Exploring building typologies through fast iterative Bayesian clustering

Alessandro Araldi, David Emsellem, Giovanni Fusco, Andrea Tettamanzi, Denis Overal

Exploring building typologies through fast iterative Bayesian clustering

Alessandro Araldi[1], David Emsellem[2], Giovanni Fusco[1], Andrea Tettamanzi[3], Denis Overal[2]

1. Université Côte d'Azur, CNRS, ESPACE,
98, Bd Herriot, BP 3209, 06200 Nice, France.
[alessandro.araldi, giovanni.fusco]@univ-cotedazur.fr

2. KCITYLABS - Kinaxia Group,
80, route des Lucioles, 06560 Valbonne, France.
[david.emsellem, denis.overal]@kcitylabs.fr

3. Université Côte d'Azur, I3S
2000, route des Lucioles, 06900 Sophia Antipolis, France.
andrea.tettamanzi@univ-cotedazur.fr

RESUME. L'identification et la description des typologies de bâtiments jouent un rôle fondamental dans la compréhension de la forme de l'espace bâti. Un nombre croissant de travaux développe et implémente de nouveaux protocoles sophistiqués de géomatique pour l'identification des typologies de bâtiments et leur organisation. Cet article partage le même objectif. Une procédure innovante, basée sur l'analyse quantitative des données est présentée ici avec, comme objectif, l'identification et la description non supervisée des types de bâtiments et de leur organisation. Après une procédure de pré-traitement spécifiquement adaptée à notre donnée source, nous développons un protocole de clustering non supervisé combinant un algorithme novateur d'inférence Bayésienne Naïve avec des approches ascendantes hiérarchiques; le tout, reposant sur cinq caractéristiques morphométriques intrinsèques de chaque bâtiment. Ce protocole permet d'identifier des groupes de bâtiments partageant des caractéristiques morphologiques similaires spécifiques ainsi que leur structure globale à différents niveaux d'agrégation. La méthodologie proposée est implémentée et évaluée dans l'espace d'étude du Département des Alpes-Maritimes, France.

ABSTRACT. The identification and description of building typologies play a fundamental role in the understanding of the overall built-up form. A growing body of works is developing and implementing sophisticated, computer-aided protocols for the identification of building typologies and their organisation. This paper shares the same goal. An innovative data-driven procedure for the unsupervised identification and description of building types and organisation is here presented. After a specific pre-processing procedure, we develop an unsupervised clustering combining a new algorithm of Naive Bayes inference and hierarchical ascendant approaches relying on five morphometric features of buildings. This protocol allows us to identify groups of buildings sharing specific similar morphological characteristics and their overall structure at different aggregation levels. The proposed methodology is implemented and evaluated in the case study of the Alpes-Maritimes Department, France.

MOTS-CLES : Bâtiments, Typo-morphologie, Clustering Bayésien Naïve, CAH

KEYWORDS: Building, Typo-morphology, Bayesian Naïve Clustering, HCA

1. Introduction

An interdisciplinary and growing body of research investigates the relationship occurring between the form and the functioning of cities (Carmona, 2019). In the typo-morphological tradition, the properties of the urban form are defined by specific spatial combinations of its constitutive elements: streets, plots and buildings (Moudon, 1997). While streets and plots show a higher inertia over time, buildings change at a faster pace depending on the specific historical, socio-economic and technological context. Moreover, while some buildings (and building types) are easily torn down and replaced, others can endure for several decades or centuries showing higher resilience to urban transformation. Both geography and urban typo-morphology developed multiple approaches and protocols for the identification and analysis of building types based on the comparative analysis of buildings. As defined by Sheer (2017, p.171) "*a building type is an abstraction, a pattern, where we observe formal similarities between one building and another even though the buildings may have different architectural expressions. [] buildings might share many common formal characteristics, but are very different in color, materials, style, and expressiveness*."

