Politecnico di Milano
Scuola di Ingegneria Industriale e dell'Informazione
Corso di Laurea Magistrale in Ingegneria Informatica

POLITECNICO
MILANO 1863

SIMBA: systematic clustering-based methodology to support built environment analysis

Relatore: **Prof. Letizia Tanca**
Correlatore: **Prof. Massimo Tadi**
Dott. Carlo Andrea Biraghi

Tesi di Laurea Magistrale di:
Emilia Lenzi, Matr. 898845

Anno Accademico 2019-2020

Contents

List of Figures

List of Tables

List of Algorithms

Abstract

Interest in searching for a sustainable development model has grown enormously over the last 50 years. The reactions of the environmental systems to the continuous extractions of the seemingly unlimited resources were so alarming that in the mid-1970s, experts agreed that the ongoing development models could not continue for long. The Sustainable Development Report 2019 presents the Sustainable Development Goals (SDGs) Index and Dashboards for all United Nations (UN) member states and frames the implementation of the SDGs in terms of six broad transformations. Despite the common interest of research, the road to achieving the SDGs seems to be still a long one. Architecture and urban planning are certainly two of the areas in which research is most devoted to achieving these objectives. Indeed, starting from the analysis of the built environment when investigating human effects on the planet is nothing new, but analysis methods are far from being defined. At the same time, the contribution that computer science can add when addressing a systematic analysis of the territory, both from a morphological point of view and as regards performance analyses, seems to have been underestimated in today's research. It is in this context that this research will fit, joining two - until now separate - worlds, the one of computer science and the one of architecture and urban planning. In particular, in this work, we present SIMBA: systematic clustering-based methodology to support built environment analysis. SIMBA has been thought of as a methodology to support the Integrated Modification Methodology (IMM) developed at the Department of Architecture, Built Environment and Construction Engineering (DABC) of Politecnico di Milano. IMM is a multi-stage, multi-layer, multi-scale, holistic, and iterative process, applied to urban components, and it allows us to evaluate the environmental performance of the city. The first stage of the process is the investigation one, in which the analysis and the synthesis of the territory are performed; this is also the phase when we apply SIMBA. Our case study is the city of Milan and its 88 NILs (Nuclei di identità Locali) on which we will perform clustering. In particular, the advances produced by SIMBA on the IMM methodology include:

- a methodology to select a reasonable but also a representative number of features when investigating the built environment;
- experimental evidence of corresponding patterns between the structural shape of the city and performances; and

- a systematic methodology to measure the distance between elements, needed when comparing different built unit.

Sommario

L'interesse nella ricerca di un modello di sviluppo sostenibile è cresciuto enormemente negli ultimi 50 anni. Le reazioni dei sistemi ambientali alle continue estrazioni di risorse apparentemente illimitate sono state così allarmanti che, a metà degli anni Settanta, gli esperti hanno convenuto che i modelli di sviluppo in corso non potevano continuare a lungo. Il Rapporto sullo sviluppo sostenibile del 2019 presenta l'indice e il quadro SDG(Sustainable Development Goals) per tutti gli Stati membri delle Nazioni Unite (UN) e inquadra l'attuazione degli obiettivi di sviluppo sostenibile in termini di sei ampie trasformazioni. Tuttavia, nonostante l'interesse comune della ricerca, la strada verso il raggiungimeto degli SDGs sembra essere ancora lunga.

L'architettura e l'urbanistica sono certamente due dei settori in cui la ricerca è maggiormente dedicata al raggiungimento di questi obiettivi. Infatti, partire dall'analisi dell'ambiente costruito quando si indaga sugli effetti che l'uomo ha sul pianeta non è una novità, ma i metodi di analisi sono ben lungi dall'essere definiti.

D'altra parte, il contributo dell'informatica nell'affrontare un'analisi sistematica del territorio, sia dal punto di vista morfologico sia per quanto riguarda l'analisi delle prestazioni,sembra essere stato sottovalutato nella ricerca odierna. È in questo contesto che questa ricerca intende inserirsi, unendo due mondi finora separati, quello dell'informatica e quello dell'architettura e dell'urbanistica. In particolare, in questo lavoro presentiamo SIMBA, una metodologia sistematica basata sul clustering, a supporto dell'analisi dell'ambiente costruito. SIMBA è stata pensata come metodologia a supporto di IMM (Integrated Modification Methodology), metodologia sviluppata presso il Dipartimento di Architettura, Ambiente Costruito e Ingegneria delle Costruzioni (DABC) del Politecnico di Milano. L'IMM è un processo multistadio, multistrato, multiscala, olistico e iterativo, applicato alle componenti urbane, che consente di valutare le prestazioni ambientali della città. La prima fase del processo è quella di indagine, in cui viene effettuata l'analisi e la sintesi del territorio ed è la fase in cui SIMBA viene applicato. Il nostro caso di studio è la città di Milano e i suoi 88 NIL (Nuclei di Identità Locale) sui quali è stato effettuato il clustering. In particolare, il contributo di SIMBA ad IMM è legato a:

- una metodologia per selezionare un numero ragionevole ma anche rappresentativo di features nell'indagine dell'ambiente costruito;

- evidenza sperimentale di modelli corrispondenti tra la forma strutturale della città e prestazioni;
- una metodologia sistematica per la misurazione della distanza tra gli elementi, necessaria quando si confrontano diverse unità costruite.

Chapter 1

Introduction

Nowadays it is clear that one of the biggest problems of our century is the impact that human beings' behaviour is having on the environment. In 1972, with the collaboration of American scholars, the Club of Rome published a report titled "The Limits to Growth" [1]. The report argued that only by stopping or at least slowing down the growth of the world's population and agricultural and industrial production would it be possible to reduce pollution and slow down the consumption and exploitation of non-renewable natural resources: minerals, oil, soil fertility. The Club of Rome's publication caused a storm, and there have been several reflections from representatives of different social classes and areas of study. Anyways, parallel with the organized environmental concern in the last 50 years, as the human population of the world doubled, the carbon emission related to the industries raised by more than twice, the planet earth surface warmed by around 0.5°. Celsius on average and its wildlife decreased by 60% [2] [3] [4] [5]. Climate change surprises us year after year, and beside severe damages to our infrastructures and resources, it brings serious and unprecedented economic and socio-political challenges. All this is happening at the golden age of humanity when man's knowledge is flowering, and global collaboration seems to be higher than ever before [5]. Sustainable development has therefore been a topic of discussion for 50 years now, but feasible solutions are a long way off. While anyone is aware of this issue by now, concrete solutions seem to be less and less feasible. In the literature, we find numerous works addressing this problem starting from the study of the built environment, but none of these seem to have led to a satisfactory conclusion. This is undoubtedly due to the infinite number of variables of this problem and its global scope. Yet, information technology, based on the processing of large quantities of data and the use of sophisticated data mining algorithms, is present in all areas. How, then, can architects and computer scientists cooperate to analyse such a complex and multifaceted problem? This is the question from which this thesis starts. The work is based on the research collaboration between the Computer Engineering Section of the Department of Engineering, Information and Bio-engineering (DEIB) [6] and the Architecture and Built Environment

and Construction Engineering(DABC) [7], at Politecnico di Milano. The goal of the thesis is to produce a fully integrated methodology, adding various degrees of innovation to the IMM (Integrated Modification Methodology) [8] procedure, created by the DABC partners to address critical urban emergencies (e.g., transport and energy problems, environmental dynamics and their impacts on carbon emission, human health, and well-being). The IMM_Lab has produced, over the years, very interesting results and advances compared to other rating systems for sustainability assessment, integrating morphology and building typology in the rating procedure, and dealing with the city as CAS. With these two actions, the procedure can consider the context in which the analysis is performed and to deal with the relations among different components of the city, which are usually hard to detect and handle. However, there is still a need for systematisation of the processes, both in the analysis and, possibly more importantly, in choosing which are the features to take into consideration while investigating the environment. To deal with these two problems, in this work we present SIMBA, a systematic clustering- based methodology, created to support the *Investigation* phase (Phase I) of IMM, which corresponds to the built-environment analysis. The choice to base the methodology on clustering is due to the ability of these algorithms to highlight similarities and differences between elements of the datasets. Since the comparison between different built environments is of fundamental importance for the analysis carried out by IMM, both between different elements and for the same element in different phases of its transformation, clustering turns out to be a particularly useful tool. Coherently with IMM, the analysis will be focused on cities. In particular, our case study will be Milan, a city divided into 88 NILs, or neighborhoods, which will be the object of our analysis. For each NIL, we have data organized in four datasets:

- *Indicators* , a dataset of the indexes used in IMM, related to performance aspects of the NIL (e.g. transport)
- *Metrics* , parameters used in NILs but also present in the literature with other names. These values represent different characteristics of the territory. Our metrics are specifically related to the ratio between built areas (Volume) and empty areas (Void)
- *Attributes* , data used to compute metrics
- *Milan* , a dataset of information retrieved from Comune di Milano regarding building performances, air pollution, populations, and services.

We will refer to these different types of data as *Dimensions* , since they represent, in our case, four different dimensions along which we will cluster. SIMBA is divided into three phases. The choice of the granularity in which we will decompose the city of Milan (NILs) and the definition of dimensions constitute the *Built Environment Decomposition* (BED). In the second phase, we will perform clustering on all the different datasets, first using a subset of features selected manually, and then extracting the most significant ones with an

entropy-based algorithm. This phase is called *First Level Clustering* (FLC). We will analyse different results, and these will be the bases for *First Level Clustering* (FLC), the last phase of SIMBA. In this phase, we will use previous results together with expert knowledge to select the most relevant *Dimensions* and features to cluster again. The contribution of this work will be the following:

- a methodology to select a reasonable but also a representative number of indicators and metrics to use when investigating the built environment;
- experimental evidence of corresponding patterns between the structural shape of the city and performances which, in our case, are respectively represented by metrics and indicators;
- a systematic methodology to measure the distance between elements to apply when using a system rating for sustainability assessment.

To do so, we will follow the following structure of the thesis:

– Chapter 2 presents the problem of sustainability, the role of cities, the open challenges, and goals of the work. It also provides a review of the state of the art, presenting the different rating systems and the IMM procedure in all its phases and elements;

– Chapter 3 reports all the notions related to Data Mining and Machine Learning needed to fully understand SIMBA, particularly focusing on clustering definition and usage;

– Chapter 4 describes all the phase of SIMBA and the setting needed for the experimental part;

– Chapter 5 presents the results of all the experiments carried out in the experimental phase;

– Chapter 3 comments more in-depth the obtained results and compares and evaluates all of them;

– Chapter 7 presents the limitations of our approach, together with interesting paths for future works; and

– Chapter 8 exhibits our conclusions and a resume of the impact of our work.

Chapter 2

State of the art and motivations

In this chapter, we present the motivations lying behind our research work. After presenting the topic of sustainability assessment and the actual state of the art, we will describe the IMM theory in all its phases and components involved. Finally, we summarize the goals of our work and the challenges to be faced in achieving such objectives.

2.1 Rating systems for sustainability assessment

The problem of sustainability in building environment has been broadly addressed in the literature. However, many issues are still open, and different approaches have deficiencies in several aspects. We now want to highlight the main problems raised, and the different solutions adopted during the years.

The first problem is the scale at which the analysis of the environment is carried out. More than two decades ago, in fact, the problem was addressed with rating tools for buildings. Although there are high demand and attention to green buildings, it has proved insufficient to guarantee the sustainability of the built environment ([9]; [10]). One of the main critiques of sustainability assessment on the building scale has been its inability to capture what makes a built environment sustainable for its citizens [11]). For this reason, in the last fifteen years, a good number of rating systems at the scale of urban communities also referred to as the neighborhood or district scale, have been introduced to overcome these limits ([12]). Sustainability assessments at the community or city level are proving to be much more than the summation of individual green buildings and infrastructures ([13]; [14]). This is a first important concept we will return on later, that the switch to a larger scale cannot be considered simply as the aggregation of sustainable objects, as scaling up results in complex interactions ([12]) All these methods consist in long lists of indicators with different weights, and benchmark values. Even if every system has its own weights and categories, the main common topics that can be identified are location, planning, transportation, management, biodiversity, economy, and well- being. [15] criticized

them stating that there is no quantitative evidence that a high- rated community emits less carbon than a lower-rated one. Considering the unscientific selection of the criteria, their weights and the benchmarks, this critic is difficult to overcome. Moreover, the aggregated level of assessment, which synthesizes the evaluation in one single rate, reduces the ability to deliver a robust and transparent output ([14]), [16]).

Another limitation of existing systems is the adoption of a static perspective. In these systems, the assessment is a process realized once at the beginning of the urban community development. However, recent definitions of sustainability have encouraged looking at this as a moving target, showing that assessments done at a single time are not sufficient (Brandon and Lombardi 2011). In fact, continuous evaluations should be encouraged, in a way that sustainability assessments become an interactive process, which could be used to map the evolution of urban development ([17]; [18]).

What appears as the main limitation of rating systems in general, and in particular at the district scale, is the difficulty of taking into consideration the physical properties of the context. Context involves a whole host of relevant psychological, social, cultural, economic, geographic, physical, ecological, and technological dimensions of a situation ([19]). Thus, a significant result in one place may or may not be significant in another, due to variations in the cultural and social context ([20]). The attempts to include this dimension produce, on the other side, hardly measurable results, often subjective and not replaceable in different contexts. The rating system clearly measures performances through a set of indicators that risk being disconnected from the reality they represent.

A way of considering context cold be to integrate morphology and building typology with the rating system, but even if morphological characteristics deeply influenced the environment, they introduce another problem that we will face when we will talk about the Complex Adaptive System.

2.2 Cities as CAS

The choice of the city as a starting point to tackle a complex problem and, as mentioned above, covers the most varied areas, is justified by more than one element. Currently, approximately 80% of the global primary energy is consumed in urban areas, and cities are responsible for emitting more than 70% of the total world's greenhouse gases and consuming 60% of disposable water. Nonetheless, cities are the economic engine of the world [21]. Moreover, cities can be defined as CAS. This definition finds its reason in different works, including the "Multi-Scale Modelling Approach for Urban Optimization: Urban Compactness Environmental Implications" by Carlo Andrea Biraghi [22], carried out within the IMM project mentioned above and the starting point for our research. Cities are complex systems since they are composed of interconnected heterogeneous elements that, as a whole, exhibit one or more performances, and, having been proven their capability of learning from the past, they can be defined as adaptive [22].

Starting from these premises, it seems evident that the urban area represents a rich and already complex and interesting representation of the issue we want to focus on. Moreover, scaling well w.r.t comparisons between different samples and data retrieval, cities seem to be perfectly suitable for a conceptual representation, model realization, and model evaluation from a technological point of view.

Sustainability in an urban environment, for its part, does not only depend on environmental performance. The current COVID-19 emergency is a prime example of this. Italian and international studies have analysed the correlation between the spread of the virus and the air quality of different cities. The rapid COVID-19 infection spread observed in selected regions of Northern Italy is supposed to be related to PM10 pollution due to airborne particles able to serve as a carrier of pathogens [23]; moreover, "Patterns in Covid-19 death rates generally mimic patterns in both high population density and high [particulate matter] PM2.5 exposure areas," the Harvard University report says [24]. Needless to say, other environmental parameters such as temperature and relative humidity may represent key the factors in activation and persistence of viruses in the atmosphere [23]. Furthermore, if on the one hand, the modification of the built environment seems to play a role in preventing the virus, on the other one the trend of the virus itself will lead to profound modifications of the urban environment and will highlight characteristics and criticalities. The COVID-19 emergency, from this point of view, is therefore proof of how urban emergencies are characterized by multiple factors and of how the problem of urban performance itself cannot be separated from the one of urban health.

Another aspect, that makes the city interesting but at the same time a complex case study, is that cities are not static objects and evolve through time. They grow, shrink, merge, can be destroyed or abandoned for many different reasons. The actual conformation of existing cities is only a point on the timeline of their evolutionary path. On this timeline, it could be possible to see the disappearance of a city even with the persistence of its physical consistency [25]. This can be explained by the phenomenon of Synekism, the union of several small urban settlements under the rule of a "capital" city and their absorption into a unique composite urban fabric with time. Former towns or villages can so become neighbourhood of a larger expanding city. This makes meaningful considering a city as an entity that inherits the properties of all the other kind of smaller settlements that could be part of it [22]. It is difficult to give a precise and unique definition for cities, in fact there are many found in the literature, depending on the context in which they are given. Merging all the definitions we can identify the fundamental components of a city. The city is a spatially defined area (Void), separated but connected to the countryside, made up of a group of houses and other public buildings (Volume, Function), where communities of citizens (Agents) live, move (Link) and recognize themselves in its name. We also notice that the scale of a city, or its hierarchical position in a region, can be explained by different aspects related to the above-mentioned components. A city is usually more densely built and populated (Volume), more accessible (Link), larger (Void), and host higher-level functions than the surrounding settlements. Moreover, as we said, cities can be defined as CAS.

Cities, considered as Complex Adaptive Systems, have the opportunity to learn from other well-performing systems. The Complex System Theory highlights some structural features of a system that can improve its performances. These are Connectivity, Complexity, and Compactness. Connectivity represents the topological integration of system elements. It is intrinsically related to the relational network existing between the elements of the system and the resulting hierarchy. Complexity refers to the number and the diversity of both elements and connections. A system with heterogeneous elements and a great number of connections have a high level of Complexity. Compactness is the more tangible and physical dimension. While the other two are represented by the relational network describing an abstraction of the system functioning, Compactness deals with the Euclidean space. It is materially built on the interaction between Volumes and Voids and deals with the reciprocal position of elements in space, no matter at which scale [22]. As we already explained in the previous section, only the comprehension of the structure of the city allows us to optimize it and improve its performances. On the other hand, even knowing from existing literature about sustainability implications of compactness, complexity, and connectivity, the problem of defining these three concepts remains open.

2.3 IMM

IMM is a multi-stage, multi-layer, multi-scale, holistic, and iterative process, applied to urban components, and it allows for evaluating the environmental performance of the city (or parts of it). It investigates the relationships between urban morphology and environmental performances by focusing mostly on the subsystems characterized by physical characters and arrangement. It also highlights the need for acting not only on the physical properties of units (architecture), but also on the operation of the urban system considering functions, services, transportation, resource management and everything possibly affecting citizens' behaviour, in an ecological perspective . The IMM methodology is aligned with the 17 Sustainable Development Goals Figure 2.1 promoted by the United Nations, indeed, the main object of this design process is to address a more sustainable and better performing urban arrangement. In IMM, the built environment is considered as a CAS, in which each part or component is structurally co-related with others: thus, a mere local modification starts a chain-reaction and ultimately a structural change of the entire system. With an extremely high level of complexity, cities are always in the state of transformation. The forceful dynamism within their arrangements produces multi-layered reactions for any single action. Rest on the form of adjustments, the whole CAS changes in a long time, brief time, or immediately and all these levels of time-related transformation take place simultaneously. From this perspective, cities are ever- changing entities, and transformation is a continuous process. Though, there are specific patterns of transformation in each specific context, which are inherent to the particularities of that very system. In other words, if two different urban systems undergo similar intervening actions, their reactions would not be the same, and

therefore, the transformation results in any given period would be undoubtedly different. Thus, to plan for any modification on an urban system, it is fundamental to learn about that system's structure. Accordingly, IMM focuses on the systemic arrangements of the built environment and proposes holistic procedures to transform the urban systems into better performing entities based on the unique qualities that each context offers. Rendering the CAS's nature, a mere local action accrued in an individual subset will produce a chain reaction within the network of its elementary parts and trigger a process that consequently leads to the global change of the entire system. That is, system agents adapt themselves in response to the complex network of reactions arisen from individual changes. In IMM, the emergence process of interaction between elementary parts to form a synergy is named *Key Categories* . This is the first element we need to focus on. Key categories are the products of the synergy between elementary parts, thus, a new organization that emerges not (simply as) an additive result of the proprieties of the elementary parts [8] . We will later go back to the definition of *Key Categories* and we will list the ones defined in IMM. What we want to highlight now is that the description of the build environment provided by the IMM procedure does not consider the city as the simple sum of its part. According to this view, the city is not solely a mere aggregation of disconnected energy consumers and the total energy consumption of the city is different from the sum of all of the buildings' consumption. This considerable gap between the total energy consumption of the city and the sum of all consumers is concealed from the urban morphology and urban form of the city [8].

Figure 2.1: SDGs

2.3.1 IMM phases

To address its scope, the IMM procedure is organized in a nonlinear phasing process involving the following structure:

- Phase I. *Investigation*: Analysis and Synthesis

- Phase II. *Assessment and Formulation*
- Phase III. *Intervention and Modification*
- Phase IV. *Optimization*

In Figure 2.2 we illustrate the process. We will briefly describe each phase and then, focusing on Phase I, we will give some definitions of the elements interacting.

