SPATIAL PATTERN ANALYSIS THROUGH DISTRIBUTION METRICS

C. A. Biraghi [1*], E. Lenzi [2]

[1] Department of Architecture, Built Environment and Construction Engineering, Politecnico di Milano, Via Giuseppe Ponzio 31, 20133 Milano, Italy - (carloandrea.biraghi, emilia.lenzi)@polimi.it
[2] Dipartimento di Elettronica, Informazione e Bioingegneria, Politecnico di Milano, Via Giuseppe Ponzio 34/5, 20133 Milano, Italy

Commission IV, WG IV/3

KEY WORDS: Spatial patterns; Image analysis; Multi-Metrics; Feature selection; Clustering; Neural networks; Urban morphology;

ABSTRACT:

Moving from the controversial results on the link between urban structure and performance aspect, this article wants to encourage the development of the independent research on urban structure, and more generally on spatial patterns, at different scales to enable future further correlations with a wider set of performance aspects (environmental, social, economic, medical). The work also exploits the potential of several unsupervised learning algorithms, whose performance and power are increasingly promising and whose use is becoming more widespread in different fields; but for which there are still many challenges concerning the correct application in urban areas and the interpretability of the results. We propose an approach for the creation of new spatial attributes and metrics (features) aiming to quantitatively describe the qualitative distribution of objects (e.g., buildings) in a 2D space. It explores an incremental bottom-up process for the creation of groups of objects (e.g., urban patches) and the evaluation of their physical properties alone and in respect with a sample area at each iteration. The process consists of 7 phases: data preparation, data processing, parameters collection, feature calculation, feature selection, clustering, results comparison. The results can be mainly divided in two. First, the feature selection allowed to extract a minimum set of non-redundant, valid, and consistent features that can explain qualitative distribution aspects of spatial patterns. Second, the comparison between feature-based and neural network clustering, gave useful insights for a preliminary understanding of unsupervised learning techniques internal mechanisms.

1. INTRODUCTION

Given the crucial role played by cities in the evolution of the climate change scenarios, understanding how to mitigate their environmental impact is more and more urgent. One existing approach, widely discussed in the contemporary literature, is investigating the relationship between urban structure and environmental performances of urban systems (Grosvenor, 2015). Many studies considered specific morphological aspects (density, building footprint, plot, or network properties etc.) trying to verify their correlation with specific climatic/energetic ones (heat island, ventilation, PV potential etc.). The results, often partial and controversial, ended up in confusing the domain of urban morphology investigation (Alberti, 1999). This article wants to encourage the development of the independent research on urban structure, and more generally on spatial patterns, at different scales to enable future further correlations with a wider set of performance aspects (environmental, social, economic, medical etc.). The present study moves from the urban studies domain but has a broader interest for all those disciplines dealing with the analysis of the distribution of objects in space, no matter at which scale. Milan municipality, together with ideal samples specifically generated for this purpose, is taken as a case study to test a novel approach for investigating spatial patterns.

Among the different components of urban systems (Tadi et al., 2020), it has been decided to focus on the relationship between Volumes and Voids analysing figure ground maps of urban fabrics. As the input data are binary images, the proposed method can be easily exported to other domains with an interest on image analysis, sharing techniques among different domains as done in previous urban related studies (Adolphe, 2001; Biraghi et al., 2019).

Image analysis consists of processing an image into key components to extract meaningful information. The field of image analysis has grown incredibly fast in the last years and has undergone a dramatic change: once, most of techniques and algorithms were built upon a mathematical/statistical description of images while, nowadays, machine learning methods are much more popular (Goodfellow et al., 2016). Unfortunately, most of the algorithms used, such as neural networks, and more in general unsupervised learning techniques as clustering, are generally perceived as being 'black boxes'. While these algorithms are very powerful and allow to analyse huge amounts of data and all possible image features, it is extremely difficult to document how specific decisions are reached and, above all, what features of the analysed image led to that decision (Qiu & Jensen, 2010). This type of algorithm therefore has the advantage of performing operations that are impossible for the human mind, but, precisely because of its complexity, it is not able to fully explain how it works. In this framework, this study proposes an approach for the creation and validation of new spatial attributes and metrics (features) aiming to quantitatively describe the qualitative distribution of objects (e.g. buildings) in a 2D space.

The topic of building distribution is not widely discussed in the literature of urban studies, probably for the difficulty of its objective characterisation, for the absence of an evident straightforward link with performance aspects, and for the difficulty of choosing the proper scale or morphological unit for its investigation. This study wants to move the first step in this promising direction to enrich the set of tools in the hands of urban designers and decision makers for understanding complex systems as cities. The results obtained could enlarge the existing pool of spatial metrics for different disciplinary domains and to suggest an approach to potentially open neural networks black boxes.

[*] Corresponding author

This contribution has been peer-reviewed.
https://doi.org/10.5194/isprs-archives-XLIII-B4-2022-69-2022 | © Author(s) 2022. CC BY 4.0 License.

2. METHOD

This study explores an incremental bottom-up process for the creation of clusters, the evaluation of their physical properties alone and in respect to a sample boundary area at each iteration. The final purpose is of extracting features able to numerically characterize differences among spatial patterns. The process can be seen as a physical clustering of polygons at different progressive distances and will be explained in detail in the data processing section (2.2).

A cluster can be defined as a group of similar elements positioned or occurring closely together. (Encyclopaedia Britannica). The closeness of the objects belonging to a cluster may be either topological or physical, but the second meaning is more interesting for the application to this branch of urban studies. In fact, especially in regional studies, clusters are used as a representation of spatially close objects, considered unique new objects not totally corresponding to the sum of their individual consistency. The most common cluster in urban studies are urban patches that have been widely used to describe urban form at city and regional scale (Frey 1999; Huang, Lu, and Sellers 2007). McGarigal (2004) defines patches as discrete areas of homogeneous environmental conditions. Alternatively, they can be seen as groups of elements within a certain mutual distance. According to Frankhauser (2004), the criteria used for their definition is not strongly scientifically based and the choice of the threshold distance is rather arbitrary. What is clear is that macroscopic patterns emerge from the interaction of the systems low-level (microscopic) adaptive agents (Brownlee 2007), as, in this case, buildings. The uncertainty related to the definition of these objects suggested a change of perspective trying to explore their continuity in space, explaining them as one step of a continuous bottom-up clustering process using buildings as input geometry. Figure 1 shows the portion of Milan at 8 different resolutions (or zoom level), corresponding to process iterations that will be later explained in detail.

