SiMBA: Systematic Clustering-Based Methodology to Support Built Environment Analysis

(Discussion Paper)

Carlo Andrea Biraghi[1], Emilia Lenzi[2], Maristella Matera[2], Massimo Tadi[1] and Letizia Tanca[2]

[1]*Politecnico di Milano - Dipartimento di Architettura, Ingegneria delle costruzioni e Ambiente Costruito*
[2]*Politecnico di Milano - Dipartimento di Elettronica, Informazione e Bioingegneria*

Abstract

The general interest in sustainable development models has grown enormously over the last 50 years, and architecture and urban planning are certainly two areas in which research on the topic is most advanced. At the same time, the contribution of computer science for a systematic analysis of the territory, both from a morphological point of view and as regards performance, seems to have been underestimated in today's research. In this context, our research aims to joining the two - until now separate - worlds of computer science, and architecture and urban planning. In particular, in this work we present SIMBA: Systematic clusterIng-based Methodology to support Built environment Analysis. SIMBA aims to enhance a consolidated analysis methodology, the Integrated Modification Methodology (IMM), through the integration of advanced analysis methods for the extraction of relevant patterns from built environment data. Using the city of Milan as a case study, we will demonstrate the possibility for SIMBA to be generalised to the analysis of any built environment.

Keywords

Clustering, Build environment, Methodology, Data mining, Multidisciplinary, Sustainability

1. Introduction

One of the biggest problems of our century is the impact that the behaviour of human beings is having on the environment. This impact, under the same conditions for the efficiency in resource management, is obviously greater in more densely populated areas. According to [1] 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. This is why the problem of sustainability in built environments - defined as every human-made space - has been broadly addressed in the literature. However, many issues are still open, and different approaches have deficiencies in several respects, i.e., the definition of the scale at which the analysis is conducted, the introduction of the context in the analysis, the usage of huge amount of data and features to describe the built environment.

SEBD 2022: The 30th Italian Symposium on Advanced Database Systems, June 19-22, 2022, Tirrenia (PI), Italy
carloandrea.biraghi@polimi.it (C. A. Biraghi); emilia.lenzi@polimi.it (E. Lenzi); maristella.matera@polimi.it (M. Matera); massimo.tadi@polimi.it (M. Tadi); letizia.tanca@polimi.it (L. Tanca)

In this paper we present SIMBA: a systematic clustering-based methodology to support built environment analysis. SIMBA has been conceived as a framework extending and enhancing IMM (Integrated Modification Methodology)[2], an innovative design methodology for the evaluation and the improvement of environmental performances of the city (or part of it). As described in [2] IMM has already proved its effectiveness in the field of sustainability evaluation; however, many challenges are still open and SIMBA proposes to support this methodology in the phase of built-environment analysis. Our case study is the city of Milan which, with its 88 NILs (Nuclei di Identità Locale), has been already the subject of many applications of IMM. In particular, the aims of the conducted experiments were (i) to enrich the phase of the built environment analysis through the use of data clustering techniques, and (ii) to systematise the basic steps of the analysis process by capitalizing on the support gained by data analysis. In particular, SIMBa aims to: (i) produce a methodology to select a reasonable, yet representative number of features when investigating the built environment; (ii) find experimental evidence of corresponding patterns between the structure of the city and its performance; (iii) produce a systematic method to measure the distance between elements, needed when comparing different built unit; (iv) promote a human-in-the-loop paradigm [3] where domain experts can intervene and refine the analysis process. With these objectives in mind, we have chosen to base the methodology on clustering techniques, that would support the identification of patterns in the built environment data, which are unlabeled, and highlight which of the input elements are more similar to each other. At the same time, such algorithms have been useful to single-out similarities and differences between the different elements, as well as the conceptual distance between them.

