cluster 3.2	1.3	722.028	754.172	544.714	518.649
Cluster 3.3	1.3	677.778	702.255	511.122	485.843
Cluster 3.4	1.3	696.954	720.821	526.009	499.058
Cluster 3.5	1.3	662.475	683.036	498.463	473.389
Cluster 4.0	1.3	684.523	708.531	515.565	489.844
Cluster 4.1	1.3	662.840	684.384	498.855	473.871
Cluster 4.2	1.4	719.958	743.269	542.054	514.792
Cluster 4.3	1.3	664.210	683.941	499.543	474.000
Cluster 4.4	1.3	688.636	712.577	518.242	492.308
Cluster 4.5	1.3	674.218	697.804	508.425	483.019
Cluster 5.0	1.3	675.491	695.878	507.313	481.524
Cluster 5.1	1.3	648.941	671.776	488.741	464.325
Cluster 5.2	1.4	709.389	732.226	532.927	506.707
Cluster 5.3	1.3	687.324	711.459	517.012	491.244
Cluster 5.4	1.3	676.968	705.316	511.578	487.418
Cluster 5.5	1.3	692.161	724.033	522.506	498.150

Table 5.6: Dimensioning PV power supply through the cluster - Macrocells Cairo

and Cairo, respectively. The reasoning done in the previous paragraph can be done also in this case. The principal difference between the two cases is in the dimension of the plants. Milan stays in the range between 3.3 and 3.5 kWp, Oslo between 6.9 and 7.5 kWp, and Cairo between 1.0 and 1.1 kWp. The other difference is in the values of the plant dimension, which is smaller.

Name	Best [kWp]	Spring [kWh]	Summer [kWh]	Autumn [kWh]	Winter [kWh]
Cluster 0.0	3.5	152.347	153.217	99.389	98.341
Cluster 0.1	3.4	147.480	148.454	95.745	94.844
Cluster 0.2	3.5	151.632	152.609	98.297	97.322
Cluster 0.3	3.4	149.009	149.954	96.752	95.879
Cluster 0.4	3.5	153.934	154.742	100.715	99.587
Cluster 0.5	3.5	151.240	152.217	97.759	96.752
Cluster 1.0	3.5	151.370	152.334	98.336	97.453
Cluster 1.1	3.4	147.400	148.343	95.909	94.961
Cluster 1.2	3.5	152.917	153.855	99.546	98.580
Cluster 1.3	3.5	151.853	152.680	99.201	98.150
Cluster 1.4	3.5	151.426	152.392	98.358	97.383
Cluster 1.5	3.5	152.634	153.683	99.081	98.306
Cluster 2.0	3.5	150.042	150.997	97.472	96.578
Cluster 2.1	3.3	145.078	146.040	94.129	93.258
Cluster 2.2	3.5	151.831	152.824	98.374	97.372
Cluster 2.3	3.4	149.160	150.112	96.825	95.932
Cluster 2.4	3.5	151.481	152.441	98.354	97.435
Cluster 2.5	3.4	148.480	149.438	96.419	95.526
Cluster 3.0	3.5	150.463	151.457	97.606	96.665
Cluster 3.1	3.3	141.014	141.945	91.515	90.641
Cluster 3.2	3.5	154.004	154.912	100.157	99.067
Cluster 3.3	3.4	148.288	149.254	96.209	95.302
Cluster 3.4	3.5	152.196	153.150	98.737	97.738
Cluster 3.5	3.4	147.779	148.732	95.938	95.059
Cluster 4.0	3.4	149.682	150.660	97.130	96.195
Cluster 4.1	3.4	146.146	147.103	94.863	93.994
Cluster 4.2	3.5	153.502	154.487	99.833	98.815
Cluster 4.3	3.4	148.922	149.886	96.692	95.824
Cluster 4.4	3.5	150.793	151.761	97.910	97.001
Cluster 4.5	3.4	147.748	148.703	95.905	95.003
Cluster 5.0	3.5	150.869	151.829	98.068	97.169
Cluster 5.1	3.3	143.574	144.529	93.185	92.314
Cluster 5.2	3.5	152.436	153.442	98.836	97.951
Cluster 5.3	3.4	149.112	150.082	96.800	95.908
Cluster 5.4	3.5	152.798	154.067	99.174	98.387
Cluster 5.5	3.4	150.005	150.948	97.682	96.704