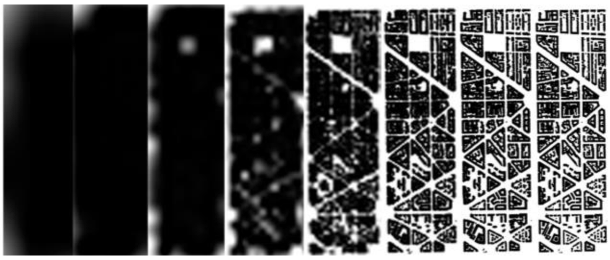

Figure 1. Resolution-clusterization analogy. From building to urban patch bottom-up process

The proposed analysis could be performed using both raster and vector data. The latter have been chosen for the need of having a more accurate representation of the objects. The process can be summarised in 7 phases: data preparation, data processing, parameters collection, feature calculation, feature selection, clustering, and neural network comparison. Each of them is presented in a dedicated subsection.

2.1 Data preparation

For this study, different kind of samples have been used to better explore the limits and the potential of the proposed methodology. They all have two inputs, one square cell (sample area) and several polygons within it (objects) and can be divided in two main groups: ideal samples and real urban samples.

Ideal samples are arrangements of black shapes on a white 100x100 unit cell and can be divided into manual and random ones. Manual samples have been drawn by the authors specifically to show extreme behaviours of the different parameters considered that will be presented in section 2.3. They consist in 25 layouts, significantly different from each other, all made of 16 black pixel-like squares of the same size (Figure 2).

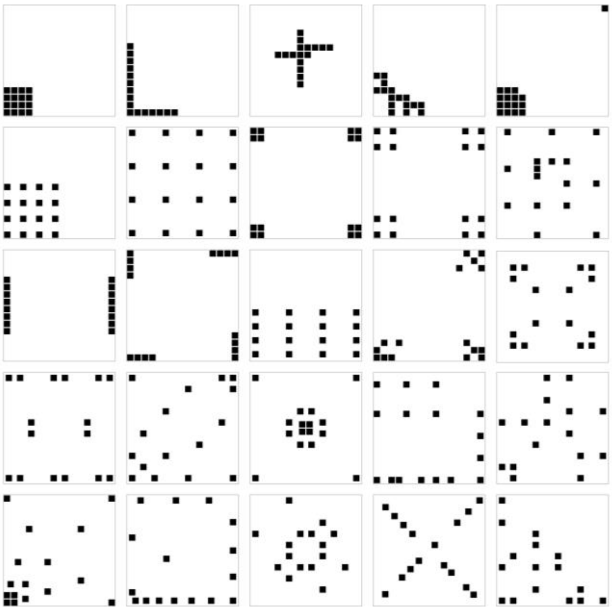

Figure 2. Manual samples

Random samples are obtained starting from a regular grid of 100 points covering the whole cell and selecting randomly a certain number of points from it. Given the novelty of the proposed methodology, different number of points were taken to test it on different percentage of coverage (black area over white area). This was done to create the conditions, for a subset of experiments, for the exclusion of the most simple and common features in the following steps of the process. In this way, the algorithms were forced to consider fewer common features. Separate experiments were made using 16, 25, 36 and 49 points used to generate first simple squares, and then applying random rotation and scaling to them keeping the total black area and the number of objects as constant. For both cases, 250 layouts were produced for each quantity, obtaining a total of 2000 ideal samples.

To overcome the rigidity of the initial grid, another set of 2000 samples was generated using a different method. For every layout, the cell was randomly populated with a random number of points (domain: 16 to 49) used to generate polygons of random number of sides (3 to 6), randomly rotated (0 to 2Pi) and scaled (0.5 to 12) with their centroids as centre of rotation and scaling. This new set of samples has no more constant number of features neither coverage value, being so closer to the variety that might be found in real cases of different domains.

Real samples are extracted from the city of Milan first creating a grid of 200x200m cell size covering the whole municipality, then keeping only those cells including at least one building. Buildings were derived from the INSPIRE compliant "Volumetric Unit" layer of the Topographic Database (DBT), downloaded from Lombardy Region Geoportal. This layer was first dissolved, then filtered to eliminate volumes with an area smaller than 100sqm, and finally simplified using the Douglas-

Peucker method (Douglas & Peucker, 2006) with a tolerance of 2m to reduce the computational time of the following steps. The objects input resulted from the intersection between buildings and the previously obtained sample area. Figure 3 shows one sample layout for each of the sets that include manually drawn, randomly generated and real urban cases.

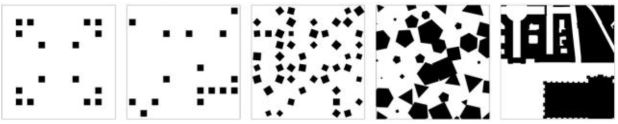

Figure 3. Sample layout of the different sets. From left to right: Manual; Random 16; Random 49 rotated and scaled, Random full; Milan (Duomo and Galleria Vittorio Emanuele II)

This operation was done using QGIS 3.16 while all the ideal samples and the steps from 2 to 4 of the methodology (data processing, parameters collection and feature calculation) were done using Grasshopper, the visual programming interface of McNell Rhinoceros 6. All the algorithms are available in a dedicated shared folder.

		layouts number	Object number	Void area	Void number	Point Distance	Cipq
Ideal	Manual	25	16	0.95	1	17.16 to 115.27	0.049
	Random 16	500	16	0.95	1	42.05 to 76.21	0.049
	Random 25	500	25	0.92	1	45.68 to 71.01	0.031
	Random 36	500	36	0.88	1	48.97 to 69.43	0.022
	Random 49	500	49	0.84	1	51.64 to 65.55	0.016
	Random full	2000	16 to 45	0.1 to 0.92	1 to 15	49.78 to 68.99	0.02 to 0.07
Real	Milan	3382	0 to 56	0.18 to 1	1 to 18	4.31 to 239.91	0.01 to 0.8

Table 1. Size of the different sets with the range of values for the V0 statistics of all the parameters. H-index is not considered as V0 is not applicable as always equal to 1

2.2 Data processing

The core of this study is the data processing phase. The proposed processing consists in buffering the input objects by an incremental distance value until specific parameters requirements are satisfied (see section 2.3). The idea moved from the study of Tadi et al. (2017) on urban porosity, where a metric called building distribution factor (BDF) was proposed with the goal of characterising buildings' arrangement. That metric was substantially the reciprocal of a density-based spatial clustering (DBSCAN) at a given distance (20m) performed on buildings' centroids. Even if the results were promising, the main limit of that approach were two: building shape was not considered and the distance was arbitrarily defined.