Building types can result from a given time period, in a specific regional context with certain easily recognizable stylistic patterns (i.e. the English terraced house; the American bungalow house; the Parisian Haussmann apartment-building). They could also be more overarching, grouping together similar buildings produced in different time periods, but sharing consistent common features. Different information at different description levels can be used for the identification and description of building types. The combination of specific features can thus result in a comprehensive building typology, i.e. an organization of building types having a given logic and inner coherence. Three levels of details can be distinguished: i) aesthetical and stylistic features (such as façades materials and composition, colours, etc.), ii) the internal organisation of the building (including some structural considerations) iii) the overall external hull of a building (shape, footprint area, height etc.). Nonetheless, formal definitions of building types, categories and structures describing the whole building stock of a large urban region are extremely rare (Orford and Readcliffe 2007). Similarly, there is no agreement about the definition of which combination of formal characteristics is required for the identification of building types and their differentiation, since different sets of features underlie the definition of each building type. When the goal of the study is the identification of building types over a large urban region, databases encompass few features, often limited to the simple geometrical form. The study of the external building hulls allows typifying what might be called the *skeletal form* of the building.

Two main approaches are currently used for the identification of building types (Hecht et al. 2015). *Knowledge-based approaches* have been proposed based on the study of small but exhaustive datasets including building features at each of the three detail levels. These approaches are based on expert knowledge of the relevant building types and features in a given study area but are hardly adaptable to other regions, or in the presence of a poorer description of buildings. Computer-aided protocols have been

developed since the 70s, especially by the Centre for Land Use and Built Form Studies (LUBFS) at Cambridge University (Steadman 2016). Thanks to more recent data analysis developments, together with the increasing computing power availability, systematic and quantitative protocols have been developed: *Data-based approaches* have been proposed especially from digital cartography, with the goal of cartographic generalisation or for the identification of urban structures, building detection and building pattern recognition. While the study of building typologies with expert-based protocols and based on highly detailed and historical datasets finds its origin in the urban typo-morphology tradition since the 1950s, the identification of building types from footprint-based data is a more recent field of study (Hecht 2015, 2016).

Two subgroups might be further specified in data-based approaches. Supervised protocols (similarly to knowledge-based analysis) require prior knowledge of the target groups we want to identify within the dataset: features are attributed to each group based on similarity rules. Examples of classificatory approaches can be found in Orford and Radcliffe 2007, Hech et al 2015, Hartmann 2016, etc. Unsupervised approaches encompass clustering protocols where the identification of groups is based on algorithms looking for internal similarity among features and without prior knowledge of the target groups or user intervention. Clustering protocols automatically determine natural partitions (clusters) from the input arising from the specific data structure without imposing a predefined identification of the classes. Examples can be found in the works of Schirmer and Axhausen 2015, Perez et al. 2019, etc. In fact, expert-based knowledge is never completely absent. In supervised protocols, it is required at the beginning of the analysis to define the target groups, their numerosity and their overall organisation. In clustering approaches, it is needed for the interpretation of outcomes. The former allows the analyst to better identify specific predefined building typologies, while the latter allows a more exploratory analysis where natural groups emerge from the data structure and are later interpreted and related to the specific characteristics of the study region. When focusing on clustering approaches, group identification is also influenced by the underlying algorithmic rules. As discussed in Fusco and Perez (2019) most of the traditional approaches (such as K-means) impose the sphericity of clusters (i.e. intra-cluster homogeneity) on all the descriptive variables, which could not always be coherent with the complexity of the context under study. Bayesian Clustering (BC) allows us to overcome these limitations. Still, as for most of the clustering approaches, BC, even when using Naïve clustering models, imposes other kinds of constraints and can be particularly time-consuming when exploring possible solutions in parameter space. This paper presents a specific protocol aiming to identify building typologies of a large urban region and their overall organisation combining unsupervised, incremental Naïve Bayesian and Hierarchical clustering approaches. The protocol is implemented on a few selected features describing the morphological characteristics of building hulls. The paper is organised as follows. Section 2 describes the proposed clustering protocol. In Section 3, the protocol is implemented to the study area of the Alpes-Maritimes Department, France, and its results are described and interpreted. Section 4 closes the paper with a discussion on limitations of the protocol, improvements and future perspectives.

2. Method

2.1. Data preprocessing

The protocol presented in this paper is implemented on French Department of Alpes-Maritimes (Section 3.1). In France, the BD TOPO® by the National Geographic Institute (IGN) is an exhaustive dataset of metrical precision providing the information on building footprint and height, retrieved from satellite or aerial imagery. Since April 2019, a new version of this dataset has been released (BD TOPO®, V3.0): its main novelty consists in the combination of the original data with information from the national computerized cadastral plan (MAJIC). Several new features are made available such as the number of dwellings, the age of construction and the number of stories. Nonetheless, this dataset has some limitations, among which three are of concern for our research, requiring specific pre-processing protocols.