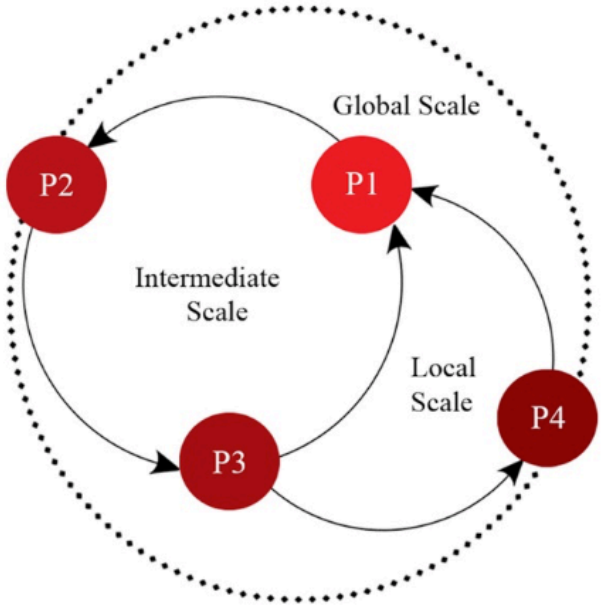


Figure 2.2: IMM - Phases

The process is non-linear. This is reasonable since it must be coherent with the nature of transformations in the built environment that have a non-linear form similar to a spiral. As a complex system, indeed, the built environment is always subjected to change and, at any stage, it produces new thresholds of transformation ([12]). Moreover, the proposed procedure must be accurate enough to cross through the integration of the built environment's different scales and flexible enough to maintain the capacity of the system to learn from itself ([12]). In the first phase, the built environment system is broken down into its subsystems, and the relationship between those parts is investigated. Accordingly, the performance of the system is evaluated in the second phase, and intervention plans are formulated. In this last mentioned stage, the Design Operating Principles (DOP) play a fundamental role as tools/instrument used to arrange the structure of the CAS. In the third phase, design/modification scenarios are tested with the same means that the actual context was investigated, which means,

a circular manner is used until the transformed context is predicted to be acceptable in arrangement and evaluation. The last stage is dedicated to overall optimization through the definition of local retrofitting strategies. Basically, the new form of the CAS is compared with the old one using the same procedures applied in Phase II. Moreover, as we said before, transformation is an endless process, and so this new configuration will become a context for other transformations. Having described the phases of the procedure, we now want to to define its constituent. More precisely, in the first phase, Investigation, the actual state of the system is dismantled into its Components (Volume, Void, Network, Type of Uses) and reassembled into Key Categories (KCs) in order to assess the previously seen system Determinants (Compactness, Complexity and Connectivity) with the goal of achieving an efficient urban form. Let's then see some definitions of the over mentioned elements.

2.3.2 IMM elements

In Figure 2.3 we can see how IMM elements interact. Before to comment the process, we want to give a definition of all of them. All the definition we will provide are consistent with those provide by Carlo Biraghi in his work "Multi-Scale Modelling Approach for Urban Optimization: Urban Compactness Environmental Implications".

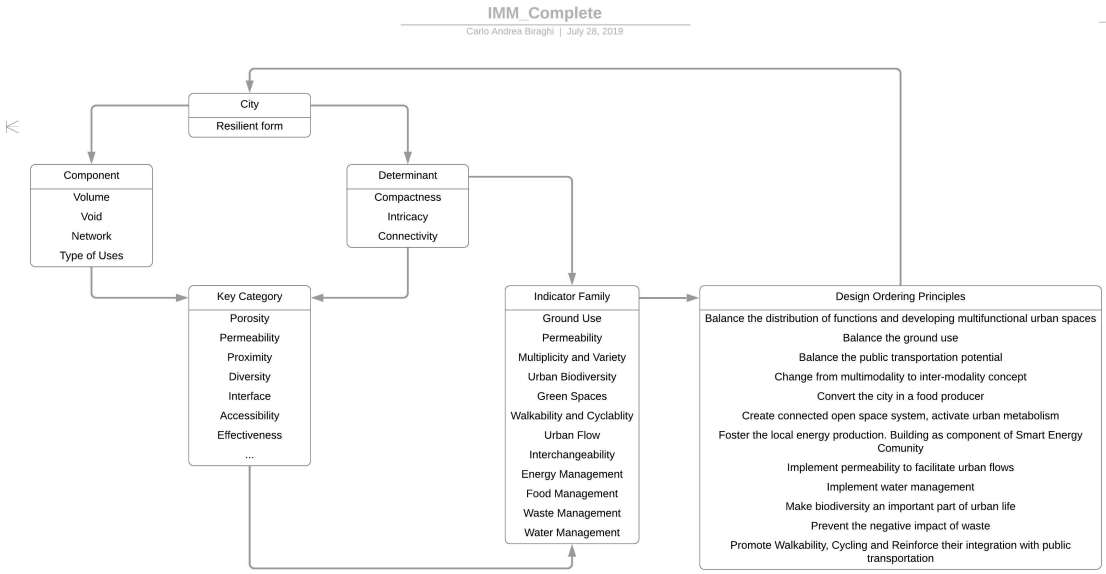


Figure 2.3: IMM - Complete

1. COMPONENTS

   We find different definitions in literature of components. Summarizing all of them

we understand that despite their heterogeneity, cities can be dismantled into four components, whose unpredictable and continuous interaction over times gave birth to contemporary urban areas. These components are Volume (the built part), Void (empty spaces), Function (activities performed by citizens) and Network (networks of different modalities) ([26]). These four elements are the most important ones to understand morphology ([27]). Concluding, components are the least set of elements to be considered when dealing with urban environment. People are agents whose behaviour is affected by the configuration and interaction of these components; with their lives, they affect and reflect the performances of the city

2. KEY CATEGORIES: 1st level of integration

   We have already introduced Key Categories, mentioning their representing role of the synergy between parts of the built environment. This synergetic integration between Key Categories mostly describe the configuration of the CAS (Figure 2.4).

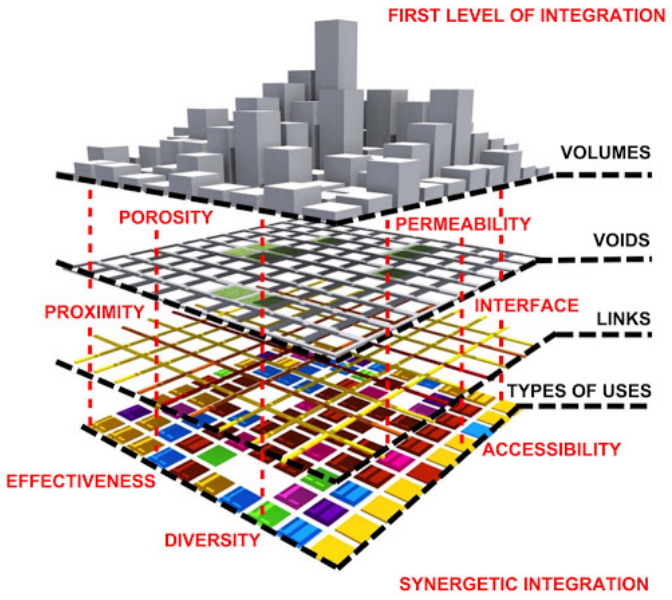


Figure 2.4: IMM - Levels

More precisely, Key Categories used in IMM (up to now) are:

- *Urban Porosity*: the spatial relationship between urban built-ups and voids;
- *Proximity*: the structural relationships driven by the distances between basic land-uses;

- *Diversity*: the structural relationship derived from the different typologies of land-uses;
- *Accessibility*: the mobility patterns driven by dynamic characteristics of origins and destinations;
- *Effectiveness*: the static effect of urban characteristics on the functioning order of mobility systems;
- *Interface*: the characteristics of the street network that influence overall connectivity;
- *Permeability*: the relationship between the street network and spatial component influencing overall connectivity.

Being complex concepts Key Categories are not represented by simple number but IMM uses six quantitative metrics. As we can see, Key Categories only illustrate the structural characteristic and not necessarily the performance. In IMM, like most of the scientific methodologies, the tools for performance evaluation are indicators ([12]).

3. METRICS

   IMM aims at providing quantitative measures (metrics) that can pinpoint significant features of the spatial organization of the urban elements in order to characterize the concept of Key Categories ([12]). Metrics, indeed, describe properties of the sample area. It's possible to create almost an infinite number of metrics even if many could result as redundant because built on the same parameters ([22]). We will give some examples of metrics when we will describe the available dates. We will see in particular the ones related to porosity and permeability.

4. INDICATORS OF PERFORMANCE

   Key Categories are used to describe urban structure and its potential way of operating. In other words, each of them draws a partial picture of potential system behaviour that is not necessarily representative of the real flows of city users. A straight street between two points represents the most direct and fast way to connect them (Permeability) but it will not necessarily be the most used by walking people. Aspects as the presence of activities on the ground floor (Proximity), of public transportation stops (Accessibility) or shading by volumes in hot climates (Porosity) can affect people choice suggesting taking more tortuous (Permeability) but integrated (Interface) streets. The comprehensive configuration of the CAS is mostly described by the correlation between the different subsystems. This distance between structure and performances make necessary the use of indicators to understand the real system behaviour and how it can be explained by system structure. It may happen that unpredictable factors affect system agents generating unexpected behaviours. Each indicator may be representative for more than one KC, according to the same logic that links KC to metrics. This

relationship is mediated by families of indicators representative of the DOP and by the three Determinants plus a fourth category of Management. The list of indicators is open and even if currently counts more than one hundred records, it could easily and hopefully reach a much higher number. Not all the indicators can be calculated in every context according to data availability, but this does not represent a limit. In fact, even just few indicators for every family may be enough to evaluate a transformation.

5. DETERMINANTS: 2nd level of integration

   In the previous paragraph we have seen some concepts largely recognized as features of a well performing system as Compactness, Complexity and Connectivity. IMM accepts this three substituting the word Complexity with Intricacy to avoid the contradiction of assessing Complex Systems directly through Complexity. They are called Determinants as they decisively affect the nature of city and are the result of the second level of integration, based on KCs that are the result of the 1st one, based on components. They are inevitably more complex categories that can be hardly understandable and representable in a synthetic nor simplified way. Their full understanding will be possible only after testing the methodology on multiple case studies and alternative transformation scenarios in order to let emerge some patterns that link them to more simple objects as the one that, as architects or urban designers, we are asked to deal with like buildings, streets, squares and parks, simple instances of the Volume and Void Components.

6. DESIGN ORDERING PRINCIPLES (DOP)

   DOP are essentially a set of actions that designer can perform in order to improve the current system behaviour. They are used during the Assumption & Formulation phase (2) when the designer interprets the results of the diagnostic. These actions are taken from the literature and are generic enough to let the designer freely move in them, and specific enough to guide it in a positive direction. Here the list of the DOP in alphabetical order:

   (a) Balance the distribution of functions and developing multifunctional urban spaces;
   (b) Balance the ground use;
   (c) Balance the public transportation potential;
   (d) Change from multi-modality to inter-modality concept;
   (e) Convert the city in a food producer;
   (f) Create connected open space system, activate urban metabolism;
   (g) Foster the local energy production. Building as component of Smart Energy Community;
   (h) Implement permeability to facilitate urban flows;

(i) Implement water management;

(j) Make biodiversity an important part of urban life;

(k) Prevent the negative impact of waste;

(l) Promote Walkability, Cycling and Reinforce their integration with public transportation.

Even if no one could disagree with none of this principle, it's clear that their positive impact strongly depends on the context where they are applied. They can't harm but they can be useless in some extreme cases. What makes them interesting and different from a checklist of a classical rating system is their dynamic nature. The importance, hierarchy or weight of each of these principles varies according to the case study and is determined by the analysis of KCs. So, the list rearranges every time in order to set some priorities among this goal.

7. ATTRIBUTES

Lastly, we need to define attributes. The IMM procedure is applied to Cities as we said many times, and it uses the above defined elements to perform a complete investigation of the build environment. To compute the elements the procedure needs of course input data or rather, attributes. Attributes could be both geometrical properties and additional information. Some of these data can correspond or be used to compute attributes. In the next chapter we will see some example to clarify this definition. Additional attributes can be obtained by numerical or spatial operation on the existing ones. Attributes are the ingredients for the construction of metrics. Finally, as we said, in fig. 2.3 is represented the interaction between IMM's elements.

The scheme can be read in two directions because of the iterative nature of the approach. Going clockwise city performances are determined, as agreed in the literature, by its level of Compactness, Complexity and Connectivity, called determinants. These, in order to be investigated, need to be dismantled into more targetable urban system properties, the KCs. The number of KCs inside IMM is currently seven but is not fixed. The KCs are the result of the integration of the Components (Volume, Void, Function and Network), the basic constituents of every city. The components are the result of the dismantling process of cities and are so the starting point of the counter clockwise process of urban diagnostic, articulated in KCs and completed by the measurement of performances through the Indicators. The complete scheme shows the relationship between all the components and the KCs, and between them and system determinants, resulting extremely hard to be represented in a comprehensible way given the complexity of the relationships between the elements. Every KC, as we said, can then be associated with more than one indicator as well each indicator may be representative for more than one KC, according to the same logic that links KC to metrics. On the other hand, each indicator is associated to one of the DOP families.

2.3.3 IMM Data Flow

Once described phases and elements of IMM, we now want to focus on the data flow in the methodology to analyse which are the lacks and where Data Mining can fit to help and improve the process.

In Figure 2.5 we show the actual IMM data flow.

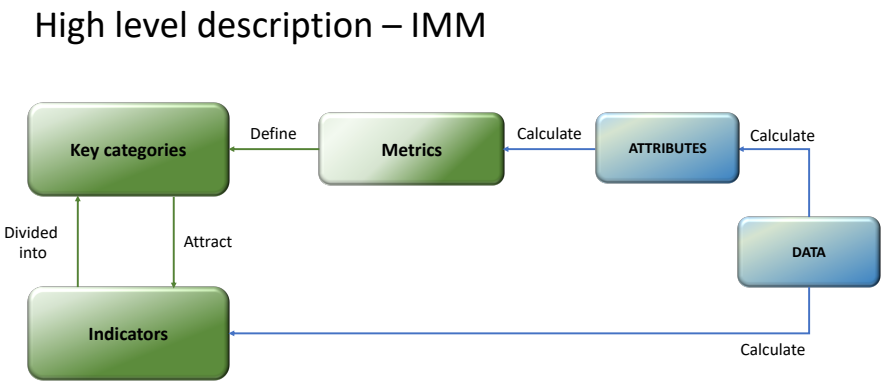


Figure 2.5: IMM - Data Flow

We have already discussed the meaning of different elements in section 2.4.2 but now we want to focus on how they interact. In particular we can see that, starting from raw data they can have different kind and level of aggregation.

- First level of aggregation: data and attributes;
- Second level of aggregation: metrics and indicators
- Third level of aggregation: KC and DOP families

More precisely, from data we compute attribute and from attributes metrics and then, some of these metrics will define key categories. On the other side of the graph we can see that we can calculate indicators from data, and sometimes these are also calculating starting from metrics. The list of possible indicators is still opened, and it can vary every time the framework is applied. Anyway, each indicator can be part of one DOP family.

Every element in the flow has a role and is useful to analyse the city from a specific point of view. Unfortunately, this the good and bad of the procedure. Indeed, if on the one hand this give us a deep description of cities and grasps all the different characteristics of them as CAS (Complex Adaptive System), on the other one it makes the procedure difficult to be applied systematically. Moreover, having such a long and opened list of features makes almost impossible to create a model that generalize well. This last one observation brings

to light a well know problem in the Machine Learning/Data Mining field: Bias Variance trade-off. This term indicates the fact that while we are building a model, we have to choose between making it very accurate or to able to generalize well. In the first case we will need lot of features, but it will lead us to have high variance, while, in the second case, there could be the risk of building a too simple model with few features and high bias. We will not go into the details of this issue. We just want to point out that in our case we are facing the same problem. We want to have a complete and accurate description of the different cities we need to analyse, but at the same time we want to produce a standard procedure to analyse and model them.

It seems reasonable at this point to try to overcome this issue using data mining techniques.

2.4 Goals and challenges

Summarizing, claiming sustainability in urban environments requires a comprehensive understanding of cities as complex adaptive systems and a clear identification of the roles played by the various sub-systems. However, current trends and design methods greatly simplify analytical approaches and practically deal with the subsystems (sectors) as independent entities by neglecting the importance of phenomena resulting from their interconnections at different scales. In this scenario, we think computer science can add significant improvements in the analysis of the built environment. Indeed, many challenges are still opened in urban optimization as we said, and for some of them, data mining techniques can become efficient support tools. In particular what we want to do in this work is to verify whether clustering techniques bring to significant results when used in the diagnostic phase of IMM. The main issue regarding IMM is the need of systematization in all its aspect. In particular the main challenges regard:

- Select a manageable but at the same time descriptive number of features to describe the built environment;
- Compare different built environments or the same built environment in different stages of its transformation;
- Provide a rich and explainable representation of the relationship between different part of CAS.

To overcome this issue we will use SIMBA, systematic clustering-based procedure to support the built environment analysis. The choice to base the methodology on clustering is due to the ability of these algorithms to highlight similarities and differences between elements of the datasets. This is the base to allow comparability among samples and it also allow us to investigate on which are the characteristic which influence more the process. SIMBA methodology is thought to be a support tool to the IMM's Investigation phase.

To prove its efficiency our case study will be the city of Milan, and in particular, we will analyse the 88 Nils (Nuclei di Identità Locale) of the city. This will let us consider the problem at a finer granularity then the whole city, but at the same time, leave the possibility of reassembling the components for future analysis on the whole CAS, without reducing the totality of the effects of the subsystems to the mere sum of them. We will provide a complete methodology including the data gathering and cleaning, the exact algorithm to apply, and the evaluation metrics.

Chapter 3

Theoretical background

3.1 Machine Learning and Data Mining

Machine Learning (ML) is a field of Artificial Intelligence (AI) that provides systems with the capability to automatically learn and improve from experience without being explicitly programmed ([28]) Machine learning focuses on the development of computer programs that can access data and use it to learn for themselves ([29]).

> "A computer program is said to learn from experience E with respect to some class of tasks T and performance measure P, if its performance at task in T, as measurable by P, improves with experience" [30]

The process of learning starts with observations or samples, such as examples, direct experience, or instruction, to search for patterns in data and make better decisions in the future based on the samples that we provide. ([31]). The purpose is to allow computers to learn automatically without human intervention or assistance and to adjust actions accordingly.

Technically, ML is the systematic study of algorithms and statistical models that computer systems use to accomplish a specific task without using precise instructions, relying on patterns and induction instead. Machine learning algorithms build a mathematical model based on sample data, known as training data, to make predictions or decisions. Several definitions of what data mining is have been used, e.g., “automated yet non-trivial extraction of implicit, previously unknown, and potentially useful information from data”, “automated exploration and analysis of large quantities of data in order to discover meaningful patterns”, “computational process of automatically extracting useful knowledge from large amounts of data”.

For what concerns Data Mining (DM), several definitions have been used, e.g., “automated yet non-trivial extraction of implicit, previously unknown, and potentially useful information from data”, “automated exploration and analysis of large quantities of data in order to

discover meaningful patterns", "computational process of automatically extracting useful knowledge from large amounts of data". All definitions are all roughly equivalent to each other. They all agree on the main aspects of data mining, which are: (i) huge quantity of data that (ii) should be analysed so as to (iii) extract what is called "knowledge", or "useful information", or "patterns", i.e.,(iv) something that can be processed and profitably exploited by human beings. ([32])

There are some who regard data mining as synonymous with machine learning. There is no question that some data mining appropriately uses algorithms from machine learning ([33]), however, there are some differences between the two of them. First of all, while data mining relies on human intervention and decision making, with machine learning, once the initial rules are in place, the process of extracting information and 'learning' and refining is automatic. In other words, as we already said, the machine becomes more intelligent by itself. This difference implies that machine learning performs well when we have little knowledge about the problem or, more precisely, what we are looking for in the data. On the other hand, machine learning has not proved successful in situations where we can describe the goals of the mining more directly. In these cases, data mining performs better. More practically, Data mining is used on an existing dataset (like a data warehouse) to find patterns where the 'rules' or patterns are unknown at the start of the process. Machine learning, on the other hand, is trained on a 'training' data set, which together with some rules and variables, teaches the computer how to make sense of data, and then to make predictions about new data sets. Clearly, there are some distinct differences between the two. Yet, as businesses look to become more and more predictive, we may see more overlap between machine learning and data mining in future. For example, more businesses may seek to improve their data mining analytics with machine learning algorithms ([34]). For this reason, also the application of the two fields often overlaps. Some examples of applications are: Costumer segmentation, price prediction, fraud detection and so on.