The first issue was easily solved by directly applying a buffer to buildings as polygons dissolving the result and then converting from multipart to single part features. The second one by repeating this operation starting from a distance of 1m and increasing it by 1m until all the buildings were finally gathered in a single cluster. Figure 4 presents a visual application of the incremental buffer process to an urban area, highlighting just selected distance thresholds.

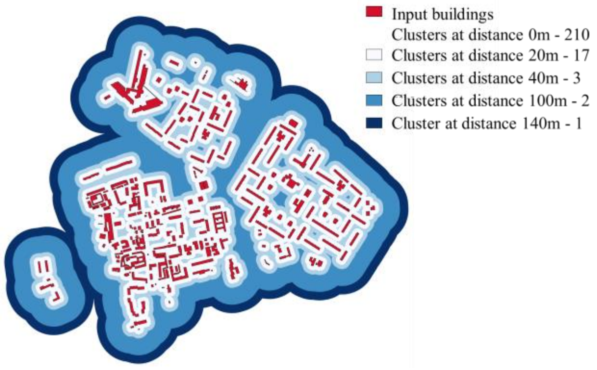
🟥 Input buildings
Clusters at distance 0m - 210
□ Clusters at distance 20m - 17
Clusters at distance 40m - 3
Clusters at distance 100m - 2
Cluster at distance 140m - 1

Figure 4. Incremental buffer applied to a urban portion of Milan. At 140m (limit distance) all input buildings (210) are gathered in a single cluster (Biraghi, 2019).

Plotting these values on a graph, with the number of iterations on the x axis and the number of clusters on the y axis, highlighted the peculiar nature of this discrete function (monotonically decreasing) and the fact that its trend was significantly different between urban contexts (Figure 5).

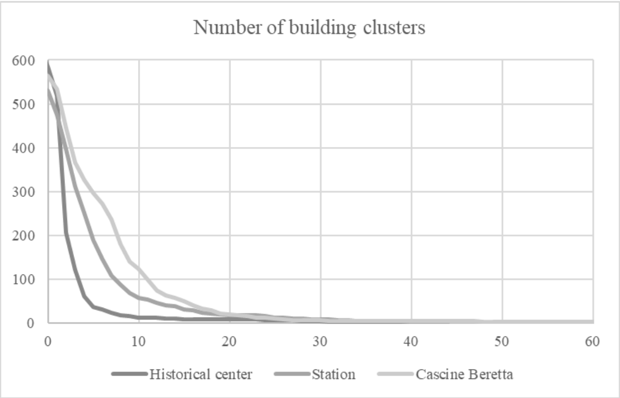

Figure 5. Graph showing the reduction in clusters number at every buffer iteration (Biraghi, 2019)

One key element of this process is the incremental buffer distance and its relationship with the sample area cell size. In fact, modifying it, the number of iterations required to reach limit values changes, and so does the number of clusters at each iteration. Given the transferability of this approach to different disciplinary domains, it is better to generalize distances defining a unit (*u*) that, case by case, can correspond to different lengths. In the ideal samples presented, e.g., the buffer distance is equal to 1/100 of the sample area cell side. To maintain the same proportion, in Milan samples, measuring 200m per side, a buffer distance of 2m was adopted. It is important to define *u* as a distance where almost at each iteration something happens. In an urban context, e.g., adopting a *u* of the order of cm or mm would increase considerably computational time without adding any value in term of detail or accuracy.

2.3 Parameters collection

The cluster number graph suggested to search for other possible parameters to be investigated, and similarly plotted, to describe objects distribution aspects that were not covered by the simple count of clusters at each iteration. In this study, 5 additional parameters, for a total of 6, are considered and here described. Figure 7 presents a comparison between 5 different manual samples (Figure 6) on the graph of each of the parameters.

Object number (On) has already been introduced and consists in counting the number of objects at each iteration. At each iteration, their value is equal or lower than that of the previous one. Another interesting property is that, at different distances, all layouts end up being grouped in a single cluster. The lower extreme of this discrete function is so 1. Its graph (Figure 7.a) is a monotonically decreasing function and is characterized by the number of steps, their entity and by how fast the limit value of 1 is reached.

Void area (Va) is based on the ratio between the total void fraction over the total sample area. The value at the first iteration is the reciprocal of the coverage while, at the limit distance, it turns to 0 because the whole sample area is filled by the buffered objects. Its graph is a continuous monotonically decreasing function too (Figure 7.b). Layouts with a homogeneous distribution of objects on the sample area will saturate earlier the sample area while layout concentrated in a small portion require more iterations. It is also affected, as can be imagined, by the number of objects in the sample area.

Void number (Vn) works as the object number parameter but focuses on its negative. The voids are objectified and counted at each iteration until there's no void space left. Its limit distance consequently coincides with that of Void area parameter. In the built environment domain, the voids at iteration 0 represent closed courts, an important morphological element for characterizing the urban fabric, while at different iterations, they can highlight semi-closed courts, squares, or potential public spaces. Considering that at each iteration some existing void can disappear, being filled by the growing buffer, and some other can appear, cropped by the buffer in between neighbouring objects, its trend is not monotone (Figure 7.c).

Point distance (Pd) represents an experimental way of determining the position of objects in the space. It consists of the average distance of all objects from a point located at u=0.2 and v=0.1 of the cell domain. Small values indicate the objects located in the bottom left quarter of the image, high values those in the top right one and intermediate values are for objects located on the quarter of circle connecting top left and bottom right quarter. Its function is continuous but not monotonic (Figure 7.d) and it's affected by the changes in the object number as new aggregations can have a different centroid respect to the parent ones. It stops changing significantly when a single object includes all the input ones.

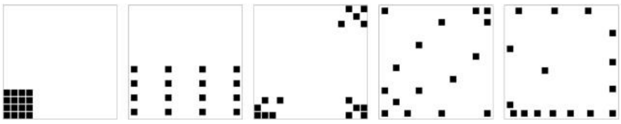

Figure 6. Manual samples compared in Figure 4. From left to right: 0; 12; 13; 16; 21.