2. IMM Methodology and data retrieved

The IMM methodology aims at improving the environmental performance of the city by modifying its structural characteristics. The main elements of the methodology related to the structures are *Attributes* (e.g., the height of a building) and *Metrics* (e.g., Building Density (BD) equal to the total number of buildings in an area divided by the total sample area). *Attributes* represent immediately measurable characteristics while *Metrics* result from calculations. Moreover, from *Metrics*, IMM defines the *Key Categories*, used to represent the products of the synergy between elementary parts of the city. At the moment, the IMM group has defined 7 *Key Categories*, but for the sake of simplicity in our experiments, we will consider only Permeability (the spatial relationship between urban built-ups and voids) and Porosity (the relationship between the street network and the spatial components that influence the overall connectivity). The *Key Categories* are used to express morphological characteristics, while the tools for performance evaluation are called *Indicators* (e.g., Public Transportation Stop Density) organized into Design Ordering Principle (DOP) families, i.e., families of actions that designers can perform to improve the current system behaviour [2]. According to these definitions, we analyze the data related to (i) *Indicators*, (ii) *Metrics*, and (iii) *Attributes*, and we also perform the same experiment using directly some raw data coming from the Municipality of Milan, which are mostly related to air pollution, building characteristics, population, transport, and services for each NIL. The structure of each dataset is summarised in Table 1.

Table 1
Datasets

Dataset	Samples available	Number of features
`Indicators`	88	25
`Metrics`	86	59 (52 Porosity + 7 Permeability)
`Attributes`	86	52
`Milan`	88	31

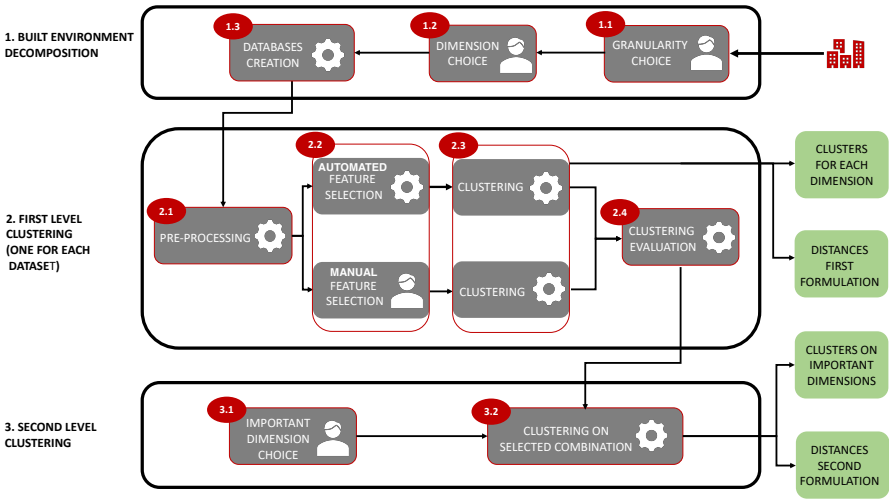


Figure 1: The SIMBA workflow.

3. Methodology and results

In this section we present the SIMBA methodology together with the experiments we carried out. Fig. 1 represents the methodological workflow. As you can see, the process takes as input a built environment of any kind: a town, a city or even a country, since it can be generalized to any input, as long as it is possible to identify comparable units in it. The case study discussed in this paper focuses on Milan. The methodology is composed of three main phases: (i) Built Environment Decomposition (BED phase); (ii) First-Level Clustering (FLC phase, one for each dataset); (iii) Second-Lelvel Custering (SLC phase). The following sections will describe each phase together with its application to our case study.

3.1. BED phase setting

The Built Environment Decomposition phase is useful to make the analysis manageable through Data Mining algorithms, since it aims at reducing the number of features in the single analysis

by splitting all the available data into distinct datasets according to different *Dimensions*.

In this phase, we first need to define the granularity at which the analysis has to be conducted (step 1.1). In other words, we need to choose which are the samples we want to cluster, and in our case study these samples are the 88 NILs. Although the definition of NIL is specific to the city of Milan, this does not affect the generalization potential of the methodology, since similar divisions (possibly at different scales) can be found in other cities, and therefore step 1.1 can be carried out. After the granularity, we need to define the *Dimensions* (step 1.2) and retrieve the data according to them (step 1.3). The *Dimensions* represent the different aspects we want to analyse. They can be of any type and category, e.g., environmental performances, morphological characteristics, demographic data and so on, but they remain abstract concepts for us: they have to work as simple guidelines to be used while looking for data to retrieve. In fact, it is in the datasets that the *Dimensions* are represented. As far as our experiments are concerned, we chose to focus on all the *Dimensions* at our disposal, as our four datasets include both performance data and structural characteristics. Consequently, the datasets we create for step 1.3 correspond to the ones we already defined in Section 2. Note that in some experiments we have considered all Metrics together, in others Permeability and Porosity separately.