3.1.1 Machine Learning paradigms

We will now present different machine learning paradigms. We will talk about machine learning and not data mining since historically they were defined in this field. However, the concepts we will present have been inherited by the data mining. Machine learning algorithms are often categorized as supervised or unsupervised.

- Supervised machine learning algorithms can apply what has been learned in the past to new data using labeled examples to predict future events. Starting from the analysis of a known training dataset, the learning algorithm produces an inferred function to make predictions about the output values. The system is able to provide targets for any new input after sufficient training. The learning algorithm can also compare its output with the correct one, intended output and find errors in order to modify the model accordingly.

- In contrast, unsupervised machine learning algorithms are used when the information used to train is neither classified nor labeled. Unsupervised learning studies how systems can infer a function to describe a hidden structure from unlabelled data. The system doesn't figure out the right output, but it explores the data and can draw inferences from datasets to describe hidden structures from unlabelled data.

- Semi-supervised machine learning algorithms fall somewhere in between supervised and unsupervised learning, since they use both labeled and unlabelled data for training – typically a small amount of labeled data and a large amount of unlabelled data. The systems that use this method are able to considerably improve learning accuracy. Usually, semi-supervised learning is chosen when the acquired labeled data requires skilled and relevant resources in order to train it / learn from it. Otherwise, acquiring unlabelled data generally doesn't require additional resources

- Reinforcement machine learning algorithms is a learning method that interacts with its environment by producing actions and discovers errors or rewards. Trial and error search and delayed reward are the most relevant characteristics of reinforcement learning. This method allows machines and software agents to automatically determine the ideal behaviour within a specific context in order to maximize its performance. Simple reward feedback is required for the agent to learn which action is best; this is known as the reinforcement signal. [35]

We will now on focus or analysis on unsupervised learning algorithms and in particular Clustering algorithms.

3.2 Clustering

Clustering is one of the most useful tasks in data analysis. The goal of clustering is to discover groups of similar objects and to identify interesting patterns in the data. Typically, the clustering problem is about partitioning a given data set into groups (clusters) such that the data points in a cluster are more similar to each other than points in different clusters. The objects are typically described as vectors of features (also called attributes). Attributes can be numerical (scalar) or categorical. The assignment can be hard, where each object belongs to one cluster, or fuzzy, where an object can belong to several clusters with a probability. The clusters can be overlapping, though typically they are disjoint. A distance measure is a function that quantifies the similarity of two objects. ([36]).

We can define different methodologies for clustering, according to the set of rules used to define "similarity" among data points:

- Connectivity clustering: as the name suggests, these models are based on the notion that the data points closer in data space exhibit more similarity to each other than the data points lying farther away. These models can follow two approaches. In the

first approach, they start with classifying all data points into separate clusters and then aggregating them as the distance decreases. In the second approach, all data points are classified as a single cluster and then partitioned as the distance increases. Also, the choice of distance function is subjective. These models are very easy to interpret but lacks scalability for handling big datasets. Examples of these models are hierarchical clustering algorithm and its variants ([37]).

- Centroids based clustering: these are iterative clustering algorithms in which the notion of similarity is derived by the closeness of a data point to the centroid of the clusters. K-Means clustering algorithm is a popular algorithm that falls into this category. In these models, the no. of clusters required at the end have to be mentioned beforehand, which makes it important to have prior knowledge of the dataset. These models run iteratively to find the local optima ([37]).

- Distribution clustering: These clustering models are based on the notion of how probable is it that all data points in the cluster belong to the same distribution (For example: Normal, Gaussian). These models often suffer from overfitting. A popular example of these models is Expectation-maximization algorithm which uses multivariate normal distributions ([37]).

- Density based clustering: These models search the data space for areas of varied density of data points in the data space. It isolates various different density regions and assign the data points within these regions in the same cluster. Popular examples of density models are DBSCAN and OPTICS ([37]).

Regardless of the specific methodology followed, a common problem of clustering algorithms is the management of databases high dimensional datasets. In general, having too many features in a data mining model can cause overfitting, i.e., obtaining a very low error during training, but very high error during testing. In the case of clustering, we do not have a training and a test set, so we do not run the risk of overfitting. The problem is however that in high dimensions, almost all pairs of points are equally far away from one another and this makes it impossible to cluster. Looking at the literature, while there are many techniques to overcome this problem in supervised learning, for unsupervised learning the problem of feature selection is still basically untouched.

3.3 Hierarchical clustering

Hierarchical clustering is an algorithm that builds a hierarchy of clusters. In Figure 3.1are shown the two approaches that can be followed depending on the problem we need to solve:

- Agglomerative approach: starting with individual clusters, at each step, merge the closest pair of clusters until only one cluster (or k clusters) left.

- Divisive approach: starting with one cluster, at each step, split a cluster until each cluster contains a point (or there are k clusters) left.

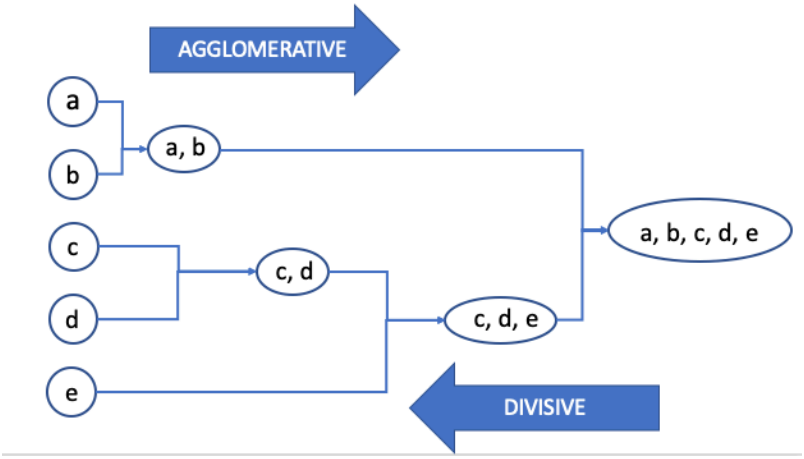


Figure 3.1: Hierarchical clustering.

We will now describe the approach we will use in our experiments, i.e. Agglomerative Hierarchical Clustering

3.3.1 Agglomerative Hierarchical Clustering

Agglomerative Hierarchical Clustering, as the name suggests, is an algorithm that builds hierarchy of clusters. In Algorithm 1 we show the pseudocode of the algorithm.

Algorithm 1: Agglomerative hierarchical clustering

Input: $frequencies\_clusters, max\_val, number\_of\_nils$

$C = \{C_i = \{x_i\} | x_i \in D\}$

$\Delta = \{\delta(x_i, x_j) : x_i, x_j \in D\}$

while $|C \ = k|$ **do**

  Find the closest pair fo clusters $C_i, C_j \in C$

  $C_{ij} = C_i \cup C_j$ // Merge the clusters

  $C = (C\{C_i, C_j\} \cup C_{ij}$ // Update the clustering

  Update distance matric Δ to reflect new clustering

This algorithm starts with all the data points assigned to a cluster of their own. Then two nearest clusters are merged into the same cluster. In the end, this algorithm terminates when there is only a single cluster left. The results of hierarchical clustering can be shown

using dendrogram. To better understand the procedure, we will follow the example carried out by [38]. The dendrogram can be interpreted as showed in Figure 3.2:

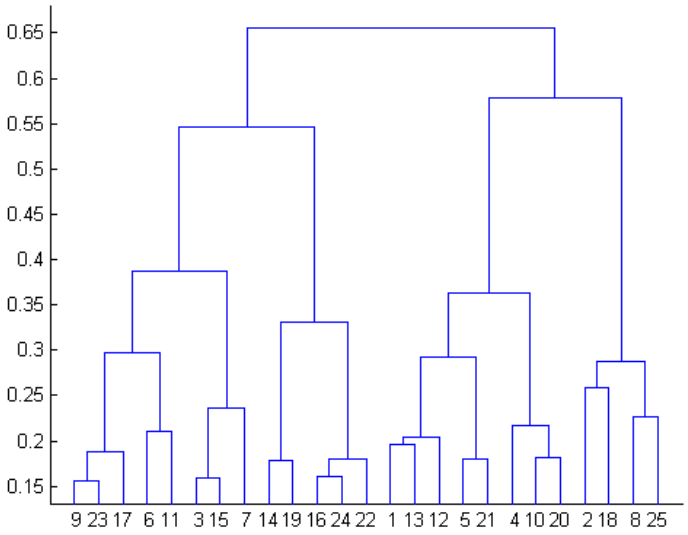


Figure 3.2: Dendrogram example.

At the bottom, we start with 25 data points, each assigned to separate clusters. The two closest clusters are then merged until we have just one cluster at the top. The height in the dendrogram at which two clusters are merged represents the distance between two clusters in the data space. The decision of the number of clusters that can best depict different groups can be chosen by observing the dendrogram. The best choice for the number of clusters is the number of vertical lines in the dendrogram cut by a horizontal line that can transverse the maximum distance vertically without intersecting a cluster(link di prima). In the above example, the best choice is number of clusters equal to 4 as the red horizontal line in the dendrogram in Figure 3.2 covers maximum vertical distance AB.

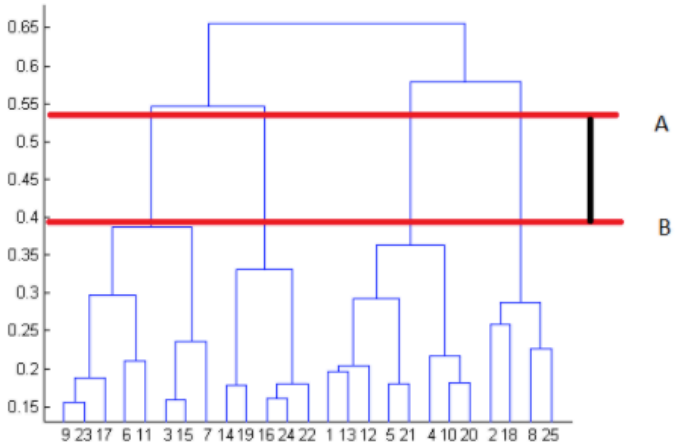


Figure 3.3: Number of clusters from dendrogram.

Advantages of hierarchical clustering are:

- deterministic results independent of initialization and reproducible;
- high level of precision;
- explainable results;and
- need to define number of clusters a priori.

On the other side, the main drawback is the incapability to handle big data. This is because the time complexity of hierarchical clustering is quadratic i.e. $O(n^2)$.

3.4 Evaluation techniques for clustering

Despite clustering problems have been studied extensively over the years, some challenges remain open, first of all how to measure their quality. One can say a good cluster is the one providing a high intra-cluster similarity and a low inter-cluster similarity, but this is not sufficient. In the following sections we will describe different metrics used to evaluate clustering algorithms performances.

3.4.1 Internal clustering evaluation

With internal metrics or internal evaluation techniques, the clustering is summarized to a single quality score. Typical objective functions in clustering formalize the goal of attaining

high intra-cluster similarity (samples within a cluster are similar) and low inter-cluster similarity (sample from different clusters are dissimilar). This is an internal criterion for the quality of a clustering. The indexes used in the literature are many. Here we present two of them.

- Davies-Bouldin index

  The Davies-Bouldin index is an internal evaluation scheme, where the validation of how well the clustering has been done is made using quantities and features inherent to the dataset.

$$DB = \frac{1}{n}\sum_{i=1}^{n}\max_{j=1}(\frac{\delta_i + \delta_j}{d(c_i, c_j)}) \tag{3.1}$$

  Where n is the number of clusters, c_x the centroid of cluster x, δ_i the average distance of all elements in cluster x to centroid c_x, and $d(c_i, c_j)$ is the distance between centroids c_i and c_j . Since algorithms that produce clusters with low intra-cluster distances (high intra-cluster similarity) and high inter-cluster distances (low inter-cluster similarity) will have a low Davies–Bouldin index, the clustering algorithm that produces a collection of clusters with the smallest Davies-Bouldin index is considered the best algorithm based on this criterion.

- Dunn index

  The Dunn index aims to identify dense and well-separated clusters and it is defined as the ratio between the minimal inter-cluster distance to maximal intra-cluster distance. For each cluster partition, the Dunn index can be calculated by the following formula:

$$D = \frac{\min_{1\leq i<j\leq n} d(i,j)}{\min_{1\leq k\leq n} d'(k)} \tag{3.2}$$

  where $d(i, j)$ represents the distance between clusters i and j, and $d'(k)$ measures the intra-cluster distance of cluster k. The inter-cluster distance $d(i, j)$ between two clusters may be any number of distance measures, such as the distance between the centroids of the clusters. Similarly, the intra-cluster distance $d'(k)$ may be measured in a variety way, such as the maximal distance between any pair of elements in cluster k. Since internal criterion seek clusters with high intra-cluster similarity and low inter-cluster similarity, algorithms that produce clusters with high Dunn index are more desirable.

3.4.2 External clustering evaluation

Generally speaking, in external evaluation, clustering results are evaluated based on benchmarks or gold standards. Such benchmarks consist of a set of pre-classified items, often

created by (expert) humans. Thus, the benchmark sets can be thought of as a gold standard or evaluation. These types of evaluation methods measure how close the clustering is to the predetermined benchmark classes. In case of labeled data, it is also possible to evaluated clusters based on the available original classes of the dataset to measure how they fill the shape of the original datasets. External evaluation can also be measured with several different indexes. Here we present Purity.

- Purity To compute the purity, each cluster is assigned to the class which is most frequent in the cluster, and then the accuracy of this assignment is measured by counting the number of correctly assigned documents and dividing by N. Formally it is calculated by the following formula:

$$\frac{1}{n}\sum_{m\in D}|m\cap d| \tag{3.3}$$

Where M is the number of clusters and D is the set of classes and N number of data.

3.4.3 Choose the number of clusters

Regardless of the technique used to evaluate them, performances of clustering algorithms depend on the number of clusters we choose. We talked about this problem when we described the dendograms, but there is another widely used technique, known in literature as Knee-Elbow analysis. This techniques consists in plotting the WSS(Withing clusters Sum of Square) together with the BSS(Between clusters Sum of Square) for every clustering and look for a knee(elbow) in the plot that show a significant modification in the evaluation metrics [39]. To better clarify the method, we show an example in Figure 3.4.

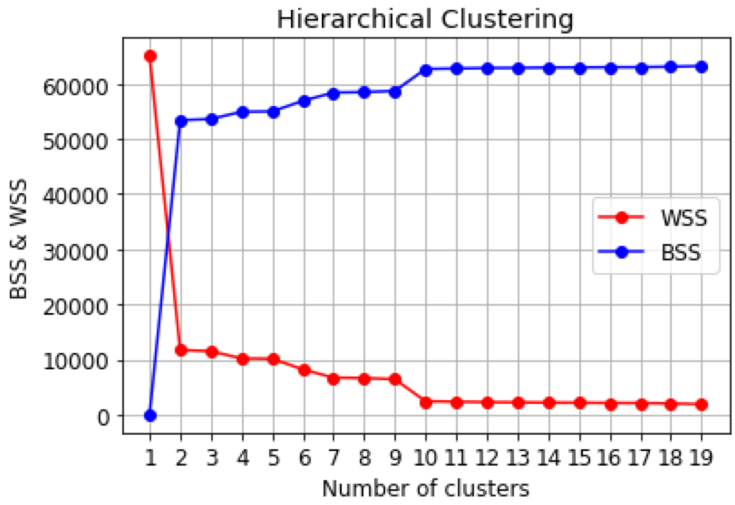


Figure 3.4: Knee-Elbow example.

In the situation showed in Figure 3.4 we can see one first big knee(elbow) for number of clusters equal to 2 and a second smaller one for number of cluster equal to 9. Anyway, even if the WSS(BSS) trend suggests to cluster using 2 or 9 clusters, this is only an indication. Other elements, like for example the presence of some outliers, may influence the result to.

Chapter 4

SIMBA methodology

In this chapter we will give an overview of the approach we use. We will describe each step of our methodology, and then we will present the application of SIMBA to our case study justifying choices made during each step.

4.1 SIMBA flow

In these following sections we want to present SIMBA describing all its phases and its contribution to the IMM methodology. In Figure 4.1 we show the entire flow of SIMBA.

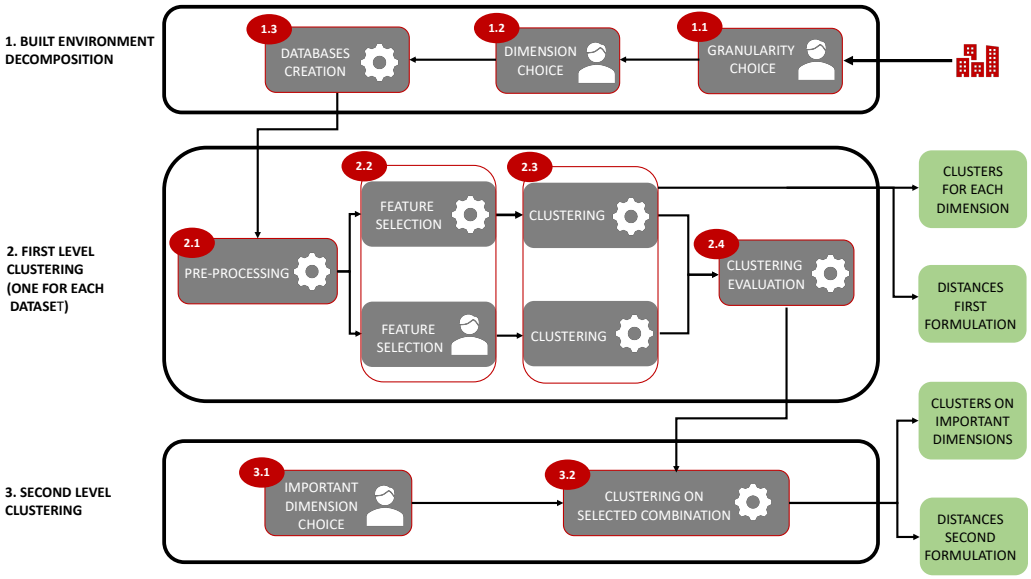


Figure 4.1: SIMBA flow.

The process is composed of three phases:

1. BUILT ENVIRONMENT DECOMPOSITION (BED phase)
2. FIRST LEVEL CLUSTERING (FLC phase, one for each dataset)
3. SECOND LEVEL CLUSTERING (SLC phase)

The input is a built environment of any *Dimension* . In Step 1.1, the granularity at which the analysis is to be conducted must be chosen. In other words, we need to choose which are the samples we want to cluster. According to our case study the input will be a city. As we said, this is reasonable since cities can be considered to be composed of many elements, but the effects on the environment cannot be considered as the mere sum of them. However, even cities are interesting case studies, the methodology can be generalized with any input, as soon as it is possible to identify comparable units in it. After the granularity, we need to define *Dimensions* (1.2) and retrieve data according to them (1.3). *Dimensions* represent the different aspects we want to analyse. They can be of any type and categories i.e. performances, morphology characteristics, demographic data and so on. As SIMBA is applied to the IMM methodology, we can define *Dimensions* as a parallel concept to *Key Categories*. The difference between the two is that while KC are identified by six or seven numerical metrics in IMM, *Dimensions* remain abstract concepts for us. They are simple guidelines while looking for data to retrieve. Is in the datasets indeed, that *Dimensions* are represented. Once decomposed the input, we enter in the FLC phase where, for each dataset, after a pre-processing step (2.1) we perform clustering. Before applying the selected algorithm to cluster (2.3) we perform feature selection in Step 2.2. This is done both manually, using a set of features selected by experts for each *Dimension* , and automatically, using an entropy-based algorithm. Comparing the obtained clusters with the Manual and the Automated procedure, we evaluate our results. In the third and last phase, the SLC one, we combine the evaluation of the obtained results together with the IMM expert's knowledge and needs (3.1), to select which are the *Dimensions* , and thus, features, we want to use in the *Second Level Clustering* . The outputs of the procedure are:

- clusters for each ⋮
- distances between elements for each *Dimension*
- clusters and distances calculated combining only the selected *Dimensions* , using the features selected for each one of them in the FLC phase.

In the next sections we will give an overall description of the setting of each phase of SIMBA, by referring to our specific case study: Milan and its NILs. This will make easier to understand experiments result that we will show in Chapter 5.

4.2 BED phase setting

In the Built Environment decomposition settin,g as we already explained, we have to define:

- the granularity of the analysis i.e. cities, districts, blocks;
- the *Dimensions* we are interested in;
- the datasets that will represent the *Dimensions* . In other words, which features for each *Dimension* are available to us.

The decomposition phase is useful to make the analysis manageable in a Data Mining sense since it allows to:

- increase the number of samples available. Starting from one input we produce different number of samples according to the granularity we choose;
- decrease the number of features in the single analysis. Splitting all the available data in different *Dimensions* to analyse separately, indeed, we have less features in each analysis.