The first issue is related to the building footprint definition. Part of the building dataset, enriched by the cadastral plan information, have a detailed definition of their footprint: while one (or more) polygon(s) corresponds to the main built-up body(ies), a number of extensions (such as terraces, loggias, porches, etc.) are separately modelled as adjoining polygons. These extensions are identified by a specific feature attribute, namely "light structure". The same attribute is also associated with independent structures such as greenhouses, garages, small and large industrial shads. Another share of buildings is defined from satellite/aerial based methods: in this case, the delineation of the footprint is defined with imagery detection algorithms where the footprint corresponds to the external demarcation of the overall built-up structure, therefore including both the main building body and all extensions. Thus, a harmonization protocol has been specifically developed in order to re-aggregate buildings made by several constituent parts. The second issue is related to the definition of the functional specialization attribute. The feature "Building Nature" provides the distinction between several specialized structures (i.e. religious, industrial, agricultural, etc.). When the '*overall architecture or aspect of a building does not reveal with exactitude its function*' (ibid.) the building is classified as undifferentiated. After a manual assessment of this field on a subspace encompassing about 15 thousand buildings, we observed that the specialization attribute is always correctly assigned, although limited to 5.1% of the actual functional specialized buildings (overall accuracy 54.2%). Thus, this feature is enriched through another IGN BD TOPO layer, namely 'Activity Zones', where specialized buildings are retrieved and collected from other national authoritative sources. A set of specific rules and filters are defined and implemented to associate the specialized function to the original building dataset: on the same 15-thousand features subset, the resulting enriched definition of the field shows 86.8% accuracy score in identifying specialized buildings. This allows us to filter them and implement our clustering protocols only on ordinary residential or mixed function buildings. Indeed, as discussed in the typo-morphological literature, specialized buildings (industrial, commercial, public, etc.) have often specific and extreme morphological properties values (Caniggia et Maffei 2001) introducing important outliers in our dataset, biasing the final outcomes.

Once the subset of non-specialized buildings is redefined, morphometric descriptors can be implemented. Three indicators are directly computed from building footprint: Surface (S), Elongation (E) and Convexity (C). One indicator, Topological Contiguity (TC), is defined as the number of neighbours within a continuous built-up unit of adjoining neighbours. Finally, the building Height (H) is provided by the original dataset. These five indicators represent a set of minimal description of the building form obtained by a simplified 3D dataset (LOD0+ of the CityGML data model in Biljecki et al., 2016). Nonetheless, building Height (as for several features of the latest BD Topo V0.3) can be unknown. As we will see in the next section, the clustering protocol developed in this work is specifically conceived to deal with partial missing information. As previously introduced, BC algorithm is based on a probabilistic framework. This approach requires qualitative or categorical data. A discretization of the five morphometric descriptors is therefore necessary: for our case study the discretisation was obtained through a mix of univariate data analysis and expert knowledge. This last pre-processing step produces a first reduction of the overall complexity of the original data: in our specific case study, for example, it allows us to pass from about 300 thousand building units to 2.2 thousand building tuples, each one corresponding to a specific combination of our features modalities.

2.2. Iterative Naive Bayesian Inference Agglomerative Clustering

Bayesian inference is a powerful probabilistic option for quantitative and qualitative multivariate data clustering using simple model architectures as the Naïve Bayesian Classifier, where the cluster variable is conceived as the common parent of all the other variables, and conditional independence among them is assumed knowing the cluster variable (Duda and Hart 1973). EM algorithm is normally used to identify an optimal clustering solution in terms of log-likelihood, for a given number of clusters (Dempster et al. 1977). Exploring solutions with varying number of clusters can be done with a random walk in solution space, using a clustering score combining log-likelihood and a penalization for model complexity, i.e. number of clusters (like in Fusco and Perez 2019). McCaffrey (2013) offers an interesting alternative to the EM algorithm for BC: Iterative Naive Bayesian Inference Agglomerative Clustering (INBIAC). So far, the implementation of the INBIAC algorithm can be only found in Carneiro et al. (2015) for credit card fraud detection. INBIAC is a much faster algorithm than EM as it replaces recursive batch inference of cluster assignments for all records to an iterative assignment of individual records which are randomly extracted from the database and assigned to the highest likelihood cluster at that given moment of the clustering procedure. The higher speed of the INBIAC algorithm can be used to perform a higher number of clustering solutions.

