Classification and clustering of buildings for understanding urban dynamics

Joan Perez, Giovanni Fusco, Yukio Sadahiro

Classification and clustering of buildings for understanding urban dynamics

A framework for processing spatiotemporal data

Perez Joan[1], **Fusco Giovanni**[1], **Sadahiro Yukio**[2]

1. Université Côte d'Azur, CNRS, ESPACE, Nice, France

2. Department of Urban Engineering, University of Tokyo, Tokyo, Japan

ABSTRACT. This paper presents different methods implemented with the aim of studying urban dynamics at the building level. Building types are identified within a comprehensive vector-based building inventory, spanning over at least two time points. First, basic morphometric indicators are computed for each building: area, floor-area, number of neighbors, elongation, and convexity. Based on the availability of expert knowledge, different types of classification and clustering are performed: supervised tree-like classificatory model, expert-constrained k-means and combined SOM-HCA. A grid is superimposed on the test region of Osaka (Japan) and the number of building types per cell and for each period is computed, as well as the differences between each period. Mappings are then performed, showing that building types have specific locations and dynamics. In some extreme cases, a specific building type can even gradually replace a type on a declining dynamic. Questions of data preparation, and clustering validation are also dealt with, underlining the interest of assessing the spatial distribution of clusters.

RÉSUMÉ. Cet article présente différentes méthodes mises en œuvre dans le but d'étudier la dynamique urbaine au niveau des bâtiments. Les types de bâtiments sont identifiés dans le cadre d'un inventaire vectoriel complet des bâtiments, couvrant au moins deux périodes. Tout d'abord, des indicateurs morphométriques de base sont calculés pour chaque bâtiment : surface, surface au sol, nombre de voisins, allongement et convexité. En fonction des connaissances d'experts disponibles, différents types de classification et de regroupement sont effectués : modèle classificatoire supervisé de type arborescent, k-means sous contrainte d'experts et SOM-HCA combiné. Une grille est superposée à la région test d'Osaka (Japon) et le nombre de types de bâtiments par cellule et pour chaque période est calculé, ainsi que les différences entre chaque période. Des correspondances sont ensuite établies, montrant que les types de bâtiments ont des emplacements et des dynamiques spécifiques. Dans certains cas extrêmes, un type de bâtiment spécifique peut même remplacer progressivement un type dont la dynamique est en déclin. Les questions de la préparation des données et de la validation des regroupements sont également abordées, soulignant l'intérêt d'évaluer la distribution spatiale des regroupements.

KEYWORDS: Spatial analysis, spatiotemporal data, classification, clustering, self-organizing map.

MOTS-CLÉS : Analyse spatiale, données spatio-temporelles, classification, regroupement, carte auto-organisatrice.

1. Introduction

Cities are in a perpetual flow of transformations with urban areas that constantly make, unmake and remake themselves. Different models are discussed in the literature, mostly focusing on the benefits and/or drawbacks of the increase (urban sprawl, compact cities, regeneration, etc.) or decrease (shrinking cities, perforation, etc.) of building densities. If it is acknowledged that all locations do not enjoy equal potential to adapt and evolve to changing densities, a factor that remains overlooked regarding this "potential" is the role of inherited urban forms. Yet, occurrences of areas having difficulties in redefining themselves are often closely related in terms of urban form. Of course, even if the form factor is often neglected, as compared to, for example, economic and social factors, this issue is nothing new. The numerous criticisms of modernist architecture (Jacobs, 1961; Salingaros, 2005, 2006; Paquot, 2019), which brought large-sized and specialized buildings, unable to evolve with the needs of the cities, are a good illustration of the link between urban life and urban form. What has changed more recently however, are the possibilities to explore this link through: (1) the ever-increasing availability of multi date large-sized datasets related to building stocks, such as the Japanese Zenrin Maps ® or the French BD TOPO®, and (2) an increase of computing power that allows running complex algorithms on the former. New analyses in the literature include the classification of building footprints by a Random Forest Classifier (Hecht *et al.*, 2015) or unsupervised Bayesian clustering for building types identification (Perez *et al.*, 2019a). This paper seeks to emphasize on different needs and precautions that shall be taken while working on the link between urban form and space-time evolutions of urban areas and, more specifically, on evolutions of building types. Limited, but yet interesting characteristics of buildings can be calculated using only building databases, whenever a vector description of footprints and heights are available. Based on the availability of expert knowledge, different types of classifications and clustering are performed to obtain building types consistent between different time periods. The first classification algorithm is a tree-like classificatory model that contains a series of conditional control statements implemented by the expert. The second model is a constrained k-means clustering, where the expert knowledge is used to split the inputs prior to the clustering. The last model is a self-organizing map clustering allowing classifying the inputs into different building family types without any expert knowledge and in an unsupervised fashion. Suggestions and recommendations regarding spatiotemporal data used as inputs within machine learning algorithms are provided throughout the paper. In addition, implementation of the procedures, choice of the number of clusters, optimization and validation are also discussed in detail. From a thematic point of view, specific periods of constructions and architectural styles can sometimes match with identified building types, thus allowing studying the spatiotemporal evolutions of specific types in the urban landscape. The evolution patterns of building types reflect changing demographic, social and functional factors behind city-making. They can be quantified and used as a reliable source of information for understanding the recent urban history, but also for urban planning and policy-making institutions and authorities.

The paper is organized as follows. Section 2 provides a literature review on the evolution of building types in contemporary cities, and on the methodologies that focused on the spatial evolution of the urban landscape. Section 3 presents the data requirement to apply the different methods discussed in this paper. It also presents the test region and the computation of basic morphometric indicators for building footprint data, which are going to be used to perform the classification and clustering methods. Section 4 applies and details the algorithm of the three different models, *i.e.* atree-like classificatory model, a double k-means and a self-organizing map. Section 5 maps and analyzes the geographical results, with the help of a grid superimposed over the test region that counts the number of each building type per cell and their temporal evolution. A final algorithm that produces thematic maps is also detailed in this section. Section 6 concludes the paper.