3.2. FLC phase setting

Once we have our datasets ready, we perform First-Level Clustering for each of them. This phase is needed to:(i) prepare the datasets for clustering; (ii) select the important features; (iii) perform clustering; (iiii) evaluate each obtained cluster. Outputs of this phase are:(i) different clusterizations for each dataset; (ii) distances between the NILs for each dataset.

These outputs are useful to investigate patterns in each dataset and so for each *Dimension*. For what concerns the setting for this phase, note that the final choice of the techniques used for pre-processig, the algorithm chosen for feature selection, and the clustering algorithm itself, were dictated by the nature of the data at our disposal. On the other hand, what is also important to underline is that, once the process has been formalised, even if the specific setting is dependent on the input data, the procedure remains unchanged. For what concerns the pre-processing phase, we mostly had to manage missing values and perform some feature engineering. We strictly collaborated with the experts in this step (2.1) to preserve the data semantics as much as possible. As far as step 2.2 is concerned, for the manual feature selection we relied totally on the experts, asking them to select no more than 8 features for each dataset; for the automated case instead, the selection was conducted using the entropy-based algorithm presented in [4]. For clustering (step 3.3), we used the Agglomerating Hierarchical Clustering algorithm [5]. The choice was dictated firstly by the few samples at our disposal (only 88 NILs), and secondly by the need to choose a criterion for selecting the number of clusters in each experiment. For this parameter, in fact, no indications were given to us by the experts, as this type of experiment is new to them and is mainly exploratory in nature. For this reason, to determine the value of the parameter `n_clusters` we relied on the dendrograms and the Knee-Elbow graph we obtained for each dataset [6],

The last step of this phase is dedicated to the evaluation of the clusters, which is still a tricky part in this field and most of the times is strongly application-dependent. Firstly, we tried to have an absolute evaluation of each cluster result using internal metrics such us the Dunn and

Davies-Bouldin [6] indexes, but the problem with these 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 results obtained with the manual and the automated choice of the features. In this way, we use the results obtained from the manual feature selection as a proxy of ground truth, even if they derive from a semiautomatic process.

To do so, we first defined the `comparison_matrix`, a matrix used to store the number of common elements among the clusters obtained by selecting different set of features. For each experiment (dataset) we computed the `comparison_matrix` between the two approaches (the manual and the automated one). The pseudo code for the computation is shown in Algorithm 1 below.

Algorithm 1: 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 $comparison\_matrix$

In 1, the `comparison_matrix` CM is a matrix of dimension MxM, where M is the number of NILs for the dataset. For two NILs i and j, CM [i][j] $= 2$ if the two methods put the NILs i and j in the same cluster, CM [i][j] $= 1$ if only one method puts the two NILs in the same cluster, and CM [i][j] $= 0$ otherwise. This means that the positive cases correspond to the values 0 and 2 since they mean that, in the two approaches, the clustering has produced the same result for that pair of NILs.