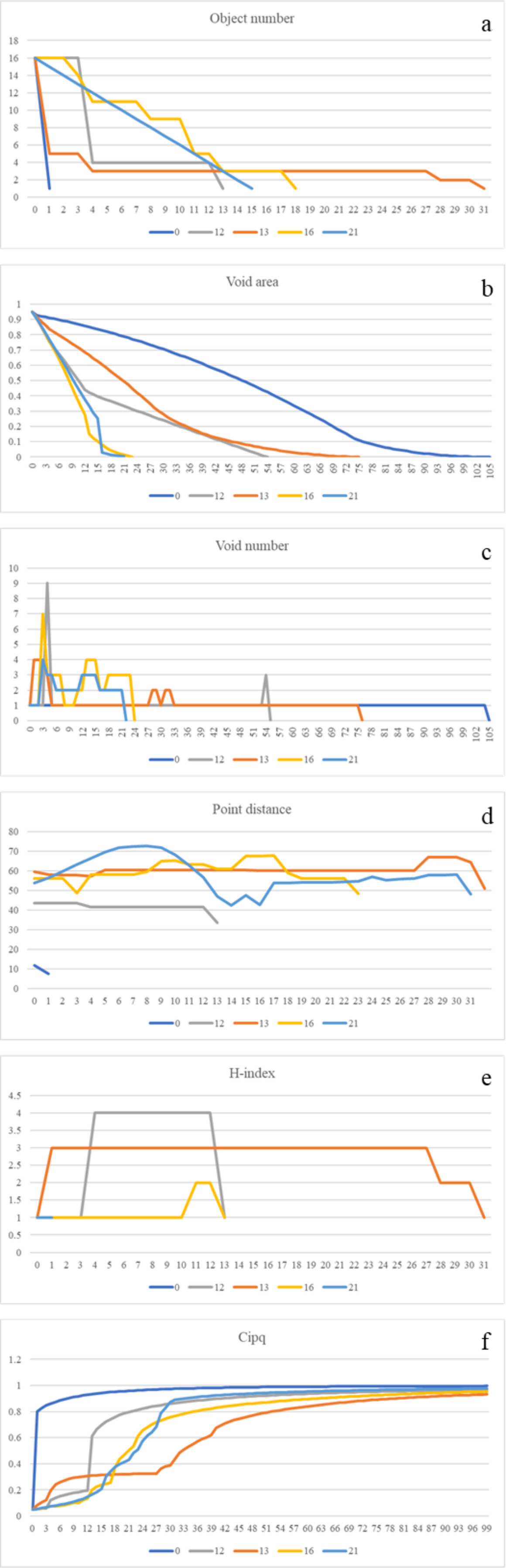

Figure 7. Graphs comparing 5 manual samples for the 6 available parameters. Each graph shows a single parameter (y axis) plotted for all the iterations (x axis) up to reaching its limit value. a) Object number; b) Void area; c) Void number; d) Point distance; e) H-index; f) Cipq

H-index (Hi) is a commonly accepted metric to evaluate the scientific production introduced by Hirsch (2005). It can be defined as the number of papers (h) that have received at least h citations. To adapt this concept to spatial analysis papers have substituted by the objects at the current iteration and citation by the input objects (iteration 0). This information is somehow complementary to the simple object number parameter and the two are not correlated. The maximum value, that is not necessarily reached by all the layouts, is the square root of the input object number. For 16 objects, e.g., no more than 4 clusters of 4 objects each can be identified (Figure 7.e).

Cipq (Ci), is the acronym of Compactness isoperimetric quotient (Osserman, 1978) and is a measure of how a 2d shape differs from a circle, the most compact shape. It was selected among the number of existing compactness measures (Bribiesca, 2008) because of its direct relationship with the proposed data processing method. If fact, the result of the buffer (with round corner) at an infinite number of iterations is always a circle, no matter the number and the shape of the input objects. More compact layouts will so arrive earlier to values close to 1 while linear or concave ones require higher number of iterations. Given that the value of 1 is the limit to infinite, to reasonable number of iterations for stopping the process must be found. Looking at Figure 7.f it can be noticed that its graph is a continuous monotonically increasing function, and that already at values in between 0.7 and 0.9 significant differences among layouts emerge.

2.4 Feature calculation

It was then clear that the richness of information enclosed in the graph couldn’t be compressed in a single metric, no matter its complexity. At the same time, visually comparing the graphs was not the ideal solution for the development of a scientific method and something more measurable was needed. Given the explorative nature of the approach, arbitrarily select one or more measures to describe each graph didn’t seem the optimal solution. For this reason, a wide set of statistics was used to extract from the graphs their content and to represent it into a collection of features, to be considered as attributes of the layout that generated the graph. In a following step (see section 2.5) these features will be processed to objectively determine their relevance for the purpose of the study. A total of 25 statistics was considered including commonly used and more experimental ones inspired by the work of Holzer et al. (2013) on constrictivity. Table 2 collects a synthetic description of them, and the acronym used in the rest of the paper.

As not all of them were meaningful or even applicable for all the parameters, only the combinations of parameters and statistics able to generate potentially useful features have been considered (Table 5). The result is a total of 106 features to be tested for correlation and for their ability to characterize peculiar aspects of a spatial pattern.

In the first attempts, some features presented null values for certain layouts. The reasons were mainly: the number of iterations performed was not high enough; Calculation method of certain statistics; impossibility to compute certain stat (es. Mode with no duplicate values and the related statistics).

The first issue was solved by increasing the number of iterations from 50 to 150. To avoid this risk, a general principle to be adopted is defining the iterations limit as the max distance among two points of a sample area divided incremental buffer distance.

V0	parameter value at iteration 0
Dl	limit distance after which parameter become constant
V=D	value at which the parameter is equal to the number of iterations
Vsum	sum of the parameter values for all the iterations
Stepsum	sum of all the unique parameter values for all the iterations
Dstepmax	distance at which there's the maximum step in the parameter value
Stepmax	maximum step in the parameter value (difference between consecutive iterations)
VD5	parameter value at iteration 5
VD10	parameter value at iteration 10
VD25%	parameter value at 25% of the limit distance (rounded)
VD50%	parameter value at 50% of the limit distance (rounded)
DV25%	distance at which the parameter value is the 25% of V0
DV50%	distance at which the parameter value is the 50% of V0
Mode	parameter value that appears more frequently (Karl Pearson, 1895)
ModeCount	number of time that the mode value occurs
D0Mode	first distance at which mode value occurs
D1Mode	last distance at which mode value occurs
Mean	sum of the parameter values divided by their number
Max	largest parameter value occuring
R2Lin	coefficient of determination R^2 for a linear regression
R2Log	coefficient of determination R^2 for a logarithmic regression
mLin	slope of the fit line of the linear regression
LinLog	R2Lin - R2Log + 0.5. Values above 0.5 are better represented by a line, value below by a log function
Gini	Gini index is a measure of statistical dispersion (C. Gini, 1912)
HHI	Herfindahl-Hirschman Index is a measure of market concentration (A. O. Hirschman and O. Herfindahl, 1945)