Table 5.7: Dimensioning PV power supply through the cluster - Microcells Milano

Name	Best [kWp]	Spring [kWh]	Summer [kWh]	Autumn [kWh]	Winter [kWh]
Cluster 0.0	7.3	157.405	157.145	62.699	61.902
Cluster 0.1	7.2	152.607	152.421	60.325	59.779
Cluster 0.2	7.3	156.880	156.662	61.919	61.287
Cluster 0.3	7.2	154.196	153.994	60.965	60.410
Cluster 0.4	7.3	158.925	158.631	63.496	62.571
Cluster 0.5	7.3	156.455	156.208	61.571	60.847
Cluster 1.0	7.3	156.589	156.437	61.782	61.244
Cluster 1.1	7.1	152.469	152.254	60.487	59.884
Cluster 1.2	7.4	158.201	157.964	62.810	62.179
Cluster 1.3	7.2	156.924	156.659	62.599	61.773
Cluster 1.4	7.3	156.583	156.393	61.963	61.335
Cluster 1.5	7.4	157.978	157.756	62.405	61.862
Cluster 2.0	7.3	155.246	155.045	61.445	60.856
Cluster 2.1	7.0	150.153	149.972	59.317	58.796
Cluster 2.2	7.4	157.005	156.808	61.952	61.336
Cluster 2.3	7.2	154.330	154.140	61.030	60.464
Cluster 2.4	7.3	156.722	156.499	61.983	61.388
Cluster 2.5	7.2	153.627	153.432	60.748	60.185
Cluster 3.0	7.3	155.652	155.441	61.494	60.913
Cluster 3.1	6.9	145.937	145.766	57.695	57.199
Cluster 3.2	7.4	159.177	158.929	63.179	62.362
Cluster 3.3	7.2	153.471	153.279	60.619	60.049
Cluster 3.4	7.4	157.366	157.152	62.272	61.669
Cluster 3.5	7.2	152.933	152.749	60.458	59.922
Cluster 4.0	7.2	154.895	154.701	61.198	60.627
Cluster 4.1	7.1	151.252	151.067	59.765	59.246
Cluster 4.2	7.4	158.793	158.567	63.014	62.380
Cluster 4.3	7.2	154.118	153.931	60.936	60.400
Cluster 4.4	7.3	156.016	155.828	61.709	61.126
Cluster 4.5	7.2	152.894	152.708	60.438	59.900
Cluster 5.0	7.3	156.093	155.892	61.810	61.233
Cluster 5.1	7.0	148.598	148.415	58.733	58.214
Cluster 5.2	7.5	157.955	157.732	62.496	61.963
Cluster 5.3	7.2	154.270	154.088	60.985	60.444
Cluster 5.4	7.4	158.145	158.158	62.382	62.130
Cluster 5.5	7.2	155.091	154.878	61.575	60.930