Just like EM, INBIAC results are sensitive to the clustering initialization. In EM, initialization implies randomly assigning all records to clusters. In INBIAC, a k-cluster solution needs the use of k records as initial cluster seeds. McCaffrey (2013) proposes a preliminary phase of seed initialization, randomly choosing k seeds and finally keeping the set of seeds with maximum Hamming distance. In our algorithm we improved McCaffrey's protocol in several respects. Firstly, to better represent the

ordinal structure of our data, we used Manhattan distance instead of Hamming distance, after normalizing for the cardinality of the ordinal values. Secondly, the usual Laplacian smoothing in the initialization of conditional probability tables, is reinforced by a further smoothing on the ordinal values which are contiguous to those of the seeds. After seed selection, all remaining records are assigned one by one to the cluster having maximum log-likelihood, and the cluster conditional probability tables will be updated after each record assignment. Iterative random removal of individual records from their current clusters to be re-evaluated and reassigned to a better cluster can thus begin. Given the Naïve model architecture and the resulting additive formula of model log-likelihood, the local optimization of log-likelihood in record assignment to clusters produces a global log-likelihood improvement for the whole model. More precisely, iterative record clustering is done in two phases. In the first phase, cluster priors are considered equal for all k clusters. In a second phase, cluster priors are determined from the current cluster probability distributions, allowing a more precise calculus of posterior probabilities in Bayes' formula.

A further improvement of McCaffrey's original INBIAC algorithm has been the treatment of missing values. Under the Missing at Random assumption, likelihoods and posterior probabilities of cluster assignments are calculated only on the observed values, but missing value imputation is later performed based on the most probable values within the assigned cluster. Imputed values are iteratively erased and re-imputed within the INBIAC procedure, and the final log-likelihood of the clustering solution includes the contribution of imputed values. The clustering iterations within INBIAC stop when no record can be reassigned to a different cluster. Finally, buildings are weighted by their footprint surface in the algorithm, giving the same importance to each m^2 of built-up surface (the clustering solution would otherwise be biased by the overrepresentation of small buildings). The implementation of the protocol for several numbers of clusters k, ranging in a user-defined interval, allows to explore different clustering solutions. The optimal clustering solution(s) can be then selected through a simple graphic method: similarly to the well-known *elbow method*, the log-likelihood loss scores are plotted across different numbers of clusters k (Fig.1.a-b). The presence of strong variations might reveal structural changes in the data clustering corresponding to suboptimal solutions at a given number of classes k.

2.3. HCA

The implementation of the INBIAC protocol, allows us to identify one (or more) optimal solution(s) based on the optimization of specific scores and descriptive parameters. Nonetheless, independently from the specific set of parameters used for the model evaluation, the selected clustering(s) solution(s) would always provide a specific partition of the original dataset defined for a given number of groups k. It is thus interesting to study the variation of building clustering across k. Buildings constantly grouped in the same clusters could reveal stronger structural patterns within the data. We thus implement an agglomerative Hierarchical Clustering Analysis (HCA) using as input the INBIAC best outcomes for each k clustering solution in the interval explored. The rationale underlying this methodological choice is the