Table 2. List with the acronym and the description of the statistics considered

The second issue was manually fixed by replacing NULLs with the proper value. In future implementations, automatic value to assign in case of NULL can be set for specific features (mostly referring to H-index parameter). The third issue was solved by discarding the layouts still presenting NULLs. The missing ratio for the different features in different datasets was quite low (between 0.05% and 14.5%) so we preferred to keep all the features (on which our analysis is focused) and delate some samples. The complete list of features with the correspondent number of NULLs for each set is available in the online shared folder.

2.5 Feature selection

The feature selection phases, as well as the following clustering one (section 2.6), was performed using the recently developed QGIS Hierarchical Clustering plug-in (Folini, Lenzi and Biraghi, 2022), and can be divided in two steps. Step 1 eliminates those features with a mutual correlation higher than 0.8, keeping only the one with the lowest average correlation, and also those features resulting as constant or quasi-constant (frequency ratio = 19; percent of unique values = 5; https://topepo.github.io/caret/pre-processing.html#nzv). As the plug-in works with shapefiles as inputs, the preliminary output of this phase is a shapefile where only the selected features are kept, removing from the attribute table all the others.

Step 2 of this phase moves from this result and applies an entropy-based algorithm for feature ranking and selection (Dash & Liu, 2000) to further select a smaller subset of features. The

algorithm is based on the idea that datasets with a higher entropy - disordered points, not all at the same distance from each other - tend to form clusters more easily. The features are therefore ranked according to their impact on the value of the initial entropy, and only those whose elimination increases it (and so worsens the dataset for clustering) are preserved.
In the results section the selected metrics for the various experiments on the different sets (introduced in Table 1) are presented.

2.6 Clustering

The features selected are then used as dimensions for a clustering process using different algorithms. The goals of the clustering are mainly two. On one side examining the quality of the clustering based on the selected features and visually compare layouts belonging to different clusters. The results of this preliminary informal evaluation were considered satisfying by the authors and in line with the expectations. On the other, producing a result comparable to that of the neural network (section 2.7) to evaluate the correspondence between the two and consequently determine potential links between the selected features among the sets described above and the completely unsupervised feature selection performed by the network.
Considering that, the Neural network algorithm was applied only to the case of Milan real samples. Being this the only case where different layouts were printable on a map and representing a context well known by the authors, the result was visually interpreted looking for macroscale patterns. The ideal cases were mainly used to analyse the features selected and validate the results obtained.
The clustering algorithms used were Hierarchical and K-means (Zaki & Meira, 2014) defining clusters number looking at the Between Sum of Squares (BSS) and Within Sum of Squares (WSS) diagrams. For the case of Milan, possible clusters number was derived from existing well-routed classifications as Local climate Zones (LCZ), Land Use of Agricultural and Forestry Soils (DUSAF), and local urban planning tools as the Milano territorial governance plan (PGT). In LCZ, 10 of the existing classes refer to the built-up area, and this number can be reduced to 5 ignoring building heights (compact, open, lightweight, large, sparsely). DUSAF also presents 10 classes, reducible to 7 by neglecting some functional distinction and even to only 3 at the lower level of detail. In table R03 of Milan PGT the Morphological norms clearly subdivide the urban fabric in 3 macro classes (NAF, ADR and ARU) that can be more ambiguously detailed in 5 or even 7 classes. 10, 7 and 5 clusters values were so considered as cluster numbers.

2.7 Neural network

This last step aims to further evaluate the clusters obtained in the previously described feature space, by comparing them with the one obtained through an even more unsupervised approach. To do this, we used the previously trained VGG16 model (Simonyan & Zisserman, 2014) available in Keras (https://keras.io/api/applications/) for the feature extraction. Having relatively few images available (less than 4000 samples) we chose the second fully connected layer as output layer to reduce dimensionality. To make the results comparable, we then used K-means as the clustering algorithm.

3. RESULTS

The last three phases presented in the method chapter produced results that are here presented and discussed in homonymous subchapters, following their order.

3.1 Feature selection

Given the number of features considered and the sets tested, the complete table with all the results can't be included in the paper but is publicly available online. In this section, a summary of the filtering due to Step 1 and Step 2 is presented with the purpose of highlighting how parameters were affected (Table 3). It can be noticed that no features were finally selected in all the experiments after both the steps and only 3 (On_DV25%; Hi_Dl; Hi_VD50%) were maintained on 5 out of 7 experiments. This can be explained by the huge heterogeneity of the samples, especially including manual samples, where many features were removed as constant, and only 7 features were finally selected.

		On	Va	Vn	Pd	Hi	Ci
	Input features	24	14	19	18	15	15
Step 1 Correlation + Variance	Features selected once or more	23	13	19	16	15	11
	Features selected 70% of the times	12	2	15	11	12	2
	Features always selected	1	0	7	3	3	0
	Features average selction rate	0.57	0.36	0.78	0.63	0.74	0.33
Step 2 Entropy	Features selected once or more	21	8	11	11	14	9
	Features selected 70% of the times	3	0	0	0	2	0
	Features always selected	0	0	0	0	0	0
	Features average selction rate	0.31	0.09	0.11	0.09	0.35	0.10

Table 3. Summary of the features selected considering all the experiments together.

Table 4 presents the number of features selected after Step 1 and Step 2 for the different experiments.
Excluding manual samples, the feature selection is quite homogenous for all the parameters at step 1, except for Ci in Random full samples and Pd in Milan samples. The first can be explained by the behaviour of the populate geometry function used that has a random but quite homogenous distribution. At step 2, the most interesting aspects to be noticed are the only one feature selected for Hi parameter (Hi_DStepmax) on Milan samples. This can be due to the fact that the input objects in the real cases can be both isolated buildings and a single geometry representing the aggregation of more buildings, as in the case of attached block houses. Using single buildings as input is a possibility that must be explored but was not considered for the unreliable way in which they're mapped in the selected context.