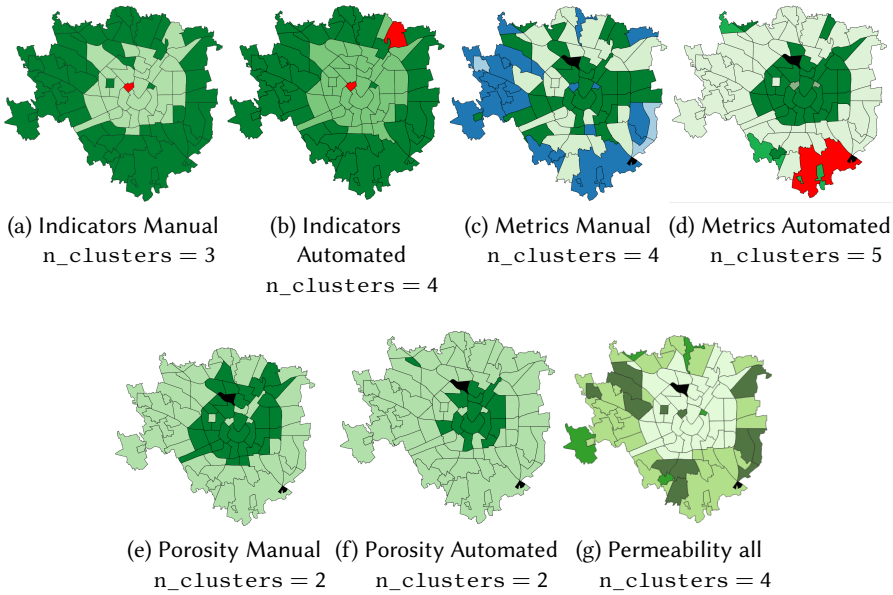
Finally, to evaluate each experiment, we defined the measure `score`, whose computation is shown in Algorithm 2.

Algorithm 2: 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 $score = \frac{good}{number\_of\_nils}$

In a given experiment, for each pair of NILs, `score` counts how many values 0 or 2 occur in the corresponding `comparison_matrix`, and normalizes this number w.r.t. the number of NILs present in that experiment (88 or 86). Assuming that a good result for our experiment is that the two clustering runs group all the NILs in the same way, in the Manual and in the Automated case, `score` can be seen as an accuracy measure for the procedure. In addition, this measure provides a direct and simple comparison between Manual and Automated cases.

Therefore, the `score` provides a numerical evaluation of the clusters obtained, and allows the procedure to be repeated in any proposed scenario. However, given the nature of the experiments and the absence of any real ground truth, a qualitative, although less formal, evaluation by the experts cannot be ignored. For this reason, in Figure 2 we report the most relevant results directly on the map of Milan. Each map is divided into NILs coloured according to the cluster they belong to. The colours have also been chosen to trace which clusters are



(a) Indicators Manual n_clusters = 3 (b) Indicators Automated n_clusters = 4 (c) Metrics Manual n_clusters = 4 (d) Metrics Automated n_clusters = 5

(e) Porosity Manual n_clusters = 2 (f) Porosity Automated n_clusters = 2 (g) Permeability all n_clusters = 4

Figure 2: Some results for the FLC phase.

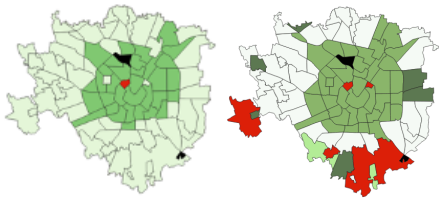
Table 2
Scores

Dataset	A Score
Indicators	76,34
Metrics	49,90
Porosity	49,72
Attributes	46,67
Milan	73,86

closest to each other; the outliers are always coloured red, therefore, in the figure, they represent a single cluster, but they contribute individually to increase the n_cluster parameter which, as we note, varies among the different experiments. In addition, black is used to highlight the NILs that are missing in some of the datasets. In table 2 we also report the scores obtained for all the datasets.

Figures 2a and 2b represent the clusters obtained by applying the Manual and the Automated algorithms to the Indicators dataset. What emerges from these maps is that the two algorithms provide almost the same results. Indeed, one might think that indicators only highlight macro-differences among NILs, but the results have been judged by the experts to be perfectly coherent with the way DOP families describe NILs. Moreover, as shown in 2, the score is rather close to the number of NILs (88).

Looking at the results related to the Metrics dataset (Figure 2c and 2d) we note that, in this case, the two algorithms (Manual and Automated) perform really differently. What emerges is that the features selected manually can express finer differences, while the automated algorithm



(a) Manual `n_clusters` = 3 (b) Automated `n_clusters` = 9

Figure 3: Results for indicators and metrics.

only separates the central NILs from the more peripheral ones, with some exceptions in the very peripheral areas. This is probably due to the fact that in the Manual case experts selected both Porosity and Permeability metrics to balance their effects, while the entropy-based algorithm mostly selected porosity-related ones. This is probably due to the imbalance between the number of metrics related to Porosity (52) and those related to Permeability (7) that affected the automated selection of the features.