2. Spatial analysis and urban form transformations

As an increasing percentage of the world population comes to live in urban areas, the massive urbanization of human societies appears as both a source of problems and a possible solution in building sustainable urban futures (Coutard *et al.*, 2013). The key concepts in play regarding urban sustainability are expanding to more global issues, from environmental challenges to human-centered approaches in planning and architecture. The academic literature is indeed highlighting on a regular basis that cities, urbanization, planning models, etc. have to be more sustainable and more human-centered. The urban of tomorrow should be, according to Sennet (2018), porous and irregular, with a system yet to be designed of shell and type-form where "*The shell is empty; the type-form is, as it were, the snail inside. There is a content within which both limits and encourages change*" (Sennett, 2018, p. 92). The fight against low-density automobile-dependent residential housing (Newman and Kenworthy, 1999) as well as the increase of the number of cities sustaining processes of urban decline (*e.g.* shrinking cities) are no strangers to the fact that planners, architects, and city thinkers in general are seeking out new sustainable solutions from and within pre-existing urban landscapes and configurations rather than developing new models from the ground. Yet, the fact that cities are in a perpetual flow of transformations is nothing new, it is even a major characteristic of urban spaces according to several authors. Lynch stated in 1981 (p.116) that a good city form is one in which "*a complex ecology is maintained while progressive change is permitted*", while Rossi (1982, p. 55) argues that "*the city is something that persists through its transformations*". To summarize, transformation is an intrinsic characteristic of what is urban and, at a global scale, the world is increasingly urban. Yet, the transformation potential of urban areas is not the same everywhere: pre-existing forms matter. Planning models and architectural concepts are for example often related to specific construction periods and needs. The most striking example of this phenomenon is undoubtedly the modernism period, which brought rigid geometries and layouts often criticized for not being able to evolve in concert with contemporary urban issues (*e.g.* New Urbanism model). It thus brings interesting questions, such as what kinds of forms are more likely to be transformed and adapted, namely through building

substitutions, as opposed to more static ones? The underlying hypothesis is that physical forms (single-family homes or collective, modernist or traditional architecture, small or large-size) impose constraints on the socioeconomic process of urban transformation. As geographers, we also know that space matters: where are those forms located within a city? Where are theydeveloping and/or disappearing? Urban morphology, which focuses on the study of urban forms and transformations, associated with quantitative approaches in spatial analysis are perfectly suited to shed a new light on these questions.

In the field of urban morphology, countless studies about the physical transformation of the cities have been conducted. Those studies follow several schools of thoughts, such as the one of M.R.G. Conzen, or Saverio Muratori and Gianfranco Caniggia, to name but just a few, and mainly focus on urban landscape, fabrics and typo-morphological changes at a micro level. However, the quantitative revolution that stepped in social sciences, notably in geography during the 1970s (Chorley and Haggett, 1967), did not reach the field of urban morphology until around the 2000s. The increase of the number of quantitative studies within the annual proceedings of annual International Seminar on Urban Form (ISUF, 2021) reflects this trend, while the recent book chapter by Larkham (2019), entitled *Extending Urban Morphology: Drawing Together Quantitative and Qualitative Approaches* gives a comprehensive overview of the evolution of quantitative studies within this field. Quantitative analytical protocols are today rapidly developing bringing together the fields of urban morphology and of spatial analysis within the encompassing approach of urban morphometrics. Beyond Larkham's chapter, a more thorough presentation of some of these protocols is offered by the whole book edited by D'Acci (2019) as well as by the special issue on urban big data analytics and morphology in *EPB – Urban Analytics and City Science* (Behnisch *et al.*, 2019). New methods are for examples proposed for the analysis of the urban fabric, traditionally defined by urban morphology as the combination of built-up forms, plot patterns and street networks (for examples Oliveira and Medeiros, 2016; Araldi and Fusco, 2019; Berghauser Pont *et al.*, 2019; Fleischman *et al*., 2021).

In this respect, the quantitative analyses carried out within this paper have a more limited focus. They will concentrate on the morphometrics of building types and on the spatial and temporal analysis of their distribution within a large metropolitan area. Related works are here Hartmann *et al.* (2019) on the evolution of German building stock or Kollwitz *et al.* (2022) on the evolution of building types in Vantaa (Finland) from an urban metabolism point of view. As defined by Case-Scheer (2015, 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"*. However, as the Italian school of typo-morphology shows, building types are a fundamental component of the urban fabric. All schools of urban morphology also agree on the fact that different types of the urban fabric define morphological regions, which have specific spatial structures (center-to-periphery gradients, sectors, etc.). We can thus assume that the building types we are going to identify will necessarily have a spatial structure and could also present different temporal dynamics.

In order to focus on the quantitative and spatial dynamics of building types, one needs spatiotemporal data which can be found within exhaustive GIS building inventories, such as the Japanese Zenrin Maps® or the French BD TOPO®. On the one hand, major transformations in urban contexts are usually observed along several decades, or even centuries, while accurate GIS building inventories have at most a multi-temporality of a few decades. For example, the Version 1.0 of the French BD TOPO ® only dates back to 1994, and to 1997 for the Japanese Zenrin Maps®. On the other hand, the lifespan of a given building depends on several factors, such as the function for which it was constructed, while the average lifespan for all types also reflects cultural preferences regarding how the urban landscape is conceived and maintained. In certain cultures, such as in western and northern Europe, buildings are renovated, while in others, such as in Japan, the privileged model is the destruction-reconstruction of new buildings. For example, single-family houses have a lifespan of only 30 years in Japan (MLIT, 2007), which is in sharp contrast with most studies that assume a building lifespan within a range of 35 to 120 years in Europe, North America and Australia (e.g. Islam *et al.*, 2015; Marsh, 2016). For both these reasons, quantitative studies in urban morphology that make use of GIS building inventories are mostly focusing on identifying patterns at a given point in time (Hecht *et al.*, 2015; Perez *et al.*, 2019a; Araldi *et al.*, 2023), the exception being researches making use of buildings or lots to identify growth or decline patterns (Lee and Newman, 2017; Sakamoto *et al.*, 2017; Usui and Perez, 2019; Perez *et al.*, 2020; Kollwitz *et al.*, 2022). Algorithmic techniques, coupled with spatiotemporal data and thematic knowledge related to urban form, have the capability to support urban planning. This kind of approach will be increasingly relevant in the future as accurate multi-temporal GIS building inventories become available. The following section presents a spatiotemporal dataset, a test region, and prerequisites and recommendations while working on building evolutions with such data.