	On (24)		**Va (14)**		**Vn (19)**		**Pd (18)**		**Hi (15)**		**Ci (15)**	
	Step 1	Step2	Step 1	Step2	Step 1	Step2	Step 1	Step2	Step 1	Step2	Step 1	Step2
Manual	4	1	0	0	11	0	10	3	6	3	1	0
Random 16	18	5	5	1	16	0	13	0	13	9	6	0
Random 25	18	12	6	3	15	1	13	1	15	8	6	0
Random 36	15	12	6	0	15	2	13	2	11	4	6	0
Random 49	11	7	5	0	15	3	12	1	10	4	7	3
Random full	16	9	8	4	18	6	12	0	13	8	2	1
Milan	14	7	5	1	13	3	7	4	10	1	7	7

Table 4. Number of features selected after Step 1 and Step 2 for the different experiments.

To make a further comparison among all the features, only the Random full and the Milan experiments have been considered.

This choice is due to the fact that, as previously explained in section 2.1, the other datasets (manual, random 16-25-36-49) were created keeping some features constant, namely, the combination of all parameters except Pd with the V0 statistic. In addition, the rigidity of their layouts (Figure 3) can't be compared with the variety of the other sets.

Table 5 presents all the combination of parameters and statistics that have been considered, telling, for each of them, if they were selected, after the two Steps of feature selection, twice (1.0), once (0.5) or never (0.0). Bold values indicate the features selected for the urban sample of Milan. Similar tables just referring to Step 1 and including all the experiments can be found online.

		Parameters					
		Object number (On)	Void area (Va)	Void number (Vn)	Point distance (Pd)	H-index (Hi)	Cipq (Ci)
Statistics	V0	0.5	0.0	0.0	0.0		**0.5**
	Dl	0.5	0.0	0.5	**0.5**	0.5	**0.5**
	V=D	0.0		0.0			
	Vsum	**1.0**		0.0	0.0	0.5	
	Stepsum	0.0		0.0	0.0	0.0	
	Dstepmax	**1.0**		**0.5**	**0.5**	**1.0**	
	Stepmax	0.0		0.0	0.0	0.0	
	VD5	0.0	0.0	0.5	0.0	0.0	0.5
	VD10	0.5	0.5	0.5	0.0	0.0	0.0
	VD25%	0.0	0.0	0.5	0.0	0.0	0.0
	VD50%	0.0	0.5	0.0	0.0	0.5	0.0
	DV25%	**1.0**	0.0				0.0
	DV50%	**0.5**	0.0				0.0
	Mode	0.0		0.0	0.0	0.5	
	ModeCount	0.0		0.0	0.0	0.5	
	D0Mode	**0.5**		0.0	0.0	0.5	
	D1Mode	0.5		0.0	0.0	0.0	
	Mean	0.0		0.5	0.0	0.5	
	Max			0.5	0.0	0.0	**0.5**
	R2Lin	**1.0**	0.0				**0.5**
	R2Log	0.0	**0.5**				0.0
	mLin	**0.5**	0.0				0.0
	LinLog	0.0	0.0				**0.5**
	Gini	0.0	0.5	**0.5**	**0.5**		**0.5**
	HHI	0.5	0.5	**0.5**	**0.5**		**0.5**

Table 5. Combinations of parameters and statistics used to calculate the features and their average selection rate for Random full and Milan sets. Milan features are bold.

A total of 5 features were selected for both Milan and Random full sets. Four of them belong to the Object number parameter (Vsum; DStepmax; DV25%; R2Lin) and one to H-Index (DStepmax). DStepmax statistic, describing the iteration at which the parameter value presents the bigger step, was selected for all the parameters it was combined with in the urban experiment. Similarly, also Gini and HHI were selected in couple for three different parameters (Vn, Pd and Ci) in the urban case. HHI was selected at least once (Random or Milan) for all the parameters. Ci_V0 feature was selected after Step 2 for the case of Milan while it was discarded, already at Step 1, for all the other experiments. This is due to the absence of holes within the objects in the ideal layouts opposed to the presence of courts in real agglomerations of buildings. In the urban case, at least on feature for all the parameters was finally selected, with a peak of 7 features for the On and Ci parameters. It's interesting to notice the presence of the R2Lin statistic for On and that of R2Log for Va, suggesting that the two parameters have commonly different trends. The Va parameter, included in the analysis for its capability to describe both the percentage of void space over the total one and the homogeneity of objects distribution in the sample, was selected less time than expected. This can be explained by looking at the correlation heatmap (Figure 8) of the different experiments that clearly show a high average correlation among most of its features. In certain experiments as the Random full, high correlation values emerged also with Ci features.

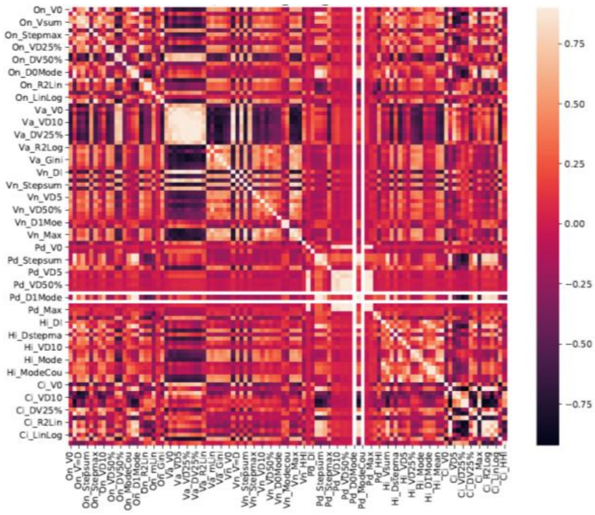

Figure 8. Correlation heatmap for Milan set

3.2 Clustering

As mentioned at the beginning, the focus of the work is on the selection of features and their ability to highlight more or less evident spatial patterns. For this reason, in the construction of the dataset and calculation of statistics, the exploration of as many parameters as possible was preferred, to the detriment of the quality of future clusters. Very often, in fact, the number of features selected by the algorithm is very high compared to the number of elements. Rather than clusters with high values of Silhouette Coefficient or Davies Boulding Index (Zaki & Meira, 2014), the results of this phase were mainly used, case by case, to have a better understanding of the possible issues. Apart from the correspondence with the Neural network clustering, that will be better explained in the next section, interesting aspects emerged while working on the manual and Milan sets.