At the same time, according to IMM, city is considered as CAS and all the important elements are taken into account.

In Figure 4.2 , the setting for the BED phase in our experiment is summarized.

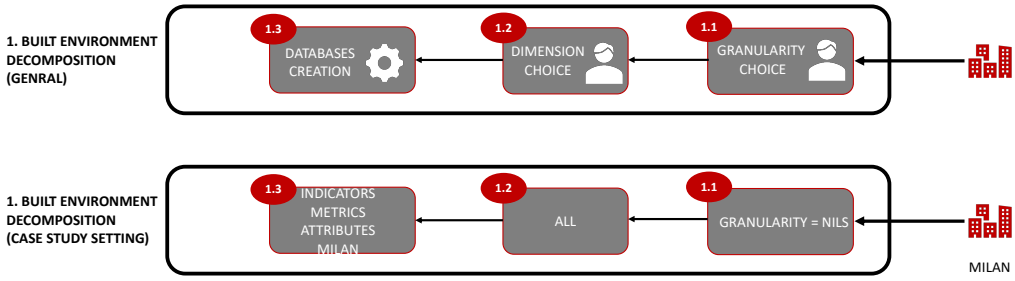


Figure 4.2: BED phase setting.

Now, to explain the setting, we want to describe each sub phase.

4.2.1 Granularity definition

As we already said, the case study will be Milan and the granularity chosen is the NIL. The NILs - Nuclei d'Identità Locale (Local Identity Units) represent areas that can be defined as neighbourhoods in Milan, where it is possible to recognise historical and design

neighbourhoods with different characteristics from each other. They are introduced by the PGT (Piano di Governo del Territorio) as a set of areas, connected to each other by infrastructures and services for mobility and green areas. They are systems of urban vitality: concentrations of local commercial activities, gardens, places of aggregation, services; but they are also 88 nuclei of local identity to be strengthened and planned, and through which small and large services can be organised. The name "NIL" is specific for the city of Milan, but this is not a limit for the generalization of the process since this same division criterion can be found in different cities around the word. Figure 4.3shows the numeration of NIL_ID map for Milan.

Figure 4.3: NIL_ID map.

4.2.2 Dimensions and datasets

For what concerns *Dimensions* , we said they are the aspects we want to consider in our analysis and thus, it is reasonable for us to consider all the possible aspects, according to the availability of data. The dataset we create correspond to the IMM's elements we already defined in Section 2.3.2 and Section 2.3.3 anyway, it worth to analyse separately each dataset available. As shown in the BED phase setting Figure 4.2 we have four different datasets:

1. *Indicators* DATASET

   As we said, indicators are performance indexes. They are grouped in DOP families which are essentially a set of actions that designers can perform in order to improve the current system behaviour. They cover different performances aspects according to the Design Ordering Principles and since we have already mentioned the meaning of each family, we do not want to describe indicators one by one since they are potentially infinite. Table 4.1 summarizes the composition of the dataset.

Table 4.1: *Indicators* dataset summary

Indicators dataset	
Samples	88
Features	25
Missing values	15

2. *Metrics* DATASET

   Metrics are quantitative measures that can pinpoint significant features of the spatial organization of the urban elements in order to characterize the concept of KC. The ones available in our dataset are related to Permeability and Porosity and represent the ratio between different built areas (Volume) and different empty spaces (Void). Again, giving a definition for each metric would be useless. They are basically the results of the combination of different elements of the NILs analysed at different scale. Most considered elements are:

   - Building
   - Courts
   - Blocks
   - Districts

   Elements are combined in different ways, so in the metrics case we do not have a categorization like for indicators. Moreover, also Area and Perimeter of each NILs is considered. Table 4.2 gives a summary of the composition of the dataset.

Table 4.2: *Metrics* dataset summary

Metrics	
Samples	88
Features	59
Missing valuse	101

3. *Attributes* DATASET

   Attributes are used to compute metrics and sometimes correspond to them (as for Area and Perimeter). They are data related to the morphological characteristic of the territory. Consistently with metrics, in our dataset we have attributes related to:

   - Building
   - Courts
   - Blocks
   - Districts

   As for metrics, we do not have a categorization for attributes. In Table 4.3 is reported the summary of the dataset.

   Table 4.3: *Attributes* dataset summary

Attributes dataset	
Samples	86
Features	52
Missing values	162

4. *Milan* DATASET

   This last dataset contains raw data provided by Comune di Milano and related to NILs. The dataset has been created by merging together three datasets:

   - *Aria*: dataset containing the PM10, PM2.5 and NO2 averaged values for the whole 2019;
   - *Buildings*: dataset of characteristic of the buildings of the NILs;
   - *Dati_ quartiere*: data related to population, transports and services in each NIL.

   The pre-processing performed to merge the datasets is described in Section 4.3.1. Table 4.4 summarizes the final elements of the dataset.

   Table 4.4: *Milan* dataset summary

Milan dataset	
Samples	88
Features	31
Missing values	0

Summarising, as it is shown in Table 4.5 , for each NIL we have 167 features. Splitting them into different datasets makes the numbers become more feasible for the clustering analysis. Despite this, considering we have at most 88 samples, we need to perform also features selection inside each dimension.

Table 4.5: All datasets

Dataset	Number of features
Indicators	25
Metrics	59
Attributes	52
Milan	31

Deciding to use separately each dataset means, as we said, to analyse different aspects of cities. Coherently with that, since inside the *Metrics* dataset we have both the metrics related to porosity and the ones related to permeability, we will also compare results obtained on datasets containing only porosity metrics or only permeability metrics. In Figure 4.4 all the different sources have the same colour to underline that we will not care about level of aggregation as we did in IMM (Figure 4.6), we will simply apply the same procedure to every dataset. In our specific case "DATA" corresponds to *Milan* dataset.

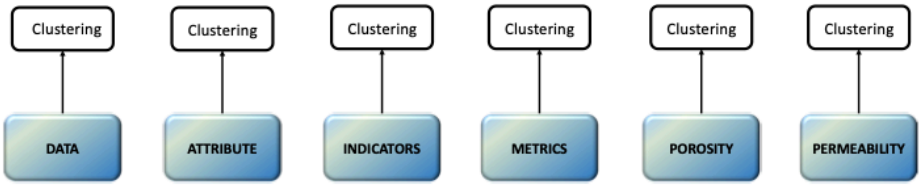


Figure 4.4: Datasets split.

4.3 FLC phase setting

The *First Level Clustering* phase is used to:

- prepare datasets for clustering;
- select the important features for each ;
- perform clustering on each dataset;
- evaluate each obtained cluster.

Outputs of this phase are:

- different cluster division for each dataset;
- distances between NILs for each dataset.

These outputs are useful to investigate patterns for each . The setting in our specific case study for the FLC phase is summarised in Figure 4.5.

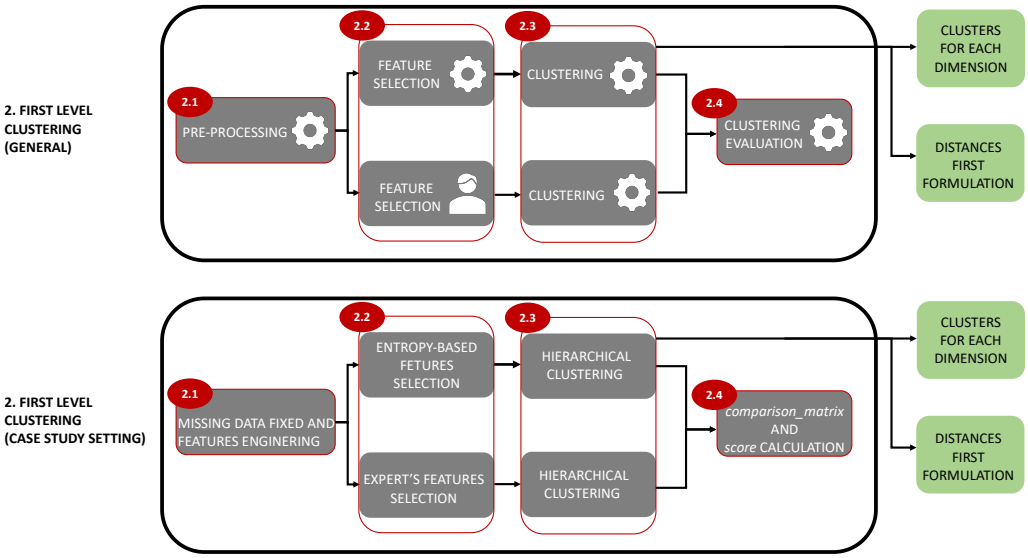


Figure 4.5: FLC Phase setting.

We will now describe more in detail the setting of each sub phase explaining the choices and assumptions behind them.

4.3.1 Pre-processing

We have already described the available datasets and we have seen, all of them have some missing values or inconsistencies between data. In this section we will describe the data cleaning actions we made to solve these issues.

First of all, we need to say that we have two distinct situations depending on the datasets. While for the data coming from Comune di Milano, we can only observe the data we have, for what concerns data regarding indicators, metrics and attributes we are completely aware about the meaning of the features and the kind of missing values we need to deal with. We therefore used different strategies in the two different cases. For what concerns indicators, metrics and attributes, we have two kind of missing values:

- **missing at random (MCAR)**: these are missing values due to the unavailability of the data itself. We cope with these missing values replacing them with the mean of the variable, the minimun values of the column or other values specified by the experts;

- **not missing at random (NMAR)**: these are missing values due to the nature of the data. In our datasets, when a feature is not applicable for a NIL, i.e. there are no green areas, the value is not set to 0 but is missing. Taking this into consideration, we substitute all of these missing values with 0.

The general approach for these datasets has been to not delate any samples since firstly, we are interested in the analysis of all the NILs and secondly, we already have few samples and we do not want to lose too much information. This is moreover a reasonable choice since the percentage of missing data is overall low, so we are not are not introducing too much noise while fixing the missing values.

For what concerns data from *Milan* we do not have missing values, but we need to perform some features engineering on dataset *Edifici*. In particular:

- we drop all the features for what we have most of the values equal to zero. This is done because the majority of values equal to zero makes the attribute irrelevant, even if zero does not correspond to a missing value but to the real value;

- originally, data in the dataset were referred to buildings. We used coordinates to refer each building to each NIL using QGis [40] and we grouped by NILs. With this method we result in having a different number of building per NIL only depending on the sampling. Since this numbers do not reflect the real number of buildings per NIL, we averaged each value to take into account how many instances, and thus buildings, of that NIL are present in the dataset.

4.3.2 Feature selection

Next step of the procedure is dedicated to feature selection. When we talked about high dimensional clustering in Section 3.2, we have stressed how much to choose a good subset of features is important to have a well performing algorithm. Moreover, we have explained that, particularly for clustering, having a high number of features can badly affect the performances of the algorithm but, most of the time, it is not trivial to identify which are good features. For all these reasons we can state with no doubts this is one of the most critical steps of our procedure. As we said, we analyse each dataset using two different approaches. Once we cluster using a subset of features provided by experts and once using a completely automated approach. One may think that a good metric to evaluate the goodness of the features selected automatically is to look at how many of them correspond to the set selected by the experts, but this is neither sufficient nor interest for us. First of

all, we need a methodology to automatically extract them, then we will clarify why we are not interested in the simple comparison with the expert set.

The methodology chosen is based on the Dash and Liu' s 2003 [41] work in entropy-based feature selection for clustering. This is why we will often call the Automated case of each experiment also "entropy-based".

To explain the algorithm, we first need to define entropy and explain the assumption behind its usage in feature ranking. Consider each feature F_i as a random variable while f_i as its value, from entropy theory we know that entropy is:

$$E(F_1, \ldots, F_M) = -\sum_{f_1} \cdots \sum_{f_m} p(f_1, \ldots, f_m) \log p(f_1, \ldots, f_m) \tag{4.1}$$

where $p(f_1, ..., f_M)$ is the probability or density at the point $(f_1, ..., f_M)$. If the probability is uniformly distributed we are most uncertain about the outcome, and entropy is maximum. This will happen when the data points are uniformly distributed in the feature space.

On the other hand, when the data has well formed clusters, the uncertainty is low and so also the entropy. As we do not have a priori information about clusters, calculation of $p(f_1, ..., f_M)$ is not direct, but we can use the following way to calculate entropy without any cluster information [41].

The definition of entropy measure given by Dash and Liu is based on the idea that, two points belonging to the same cluster or different clusters will contribute to the total entropy less than if they were uniformly separated. Similarity $S_{i1,i2}$ between two instances X_{i1} and X_{i2} is high if the instances are very close and low if they are far away. Therefore, according to Dash and Liu, entropy $E_{i1,i2}$ will be low if $S_{i1,i2}$ is either low or high and $E_{i1,i2}$ will be high otherwise [41]. The similarity measure used for numeric data is $S_{i_1,i_2} = e^{\alpha \times D_{i_1,i_2}}$ where D is the distance and it is equal to $D_{i_1,i_2} = [\sum_{k=1}^{M} (\frac{x_{i_1 k} - x_{i_2 k}}{\max k - \min k})^2]^{1/2}$ and α is a parameter calculates as $\alpha = \frac{-\ln 0.5}{\bar{D}}$, where $\bar{D}$ is the averaged distance among data points. The interval in the k^{th} dimension is normalized by dividing it by the maximum interval $(max^k - min^k)$ before calculating the distances. The definition of entropy for a dataset of N points provided by Dash and Liu resulted to be:

$$E = -\sum_{i_1=1}^{N} \sum_{i_2=1}^{N} [S_{i_1,i_2} \times \log S_{i_1,i_2} + (1 - S_{i_1,i_2} \times \log(1 - S_{i_1,i_2})] \tag{4.2}$$

which, for every couple of points X_{i1} and X_{i2}, assume the maximum value of 1,0 for $S_{i_1,i_2} = 1$ and the minimum value of 0,0 for $S_{i_1,i_2} = 0$ and $S_{i_1,i_2} = 1$.

According to this definition, we can rank the features according to their effect on the entropy. Each feature is removed in turn and E is calculated. If the removal of a feature results in minimum E the feature is the least important and vice versa [41]. In Algorithm 2 M is the initial set of features, P is the rank of the features and CalcEnt(Fk) calculates E of the data after discarding feature F_k.

Algorithm 2: Entropy-based features ranking

Input: M
Output: P
P = **for** $k = 0$, $k < |M|$, $k{+}{+}$ **do**
  $P_k = CalcEnt(M_k)$
Sort P **return** P

We apply the entropy-based feature ranking algorithm and we chose only the features that, once dropped, produce an increment of the entropy. The number of features selected in each case is different in each experiment. This is actually one of the strengths of this procedure since the goal of this work is to provide a reasonable analysis of the city characteristic which is on the one hand systematic but also independent from the available data. This is also the reason why we are not interested in the mere comparison of the sets of features while we are more focused on the evaluation of the final results. Different features selected by the algorithm have been of course analysed while doing the experiment and their reasonability have been part of the evaluation of the performances of the procedure, however, as we said multiple times, unsupervised features selection is largely untouched and the purpose of this work is not to directly evaluate the entropy based features ranking algorithm.

4.3.3 Clustering

The chosen method for clustering has been Agglomerative Hierarchical Clustering for two main reasons:

- few numbers of samples available;
- explicability

As we discussed in Chapter 3, Hierarchical Clustering techniques have several advantages including that of not having to assume a priori number of clusters, which allows us to choose a different number of clusters depending on the case. The major disadvantage is the temporal quadratic complexity in N number of elements. This problem does not arise in our case, since having 88 or 86 NILs, the complexity of the algorithm is always manageable. Moreover, this technique allows us to represent the results using dendrograms that represent points and how they are clustering, without taking into consideration number of the dimensions that in our experiments is always more than 2. Having stated this, we now describe how to set the parameters of the algorithm. Here we present a general overview. A more detailed description of the implementation choices can be found in Chapter 5, where we will report and comment the experiments results. We first standardize data using the `StandardScaler()` function provided by shikit- learn [42]. We need to standardize since we have different scales variables with different unit of measure, and we do not want to

make any assumption on the weight each variable has while clustering. In addition, we prefer standardization instead of normalization to not suppress the effect of outliers [43]. Once having the standardized dataset, containing only the features we selected (manually or using the entropy-based algorithm), we run the algorithm. We have already described how Hierarchical clustering works theoretically. For what concerns implementation we need to set three main parameters:

1. `n_clusters`: this parameter represents the final number of clusters we want to obtain. It is different in each experiment, we choose it according to dendrogram, the WSS and BSS trends and comparability among experiment. We will discuss each choice in Chapter 5;

2. `affinity`: this parameter represents the distance measure between samples. We choose to use euclideian distance;

3. `linkage`: this parameter represents how we calculate the euclidean distance between clusters. We decide to measure distance as the euclidean distance between the two farther points so, when we compare distances between clusters, we will compare the max distance between each other. This is the complete linkage method, so the parameter is set to "complete".

4.3.4 Clustering evaluation

Last step of the procedure is dedicated to the clustering evaluation. As we broadly discussed in Chapter 3, clustering evaluation is still a tricky part in this field and most of the time is strictly related to the application. Firstly, we tried to have an absolute evaluation of each cluster result using internal metrics such us Dunn and Davies Bouldin index. The problem with this metrics is that, having a small number of samples, even few distant samples in a cluster would produce a decrease in the score. For this reason, we decided to evaluate clustering results referring mostly to the comparison between the Manual and the Automated results. Indeed, we mentioned the comparison with a ground truth as one the techniques used to evaluate clustering. There is no ground truth in this case but thanks to the collaboration with ABC department we are able to compare our completely automated approach with a more guided one. To do so, we compute the `comparison_matrix` between the two approaches for each experiment. The pseudo code is show in Algorithm 3 below.

Algorithm 3: Comparison_matrix computation

Input: *cluster_results*
Output: *comparison_matrix*
$idx\_nil = cluster\_results.columns[0]$
for $i = 0, i < len(idx\_nil), i++$ **do**
  for $j = 0, j < len(idx\_nil), j++$ **do**
    $comparison\_matrix[i][j] = (cluster\_results[i] == cluster\_results[j]).sum()$
return *comparison_matrix*

`Comparison_matrix` is a simple matrix of dimension MxM where M is the number of NILs for the dataset. Taking into account the clusters results, what the matrix evaluate is how many times two NILs are grouped in the same cluster. For each approach, it checks if the two NILs have the same cluster_ID and, if they do, it increments the cell corresponding to that couple of NILs. For each experiment we have a matrix having values between zero and two:

- 0 means Nils are never in the same cluster;
- 1 means Nils are once together once not;
- 2 means they are always in the same cluster.

This means that positive cases are values 0 and 2 since they mean clustering in the two approaches has produced the same results for that couple. We then compare different experiments using the variable `score` whose computation is shown in Algorithm 4.

Algorithm 4: Score computation

Input: $comparison\_matrix, max\_val, number\_of\_nils$
Output: *score*
for $i = 0, comparison\_matrix.index, i++$ **do**
  for $j = 0, comparison\_matrix.columns, j++$ **do**
    if *comparison_matrix[i][j]* $== 0$ *or comparison_matrix[i][j]* $== max\_val$ **then**
      $good = good + 1$
return $score = \frac{good}{number\_of\_nils}$

`Score` counts how many times value 0 or 2 occur in `comparison_matrix` and it normalizes this number with the number of instances (88 or 86). Assuming that a good result for our experiment is that the clustering algorithm groups the NILSs in the same way both in the Manual and in the Automated case, `score` can be seen as an accuracy measure for the procedure. Moreover, it can be used to compare also different experiments for the FLC phase. We only need to set `max_val` equal to number of experiments we are comparing.

4.4 SLC phase setting

Second Level Clustering phase takes:

- clusters results to identify on which dataset clustering performed better;
- IMM expert's indication about the dimension needed for the analysis;

and produces two different outputs:

- clusters results on the dataset created by combining the important *Dimensions* ;
- a formulation of distances between NILs considering only selected features of the important *Dimensions* .

These outputs represent the final results of SIMBA procedure and will be used, together with the FLC ones, as inputs for the *Investigation* phase in IMM. The setting for our specific case study for the SLC phase is summarised in Figure 4.6.