Table 5.8: Dimensioning PV power supply through the cluster - Microcells Oslo

Name	Best [kWp]	Spring [kWh]	Summer [kWh]	Autumn [kWh]	Winter [kWh]
Cluster 0.0	1.0	140.207	142.075	109.583	105.977
Cluster 0.1	1.0	144.506	146.476	113.001	109.301
Cluster 0.2	1.1	143.468	145.301	111.720	107.987
Cluster 0.3	1.0	140.451	142.467	109.818	106.346
Cluster 0.4	1.1	140.926	142.684	109.772	106.420
Cluster 0.5	1.0	137.608	139.409	107.110	103.516
Cluster 1.0	1.0	139.939	141.894	109.233	106.038
Cluster 1.1	1.0	137.077	138.916	107.027	103.720
Cluster 1.2	1.0	139.143	141.077	108.583	105.180
Cluster 1.3	1.0	139.201	141.048	108.750	105.385
Cluster 1.4	1.0	138.045	139.927	107.686	104.385
Cluster 1.5	1.0	143.128	145.051	112.473	109.120
Cluster 2.0	1.0	137.341	139.183	107.387	104.160
Cluster 2.1	1.0	132.653	134.436	103.482	100.304
Cluster 2.2	1.0	141.975	143.891	110.866	107.430
Cluster 2.3	1.0	139.942	141.835	109.344	106.019
Cluster 2.4	1.0	138.840	140.741	108.455	105.153
Cluster 2.5	1.0	139.045	140.916	108.456	104.996
Cluster 3.0	1.0	134.403	136.151	104.786	101.462
Cluster 3.1	1.0	139.217	141.126	108.720	105.375
Cluster 3.2	1.0	140.717	142.634	109.685	106.299
Cluster 3.3	1.0	135.705	137.579	105.869	102.499
Cluster 3.4	1.0	139.907	141.813	108.903	105.614
Cluster 3.5	1.0	135.839	137.707	106.001	102.695
Cluster 4.0	1.0	141.428	143.356	110.474	106.970
Cluster 4.1	1.0	139.328	141.289	108.546	105.159
Cluster 4.2	1.0	138.765	140.732	108.292	104.888
Cluster 4.3	1.0	136.736	138.585	106.778	103.482
Cluster 4.4	1.0	139.075	140.894	108.435	105.033
Cluster 4.5	1.0	138.428	140.361	107.897	104.584
Cluster 5.0	1.0	133.422	135.263	103.970	100.780
Cluster 5.1	1.0	132.892	134.627	103.605	100.352
Cluster 5.2	1.0	138.318	140.275	107.637	104.373
Cluster 5.3	1.1	143.692	145.556	112.282	108.711
Cluster 5.4	1.0	144.233	146.164	112.792	109.179
Cluster 5.5	1.0	142.929	144.744	111.782	108.100

Table 5.9: Dimensioning PV power supply through the cluster - Microcells Cairo

Chapter 6

Conslusions

The objective of this thesis has been to size a photovoltaic system, for the purpose of powering a BS, through the use of clustering techniques.
The work has started with the dataset analysis deriving from 1419 areas of the city of Milan. The dataset is composed by the aggregate traces representing the traffic volume exchanged by all the BSs present in each area. In order to continue with the cluster analysis, the traces corresponding to the various areas were considered as coming from one single BS.
Traces have been considered according to their averaged weekly pattern, for the cluster analysis, and according to their annual pattern for the consumption calculation.
As far as the dimensioning is concerned, different ideas have been made, both regarding the nature of the BS, i.e. whether it is a macrocell or a microcell, and the geographical nature, i.e. the climatic and geographical band the trace belong to. This study has been carried out taking into account three cities: Milan, Oslo and Cairo to consider the temperate, Artic and Tropical zone, respectively.
The results of the sizing have been found consistent with the expected ones, as a matter of fact, an average dimension has been found for the temperate zone, a larger one for the Arctic zone and a smaller one for the tropical zone. Unfortunately, it has not been able to strongly highlight the difference, in terms of size of the photovoltaic power system, between traces that had diametrically opposed patterns. A diametrical opposed pattern refers to traces that differ from each other by the time position of the peak (a trace which shows the peak during the morning with respect to another one which shows its peak during the evening). This is due to the used consumption model, which is composed of a static part much larger than the variable part, which depends on traffic. The latter translates into a flattening of the consumption curves which show a variation during the day according to the traffic trace (variable part of the consumption equation),