Dividing the layouts of the manual set in relatively small number of clusters is a tough task both for humans and for algorithms because of their heterogeneity. A person, by doing it visually, risks to involuntarily overweight one or more feature neglecting others. Trying to cluster them considering a single parameter at the time is surely simpler, even if not trivial, but generates a different clustering for each parameter. Considering such a huge number of features together can provide not intuitive results, in contrast with visual evidence.

As there are here no reasons for choosing a specific number of clusters, the dendrogram of the hierarchical clustering was plotted to visually examine different scenarios. (Figure 9).

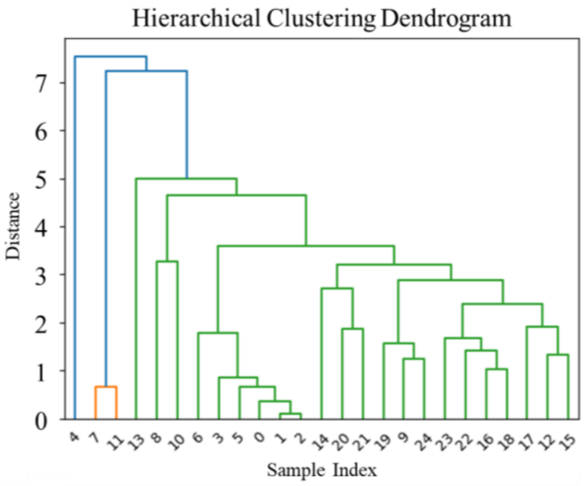

Figure 9. Dendrogram resulting from the hierarchical clustering on the manual set

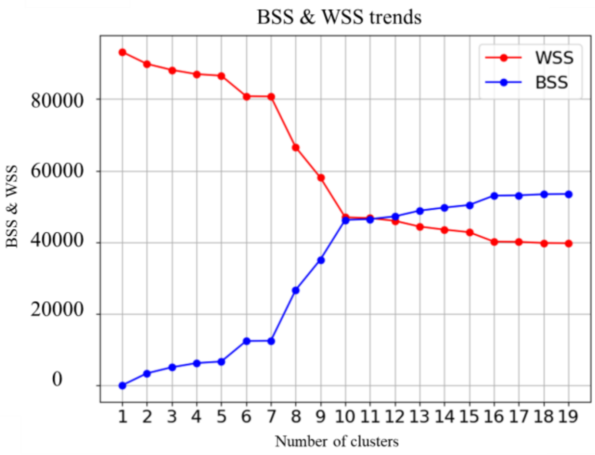

Figure 10. BSS WSS diagram of the hierarchical clustering on Milan set

From the plot we can infer that a good number of clusters could be 3. In this way we obtain quite good clusters (having low inter–class dissimilarity) but we highlight only few outliers. By increasing the number of clusters, the number of outliers also increases, but clusters with a reasonable number of elements, and significant from the point of view of our analysis, are also created. The evaluation of this type of experiment is therefore closely linked to the type of analysis to be carried out. Given the exploratory nature of our work, there is no correct number of clusters, but this procedure provides a valid tool to analyse possible results and discover patterns not visible to the human eye. In addition, an interesting finding from Milan set, is the correspondence between the number of clusters derived from the literature (5, 7 and 10, see section 2.6) and those emerging from the BSS WSS diagram of the hierarchical clustering using the selected features (Figure 9). Results obtained using K-means are described in the next section.

3.3 Neural Network

As said, the purpose of this last step is to validate the results obtained so far. For this reason, the results are presented in Table 6 in which, for different pairs of experiments, the `score` metric (Lenzi, 2020) was calculated. The score counts how many times two samples (in this case images) are clustered together in the different experiments being compared.

		Feature based					Neural network				
	cluster number	5	7	10	15	20	5	7	10	15	20
Feature based	5	1.000									
	7	0.891	1.000								
	10	0.790	0.802	1.000							
	15	0.789	0.786	0.877	1.000						
	20	0.750	0.770	0.846	0.898	1.000					
Neural network	5	0.690	0.711	0.744	0.757	0.774	1.000				
	7	0.688	0.715	0.758	0.784	0.813	0.883	1.000			
	10	0.675	0.707	0.761	0.795	0.830	0.851	0.926	1.000		
	15	0.683	0.716	0.779	0.821	0.860	0.818	0.877	0.915	1.000	
	20	0.678	0.713	0.780	0.824	0.868	0.814	0.865	0.901	0.934	1.000

Table 6. Comparison of the score values for the different clustering method (Feature based and Neural network) using Kmeans and cluster number values (5, 7, 10, 15, 20)

Looking at the values highlighted in the table, it can be noticed that, for the same number of clusters, the feature-based Kmean and the neural network based one clustered at least 69% of the images in the same way. Moreover, these numbers clearly show an increase in the score values between Feature based and Neural network clustering as the number of clusters increases (grey cells). This is somehow expected and coherent with the research assumptions. In fact, to simply classify spatial patterns in a small number of extremely heterogenous sets, few simple features, like the percentage of void space or the average objects size, are usually enough. To represent finer grain differences, more dimensions need to be considered. To conclude, it can be noticed that experiments with a minor distance in terms of cluster number correctly perform better than those with a larger one.

Apart from the calculation of the score, the results of this phase were also plotted on the map of Milan and analysed by experts. In all cases they turned out to be more than reasonable despite the lack of finetuning of the network, and the scarcity of data on the one hand and the abundance of features on the other. Figure 10 presents the Milan map with the morphology in black overlapped to the grid cells divided in 10 clusters using the feature based K-means algorithm.

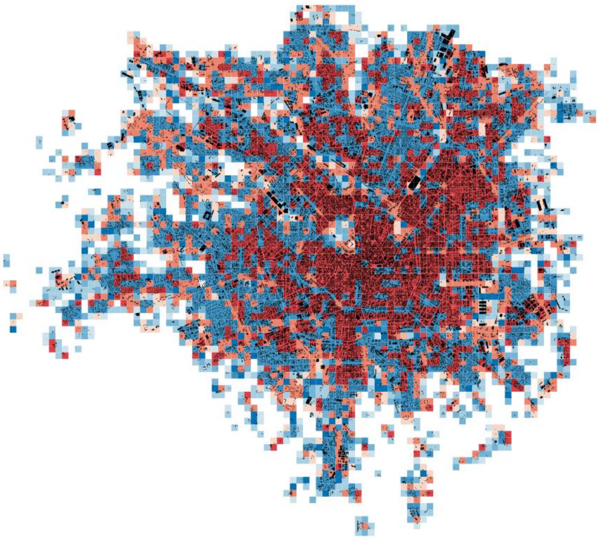

Figure 10. Milan feature based 10 clusters

Darker clusters, both red and blue, represent the more compact and densely built part of the city. Other 2 clusters, light blue, and orange, characterise cells where the compact city encounters urban voids. The other lighter fades highlight differences among cells with a predominant percentage of void. In general, commenting this map is quite hard and a cell-to-cell correspondence should be found instead. The use of larger cells may produce results of clearer interpretation at the level of the hole city.