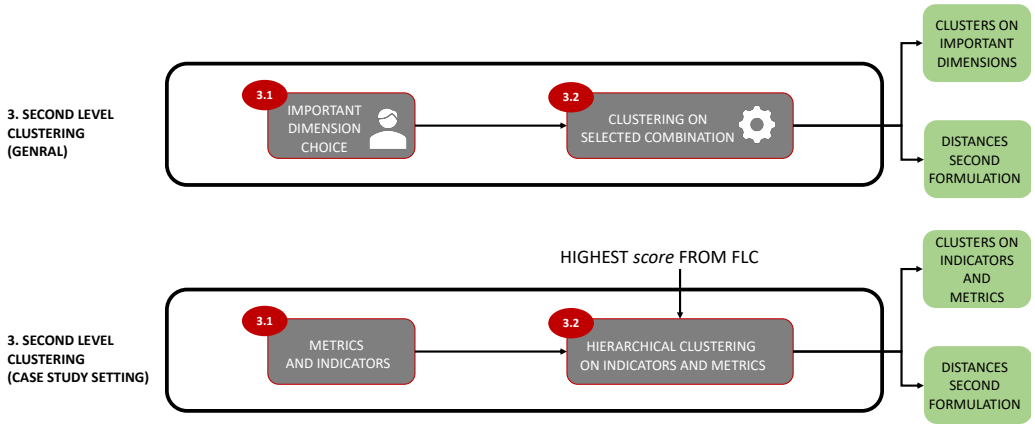


Figure 4.6: SLC phase setting.

Chapter 5

Experiments

In chapter 4 we described in depth each available dataset, and we explained the selected algorithms for feature selection and clustering.

In this section we will present the obtained results. We provide the selected features list and we will analyse the obtained clusters, looking at the dendrogram produced for each experiment, both for the manual and the automated case. The number of clusters in each case has been chosen according to the dendrogram, the knee elbow graph and the overall assumption that, since the number of NILs is less or equal than 88, a reasonable number of clusters for the problem must be less than 10. For each experiment, in the manual case the covariance among features is not taking into account. In the entropy-based case, before to run the feature selection algorithm, we drop all of those having correlation ≥ 0.8.

5.1 Experiment 1 - FLC for *Indicators* dataset

In this first experiment, we apply both the manual and the automated procedure to dataset *Indicators* .

5.1.1 Manual feature selection

The set of features provided by experts corresponds to column "Manual" in Table 5.1

Figure 5.1 shows the obtained dendrogram. Two big clusters are visible and there is one cluster composed by only one sample that we will define as an outlier. Sample number 7 corresponds to NIL 8 “PARCO SEMPIONE”. Since looking at the WSS-BSS graph in Figure 5.2 we can see a big step between number of clusters equal to 2 and number of clusters equal to 3, the agglomerative hierarchical clustering is run setting `n_clusters` = 3.

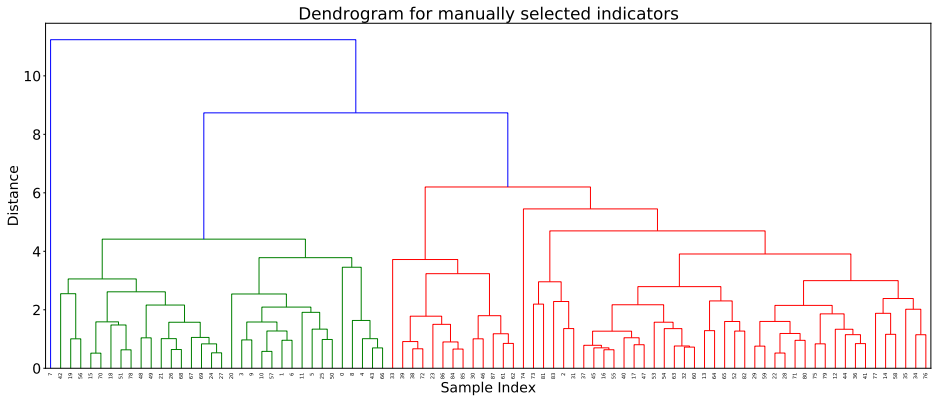


Figure 5.1: Dendrogram for clusters considering only manually selected indicators

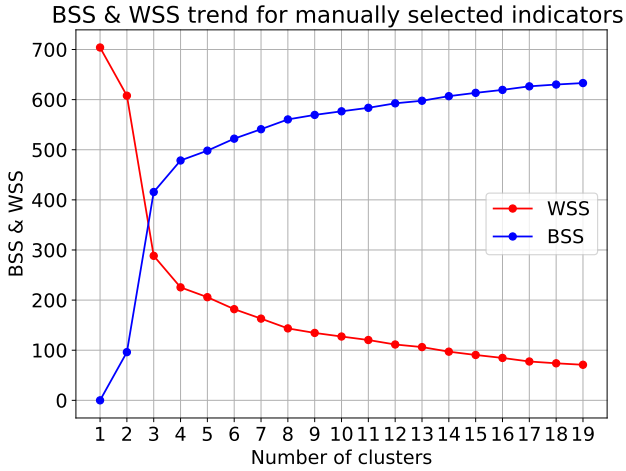


Figure 5.2: BSS and WSS trend for clusters considering only manually selected indicators

5.1.2 Automated feature selection

The set of features provided by the entropy-based algorithm corresponds to column *Automated* of Table 5.1

Figure 5.3 shows the obtained dendrogram, while Figure 5.4 shows the WSS-BSS trend. Also in this case, two big clusters are visible and there are other two clusters, composed by only one sample, that we will define as outliers. Sample number 7 corresponds to NIL 8 "PARCO SEMPIONE" and sample number 46 corresponds to NIL 47 "CANTALUPA". The

WSS-BSS graph of fig. 5.4 in this case presents more steps. We chose `n_clusters` =4 for consistency with the manual case, but also `n_clusters` = 6 and `n_clusters` = 8 could be interesting cases to analyse.

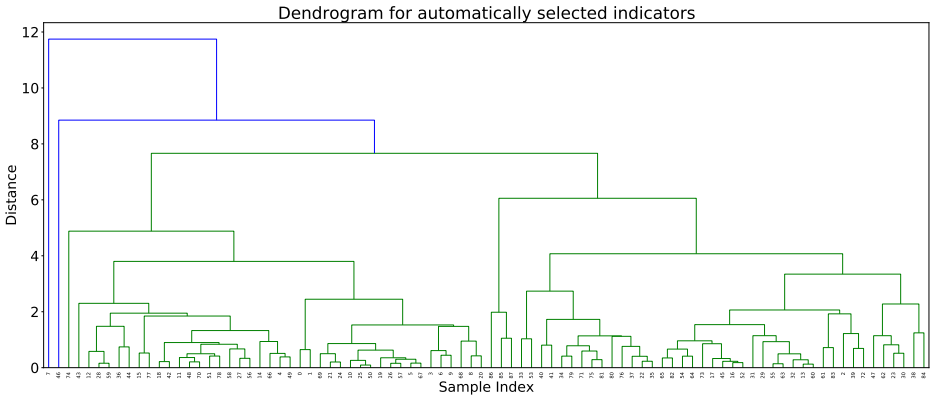


Figure 5.3: Dendrogram for clusters considering only automatically selected indicators

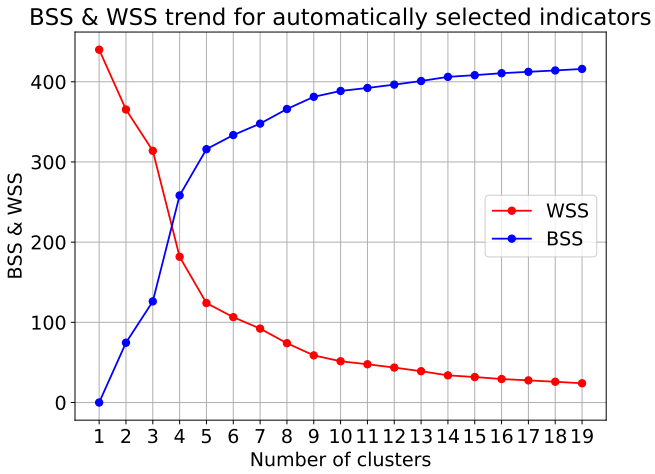


Figure 5.4: BSS and WSS trend for clusters considering only automatically selected indicators

5.1.3 Summary

In Table 5.1 we compare the different features selected using different approaches. Even though the set corresponding to the manual case and the automated one are basically

disjoint, `score` is 76,34, which means that almost all the NILs were clustered in the same way in the two approaches.

Table 5.1: Features *Indicators*

Original	**Manual**	**Less correlated**	**Automated**
VD		X	X
BD		X	
PD	X		
SCR			
BLD	X		
PAcR		X	
JHR	X	X	X
LUsh		X	
GCRt	X	X	X
GCRu		X	
TD		X	
BikeD		X	
BikeAI		X	
ND			
AxBLP		X	X
GFAc	X		
PTA			
LIPR			
NDER	X	X	
Modesh	X	X	
MMsh		X	
StopD			
LineD	X	X	
GCRa			
WAR		X	X

5.2 Experiment 2 - FLC for *Metrics* dataset

In this second experiment we apply both the manual and the automated procedure to dataset *Metrics* .

5.2.1 Manual feature selection

The set of features provided by the experts corresponds to column *Manual* of Table 5.2 and Table 5.3. Figure 5.5 shows the obtained dendrogram while Figure 5.6 shows the WSS-BSS trend. Looking at the first one, 4 clusters are evident: the green, the red, the azure and the purple. On the other hand, the knee-elbow graph does not present any relevant steps for `n_clusters` = 4 and the trend continues to grow/decrease after 4. This is probably due to the fact that, stopping the algorithm for `n_clusters` = 4, some sub-clusters are still merged at a quite significant distance. We should wait for `n_clusters` = 7 or more to have a good WSS/BSS. Since these values would be meaningless, we choose `n_clusters` = 4 coherently with the dendrogram.

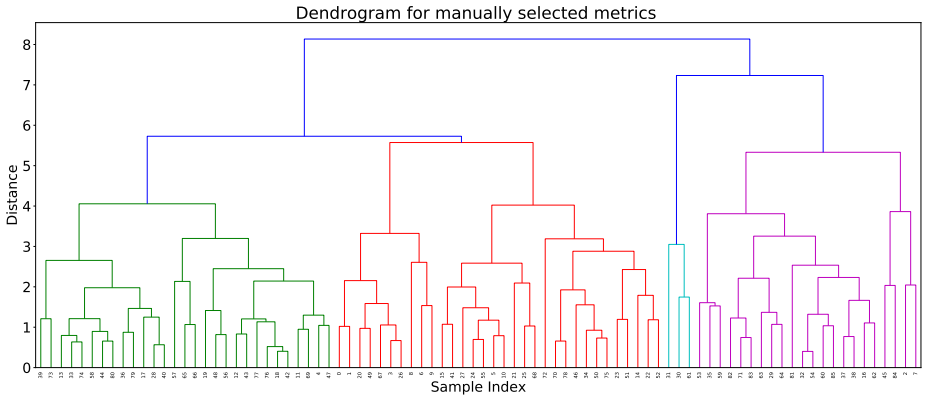


Figure 5.5: Dendrogram for clusters considering only manually selected metrics

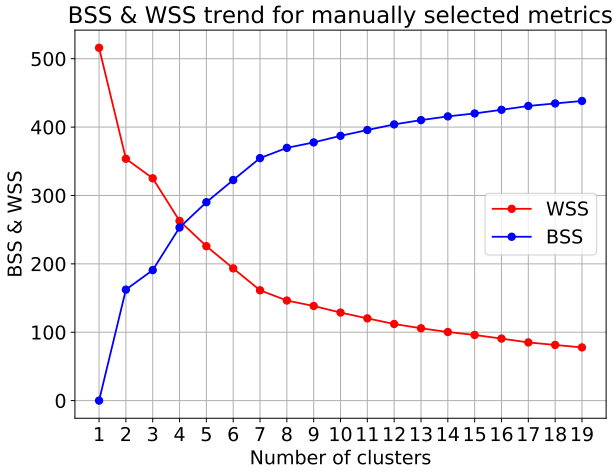


Figure 5.6: BSS and WSS trend for clusters considering only manually selected metrics

5.2.2 Automated feature selection

The set of features provided by the entropy-based algorithm corresponds to column *Automated* of Table 5.2 and Table 5.3.

Figure 5.7 shows the obtained dendrogram. In this case we have two big clusters, other two smaller ones, and one cluster composed by only one sample. NIL 85 "PARCO DELLE ABAZIE" resulted to be an outlier. The choice of `n_clusters` = 5 seems to be also coherent with the WSS-BSS graph in Figure 5.8 even though both the latter and the dendrogram suggest that `n_clusters` = 6 could be interesting to analyse too.

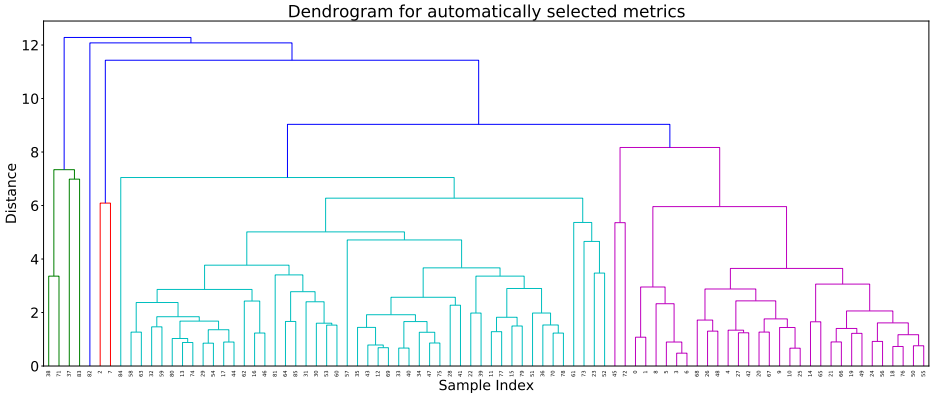


Figure 5.7: Dendrogram for clusters considering only automatically selected metrics

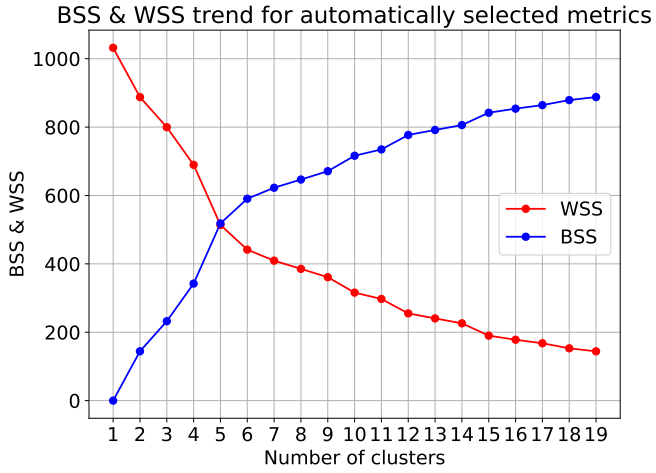


Figure 5.8: BSS and WSS trend for clusters considering only automatically selected metrics

5.2.3 Summary

Table 5.2 and Table 5.3 compare different sets of features used in different steps and approaches. As in the previous experiment, the manual and the automated approach have almost no features in common, but this time also the `score` is significantly lower, i.e. equal to 49,9.

Table 5.2: Features *Metrics* , 1

Original	**Manual**	**Less correlated**	**Automated**
BCR_G	X	X	X
FAR_G			
BCR_N		X	X
FAR_N	X		
BVR			
BBVR		X	
BCHVR		X	
BVD			
FAI			
BSR			
FSD			
CAR		X	X
BSAR		X	
B_AMBG		X	
B_PMBG		X	X
B_CMBG		X	
B_OMBG		X	
B_ACH			
B_PCH		X	X
B_ACHMBG			
BD			
B/UV		X	
B_DS/B			
BDF5		X	
BDF10	X		
UBR		X	
Apass		X	

Table 5.3: Features *Metrics* , 2

Original	**Manual**	**Less correlated**	**Automated**
P/A		X	
S/V		X	X
Cff		X	
Cipq		X	X
Cndc_B			
Cdcm		X	
Concavity		X	
SAR		X	
BLD			
IC	X		
BL_P/A	X	X	
BL_AMBG		X	
BL_PMBG		X	
BL_OMBG		X	
BL_CMBG	X	X	
AwaP		X	
VBLAR		X	
VBLPR		X	X
BL_PBLmean			
BL_BPLR		X	X
BTBL		X	X
BDSTBL		X	
CTD			
CTTB			
CTTBDS			
CTAR_N			
CTBLPR			
CTCLBPR		X	X
CT_AMBG		X	X
CT_ACH			
CT_PCH			
CT_OMBG			

5.3 Experiment 3 - FLC for porosity metrics

In this third experiment we apply both the manual and the automated procedure only to metrics related to porosity.

5.3.1 Manual feature selection

The set of features provided by experts corresponds to column "Manual" in Table 5.4.

Figure 5.9 shows the obtained dendrogram. Two big clusters are easily identifiable and also in the WSS-BSS graph of Figure 5.10 we can see a knee (elbow) for number of clusters equal to 2. Another significant division is obtained cutting the dendrogram after obtaining five clusters. We choose `n_clusters` = 2 for the result to be comparable with the next case.

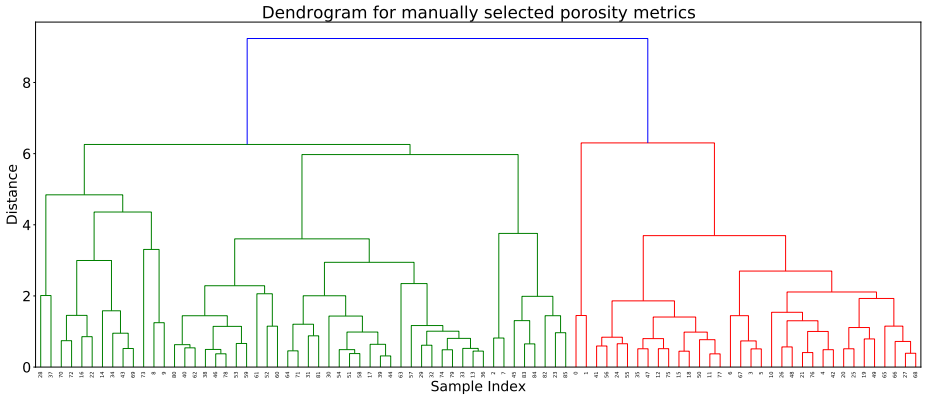


Figure 5.9: Dendrogram for clusters considering only manually selected metrics related to porosity

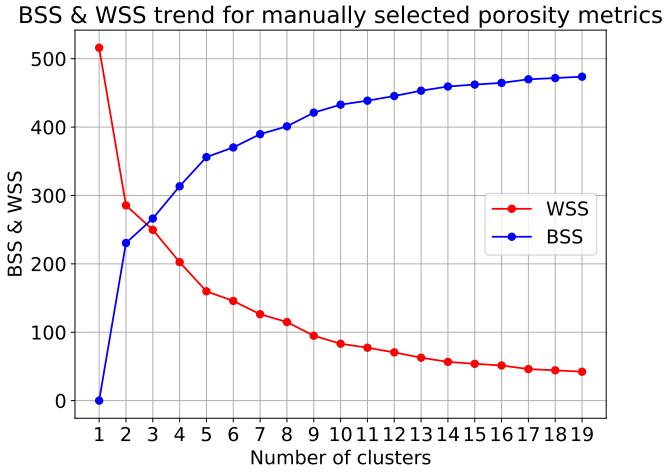


Figure 5.10: BSS and WSS trend for clusters considering only manually selected metrics related to porosity

5.3.2 Automated feature selection

The set of features provided by the entropy-based algorithm corresponds to column "Automated" of Table 5.4.

Both dendrogram (Figure 5.11) and the knee elbow graph (Figure 5.12) highlight different options. Since from the dendrogram we can see two distinct clusters as for the previous case, we prefer `n_clusters` = 2 in order to have two well distinct clusters instead of more small clusters identifying also not representative outliers.