3. Data preparation

3.1. Minimum data requirement

The bare minimum dataset required to implement different algorithmic techniques to work on building type evolutions is a GIS layer of building footprints with multiple years of consistency (at least two different periods). The gap in years between the layers is dependent upon the scope of the analysis, but also upon the specificities of the location under study. A gap of 10 years is enough in countries following a deconstruction/reconstruction model of buildings (*e.g.* discussion in section 2) while in countries where buildings are usually renovated or rehabilitated, a longer time depth may be required to observe structural changes in the distribution of building types. Each building shall be digitized as a single unit. Corrections should thus be made to GIS layers which aggregate contiguous buildings (like row houses) into single units. In addition, one attribute is required: building height, and another one is valuable, but not mandatory: building specialization. Specialization attribute allows filtering non-residential buildings, if one were to solely focus on

residential patterns. Yet, due to mixed-use buildings, the distinction between residential and non-residential buildings is hardly ever straightforward. Thus, according to the source of data, attention should be paid to how specialization is encoded.

3.2. Test region and data presentation

The test region is an area of 15 by 18 km in Japan containing central Osaka and its surroundings (Figure 1a). As stated in section 2, the peculiarities of Japan are that houses and small collective residential complexes are easily demolished to be reconstructed or to make way to new urban projects (Shelton, 2012) and that urban areas are experiencing urban shrinkage phenomena (Fujii, 2008; Buhnik, 2010). A short lifespan for buildings is an undeniable advantage for the study of the evolution of building types. Indeed, even now, high-quality datasets of building footprints are uneasy to access and, if they are, historic data of equivalent quality are usually not. Regarding urban shrinkage, demolitions without reconstructions is another interesting phenomenon to monitor over time, since the non-replacement of building could indicate perforation dynamics.

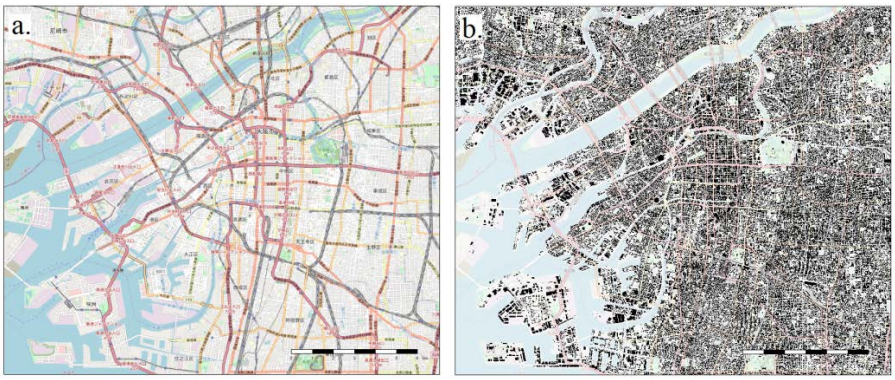

Figure 1. a. Osaka, openstreetmap. b) buildings in Osaka in 2013-14 (zmap town ii)

ZENRIN Residential Maps (Zmap TOWN II[1]), which are digital maps focusing on the building footprints throughout Japan, are extracted and compiled into a GeoPackage file for the extent of the test region in 2003/04 and in 2013/14 (Figure 1b). The original GeoPackage file only contains the two aforementioned building layers. There are 760.067 inputs in 2003/04 and 739.536 in 2013/14, thus showing a decrease of 2.7% of the raw number of buildings in 10 years. ZENRIN

1 ZENRIN is a private map information company that holds the top share in the Japanese market for local residential and car navigation maps.
https://www.zenrin.co.jp/product/category/gis/basemap/zmaptown/index.html

Residential Maps possess several attribute data from which two are of interest in this research: the height of the buildings, expressed in number of floors, and building specialization. However, no distinction is made between one- and two-story buildings(the count starts from three-story buildings). Building specialization is an attribute, encoded as follows: 1363 for collective housing, 1364 for single-family homes, 1365 for private offices and mixed-use buildings, 1200 for official and religious buildings (schools, administrative buildings, temple, etc.), etc. In this paper, we focus on residential buildings, both collective and individual housings (1363 and 1364), which concern 469.519 buildings in 2003/04 and 428.875 in 2013/14. Each layer within the GeoPackage file thus only retain the following attribute variables: an identifier, the number of floors and specialization.

3.3. Indicator computation

The first step consists in computing a basic series of morphometric indicators for each building made of: the building footprint surface (area), the total amount of usable floor area, elongation, convexityand the number of adjoining neighbors. Elongation, convexity and the number of adjoining neighbors are detailed in Perez *et al.* (2019a). They respectively provide a measure of how buildings are elongated compared to the most compact equivalent shape (a circle), how buildings have intricated or squared shapes, and if buildings are free-standing detached structures or possess adjoining neighbors. Finally, in order to smooth the lack of difference between one- and two-story buildings, floor is removed in favor of floor-area, which simply is the surface area multiplied by the number of floors for each building. The final set of variables is made of the newly calculated morphometric indicators plus specialization.