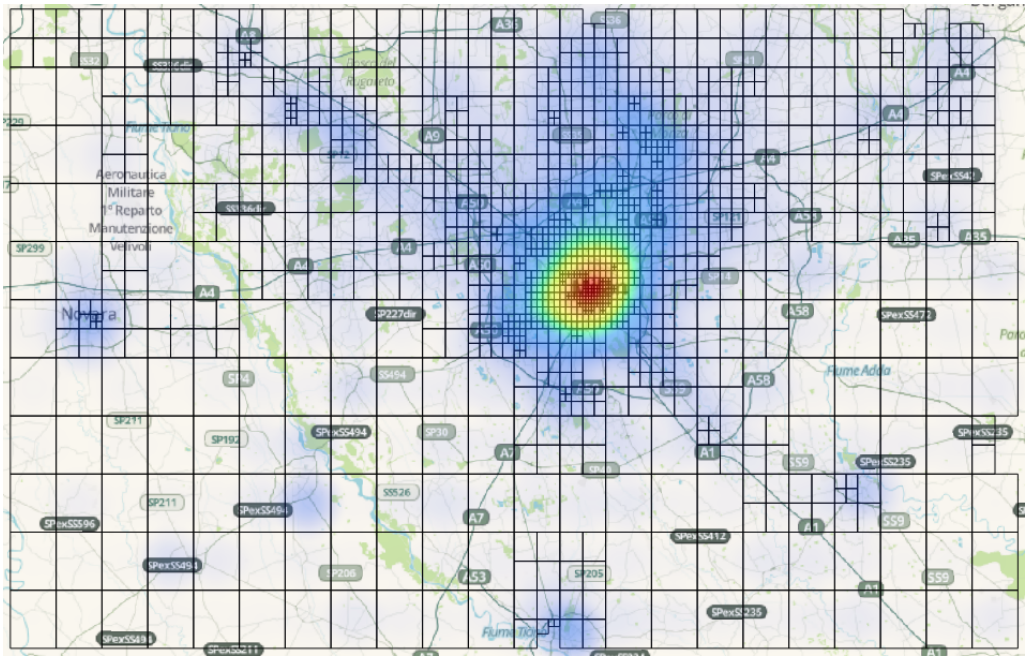
which is nevertheless not that significant with respect to the fixed part of the consumption equation.
In conclusion, it has been possible to size the systems according to the different Clusters. To this scope, a table takining into account the various climate zones and types of tower has been created. In this way, the optimal size of a potential photovoltaic system can be roughly estimated by analyzing the traffic pattern of the specific area where the installation is intended to be done.

6.1 Future Works

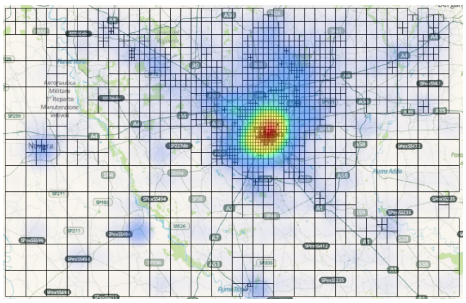
Future work may regard the application of this methodology to data not aggregated by area, that is the one related to the single BSs, knowing also whether they are macrocells or microcells.
Moreover, another type of analysis which could be carried out would be the analysis of energy savings due to the application of the same photovoltaic system to different clusters. This analysis would aims to understand to which cluster may economically be the most convenient to install that system.
Finally, the analysis carried out in this thesis work can be continued. More specifically, it would be interesting to take into account in the system, not only the instantaneous produced energy by the photovoltaic plant, but also the one coming from the insertion of a battery (powered by the extra photovoltaic produced energy).