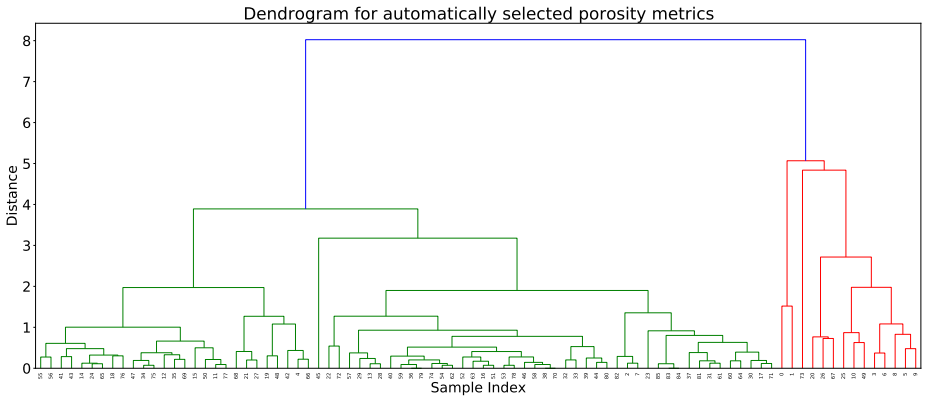


Figure 5.11: Dendrogram for clusters considering only automatically selected metrics related to porosity

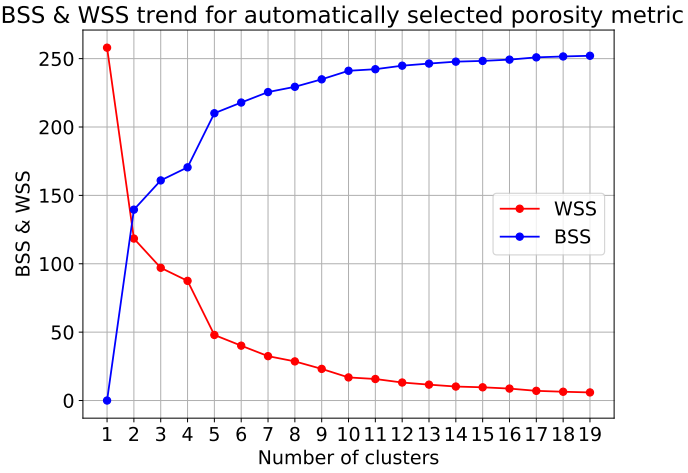


Figure 5.12: BSS and WSS trend for clusters considering only automatically selected metrics related to porosity

5.3.3 Summary

As for the other experiments, Table 5.4.shows the features used in the different cases. This time the sets are no longer disjoint, but this is because the entropy-based algorithm selects only three features. As explained, fewer features can often be an advantage in clustering. This time, in fact, the clusters produced in the case of the automated approach are well

spaced. Looking at the `score` of the experiment, however, this is, as in the case of the *Metrics* dataset, still low, i.e. equal to 46,7.

Table 5.4: Features porosity

Original	Manual	Less correlated	Automated
BCR_G	X	X	X
BCRG.1	X		
BBVR	X	X	
BSR	X		
B_AMBG		X	
B_CMBG		X	
B/UV		X	
B_DS/B		X	
BDF10	X	X	
Apass		X	
S/V		X	
Cdcm		X	
Concavity		X	
CTTB	X	X	X
CTAR_N		X	X
CTBLPR			
CT_AMBG		X	
BD_norm			
CTD_norm			

5.4 Experiment 4 - FLC for permeability metrics

For this fourth experiment we only have seven permeability related metrics. Since seven is already a manageable number of features, and the metrics were associated to the KC manually by experts, we will consider only the manual case.

5.4.1 Manual feature selection

Figure 5.13 shows the obtained dendrogram from where we can individuate four distinct clusters. To make results comparable with the other experiment, we choose `n_clusters` = 4 even though in the knee elbow graph of Figure 5.14 is clear the WSS (BSS) continues to decrease (increase) after number of clusters equal to 4.

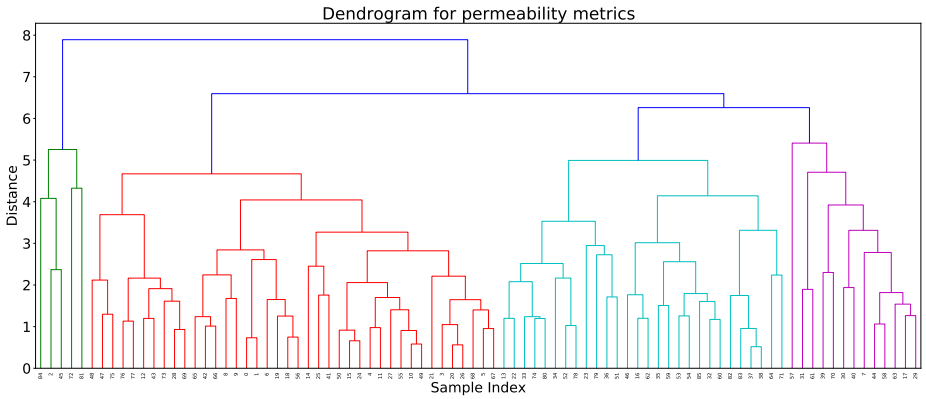


Figure 5.13: Dendrogram for clusters considering only metrics related to permeability

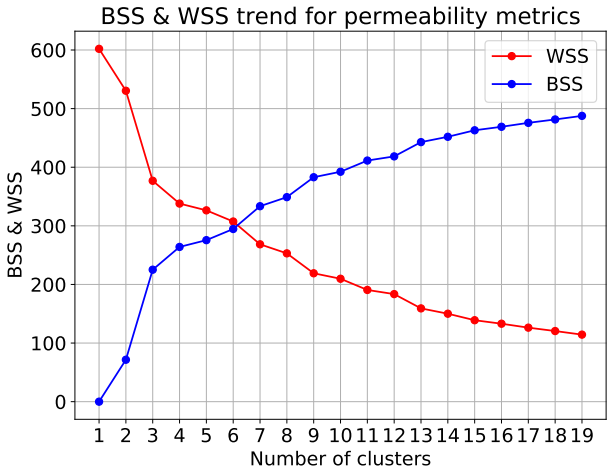


Figure 5.14: BSS and WSS trend for clusters considering only metrics related to permeability

5.5 Experiment 5 - FLC for *Attributes* dataset

In this first experiment we apply both the manual and the automated procedure to dataset *Attributes* .

5.5.1 Manual feature selection

The set of features provided by experts corresponds to column "Manual" of Table 5.5 and Table 5.6.

Figure 5.15 shows the obtained dendrogram. In this case the number of clusters is not so evident. To choose the `n_clusters` parameter we took into consideration the knee elbow graph of Figure 5.16 which suggests a good number of clusters to be between 4 and 7, again, to allow an easier comparability of the results with the precedent experimets we choose to run the algorithm with `n_clusters` = 4. By doing so, NIL 85 “PARCO DELLE ABBAZIE” resulted to be an outlier.

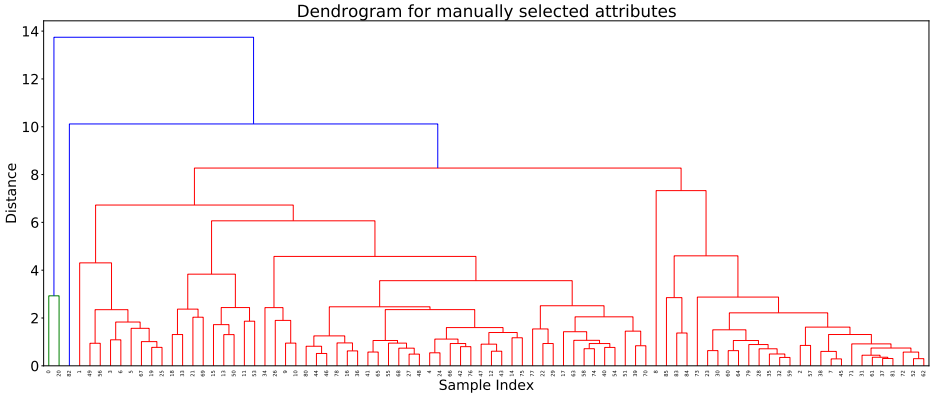


Figure 5.15: Dendrogram for clusters considering only manually selected attributes

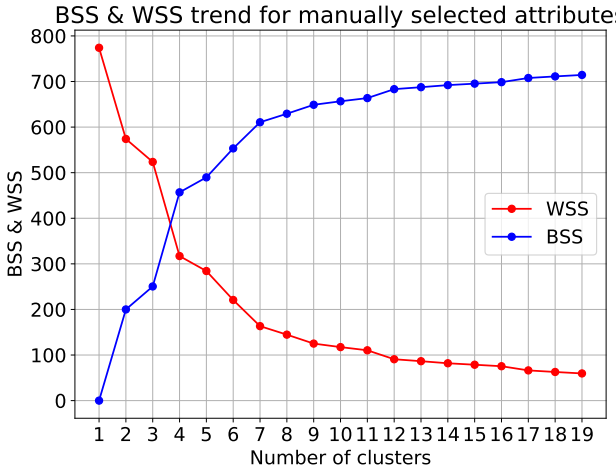


Figure 5.16: BSS and WSS trend for clusters considering only manually selected attributes

5.5.2 Automated feature selection

The set of features provided by the entropy-based algorithm corresponds to column "Automated" Table 5.5 and Table 5.6

Figure 5.17 shows the obtained dendrogram where different cluster divisions are present while the knee elbow analysis of Figure 5.18 suggests choosing a number of clusters between 4 and 7. According to these two and the manual case, we choose `n_clusters` = 4.

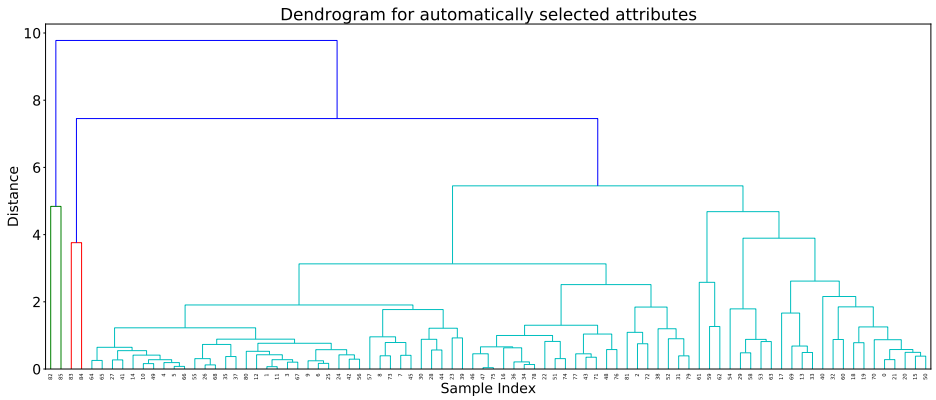


Figure 5.17: Dendrogram for clusters considering only automatically selected attributes

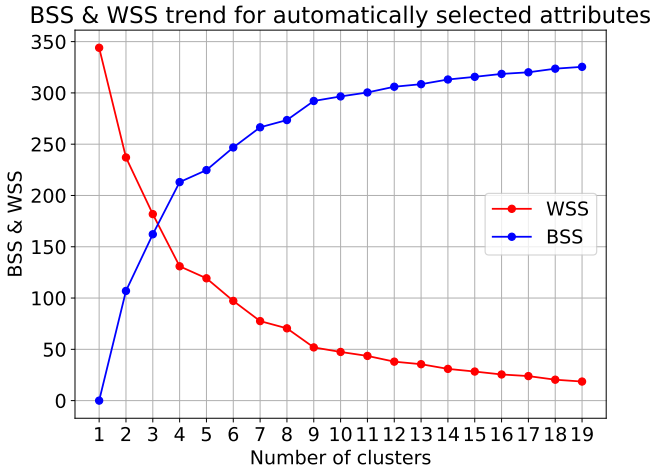


Figure 5.18: BSS and WSS trend for clusters considering only automatically selected attributes

5.5.3 Summary

We present in Table 5.5 and Table 5.6 the comparison of different set of features as we did for previous experiment. As it happens in Section 5.3 the entropy- based algorithm select a very small number of features. However, this time, neither the clustering algorithm provides significant results, nor the score is high. In fact, both in the manual and in the automated case, only a big cluster is created while the others are very small or represent just outliers, moreover, the `score` is 46,67 which is the lowest founded till now.

Table 5.5: Features *Attributes* , 1

Orginal	**Manual**	**Less correlated**	**Automated**
A	X	X	X
P			
Pop		X	
UV_N	X	X	
B_N			
B_U_N			
B_DS_N	X	X	
B_BF5_N			
B_BF10_N			
B_P			
B_DS_P			
B_A	X		
B_S	X		
B_V			
B_Vmax			
B_FA			
MBG_Dmin			
MBG_Dmax			
MBG_A			
MBG_P			
CH_A			
CH_P			
Circle_A			
A6		X	
B_Bpt_dmin			

Table 5.6: Features *Attributes* , 2

Orginal	**Manual**	**Less correlated**	**Automated**
MBG_O		X	
B_Hmax	X	X	
B_Hmean		X	
Street Area			
BL_N	X		
BL_A		X	X
BL_P			
BL_MBG_A			
BL_MBG_P			
BL_MBG_Dmi			
BL_MBG_Dma			
BL_MBG_O		X	
BL_BP_N			
BL_BP	X		
VBL_A			
VBL_P			
BL_Amean		X	X
VBL_Amean			
VBL_Pmean			
CT_N	X		
CT_A			
CT_P			
CT_CH_A			
CT_CH_P			
CT_MBG_A			
CT_MBG_P			
CT_MBG_O		X	X

5.6 Experiment 6 - FLC for *Milan* dataset

In this first experiment we apply both the manual and the automated procedure to dataset *Milan* .

5.6.1 Manual feature selection

The set of features provided by experts corresponds to column "Manual" of Table 5.7.

Below are reported the dendrogram (Figure 5.19) and the knee elbow graph (Figure 5.20) related to this first part of Section 5.6. From the dendrogram, we can see the creation of one big cluster, the green one. We decide to preserve this cluster and also the red and azure ones. Since samples 7 and 74 are merged at a great distance, we decide to split the cluster they form making them outliers. We remind that they correspond to NIL 8 "PARCO SEMPIONE" and 75 "STEPHENSON". This procedure results to make the `n_clusters` parameter equal to 5. Even though we are again in the case where we have one cluster way bigger than the others, we prefer to not split it to not create meaningless clusters. As we can see in the WSS-BSS trend, we should have a lot of clusters before to have a good division.

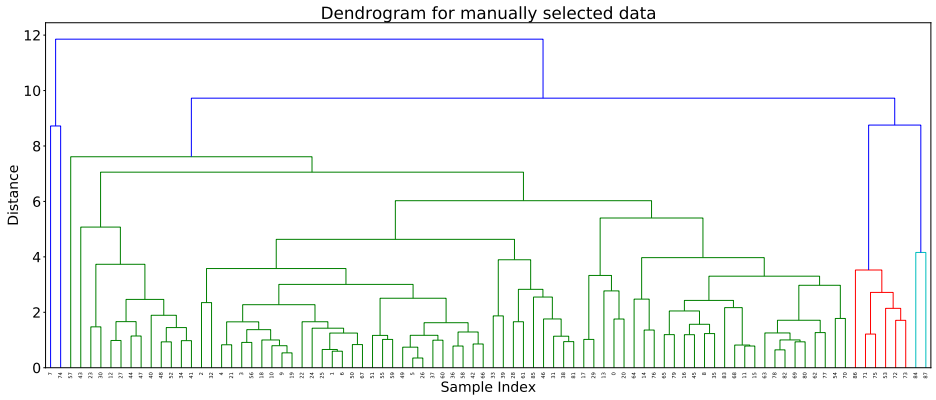


Figure 5.19: Dendrogram for clusters considering only manually selected data from Comune di Milano

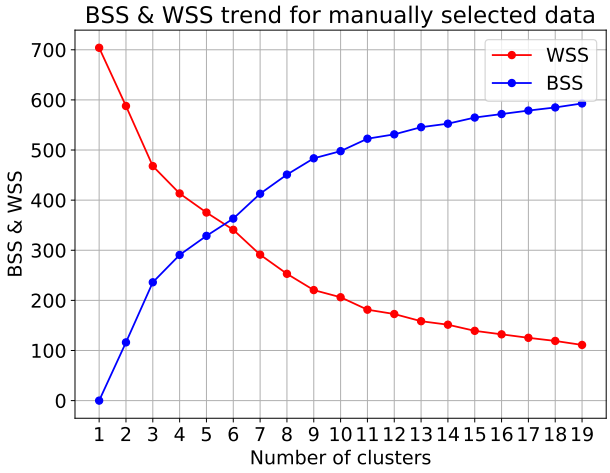


Figure 5.20: BSS and WSS trend for clusters considering only manually selected data from Comune di Milano

5.6.2 Automated feature selection

The set of features provided by the entropy-based algorithm corresponds to column "Automated" Table 5.7.

As we did for all the experiments before, we choose `n_clusters` according to the dendrogram and the BSS/WSS trend. The first one is shown in Figure 5.21 while the latter in Figure 5.22. Again, we can see that most of the NILs are grouping together. However, since this time the knee elbow graph suggests 5 to be a good value for `n_clusters`, and this is also coherent with the previous case, we choose to set `n_clusters` = 5. By doing so, NIL 8 “PARCO SEMPIONE” and NIL 75 “STEPHENSON”, resulted to be and outliers.

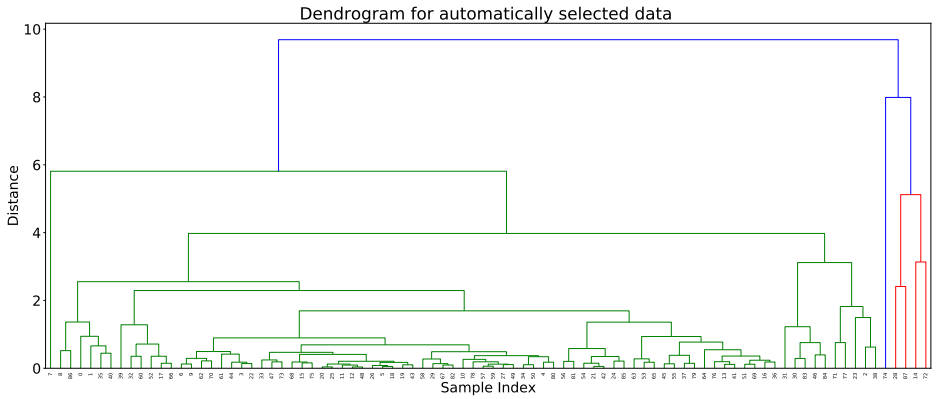


Figure 5.21: Dendrogram for clusters considering only automatically selected data from Comune di Milano

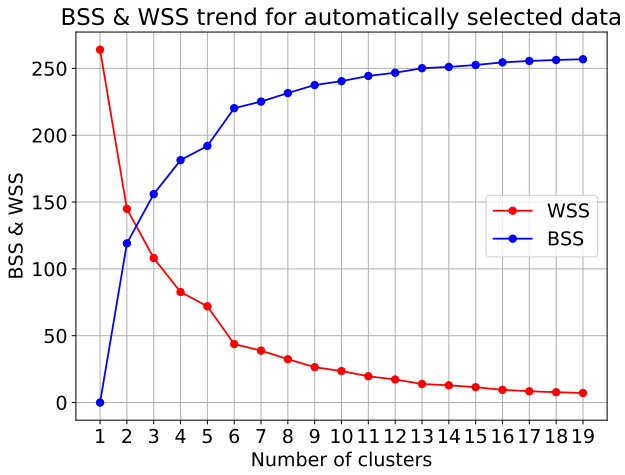


Figure 5.22: BSS and WSS trend for clusters considering automatically selected data from Comune di Milano

5.6.3 Summary

In Table 5.7 we report the comparison between set of different features used. Once again, the sets are disjoint. Moreover, in this experiment, even though the score is quite high (i.e. 73,86) this is probably due to the fact that in both cases one big clusters is created. However, even if the set of manually selected metrics presents a mix of features regarding

quality of the aria, buildings energy consumption and structural characteristic, while the one produced by the entropy-based algorithm contains only information regarding buildings, the same outliers are individuated in both cases.

Table 5.7: Features *Milan*

Original	**Manual**	**Less correlated**	**Automated**
avg(pm10_01)	X	X	
avg(‘pm2.5_01‘)	X		
avg(no2_01)	X	X	
sum(superf_utile_riscaldata)		X	X
sum(superf_utile_raffrescata)			
sum(volume_lordo_riscaldato)			
sum(volume_lordo_raffrescato)			
sum(ep_gl_nren)		X	X
sum(ep_gl_ren)	X		
sum(emissioni_co2)	X		
sum(consumi_energia_elettrica)			
sum(consumi_gas_naturale)		X	X
sum(superficie_disperdente)			
avg(uomini)		X	
avg(donne)			
avg(minori)			
avg(famiglie)			
avg(famiglie_unipersonali)			
avg(stranieri)			
avg(80_e_piu)			
avg(80_e_piu_soli)			
avg(65_e_piu)			
avg(residenti_prima_cittadinanza)			
avg(residenti_terza_cittadinanza)			
avg(residenti_seconda_cittadinanza)			
avg(scuola_primaria_numero)		X	
avg(scuola_infanzia_numero)			
avg(scuola_secondaria_di_primo_grado)			
avg(fermate_metro + fermate_linee)	X	X	
avg(area_metri2)	X	X	
avg(piste_ciclabili_m)	X	X	

5.7 Experiment 7 - FLC for indicators and metrics

In this first experiment we apply both the manual and the automated procedure to dataset Indicators.