Appendices algorithm IC (Indicators Computation) shows how indicators can be easily computed within R for a standard GIS building inventory. For the number of adjoining neighbors, we operate a small buffer (**algorithm IC**,# neighbors 1/2) with the aim of correcting buildings that shall be considered as adjoining, but are not due to low geolocation accuracy. Even if such occurrences are negligible within the ZENRIN Residential Maps, this precautionary step ought to be taken before computing the number of adjoining neighbors.

4. Residential buildings: classification of types

Once a basic set of indicators has been calculated for each input layer, the next step is to perform a classification of building types that can be applied to the different time points. To be more specific, with such a basic set of morphometric indicators, we are rather classifying types of building hulls, *i.e.* geometrical envelopes of buildings, since an accurate classification of architectural types should include data about building materials, periods of construction, internal distribution of housing units and rooms, etc.For the obtention of class labels, three options arise. First, reliable expert knowledge is available, and as such, the number and the characteristics of the relevant building types are known in advance. Second, limited

expert knowledge is available, which leads to partial information known in advance, such as the number of types sought, but not their characteristics. Lastly, no expert knowledge is available, thus leading to a situation in which both the number and the content of the building types are unknown.

4.1. Classification with expert knowledge

When expert knowledge is available, a simple and efficient method of classification is found in a supervised tree-like classificatory model. Successive tests on attributes split the original dataset into different class labels that have been defined by the expert. When the corresponding label outputs are deemed valid by the expert, the conditional control statements can be used as benchmarks for other datasets possessing similar attributes. **Appendices algorithm CEK** (Classification with Expert Knowledge) is a user defined function named *DTs* which allows classifying each row of a dataset using given variables (var1, var2 and var3). Each test follows a*if condition* then *outcome* structure. Several conditions per test are possible, and the outcome is always a label attribution. The DTs function is then applied to each row of a given dataset. In the case of the test data, the DTs function is applied on each row of both Zenrin® layers using the aforementioned morphometric indicators.

Figure 2 is a flowchart visualization of the successive tests implemented for the Zenrin® layers. The relevance of these building types in Japanese urban areas is derived from expert knowledge and literature (*e.g.* Shelton, 2012; Bonnin *et al.*, 2014; Perez *et al.*, 2019b). They are 9 different outcomes, in which 4 are for collective housings (C1 to C4) and 5 are for single-family homes (S1 to S5). C1 are high rise residential buildings following a “tower in the park” model. They have been identified by their lack of neighbors, a large floor-area surface and a limited elongation ratio. C2 are elongated residential complexes representative, amongst other structures, of *Danchi* housing (団地) which are complexes of apartment buildings built following western standards after the Second World War, C3 are adjoining narrow towers, while C4 are small-size residential complexes which often host the famous Japanese micro-apartments. S1 are very small and elongated row houses, with thresholds that have been set to precisely extract traditional wooden *Nagaya* (長屋), and their modern counterparts. S2 are small-size adjoining intricated houses, S3 are the compact counterpart of S2, which, as compared to S2, also include large-size compact townhouses. S4 are large-size detached houses, typical, amongst other things, of large-size villas and traditional Japanese wooden houses, and finally, S5 are small-size detached houses. The thresholds are applied for each period. Detailed thematic results based on a similar tree-like classificatory model are cross-analyzed with population evolution, academic literature, building bye-laws, planning regulations and fieldwork related to the metropolitan area of Osaka in Perez *et al.* (2023).

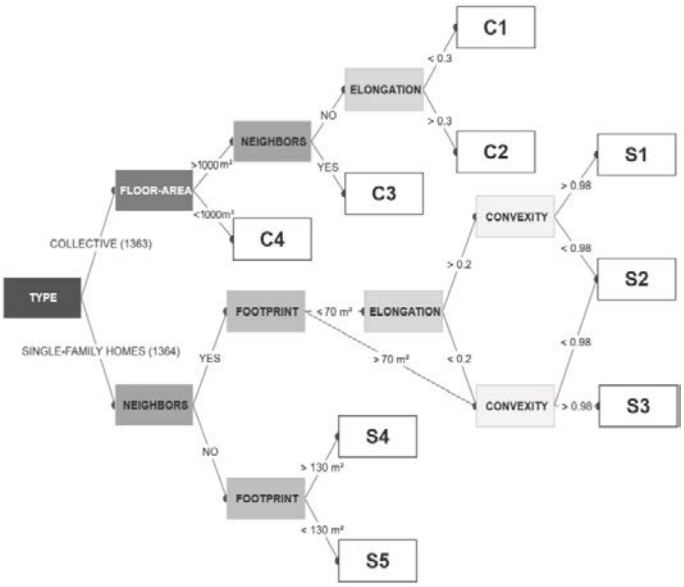

Figure 2. Visualization of the tests implemented for the Zenrin ® layers

4.2. Classification with limited expert knowledge

If expert knowledge is limited, but still available to some extent, it is possible to subset the data in a supervised fashion prior to a series of unsupervised cluster analyses. It amounts to performing a form of constraint-based clustering. For example, we could know that some inputs must not be grouped together. This could for instance be the case for collective and single-family homes, for specialized and residential buildings, or for building possessing large and small footprints. In such cases, standard unsupervised clustering techniques that minimize intra-cluster distances and maximize inter-cluster distances can be independently performed for each subset, such as K-Nearest Neighbors, hierarchical clustering, *k*-means, classification trees, random forests, neural networks, etc. Regarding the multi-temporality of the data, three possibilities appear: (1) clustering are performed independently for each time point and compared between one another using, for example, similarity indexes, (2) the underlying model of one time point is used for the training of the others time points, (3) a test sample is drawn and merged altogether from different time points, thus avoiding giving more weight to one period over the others. Finally, expert knowledge can be used to fix the number of required clusters (otherwise see Section 4.3). To illustrate the aforementioned discussion, one of the simplest techniques, *k*-means clustering, which aims at partitioning inputs into clusters such that the sum of squares from inputs to the assigned cluster centers is minimized, is applied in **Appendices algorithm CLEK** (Clustering with Limited Expert Knowledge). This algorithm performs a k-means

with a predefined number of clusters using a sample made of two random subsets from two time points. The trainings of the initial datasets are subsequently made using a k-nearest neighbor searching algorithm.