4. CONCLUSIONS

Thanks to this study, new measures for numerically characterising urban morphology, and more generally, spatial patterns, emerged. This will hopefully encourage more independent research on the purely structural component of systems, both urban and not. The results achieved are far from being exhaustive but are promising, particularly in the perspective of better understanding the behaviour of unsupervised clustering techniques.

The potential of the proposed approach of being applied, with the proper adaptations, to any kind of spatial patterns, encourages to keep exploring its potential and limits. In fact, keeping phase 2.2 as a constant, the proposed approach can be easily applied to new input data (2.1), include new parameters (2.3) and apply them additional statistics (2.4) to enlarge the set of features to be tested in the last three phases. Also, the algorithms presented in section 2.5, 2.6 and 2.7 can be substituted with alternative ones, performing similar tasks. In particular, in our case, the choice of clustering algorithms was dictated by the characteristics of the tools used, which do not provide density-based or probabilistic clusters; for the neural networks, on the other hand, due to the lack of data, it was not possible to draw the most suitable structure for the task. Eventually, even details of phase 2.2 could be modified substituting, e.g., the incremental buffer with an incremental bidimensional scale centred on each object centroid or changing the type of buffer from round to square. This last change will mostly affect the Vn values reducing the appearance of micro-voids in orthogonal layouts. To consolidate the results achieved in the urban domain, cross-comparisons between a great variety of urban contexts as well as different morphologically relevant grid size (e.g. 400x400m or 800x800m) have to be considered.

This study ultimately represents a preliminary domain-specific investigation of a promising multidisciplinary technique for analysing spatial patterns. Replicating this approach on a variety of urban contexts, as well as on samples from other domains, will allow to provide clearer evidence on the importance of specific features for explicitly or implicitly guide unsupervised clustering algorithms.

REFERENCES

Adolphe, L. (2001). A simplified model of urban morphology: Application to an analysis of the environmental performance of cities. *Environment and Planning B: Planning and Design*, *28*(2), 183–200. https://doi.org/10.1068/b2631

Alberti, M. (1999). Urban patterns and environmental performance: What do we know? *Journal of Planning Education and Research*, *19*, 151–163.

Biraghi, C. A. (2019). *Multi-scale modelling approach for urban optimization: compactness environmental implications* [Politecnico di Milano]. http://hdl.handle.net/10589/150884

Biraghi, C. A., Ceriotti, G., Porta, G., & Tadi, M. (2019). Development and implementation of a quantitative multi-metrics methodology to characterize urban Permeability. *ACEU*, *June*. https://doi.org/10.5176/2301-394X

Bribiesca, E. (2008). An easy measure of compactness for 2D and 3D shapes. *Pattern Recognition*, *41*(2), 543–554. https://doi.org/10.1016/j.patcog.2007.06.029

Dash, M., & Liu, H. (2000). Feature Selection for Clustering. *Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, *1805*, 110–121. https://doi.org/10.1007/3-540-45571-X_13

Douglas, D. H., & Peucker, T. K. (2006). Algorithms for the reduction of the number of points required to represent a digitized line or its caricature. *Https://Doi.Org/10.3138/FM57-6770-U75U-7727*, *10*(2), 112–122.

Folini, A., Lenzi, E. & Biraghi, C.A., 2022. Cluster Analysis: a comprehensive and versatile QGIS plugin for pattern recognition in geospatial data Int. Arch. Photogramm. Remote Sens. Spatial Inf. Sci.

Goodfellow, I., Bengio, Y., & Courville, A. (2016). *Deep Learning*. MIT Press.

Grosvenor, M. (2015). *Can urban planning deliver sustainable outcomes: measuring the association between urban structure and form and sustainable household behaviour*. *194*, 131–141. https://doi.org/10.2495/SC150121

Hirsch, J. E. (2005). *An index to quantify an individual's scientific research output*. www.pnas.orgcgidoi10.1073pnas.0507655102

Holzer, L., Wiedenmann, D., Münch, B., Keller, L., Prestat, M., Gasser, P., Robertson, I., & Grobéty, B. (2013). The influence of constrictivity on the effective transport properties of porous layers in electrolysis and fuel cells. *Journal of Materials Science*, *48*(7), 2934–2952. https://doi.org/10.1007/s10853-012-6968-z

Lenzi, E. (2020). *SIMBA: systematic clustering-based methodology to support built environment analysis*. Politecnico di Milano.

Osserman, R. (1978). The Isoperimetric Inequality. *Bulletin of the American Mathematical Society*, *84*(6). https://www.ams.org/journal-terms-of-use

Qiu, F., & Jensen, J. R. (2010). Opening the black box of neural networks for remote sensing image classification. *Http://Dx.Doi.Org/10.1080/01431160310001618798*, *25*(9), 1749–1768. https://doi.org/10.1080/01431160310001618798

Simonyan, K., & Zisserman, A. (2014). Very Deep Convolutional Networks for Large-Scale Image Recognition. *3rd International Conference on Learning Representations, ICLR 2015 - Conference Track Proceedings*. https://doi.org/10.48550/arxiv.1409.1556

Tadi, M., Zadeh, M. H., & Biraghi, C. A. (2020). The Integrated Modification Methodology. In *Environmental Performance and Social Inclusion in Informal Settlements A Favela Project Based on the IMM Integrated Modification Methodology* (Research f, pp. 15–37). Springer.

Tadi, M., Zadeh, M. H. M., Biraghi, C. A., & Brioschi, L. (2017). Urban Porosity. A Morphological Key Category for the Optimization of the CAS's Environmental and Energy Performances. *Journal of Engineering Technology (JET)*, *4*(3), 478–484. https://doi.org/10.5176/2301-394x_ace17.68

Zaki, M. J., & Meira, W. (2014). *Data mining and machine learning : fundamental concepts and algorithms*.

This contribution has been peer-reviewed.
https://doi.org/10.5194/isprs-archives-XLIII-B4-2022-69-2022 | © Author(s) 2022. CC BY 4.0 License.