5.7.1 Manual feature selection

According to Section 5.1.1 and Section 5.2.1, the manually selected set of features provide by experts includes:

- 8 indicators
- 6 metrics

From Figure 5.23 and Figure 5.24 is clear that a reasonable number of clusters is three. We then set `n_clusters` equal to 3 and NIL 8 "PARCO SEMPIONE" reluts again to be an outlier. Moreover, we notice the dendrogram to be very similar to the one obtained in Section 5.1.2.

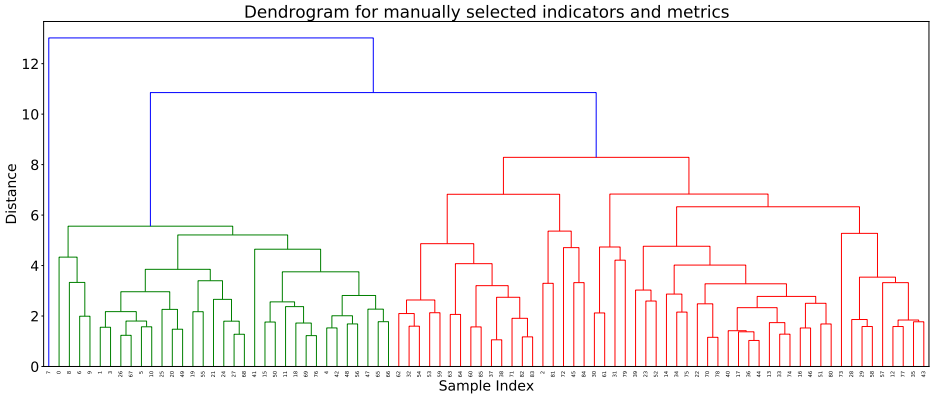


Figure 5.23: Dendrogram for clusters considering only manually selected indicators and metrics

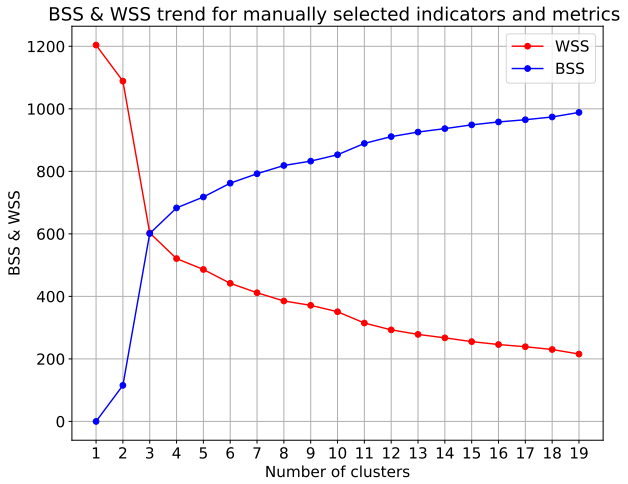


Figure 5.24: BSS and WSS trend for clusters considering only manually selected indicators and metrics

5.7.2 Automated feature selection

According to Section 5.1.2 and Section 5.2.2, the automatically selected set of features includes:

- 5 indicators
- 12 metrics

This time, both the dendrogram in Figure 5.25 and the BSS and WSS trend in Figure 5.26 suggest that there are more clusters in the dataset. We decide to set `n_clusters` = 9 since this number corresponds both to a knee(elbow) in the wss(BSS) trend, and to a cut in the dendrogram able to identify multiple clusters without create meaningless outliers.

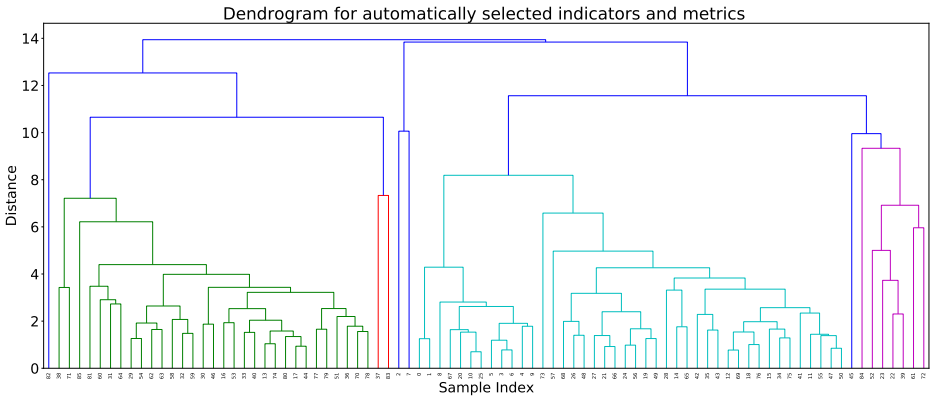


Figure 5.25: Dendrogram for clusters considering only automatically selected indicators and metrics

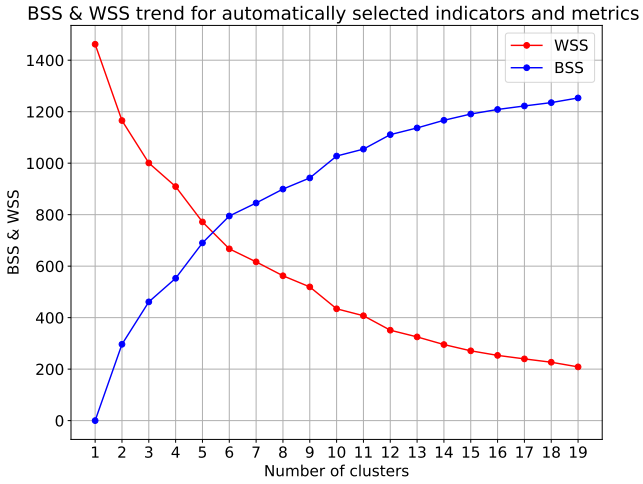


Figure 5.26: BSS and WSS trend for clusters considering only automatically selected indicators and metrics

Chapter 6

Experiment comparison and evaluation

We dedicate this chapter to compare the results we obtained to better understand the conclusions we will draw from them. First of all, we will use different graphical method to compare the results, focusing on how the clustering algorithm classify the same NILs in the FLC phase. Lately, we will evalute all the results using the `score` variable and we will comment also the results obtained in the SLC phase.

6.1 FLC results comparison

Having chosen a different number of clusters in each experiment makes hard to directly compare the result. What we need to show is:

- Which are the NILs grouped together in each experiment;
- Which are the NILs grouped together in all the experiments;
- Evaluate result for each experiment.

For the firs two, we will use Table 6.1 and Table 6.2 to show the obtained results. The columns of the table correspond to:

- A : Experiment 1 - manual feature selection results, `n_clusters` = 3;
- B : Experiment 1 - automated feature selection results, `n_clusters` = 4;
- C : Experiment2 - manual feature selection results, `n_clusters` = 4;
- D : Experiment 2 - automated feature selection results, `n_clusters` = 5;

- E : Experiment 3 - manual feature selection results, `n_clusters` = 2;
- F : Experiment 3 - automated feature selection results, `n_clusters` = 2;
- G : Experiment 4 - manual feature selection results, `n_clusters` = 4;
- H : Experiment 5 - manual feature selection results, `n_clusters` = 4;
- I : Experiment 5 - automated features selection results, `n_clusters` = 4;
- L : Experiment 6 - manual feature selection results, `n_clusters` = 5;
- M : Experiment 6 - automated feature selection results, `n_clusters` = 4.

For what concerns the colours:

- all the outliers are coloured in red;
- different shadow/colour are used to highlight different clusters trying to copy also closeness between them;
- black is used to indicate we do not have data for that NIL.

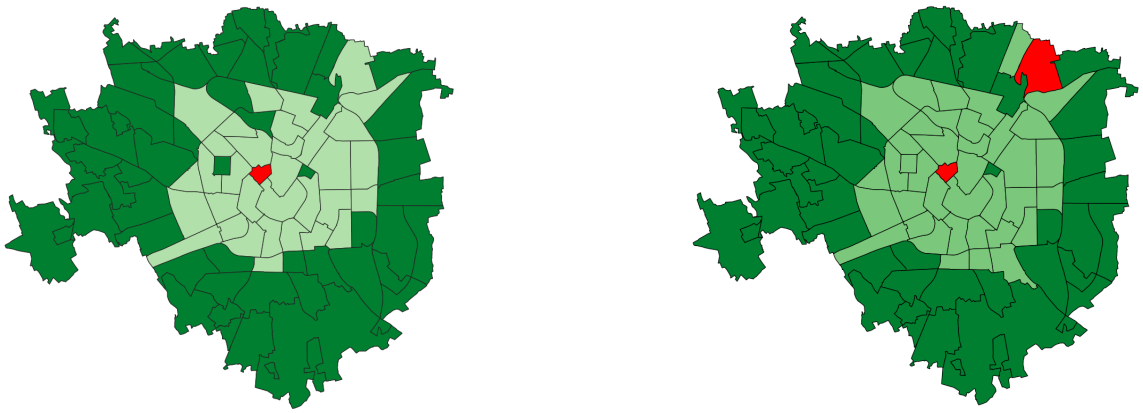
Table 6.1: Compared results FLC, 1

ID_NIL	A	B	C	D	E	F	G	H	I	L	M
1	2	2	0	0	0	0	2	3	0	0	1
2	2	2	0	0	0	0	2	0	1	0	1
3	0	0	1	3	1	1	1	1	1	0	1
4	2	2	0	0	0	0	2	0	1	0	1
5	2	2	3	0	0	1	2	0	1	0	1
6	2	2	0	0	0	0	2	0	1	0	1
7	2	2	0	0	0	0	2	0	1	0	1
8	1	3	1	3	1	1	0	1	1	4	3
9	2	2	0	0	1	0	2	1	1	0	1
10	2	2	0	0	1	0	2	0	1	0	1
11	2	2	0	0	0	0	2	0	1	0	1
12	2	2	3	4	0	1	2	0	1	0	1
13	0	0	3	4	0	1	2	0	1	0	1
14	0	0	3	4	1	1	3	0	0	0	1
15	0	2	0	0	1	1	2	0	1	0	4
16	2	1	0	4	0	1	2	0	0	0	1
17	0	0	1	4	1	1	3	0	1	0	1
18	0	0	3	4	1	1	0	0	0	0	1
19	2	2	3	0	0	1	2	0	0	0	1
20	2	2	3	0	0	1	2	0	0	0	1
21	2	2	0	0	0	0	2	3	0	0	1
22	2	2	0	0	0	1	2	0	0	0	1
23	0	0	0	4	1	1	3	0	1	0	1
24	0	0	0	4	1	1	3	1	1	0	1
25	2	2	0	0	0	1	2	0	1	0	1
26	2	2	0	0	0	0	2	0	1	0	1
27	2	2	0	0	0	0	2	0	1	0	1
28	2	2	0	0	0	1	2	0	1	0	1
29	0	0	3	4	1	1	2	1	1	0	0
30	0	0	1	4	1	1	0	0	0	0	1
31	0	0	2	4	1	1	0	1	1	0	1
32	0	0	2	4	1	1	0	1	1	0	1
33	0	0	1	4	1	1	3	1	0	0	1
34	0	0								0	1
35	0	0	3	4	1	1	3	0	0	0	1
36	0	2	0	4	1	1	3	0	1	0	1
37	0	2	1	4	0	1	3	1	1	0	1
38	0	0	3	4	1	1	3	0	1	0	1
39	0	0	1	1	1	1	3	1	1	0	1
40	0	0	1	1	1	1	3	1	1	0	1
41	0	0	3	4	1	1	0	0	1	0	1
42	0	0	3	4	1	1	0	0	0	0	1
43	2	2	0	4	0	1	2	0	1	0	1
44	2	2	3	0	0	1	2	0	1	0	1
45	0	0	3	4	1	1	2	0	1	0	1

Table 6.2: Compared results FLC, 2

ID_NIL	A	B	C	D	E	F	G	H	I	L	M
46	0	0	3	4	1	1	0	0	1	0	1
47	0	0	1	0	1	1	1	1	1	0	1
48	0	0	0	4	1	1	3	0	1	0	1
49	2	2	3	4	0	1	2	0	1	0	1
50	2	2	3	0	0	1	2	0	1	0	1
51	2	2	0	0	0	0	2	0	1	0	1
52	2	2	0	0	0	1	2	0	0	0	1
53	0	0	0	4	1	1	3	0	1	0	1
54	0	0	0	4	1	1	3	1	1	3	1
55	0	0	1	4	1	1	3	0	0	0	1
56	0	0	1	4	1	1	3	0	0	0	1
57	2	2	0	0	0	1	2	0	1	0	1
58	2	2	3	0	0	1	2	0	1	0	1
59	0	2	3	4	1	1	0	1	1	0	1
60	0	0	3	4	1	1	0	0	0	0	1
61	0	0	1	4	1	1	3	1	0	0	1
62	0	0	1	4	1	1	3	1	0	0	1
63	0	0	2	4	1	1	0	1	0	0	1
64	0	0	1	4	1	1	3	1	0	0	1
65	0	0	1	4	1	1	0	0	0	0	1
66	0	0	1	4	1	1	3	1	1	0	1
67	2	2	3	0	0	1	2	0	1	0	1
68	2	2	3	0	0	1	2	0	1	0	1
69	2	2	0	0	0	0	2	0	1	0	1
70	2	2	0	0	0	1	2	0	1	0	1
71	2	2	3	4	1	1	2	0	0	0	1
72	0	0	0	4	1	1	0	0	0	3	1
73	0	0	1	1	1	1	3	1	1	3	4
74	0	0	0	0	1	1	1	1	1	3	1
75	0	0	3	4	1	0	2	1	1	2	2
76	0	0	3	4	1	1	3	0	1	3	1
77	0	2	0	4	0	1	2	0	1	0	1
78	0	2								0	1
79	2	2	3	0	0	1	2	0	1	0	1
80	0	0	3	4	0	1	2	0	1	0	1
81	0	0	0	4	1	1	3	0	1	0	1
82	0	0	3	4	1	1	3	1	1	0	1
83	0	0	3	4	1	1	3	0	1	0	1
84	0	0	1	4	1	1	1	1	1	0	1
85	0	0	1	2	1	1	3	2	3	1	1
86	0	0	1	1	1	1	3	1	2	0	1
87	0	0	1	4	1	1	1	1	2	3	1
88	0	0	1	4	1	1	3	1	3	1	0

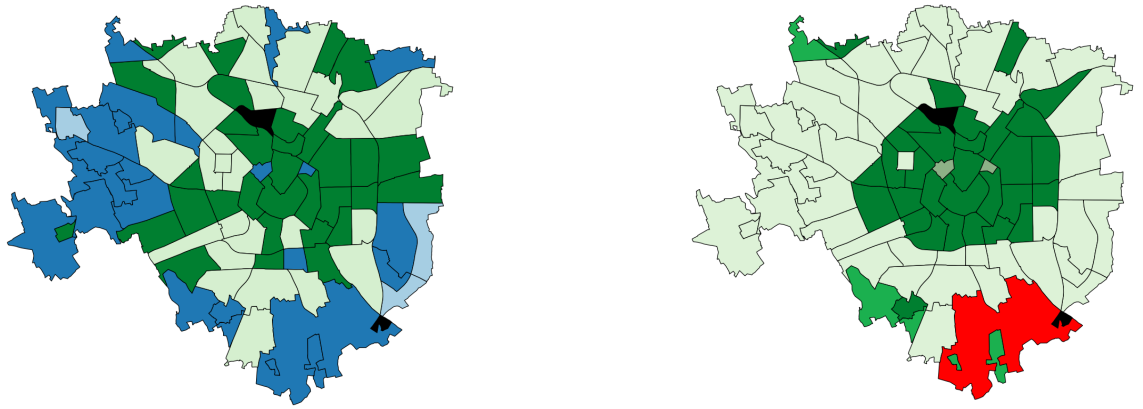
Another way to compare results is to show the clusters directly on the map of Milan. The following images represents all the obtained results. Colours are chosen with the same criterion as for Table 6.1 and Table 6.2



(a) Manual (b) Automated

Figure 6.1: Results for *Indicators* dataset.

What emerges from this maps is that the two algorithms provide almost the same results. Even though it seems indicators only highlight macro differences among NILs, the results has been judged by the experts to be perfectly coherent with the way DOP families reflect on NILs.

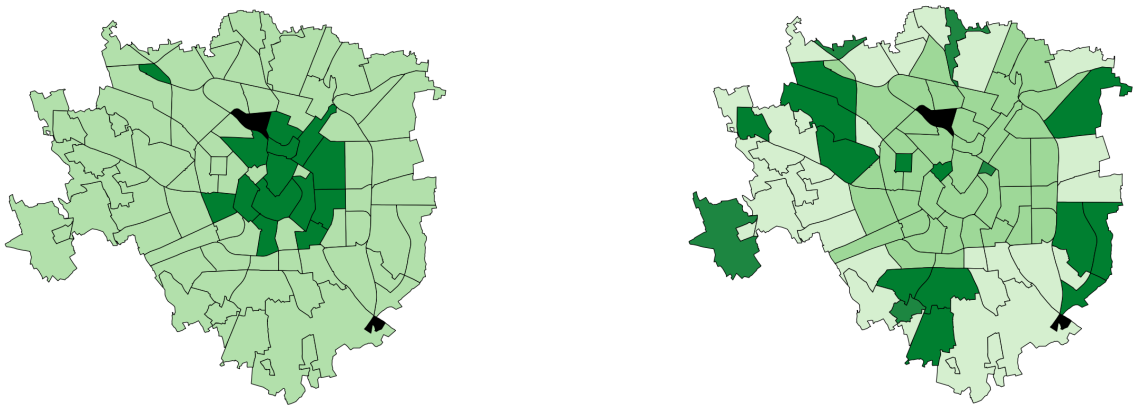


(a) Manual (b) Automated

Figure 6.2: Results for *Metrics* dataset.

In this case the two algorithms perform really differently. What emerges is that the

features selected manually are able to express finer differences, while the automated algorithm only separates NILs from the centre from the more peripheral ones, and individuates small groups very different from the others. This is probably due to the fact that in the manual case we also select permeability metrics while in the automated one we mostly have porosity related ones.



(a) Manual (b) Automated

Figure 6.3: Results for metrics related to porosity.

Probably because of the low number of clusters, the maps show a simple division between NILs in the centre and the more peripheral ones. However, this a consistent result with natures of porosity. Moreover, this result confirms the one showed in Figure 6.2(b) where the metrics selected by the entropy-based algorithm are mostly related to porosity.

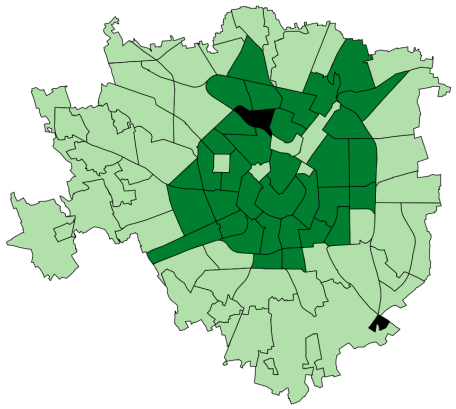
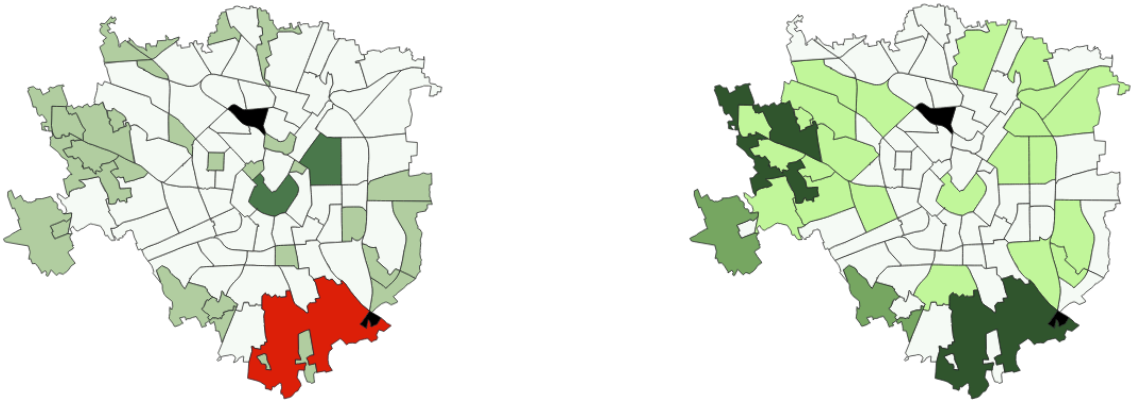


Figure 6.4: Results for metrics related to permeability.