following: the subset of n best clustering solutions can be used to partition our original dataset (or similarly, the 2.2 thousand tuples) in smaller subgroups of elements (kernels) always being clustered together independently of the number of clusters k. These kernels represent the finest partition for which the highest inter-level consensus is observed: no cluster at any level further divides the element in finer groups. Within our specific context, these kernels correspond to a highly detailed meta-cluster solution of specific building sub-types; few kernels gather most of the buildings (more precisely of the built-up surface, given our weighting scheme), and vice versa a large number of kernels encompass less built-up surface with less recurrent shapes. HCA will be here implemented with Gower's dissimilarity metric for cluster distance and Ward-linkage agglomerative principle among clusters. Implementing an HCA allows us to produce hierarchically nested groupings based on the similarities within this elementary kernel partition. Kernels are arranged in a hierarchical manner. Thus, the combination of the INBIAC and HCA protocol combines the advantages of the two protocols. On the one side, we keep the ability of probabilistic Bayesian inference to select non-spherical clusters defined with a maximum log-likelihood approach on subgroups of features. On the other, an overall hierarchical structure allows the analyst to observe the overall data clustering organization similarly to knowledge- and ontological-based classifications. Moreover, the main advantage compared to regular HCA applied on raw data, is that the outcome variability produced by the high sensitivity to the initial clustering settings is strongly reduced. This approach shares the same underlying hypothesis of consensus clustering protocols (Monti et al., 2003), where several cluster solutions are combined in order to achieve a more robust solution. Nonetheless while consensus clustering looks for similarities within a larger number of clustering solutions for a given number of clusters k, in our case we use a more 'controlled' subset of n suboptimal solutions at different levels k. Moreover, instead of implementing the same clustering protocol at two different stages of the analysis, we combine a non-hierarchical and a hierarchical protocol in the first and second stage of the analysis, respectively. Finally, our protocol provides a profile for each cluster as probability distributions of its members over the values of each clustering variable. These profiles are later used for cluster interpretation (section 3.2).

3. Application

3.1 Study Area

The protocol presented in this work, is implemented on the building stock of the Alpes-Maritimes Department in southern France. This study area is made up of a large coastal conurbation and of its alpine hinterland. The coastal conurbation of the French Riviera stretches over 60 km from the French-Italian border to the Esterel mountains. Its western section includes the cities of Cannes, Grasse and Antibes, counting 74.2, 51 and 73.8 thousand inhabitants, respectively. Nice, with its 343 thousand inhabitants represents the largest municipality of the French Riviera and the administrative centre of the Department. The enclave of Monaco and the border city of Menton have respectively 38 and 28 thousand inhabitants. Spread around these main centres, 295 thousand people find their home in smaller cities, villages and hamlets. The alpine

hinterland represents over one third of the overall departmental surface but has a population of only 15 thousand inhabitants (1% of the departmental population). The building stock of the Alpes-Maritimes shows a hyper-concentrated urban fabric along the coast, a strong sprawl in the western sector, and an urban development in the eastern sector strongly influenced by its higher topographic constraints. With the exclusion of Grasse, the other five main urban centres are located along the coastline, with other 11 municipalities, gathering 42% of the overall departmental buildings (51% building surface). When including the whole of the coastal conurbation 33% of the departmental surface encompasses almost 97% of the building stock.

3.2 Implementation and results

From the original 500 thousand polygons, the data pre-processing protocol allows us to identify about 300 thousand buildings: 30% are specialized and the remaining 70% have a residential or mixed use. Our clustering protocol will be applied here only to this larger stock of ordinary buildings. The segmentation of the five features (S, H, C, E, TC) further reduces the variability of our dataset resulting in 2.2 thousand observed combinations. Then, the INBIAC protocol is implemented exploring 100 solutions with different initial seeds, for each number of clusters k between 5 and 15 producing a total of 1.1 thousand models (Fig.1.a). The quality parameters of these solutions vary with k: loglikelihood loss is maximal for the 5-cluster solution and minimal for the 15-cluster solution. Fig.1.a shows the scatter plot of the log-likelihood loss vs the category utility score (Gluck and Corter, 1985) for all the models implemented: important gains are observed for the 6-, 8- and 11-cluster solutions, further confirmed by the elbow method applied on the log-likelihood loss scores (Fig.1.b).

According to our approach, we will nevertheless use the information derived from all the 11 best clustering solutions. Their combination further defines 280 kernels, 15 of which encompass 57% of the overall built-up surface. The implementation of the agglomerative HCA on these kernels, allows us to identify and describe the overall organization of the building types in the study area (Fig.1.c-e). The dendrogram in Fig.1.c shows the succession of cluster agglomerations along a distance axis, starting from the 280 kernels (below) and arriving to the complete amalgamation of clusters (top). The length of the segment on the distance axis during which a given k-cluster solution is present is indicative of its importance in structuring the building typology in the study area. The first four solutions showing the longest segments on the distance axis of the dendrogram correspond to 4, 10, 21 and 8 clusters (Fig.1.d). The ten-cluster solution is described as follows.