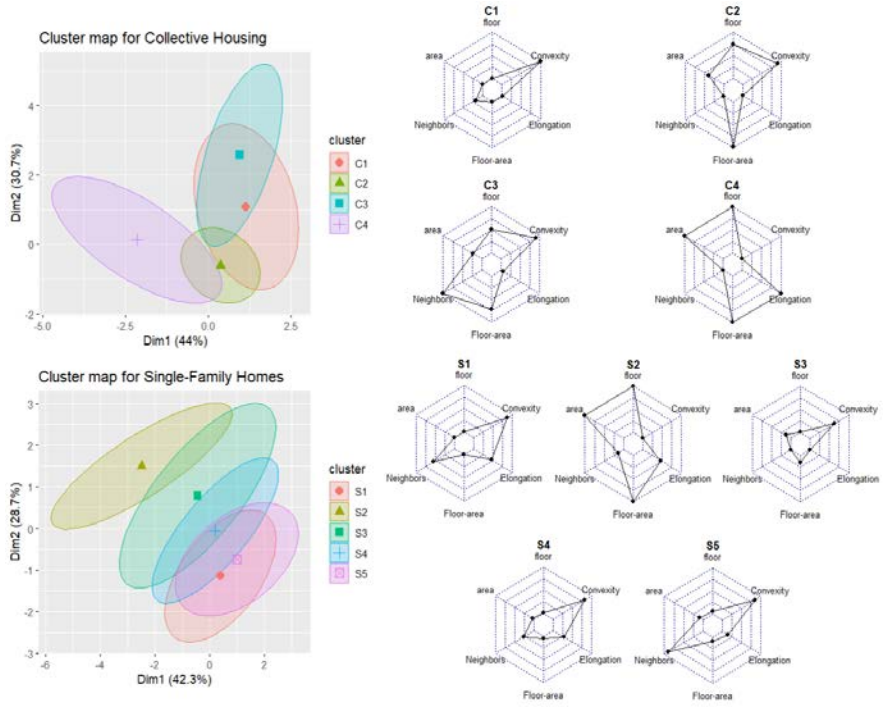

Figure 3. Cluster Map and Profiles for individual Clusters for Collective Housing (top) and Single-Family Homes (bottom)

Using the structure of **algorithm CLEK**, two different k-means are performed on the Zenrin® data, one for collective (attribute 1663) and one for single-family homes (attribute 1664). Two random subsets are draw and merged from each period for collective buildings only (sample size of 60% of the inputs), and, once again, from each period for single-family homes only (60% again). The k-means clustering are then performed for each subset (collective housing and single-family homes), with a number of clusters per model chosen *a priori*. We selected the same number of clusters as within the DTs model (section 4.1: 4 clusters for collective housing and 5 clusters for single-family homes). Before running the algorithm, data are scaled, normalized and symmetrized. The two clustering are performed on convexity, elongation, number of direct neighbors and floor-area. Once the two models are trained, for both periods, collective inputs that have not been trained are mapped to the nearest clusters of the collective model and single-family homes inputs to the nearest clusters of the single-family homes model.

The left side of Figure 3 is a possible representation of the two clustering within two-dimensional spaces made by extracting the first two principal components (package “factoextra”) along which the variation in the data is maximal. The right side of Figure 3 shows radar charts for each variable (package “fmsb”). In order to obtain striking radar charts, minimum and maximum values of the radar chart axes must be calculated independently for each model (collective housing and single-family homes). It is possible to add variables in the radar charts that have not been used during the clustering processes, such as area and floor in Figure 3.

The charts point at highly differentiated profiles within each model. Regarding collective housing, C1 are compact small-size complexes, including the Japanese micro-apartments, C2 are mid-to-high-rise compact complexes, C3 are high-to-mid-rise adjoining residential buildings, while C4 are regrouping both large-size and intricate buildings, typical of contemporary collective housing models, and Danchi housings. Regarding single-family homes, S1 are adjoining elongated houses, typical of traditional wooden *Nagaya*, S2 are large-size modern villas, S3 are intricated detached houses, S4 is close to S1, although less convex and elongated, and finally, S5 is characterized by adjoining compact houses, typical of modern townhouses. Thematically, this clustering seems robust, but it is always possible to improve the results. For example, S1 and S4, although different, share several similarities.

4.3. Classification with no expert knowledge

Finally, in the case of a total lack of expert knowledge, any widely known unsupervised clustering algorithms can be used, such as those mentioned in section 4.2. The main issues with unsupervised clustering are the assessment of the quality of the partitions, as well as the number of “interesting” partitions (Haldiki *et al.*, 2001). Since there is no way, in the absence of expert knowledge, to *a priori* determine the most suitable number of clusters, researchers tend to use heuristics, such as looking at a cutoff point regarding the improvement of the explained variance (or log-likelihood in the case of Bayesian clustering) for each additional partition (elbow method), looking at the similarity coefficients between inputs and their own clusters (silhouette), comparing the within-cluster dispersion with a null hypothesis (gap statistic), etc. However, these methods often yield different results. To cope with this issue, it is possible to look at the consensus of different methods regarding the most suitable number of clusters for partitional clustering methods (Charrad *et al.*, 2014).