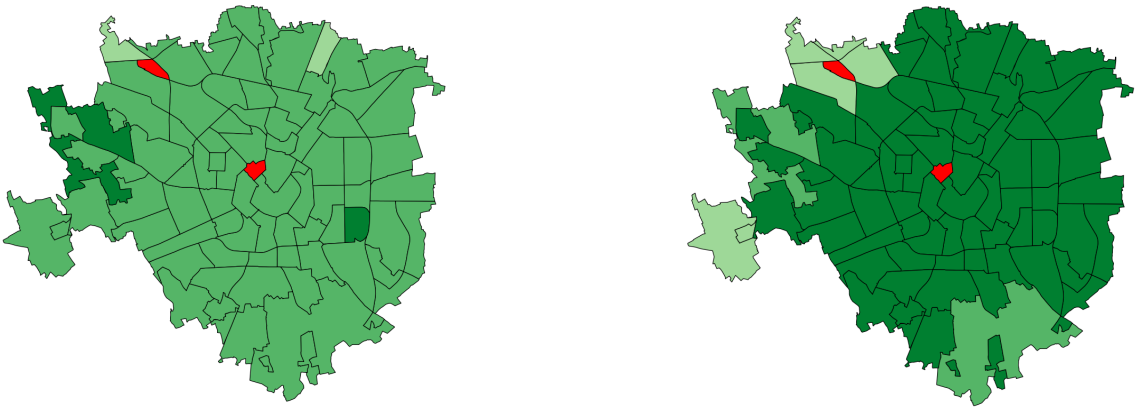
Once again, NILs of the centre are grouped together, and this perfectly reasonable. Moreover, also this result confirms that adding permeability metrics as variables contributes to create more clusters in the peripheral areas.



(a) Manual (b) Automated

Figure 6.5: Results for *Attributes* dataset.

From the images we have the prove that the two experiments preform really differently. This would probably mean that the results are strictly related to the selected features which do not correspond in the two cases.



(a) Manual (b) Automated

Figure 6.6: Results for *Milan* dataset.

Results for this experiment, as we already noticed, are not so relevant.

6.2 FLC evaluation and output

After comparing clusters, in order to provide the needed outputes for the *Second Level Clustering* , we now need to evaluate our results. In Chapter 4 we defined the `comparison_matrix` and the `score` variable as the ways to compare results. In Table 6.3 we report scores for all the datasets except the one related to permeability metrics, as we did not perform any feature selection for them. In Section 6.2 instead, the values are compared using a bar plot.

Table 6.3: Comparison score

Dataset	Score
Indicators	76.34091
Metrics	49.90698
Porosity	49.72093
Attributes	46.65116
Milan	73.86364

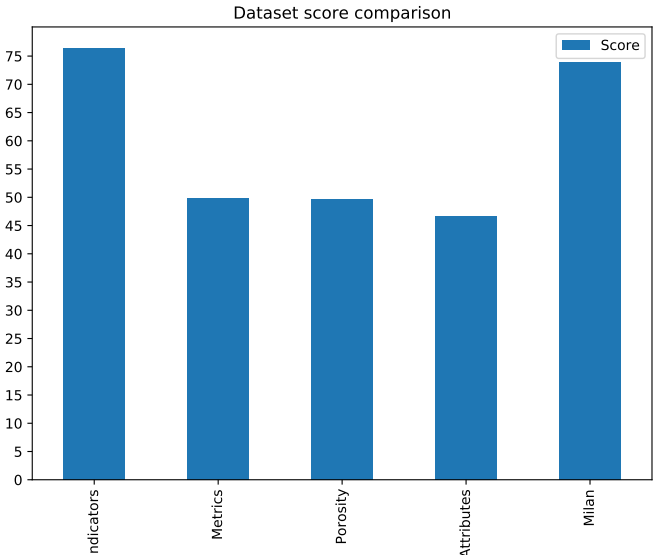


Figure 6.7: `score` values comparison

The highest `score` is obtained for dataset *Indicators* . The one for *Milan* is also pretty high but this is probably due to the fact that both in the Manual and, in the Automated case, the algorithm simply individuates outliers. Repeating the experiments with other numbers of clusters will of course provide other interesting results. Anyway, what is evident,

and is important to us, is that we obtain good performances when we cluster using indicators as features. This seems to be due to two main facts:

- Indicators are theoretically divided in DOP families and this division is preserved in our original dataset. This means that we take into consideration different performance aspects while clustering;
- Ability of the feature selection algorithm to pick indicators from different DOP families.

This makes indicators eligible as important dimension for the Second Level Clustering. For what concerns distances along each dimension, we can extrapolate it looking at the clusters results. Clusters are based on euclidean distances, so it more than reasonable to approximate distances between elements with the distances between clusters. As we said, colours in Table 6.1 and Table 6.2, and in the maps, are chosen to take into account the distance between clusters. To represent distances between each NIL we used the euclideian distance using as variables only the features selected.

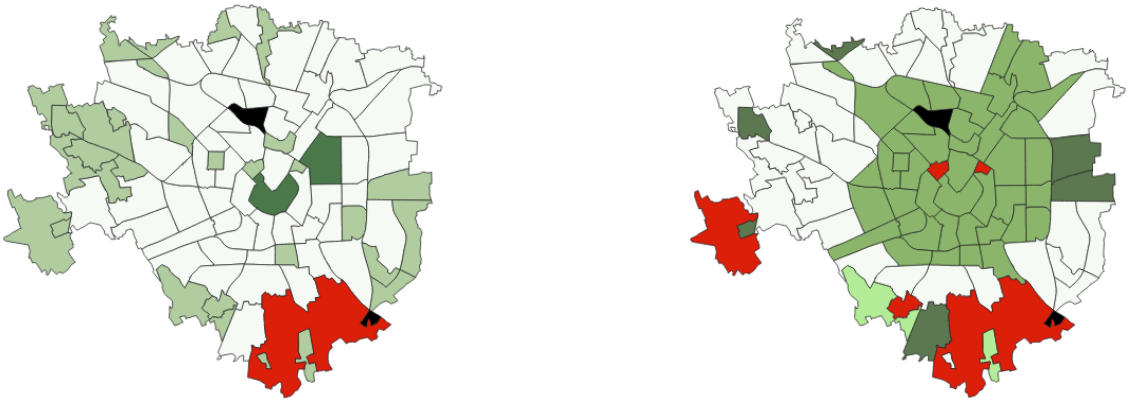
6.3 SLC evaluation and output

In the previous section we justified the choice of indicators as important dimension. Even though *First Level Clustering* performed badly on *Metrics* dataset, the second chosen dimension is "metrics" itself. This is done according to the IMM procedure, since metrics and indicators are at the same level of aggregation in the data flow Figure 2.5. However, we can see in Table 6.4, that when we compute `score` to compare the two experiments, both entirely and considering only the Manual or the Automated case, we have surprisingly good values.

Table 6.4: Comparison score indicators and metrics

Compared	Score
Exp_1 - Exp_2	35.59091
Exp_1.1 - Exp_2.1	47
Exp_1.2 - Exp_2.2	47

From Table 6.4, we can see that almost the 50% of NILs are always clustered in the same way. This is a high percentage if we think the number of clusters (without considering the outliers) is different in the two experiments. Coming to the results, even if we did not highlight an evaluation subphase in SLC, we have performed the algorithm both with the manually feature selected and the automated ones. In Figure 6.8 we report the obtained results directly on the map.



(a) Manual (b) Automated

Figure 6.8: SLC results for indicators and metrics.

From the image above and the previous ones, two important things emerge:

1. most of the NILs of the centre are always in the same cluster;
2. comparing the automated cases for Indicators, Metrics the one for indicators and metrics together, it seems that the latter is a sort of sum the previous two.

Again, we can extract the distance between NILs from the distance between clusters they are grouped in, but we we also show the distances heatmap both for the automated and the manual case.

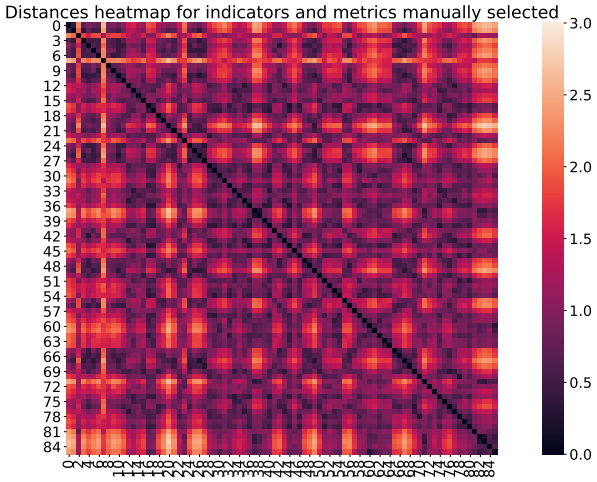


Figure 6.9: Distances heatmap for indicators and metrics manually selected

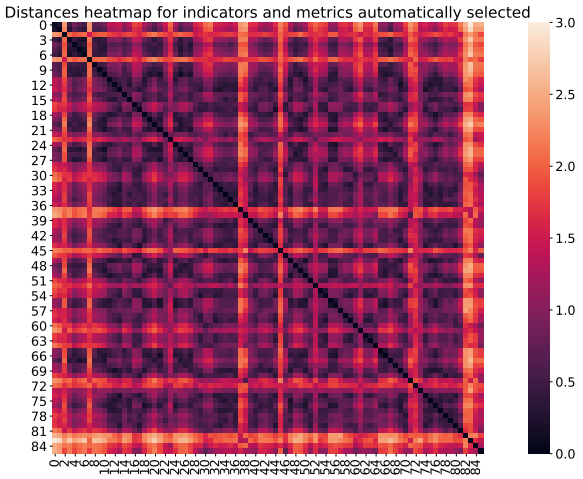


Figure 6.10: Distances heatmap for indicators and metrics automatically selected

As shown in the legends, the darker the colour, the closer the two samples are to each other. What emerges from these two heatmaps is that probably, NILs with similar ID_s will be similar to each other, and therefore close to each other. In both figures, in fact, you can see that the colour gradations follow a precise pattern, in particular they form squares. What is even more important, though, is that, broadly speaking, the two maps are similar. This means that even with differing features, the differences/similarity between elements has been preserved.

Chapter 7

Limitations and future works

The main limitation of this work is related to the lack of data. Having only 88 samples makes the results sensible to outliers. Both when looking for important features and while clustering, the presence of NILs showing unusual behaviours deeply influences the results, making it difficult to discover patters at a finer level. By considering a higher number of samples, future works could limit this effect; alternatively, if not bound by the need not to lose any sample, they could try to eliminate outliers. To increase the number of samples, futures works can:

- define a finer granularity for Milan or another city;
- compare more cities at the same time, defining the same granularity for all of them.

On the other hand, even if we have lots of features, another limitation is related to the poor variety of them. Both metrics and attributes are mostly related to porosity, and this makes it impossible to analyse different aspects of the built environment. In addition, the quality of the data retrieved from Comune di Milano is really poor. The *Aria* dataset, for example, was generated interpolating data from only seven ARPA (Azienda Regionale per la Protezione Ambientale) stations, spreading the results on different NILs. This makes the reliability of the data really low, also considering that air quality is not only due to agents strictly related to the city, and can be influenced by external factors. To better account for variety, future works can:

- add more features to try to capture more *Dimensions* ;
- looking for features related to the same *Dimension* that can be categorized, and choose a representative number for each category.

This last observation finds its reason in the results obtained for the *Indicators* dataset. This is indeed the only dataset in which we have found a categorization of the features

through the DOP families, and it was also the dataset where we obtained the best results. Moreover, regardless the results, another proof of the need for variety in features is that the entropy-based algorithm selected indicators from different families. Concluding, a last limitation is probably related to the nature of NILs themselves. The division does not always reflect the actual territorial characteristics. A proof of this is the new division carried out in March 2020 by Comune di Milano. The comparison with this new structure could be an interesting starting point for future analyses.

Chapter 8

Conclusions

In this work we presented SIMBA, a systematic clustering – based methodology to support built environment analysis. SIMBA is a support tool for architects and urban planners in understanding and analysing urban environments and the relationships between their parts. In particular, it is designed to improve the Investigation phase of IMM, a methodology developed at the Department of Architecture, Built Environment and Construction Engineer of Politecnico di Milano. By looking at the different clusters obtained and the selected features, it is possible to identify both how the different *Dimensions* , which can be conceptually associated to the *Key Categories* of IMM, interact with each other, and also to compare different units. In order to prove the effectiveness of SIMBA we have reported the results obtained by applying the methodology to the city of Milan, looking for clustering it in NILs, neighborhoods in which the city has been divided according to mainly socio-demographic and cultural criteria. The results confirmed the capability of clustering algorithms to represent conceptual distances between elements. This can be used by the scholars of the subject to calculate these distances and to avoid basing their analyses on their personal observations only. Moreover, SIMBA allows to select a manageable number of features that have been proven to be also descriptive for the built environment. The six metrics selected for each KC in IMM, nonetheless, were chosen without a systematic and objective procedure. By looking at the clusters results it is also possible to observe how the different dimensions interact with each other, to analyse the effects of their different combinations on the results and to idetify which ones influence them most. Together with these two big improvements for IMM, another interesting result provided by SIMBA is the proof, even if with some limits, of a correspondence between performances and structural patterns, emerged while we comparing our experiments. Summarizing, the results obtained by using SIMBA are:

Summarizing, the results obtained by using SIMBA are:

- a selection of reasonable but also representative number of features when investigating the built environment;

- experimental evidence of correspondence between the structural shape of the city and performance patterns;
- a systematic methodology to measure distance between elements, needed when comparing different built environment;and
- explicable results of the interaction between the different components of the city.

Bibliography

[1] D. L. Meadows, D. H. Meadows, J. Randers, and W. W. B. III, *The limits to growth.* Donella H. Meadows, 1972.

[2] Allan, James, O. Venter, S. Maxwell, B. Bertzky, K. Jones, Y. Shi, and J. E. Watson, "Recent increases in human pressure and forest loss threaten many natural world heritage sites.," *Biological Conservation*, 2017.

[3] "Living planet report - 2018: Aiming higher." `https://doi.org/10.1080/09528820802312343.`, 2018.

[4] O'Neill, B. C., B. Liddle, L. Jiang, K. R. Smith, S. Pachauri, M. Dalton, and R. Fuchs., "Demographic change and carbon dioxide emissions," *The Lancet*, 2012.

[5] Geck and Caroline, "The world factbook.," *The Charleston Advisor*, 2017.

[6] P. di Milano, "Dipartimento di elettronica informazione e bioingegneria." `https://www.deib.polimi.it/eng/news/details/850`.

[7] P. di Milano, "Dipartimento di architettura, ingegneria delle costruzioni e ambiente costruito." `https://www.dabc.polimi.it/`.

[8] P. di Milano, "Immdesignlab." `http://www.immdesignlab.com`.

[9] T. Häkkinen, "Assessment of indicators for sustainable urban construction," *Civil Engineering and Environmental Systems*, vol. 24, no. 4, pp. 247–259, 2007.

[10] R. J. Cole and M. J. Valdebenito, "The importation of building environmental certification systems: international usages of breeam and leed," *Building Research & Information*, vol. 41, no. 6, pp. 662–676, 2013.

[11] W. Rees and M. Wackernagel, "Urban ecological footprints: Why cities cannot be sustainable—and why they are a key to sustainability," *Environmental Impact Assessment Review*, vol. 16, no. 4, pp. 223 – 248, 1996. Managing Urban Sustainability.

[12] G. Masera and M. Tadi, "Environmental performance and social inclusion in informal settlements," in *Environmental Performance and Social Inclusion in Informal Settlements*, ch. 2, Springer International Publishing, 2020.

[13] A. Haapio, "Towards sustainable urban communities," *Environmental Impact Assessment Review*, vol. 32, no. 1, pp. 165 – 169, 2012.

[14] K. Mori and A. Christodoulou, "Review of sustainability indices and indicators: Towards a new city sustainability index (csi)," *Environmental Impact Assessment Review*, vol. 32, no. 1, pp. 94 – 106, 2012.

[15] M. Batty, *Cities as Complex Systems: Scaling, Interaction, Networks, Dynamics and Urban Morphologies*. New York, NY: Springer New York, 2009.

[16] J. Klemeš, "Assessing and measuring environmental impact and sustainability," *Clean Technologies and Environmental Policy*, vol. 17, pp. 577–578, 03 2015.

[17] V. Bentivegna, S. Curwell, M. Deakin, P. Lombardi, G. Mitchell, and P. Nijkamp, "A vision and methodology for integrated sustainable urban development: Bequest," *Building Research & Information*, vol. 30, no. 2, pp. 83–94, 2002.

[18] I. Lowe, "Shaping a sustainable future – an outline of the transition," *Civil Engineering and Environmental Systems*, vol. 25, no. 4, pp. 247–254, 2008.

[19] A. Churchman, "Disentangling the concept of density," *Journal of Planning Literature*, vol. 13, no. 4, pp. 389–411, 1999.

[20] A. Forsyth, "Measuring density: Working definitions for residential density and building intensity." `http://annforsyth.net/wp-content/uploads/2018/05/db9.pdf`, 2003.

[21] W. Bank, "Cities and climate change: an urgent agenda," 2010.

[22] C. A. Biraghi, *Multi-Scale Modelling Approach for Urban Optimization: Urban Compactness Environmental Implications.* Phd thesis, Politecnico di Milano, 2018,2019.

[23] L. Setti, F. Passarini, G. de Gennaro, A. D. Gilio, J. Palmisani, G. F. P. Buono, M. Perrone, A. Piazzalunga, P. Barbieri, E. Rizzo, and A. Miani, "Evaluation of the potential relationship between particulate matter (pm) pollution and covid-19 infection spread in italy," 2020.

[24] J. Wu, S. Guo, H. Huang, W. Liu, and Y. Xiang, "Information and communications technologies for sustainable development goals: State-of- the-art, needs and perspectives," *IEEE Communications Surveys & Tutorials*, 2018.

[25] M. H. M. Zadeh, *A Systemic Modeling Methodology for Evaluating Built Environment Performance: Measuring Urban Proxinity.* Phd thesis, Politecnico di Milano, 2020 – Cycle 32.

[26] M. Tadi and S. V. Manesh, "Integrated modification methodology (imm): A phasing process for sustainable urban design," 2013.

[27] M. Carmona, R. CARMONA, T. Heath, T. Oc, and S. Tiesdell, *Public Places, Urban Spaces: The Dimensions of Urban Design.* Architectural Press, 2003.

[28] C. M. Bishop, *Pattern Recognition and Machine Learning (Information Science and Statistics).* Berlin, Heidelberg: Springer-Verlag, 2006.

[29] `https://expertsystem.com/machine-learning-definition/`.

[30] T. Mitchell, *Machine Learning.* McGraw Hill, 1997.

[31] S. Houston, T. Dieckhaus, B. Kircher, and M. Lardner, *An Introduction to Nursing Informatics, Evolution, and Innovation, 2nd Edition: Evolution and Innovation.* HIMSS Book Series, Taylor & Francis, 2018.

[32] F. Gullo, "From patterns in data to knowledge discovery: What data mining can do," *Physics Procedia*, vol. 62, p. 18–22, 12 2015.

[33] J. Leskovec, A. Rajaraman, and J. Ullman, *Mining of Massive Data Sets.* Cambridge University Press, 2020.

[34] `https://bernardmarr.com/default.asp?contentID=1741`.

[35] P. Ganapathi and D. Shanmugapriya, *Handbook of Research on Machine and Deep Learning Applications for Cyber Security.* Advances in Information Security, Privacy, and Ethics, IGI Global, 2019.

[36] D. Gunopulos, *Clustering Overview and Applications*, pp. 383–387. Boston, MA: Springer US, 2009.

[37] S. Paul, P. Bhattacharya, and A. Bit, *Early Detection of Neurological Disorders Using Machine Learning Systems.* Advances in Medical Technologies and Clinical Practice, IGI Global, 2019.

[38] `https://www.analyticsvidhya.com/blog/2016/11/an-introduction-to-clustering-and-different-me`
`).`

[39] M. J. Zaki and J. Wagner Meira, *Data Mining and Machine Learning: Fundamental Concepts and Algorithms).* Cambridge University Press, 2020.

[40] QGis, "Qgis geographic information system.," 2020.

[41] M. Dash and H. Liu, "Feature selection for clustering," *ACM digital library*, 2000.

[42] L. Buitinck, G. Louppe, M. Blondel, F. Pedregosa, A. Mueller, O. Grisel, V. Niculae, P. Prettenhofer, A. Gramfort, J. Grobler, R. Layton, J. VanderPlas, A. Joly, B. Holt, and G. Varoquaux, "API design for machine learning software: experiences from the scikit-learn project," in *ECML PKDD Workshop: Languages for Data Mining and Machine Learning*, pp. 108–122, 2013.

[43] P. Bhatia, *Data Mining and Data Warehousing Principles and Practical Techniques*. Cambridge University Press, 2019.