A.1 corresponds to compact small detached single-family houses and duplexes while A.2 to compact small detached two and three stories houses; they account for 23.5 and 12.6 % of the overall built-up surface of the study area, respectively. These groups correspond to two of the most common building types of the French Riviera, known as the *niçois house*. These building types often do not exceed 150 m^2 and three stories; they represent a specific production of the early XX century and in the second after war. While A.1 one corresponds to a single dwelling house, A.2 might also include few dwellings initially intended for different members of the same family group. In

the same building type, we also find some more recent small real estate development projects, often used as second homes and for touristic purposes.

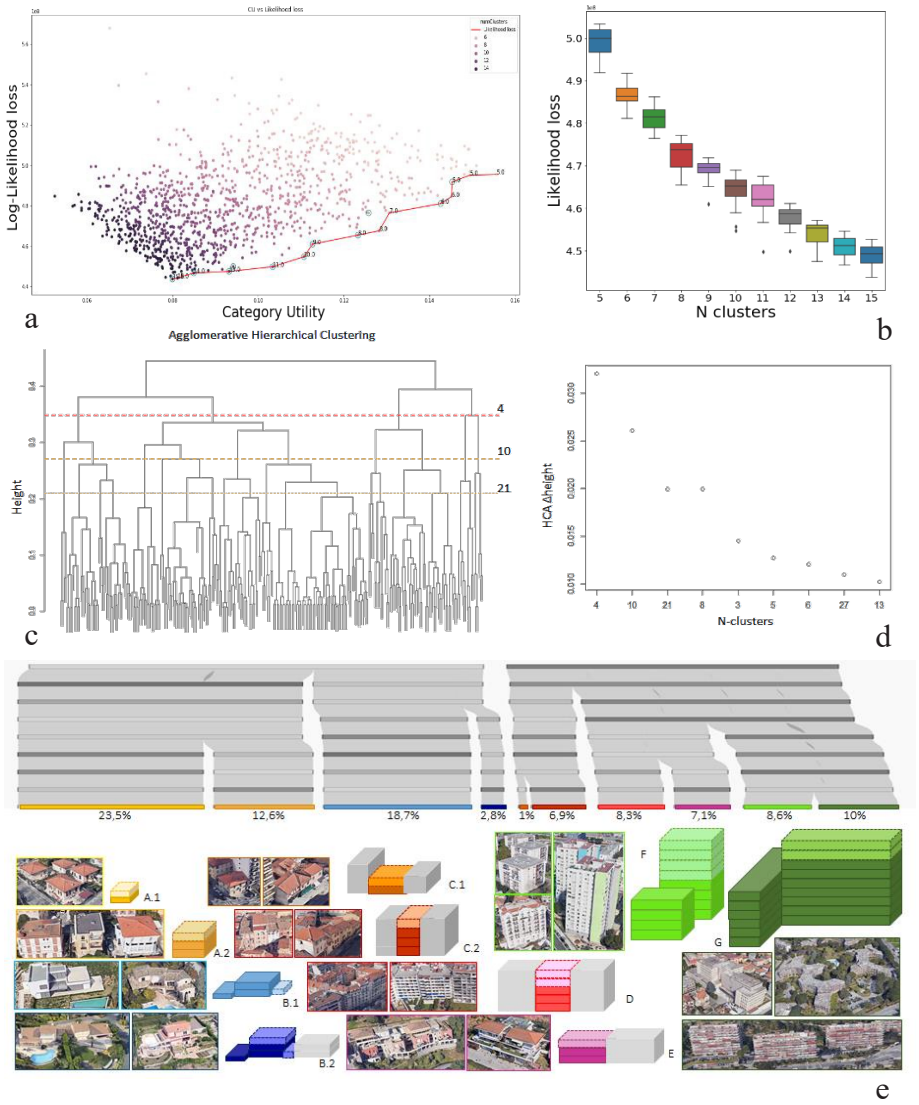

FIGURE 1. *Outcomes of the building classification: a) Scatter plot of the 1000 models scores obtained with the INBIAC clustering: log-likelihood loss vs. category utility (CU); b) Box-plots of the top 20 INBIAC models scoring the highest loglikelihood-loss scores for each number of cluster k in the interval [5:15]; c) HCA dendrogram implemented on the outcomes of the 11 best scoring INBIAC models; d) identification of the partitions corresponding to the highest dendrogram depths; e) interpretation of the 10-cluster HCA solution: the dendrogram partitions between 2 and 10 are weighted by the building surface assigned to each group.*