Figure 4 displays the consensus results (*n_clusters* function from the *parameters* package, Lüdecke *et al.*, 2020) of a random sample made of 10% of the inputs from both periods of the Zenrin® data. Since 30 indexes are compared (Elbow, Silhouette index, Duda index, Scott, etc.; Charrad *et al.*, 2014), a small sample is mandatory to avoid too much computing time. Figure 4 hints at a consensus for 2, 3, 9 and 12 cluster-solutions. An unsupervised Bayesian clustering on the Zenrin® building layer of 2013-14 only, based on the indicators detailed in section 3.3 (plus specialization) has been performed in Perez *et al.* (2019a). This clustering used an

expectation-maximization algorithm (Dempster *et al.*, 1977) to perform one thousand clustering analyses and a MDL score (combining log-likelihood and a penalization function for the growing number of clusters) to define the optimal number of partitions. In what follows, we will propose an alternative approach to unsupervised clustering based on the Self-Organizing Maps (SOM) (Kohonen, 1982) neural network. Bayesian and SOM clustering have already been compared in unsupervised approaches (Fusco and Perez, 2019).

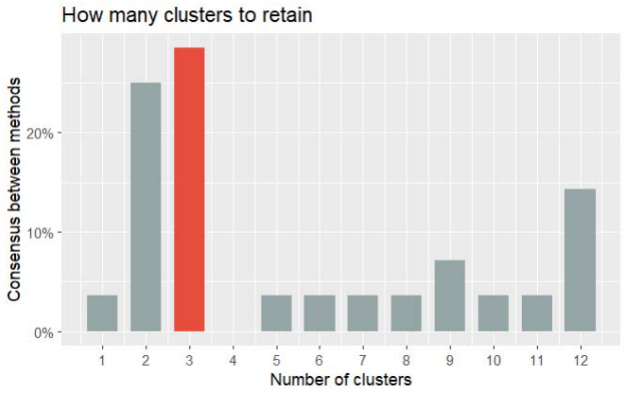

Figure 4. Consensus between 30 indexes looking at the Optimal Number of Clusters for a sample of the Zenrin ® data

Appendices algorithm CNEK (Clustering with No Expert Knowledge) performs SOM clustering using, once again, a training sample made of two random subsets from two time points each possessing 60% of the initial inputs. This time, we do not distinguish collective from single-family homes buildings. The same four indicators than for the k-means clustering are used. SOM use a neighborhood function for each input to find the Best Matching unit (competitive training). Units, also called nodes, are distributed on a two-dimensional space (map). Once the best matching unit is found, a radius parameter allows updating the neighboring nodes, thus giving topological properties to the output space. This output space, and the proximity of the nodes, can be visualized and investigated through a 2D representation of the SOM map. There are two ways to identify clusters within a SOM. First, the number of nodes within the two-dimensional space can be set to a small value. Each node is then considered as a cluster on its own. The second approach, which is the one used in **algorithm CNEK,** is to parameterize a large number of nodes for the map (225, *i.e.* 15 by 15 in our application) before segmenting the Euclidian distance matrix between all the couples of nodes with a simple hierarchical clustering (HCA). The distances between couples of nodes form a distance matrix (U-Matrix). The advantage of this approach is that it allows visualizing each variable distribution across the map. Depicted by colors, these graphical representations are called heatmaps. Since similar values are aggregated in the same areas, heatmaps provide relevant information that can be used prior to the

segmentation by the HCA. **Algorithm CNEK** also provides the mapping to the test data, and extracts the cluster correspondences between the SOM nodes and the final clusters segmented by the HCA.

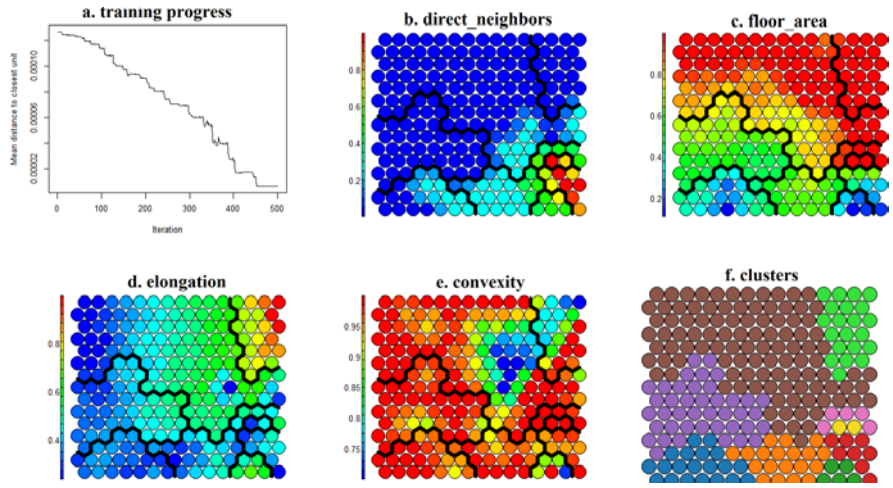

Figure 5. Self-Organizing Map outputs using a sample from both periods of the Zenrin ® dataa. Training Progress b. Heatmap for direct neighbors c. Heatmap for floor-area d. Heatmap for Elongation e. Heatmap for Convexity f. Clustering results

Figure 5 displays the outputs of a SOM clustering performed on a Zenrin® sample made of 60% of the initial inputs merged together from each period. As for the k-means clustering, data have been scaled, normalized and symmetrized. The first thing to control is that the training curve flattens, as in Figure 5a. The *rlen* parameter in **algorithm CNEK** sets the number of iterations. Figure 5b to 5e are the different heatmaps**,** showing that similar values are aggregating in the same areas. Each map shows clear patterns of minimum and maximum optima, thus clearly demonstratingthat greater weights have not been given to some variables over others. If some heatmaps do not show a clear gradient, *e.g.* if the differences in the range of valuesarelocated only in a reduced number of nodes, it is usually due to scaling/symmetrizing issues between the input variables. Local minimum and maximum optima of each heatmap are not located within the same nodes, thus demonstrating that a large number of clusters could be a relevant solution. A fully unsupervised clustering would have taken 3 as the final solution for the number of clusters (Figure 4). Yet, in order to compare the output spaces of the three methods performed in this paper, we set the number of clusters to nine, which was also a consensus possibility hinted by Figure 4. Figure 5f is the final cluster map, where each color is a different building type. The black lines are the cluster boundaries, which have been retroactively mapped to the heatmaps after running the HCA. Cluster 1 (“blue” in Figure 5f) is mostly made of different kinds of detached small