Appendix A

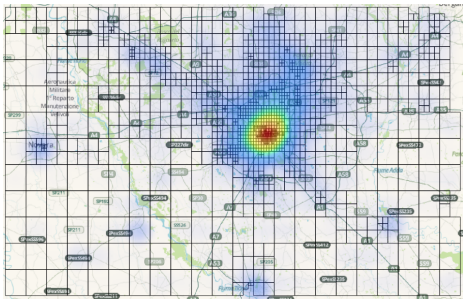
Extra Figures



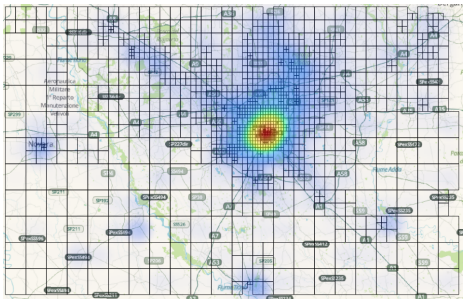
(a) Daily Traffic Consumption Map Distribution



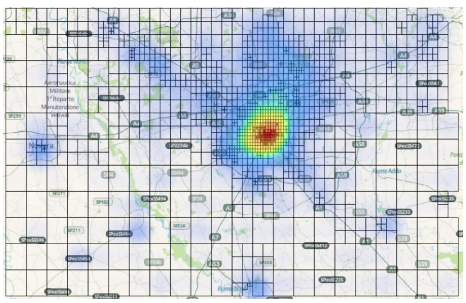
(b) 4 A.M.



(c) 8 A.M.

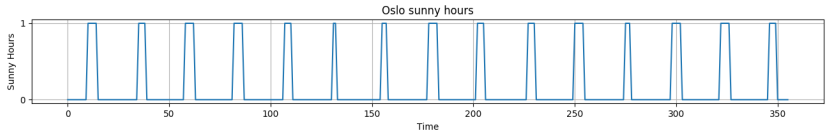


(d) 12 P.M.

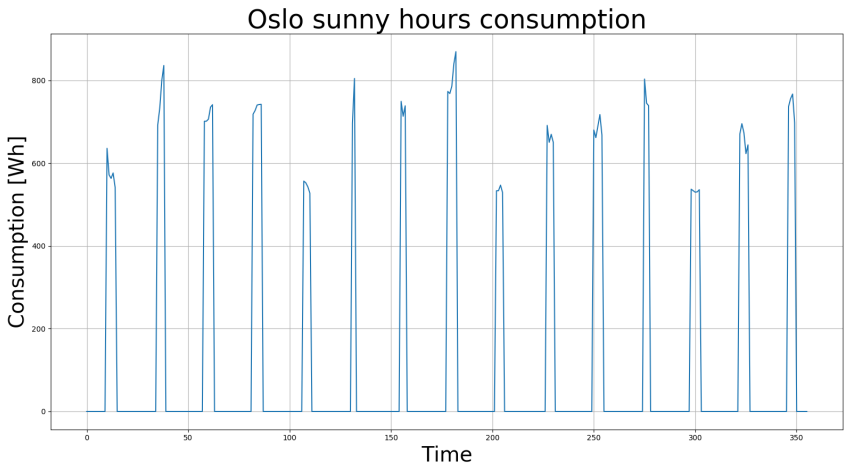


(e) 8 P.M.