B.1 Identifies articulated homes and villas while B.2 corresponds to very articulated large villas (18.7 and 2.8 % of the overall built-up surface, respectively). Both groups combine large villas and (semi) detached houses with a footprint surface mainly between 150 and 300 m^2. On the one side, we find large individual compact homes and villas produced before the '50s which have been enriched by later functional additions/extensions (garages, porches, etc.). On the other side, more recent large articulated villas resulting from the suburban sprawl of end of the XX century and the beginning of 2000. What mainly differentiate the B.2 from B.1, is a more complex shape together with higher building heights (2-3 stories). High standing villas often occupy geographic locations with specific peculiar positions such as the top of hills, capes etc.; large gated communities are mainly made of these two building types.

C1. and C.2 are small-sized (<300 m^2) mid-rise adjoining buildings (1 and 6.9% of the total building surface, respectively). Most of the medieval centres of cities and villages are made up of these groups, as well as more informal settlements along connective axes. C1 differs from C2 for its lower height (1-2 stories) and its larger surface; moreover, C1 can often be found in more recent urban fabric as a residual of a previous settlement, surrounded (and often substituted) by mid-rise buildings D.

D. Adjoining mid-size and mid-rise apartment buildings represents the most common building type characterizing city centres outside their historical core (8.3% of the total building surface). Despite their different aesthetical properties (ranging from belle epoque, art deco or more modern and contemporary stylistic features), these buildings share similar footprint size (between 300 and 600 m^2), elevated contiguity and height (above 4 stories, but for long time limited at 6 stories by cityscape rules).

E. Two building types can be found within cluster E (7.1% of the building surface). On the one side we find mid-sized and mid-rise detached and semi-detached large apartment buildings. On the other we find some huge (historical) villas and manors, often occupying the most priced locations of the French Riviera (such as capes) and immersed in residential areas made by large villas of type B.

F. Free standing mid-to-high rise towers (8.6% of the total building surface). Towers are apartment buildings whose simple compact footprint is relatively small in comparison to their vertical development. More precisely, cluster F encompasses two different building types: traditional small mid-rise towers (3 to 6 storeys and building footprint between 300 and 600 m^2) and more modern, larger high-rise towers (more than 6 storey-tall but rarely skyscrapers in our study area).

G. Freestanding very large and high-rise apartment buildings, with articulated shapes, (10% of the total building surface). Buildings of G type have a footprint surface above 1200 m^2; they have very elongated and/or composite shaped buildings; they can be freestanding or belonging to large and complex urban residential development projects. Most of the housing production is located in the close peripheries; their production only started from the late 60s and it encompasses both high standing apartment buildings and large projects developed by local and national housing programs. Some of these exceptional buildings can also be found in more central and

compact contexts (i.e. demolition and reconstruction of a large articulated buildings built over an entire block).

4. Conclusion and perspective

In this paper we presented an innovative protocol for the identification and description of building types and their overall organisation combining Naïve Bayesian and Hierarchical clustering protocols. The outcomes of this systematic and quantitative analysis allow the data-driven derivation of a system of typologies of buildings, hierarchically organised. The protocol is described and implemented for the real-world contemporary case study of the Alpes-Maritimes Department. Ten building types have been identified and described.

Several research perspectives can be outlined. From the methodological point of view, sensitivity analyses should be implemented to assess the robustness of the three main steps of the protocol presented in this paper. The first step considers the role of the variable discretisation: the same protocol should be evaluated both with a general binning method and with segmentation approaches based on the specific statistical distribution of variables observed in the (sub)region under analysis. The second and third phases correspond to the two clustering protocols: both INBIAC and HCA can be assessed considering different distance measures and evaluating their validity under different parametric conditions. Independently of the specific parametric choices, a comparative analysis should also be carried out with more traditional approaches (i.e. k-means, DBSCAN) in order to further identify relative strengths and weaknesses of the protocol here proposed, both in the specific context of building type identification but also within other thematic fields.