buildings, cluster 2 (“orange”) regroups adjoining compact buildings, cluster 3 (“green”) is made of small, elongated and adjoining buildings, which appears to be the characteristics of the wooden *Nagaya*, cluster 4 (“red”) is for small and compact buildings, cluster 5 (“purple”) displays the characteristics of mid-size detached villas, cluster 6 (“brown”) is an average cluster for mid- to large-size buildings, cluster 7 (“pink”) displays characteristics of large-size buildings, somewhat elongated and with neighbors, cluster 8 (“yellow”) is probably for high and narrow buildings (high values of floor-area, compacts and with neighbors), and finally, cluster 8 (“light green”) are detached large-size buildings, and could correspond, amongst other things, to *Danchi* (団地) housing. Test data, for both periods, are then mapped to the trained SOM.

5. Validation and spatial analysis

5.1. Validation

In supervised classification applications, such as for the tree-like classificatory model performed in section 4.1, quality metrics are of little use regarding the evaluation of the class label attributions. For classification, a good way to check the validity of the building types is to draw random samples and use a street-based urban imagery (such as Google Street View) to verify that the objects are classified as expected by the expert. We consider as true positives (TP) the buildings assessed by the expert as belonging to the class label proposed by the algorithm. Remaining buildings are considered as false positives (FP). It is then possible to calculate the usual classificationprecision metrics as (TP / (TP + FP)) * 100. Using a random sample of 100 buildings for each class label, the overall precision reaches 92.77%. The lowest score is of 85.56% for S4 (large-size detached houses), with wrongly classified outputs mostly due to digitization issues within the Zenrin® data, with a lack of neighbors when some should be present, and *vice versa*. If a class label accuracy is low, with no digitization issues in the input data, the conditional control statements have to be modified in the tree-like classificatory model.

Cluster validity for unsupervised clustering is a complex task extensively discussed in the literature (Halkidi *et al.*, 2001; Charrad *et al.*, 2014). It focuses on evaluating to which extent clusters are compact and well-separated. Yet, those values have to be compared to others in order to be appraised and, as a result, the same metrics are often used for both finding the number of clusters and cluster validity. For this reason, we recommend using clustering algorithms that allow obtaining information about the structure of the data prior to fix the number of clusters. This is exactly what SOM coupled with HCA does thanks to the heatmaps. The usual quality metrics (consensus methods in Section 4.3) can then help for both the number of clusters and cluster validity, but something far more important for cluster validity can be found in traditional spatial analysis. As stated by Halkidi *et al.* (2001), validation is based on external, internal or relative criteria. External validation is based on a “*pre-specified structure, which is imposed on a data set and reflects our intuition about the clustering structure of the data set*”. To illustrate this

statement, Halkidi *et al.* (2001) test whether the points within the clusters are randomly distributed or not (*Null Hypothesis*). Yet, since we are dealing with spatial data, the emergence of well-known spatial structures (monocentricity, polycentricity, aggregation, dispersion, etc.) for each class label, spatially mapped, can serve as a form of cluster validity. In this respect, minor changes can often improve the quality and the robustness of a clustering, such as (1) Adding/removing variables (e.g. newly computed morphometric indicators for the case of our test region); (2) Partitional clustering algorithms are sensitive to random initialization. As such, re-running the procedure with a different seed sometimes improve the results. In this regard, running several clustering with different pseudo-random seeds and computing similarity indexes (Fowlkes-Mallows, Rand, etc.) on the output spaces is even more effective, since it allows identifying a clustering robust to random initialization (Fusco and Perez, 2019). (3) Instead of partitional algorithms, other categories of algorithms could be tested, such as *hierarchical*, *density-based* or *grid-based clustering*.

5.2. Correspondences and spatial analysis

In the previous section, we performed three different methods that have been applied to the different periods associated to the Zenrin® data. To summarize, **algorithm CEK**, **CLEK** and **CNEK** are individually able to provide a class label, that can be compared between two periods. This is the only input required for running **algorithm GSC** (Grid Superimposition and Count): a class label associated to at least two spatial datasets describing the same study area at different time points. **Algorithm GSC** creates a grid with predetermined cell sizes, 250 meters in this example (closest size to the Japanese population grid census), that are superimposed on the extent of the test region. Cells that are not intersecting any inputs are filtered out of the grid. Then, an intersect is performed for each remaining cell to count the number of inputs grouped by labels. A final line of code groups the results of both periods into a single dataframe. For each cell, the input differences between the different periods are also calculated.

Using the structure of **algorithm GSC**, we calculate the number of building types per cell, per period, and the temporal evolutions of each type. To map the count and evolution of building types, we use the “tmap” package, which allows building thematic maps following a layer-based structure. **Appendices algorithm LSM** (Layered Spatial Mapping)shows a layered structurewhich can be used to produce any of the map within Figure 6 and Figure 7. Before setting the maps, we first import an OpenStreetMap background. The variable (count or evolution) is discretized into several categories, sequentially added as new layers to the map. **Algorithm LSM** allows dissolving each discretized category into a single multipart polygon. It allows plotting a single geometry per category, associated to colors (*e.g.* reversed magma color ramp for **Algorithm LSM**), thus substantially reducing computation time.