Figure A.1: Total and hourly heatmaps

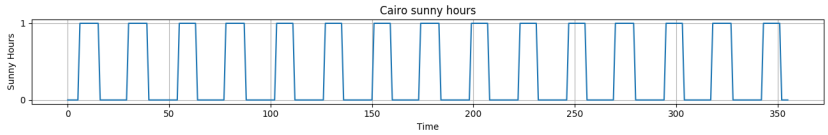


(a) Daylight hours vector

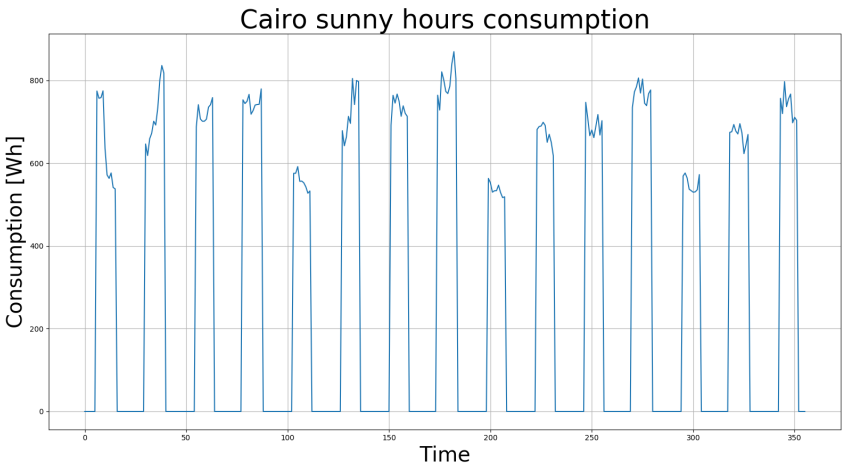


(b) Daylight consumption (partial) for a random trace

Figure A.2: Daylight hours vector and daylight consumption in Oslo



(a) Daylight hours vector



(b) Daylight consumption (partial) for a random trace

Figure A.3: Daylight hours vector and daylight consumption in Cairo

Bibliography

[1] Albert Bifet, Ricard Gavaldà, Geoff Holmes, and Bernhard Pfahringer. *Machine Learning for Data Streams with Practical Examples in MOA*. MIT Press, 2018. `https://moa.cms.waikato.ac.nz/book/`.

[2] StatPortal Open Data. *Dati Open Website*. 2016. Also available at `http://www.datiopen.it/it/opendata/Mappa_delle_antenne_in_Italia`.

[3] David L Davies and Donald W Bouldin. A cluster separation measure. *IEEE transactions on pattern analysis and machine intelligence*, (2):224–227, 1979.

[4] Telefonaktiebolaget LM Ericsson. Ericsson mobility report. 2019.

[5] Solar Power Europe. Global market outlook for solar power 2018–2022. *Solar Power Europe: Brussels, Belgium*, 2018.

[6] A Huerta-Barrientos and M Elizondo-Cortés. Optimizing the cellular network planning process for in-building coverage using simulation. *Journal of applied research and technology*, 11(6):912–919, 2013.

[7] Infrastrutture Wireless Italiane. *Inwit Website*. 2020. Also available at `https://www.inwit.it/it/copertura`.

[8] David J Ketchen and Christopher L Shook. The application of cluster analysis in strategic management research: an analysis and critique. *Strategic management journal*, 17(6):441–458, 1996.

[9] Open Street Map. *Cell Mapper Website*. 2020. Also available at `https://www.cellmapper.net`.

[10] Peter J. Rousseeuw. Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. *Computational and Applied Mathematics*, 20:53–65, 1987.

[11] The European Commission's science and knowledge service. *EU Science HUB Website*. 2015. Also available at `https://re.jrc.ec.europa.eu/pvg_tools/en/tools.html`.

[12] Maria Spano. Tecniche di validazione per il clustering di documenti. *Master Thesis at Universita' degli studi di Napoli Federico II*, 2015.

[13] Robert L Thorndike. Who belongs in the family. In *Psychometrika*. Citeseer, 1953.

[14] Greta Vallero, Daniela Renga, Michela Meo, and Marco Ajmone Marsan. Greener ran operation through machine learning. *IEEE Transactions on Network and Service Management*, 16(3):896–908, 2019.

[15] Huandong Wang, Fengli Xu, Yong Li, Pengyu Zhang, and Depeng Jin. Understanding mobile traffic patterns of large scale cellular towers in urban environment. In *Proceedings of the 2015 Internet Measurement Conference*, pages 225–238, 2015.