From the thematic-related point of view four major directions for further developments can be outlined. Firstly, this same protocol can be tested with an incremental number of descriptors of the building envelope aiming at testing and identifying the role played by individual morphometrics into the building typologies. This work might contribute to the debate on the definition of a reliable and universally accepted set of characters and variables for the identification of building envelope typologies. Moreover, internal layout and details of style, facade, roof coverage might also be included: the implementation of the same clustering approach with different levels and granularity of information can shed a new light on the relative role played by skeletal, internal and stylistic features in the identification and definition of building typologies. Secondly the protocol proposed in this work should be tested on different study areas and at larger scales, to allow a better appreciation of the methodological robustness, its computational and geographical scalability; more important, the reproducibility will permit systematic, quantitative and fine grained comparative analysis on a large number of case studies (both with national-wide datasets and with volunteered geographic large dataset such as OpenStreetMap at the international level). Thirdly, the analysis of the spatial organisation of buildings types and their relative cooccurrences represent a key factor in the definition of streetscapes, urban fabrics and morphological regions, allowing to investigate the multi-scalar

nature of cities based on the finest level of the building unit. The work here described represents indeed only the first step of an undergoing larger research aiming at understanding the building types and the urban fabrics of French contemporary cities.

Acknowledgment: This research was funded by the IDEX UCA JEDI, within the AAP Partenariat 2019 (action 6.5).

References

Caniggia G, Maffei GL (1979). *Lettura dell'edilizia di base*. Firenze: Alinea.

Carmona M, (2019) Place value: place quality and its impact on health, social, economic and environmental outcomes, *Journal of Urban Design*, 24:1, 1-48.

Carneiro EM, et al. (2015) Cluster analysis and artificial neural networks: A case study in credit card fraud detection. 2015 *12th Int. Conference on Information Technology*. IEEE, 2015.

Dempster AP, Laird NM, Rubin DB (1977). Maximum Likelihood from Incomplete Data via the EM Algorithm. *Journal of the Royal Statistical Society*, Series B. 39 (1): 1–38

Duda RO, Hart PE (1973) *Pattern classification and scene analysis*. NY: John Wiley & Sons.

Fusco G, Perez J (2019) Bayesian Clustering and SOM Neural Networks under the Test of Indian Districts. A comparison. Cybergeo, n°887, .https://journals.openedition.org/cybergeo/31909

Gluck MA, Corter JE (1985), Information, uncertainty, and the utility of categories, *Program of the Seventh Annual Conference of the Cognitive Science Society*, pp. 283–287

Hartmann A, Meinel G, Hecht R, et Behnisch M (2016). A workflow for automatic quantification of structure and dynamic of the German building stock using official spatial data. *ISPRS International Journal of Geo-Information*, 5(8), 142.

Hecht R., Meinel G. & Buchroithner M. (2015). Automatic identification of building types based on topographic databases – a comparison of different data sources, *International Journal of Cartography*, 1:1, 18-31.

IGN (2020) BD TOPO®, Version 3.0. Descriptif de contenu. https://geoservices.ign.fr/ressources_documentaires/Espace_documentaire/BASES_VECTORIELLES/BDTOPO/DC_BDTOPO_3-0.pdf

McCaffrey J. (2013) *Data Clustering Using Naive Bayes Inference*. http://msdn.microsoft.com/en-us/magazine/jj991980.aspx.

Monti S, Tamayo P, Mesirov J et Golub T, (2003). Consensus clustering: a resampling-based method for class discovery and visualization of gene expression microarray data. *Machine learning*, 52(1-2), pp.91-118.

Moudon, A. V. (1997). Urban morphology as an emerging interdisciplinary field. *Urban morphology*, 1(1), 3-10.

Orford, S., Radcliffe, J. (2007). Modelling UK residential dwelling types using OS Mastermap data: A comparison to the 2001 census. *Computers, Env. & Urban Systems*, 31(2), 206–227.

Perez J., Fusco, G., Araldi, A., & Fuse, T. (2020). Identifying building typologies and their spatial patterns in the metropolitan areas of Marseille and Osaka. *Asia-Pacific Journal of Regional Science*, 4(1), 193-217.

Scheer, B. C. (2018). Urban Morphology as a Research Method. *Planning Research and Knowledge.* Abingdon, UK: Routledge, 167-181.

Steadman P (2016). Research in architecture and urban studies at Cambridge in the 1960s and 1970s: what really happened. *The Journal of Architecture* 21, no. 2 (2016): 291-306