This is confirmed by looking at the remaining maps in Figure 2 where we note that using only Porosity-related metrics we have a sharp separation between centre and periphery NILs, while, using the ones related to Permeability, it is possible to define some clusters also in the peripheral areas. This gives us an idea of how SIMBA can be useful to study the effect of the different Key Categories separately or jointly, and this is of huge importance for the IMM group.

3.3. SLC phase setting

Second-Level Clustering is the third and last phase of SIMBA, and it takes two inputs: (i) the First-Level Clustering results, to identify on which dataset the clustering performed better; (ii) the IMM experts' indications about the *Dimensions* to focus on to continue the analysis (step 3.1); and produces two outputs:(i) the clustering results on the dataset created by combining the important *Dimensions*; (ii) a formalization of distances between NILs considering only selected features of the important *Dimensions*.

In the previous sections we showed the results obtained by clustering NILs keeping the datasets separated. This was an important step, since it allowed us to understand what are the important IMM elements to focus on, both in an automated way and in a more guided one. As we have seen, both looking at the `score` values and at the produced maps, the Indicators dataset was the one where the procedure performed best. This is probably due to the fact that the indicators themselves are well separated into DOP families and the entropy-based feature-selection algorithm can extract at least one indicator for each family. Other interesting results, as we have shown, are obtained on the Metrics dataset. This time the score is considerably worse, but the meaning of the resulting clusters in terms of morphology can't be ignored. This is actually why we decided to add this last phase to the methodology, applying clustering to a dataset composed both by indicators and metrics (step 3.2). The results are reported in Figure 3.

By comparing maps, three important things emerge:(i) most of the NILs of the city centre are always in the same cluster; (ii) looking at the manual case, this seems to show practically the

same results as the Indicator case, while the automated one contains also some of the clusters identified in the Metrics automated case. This might mean that there could be some indicators that guide the whole creation of clusters, but some of the metrics selected in the automated case end up smoothing its effect.

4. Conclusion and future works

In this work, we presented SIMBA, a systematic clustering-based methodology to support built-environment analysis, and we proved its potential as a support tool for architects and urban planners in understanding and analysing urban environments. The biggest limitation of the work is the nature of the data. On the one hand, the scarcity of the samples (only 88 NILs) makes the analysis sensitive to outliers and difficult to formally generalise; on the other hand, the excessive number of characteristics, which are often redundant and not very informative, affects the quality of the results produced. Two aspects on which future work will therefore focus are the definition of finer granularity and the selection of features. Additional efforts will be devoted to refining the human-in-the-loop approach that SIMBA wants to support, by means of explanations that can guide the domain experts (and eventually, different experts) in the selection of relevant analysis *Dimensions* in the SLC phase. In addition, to assess the actual possibility for SIMBA to be generalised to the analysis of any built environment, different datasets with different characteristics should be analyzed.

5. Acknowledgments

We would like to thank Dott. Hadi Mohammad Zadeh who has contributed to this work and its future development.

References

[1] W. Bank, Cities and climate change: an urgent agenda, 2010.
[2] M. Tadi, M. H. M. Zadeh, O. Ogut, Measuring the influence of functional proximity on environmental urban performance via integrated modification methodology: Four study cases in milan, International Journal of Urban and Civil Engineering (2020).
[3] R. Guidotti, A. Monreale, S. Ruggieri, F. Turini, F. Giannotti, D. Pedreschi, A survey of methods for explaining black box models, ACM Comput. Surv. 51 (2019) 93:1–93:42.
[4] M. Dash, H. Liu, Feature selection for clustering, ACM digital library (2000).
[5] D. Wunsch, R. Xu, Clustering, IEEE Press Series on Computational Intelligence, Wiley, 2008.
[6] M. J. Zaki, J. Wagner Meira, Data Mining and Machine Learning: Fundamental Concepts and Algorithms, Cambridge University Press, 2020.