Table 1. Rand index scores between the 3 models (random sample of 9000 inputs with classified labels for each method, 1/9 per cluster)

2003-2004				**20013-2014**			
	tree-likemodel*	k-means *	self-organizing map *	2013-2014	tree-like model*	k-means*	self-organizing map*
tree-like model	1	0.37	0.25	tree-like model	1	0.39	0.29
k-means	0.47	1	0.46	double k-means	0.49	1	0.49
self-organizing map	0.28	0.41	1	self-organizing map	0.26	0.42	1

* output space from which the random sample is taken

Various measures for comparing classification and clustering similarities exist such as the Rand (Rand, 1971) or the Jaccard index (Jaccard, 1901). Table 1 provides the Rand index scores between the models, which vary between 0 (no correspondence between clusters) and 1 (identical clusters). As expected, Table 1 highlights a decreasing similarity of the output spaces between supervised classification, constrained clustering and fully unsupervised clustering (*e.g.* 0.49 to 0.26 for 2013-14). No output space is better than another, the models focus on different building characteristics. Furthermore, similarity indexes are consistent between 2003-04 and 2013-14, thus showing that the statistical mappings of the training datasets were rather accurate (random initialization also impacts inputs that have different possibilities of cluster attributions). Appendices Table A1 provides additional information about cluster correspondences.

Figure 6 shows the count of building types per cell in 2013-14 for four selected class labels of each of the three classification and clustering models performed in section 4. For each and every cluster, different localized aggregates and/or dispersion patterns stand out, thus contributing to cluster validation for the unsupervised clustering (k-means and SOM). Osaka, like most cities in Japan, possesses a complex polycentric structure (Hanes, 2002; Perez *et al.*, 2019b). C3 (DTs), C4 (k-means) and Cluster 7 (SOM), which are mostly tall and/or large-size collective buildings, show clear patterns of aggregation in the central part of Osaka and around the satellite centers. This is especially true for C3 (DTs), which are strictly adjoining tall and narrow buildings, with aggregate patterns that stand out where the urban structure is already the most dense and compact, such as in and around the main train and JR stations (Shin-Osaka, Kyobashi, Ōsaka Abenobashi, etc.). S1 (DTs and k-means) and Cluster 3 (SOM) are in all three models small-size and elongated townhouses, which are, amongst other things, the characteristics of the wooden *Nagaya*. This building type has a spatial structure opposed to the former

clusters of tall-narrow buildings, since they are aggregated in peripheral areas. These areas are usually residential neighborhoods, dense in term of quantity of buildings, but nonetheless less compact than in the aforementioned central parts of Osaka. S3 (DTs), S5 (k-means) and Cluster 4 (SOM) will be discussed in the next paragraph. Finally, S4 (DTs), S2 (k-means) and Cluster 5 (SOM) are mid to large-size detached houses, including modern intricate villas, traditionalJapanese wooden houses, etc., mostly located in the outskirts of Osaka, particularly in the southern part.

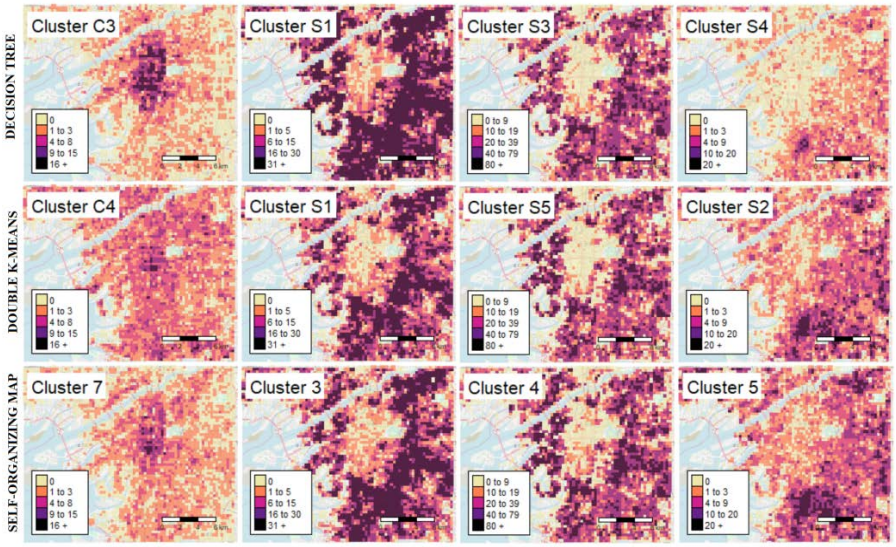

Figure 6. Count of building types per cell in 2013-14 for selected profiles of each model (tree-like classificatory model, double k-means and self-organizing map; class titles available in Table A1)

Figure 7 shows the evolution of building types per cell in between 2003-04 and 2013-14 for the very same class labels than those mapped in Figure 6. The evolution maps for tall and/or large-size buildings show hot and cold spots within and around the central part of Osaka, thus pointing at a building type which is self-regenerating according to local urban projects. The maps for elongated row houses, typical of traditional wooden *Nagaya*, point at a gradual disappearance of this traditional building type, even perhaps at a replacement if we link this disappearance with the increase of other building types in the same locations. At first sight, the clusters S3 (DTs), S5 (k-means) and Cluster 4 (SOM) appear to have some characteristics in common, such as the fact that these buildings are mostly compact, and located on similar locations. However, the evolution maps highlight a disappearance of S3 (DTs) and an increase of S5 (k-means) and Cluster 4 (SOM). This is because S1 (DTs) solely focuses on identifying very small-size traditional *Nagaya* (Appendices Table A1). As a result, mid-size and/or intricate single-family homes, but elongated, are classified within other clusters in the tree-like classificatory model, while all

small to mid-size buildings with elongated characteristics are clustered with S1 (k-means) and Cluster 3 (SOM). Thus, what is disappearing is the elongated micro-tomid-size townhouses, which are also sometimes compact and intricate, to the benefit of small to mid-size compact houses, which is consistent with the literature pointing at mini developments of detached houses on small land parcels of 50-70 m² (Asami and Niwa, 2008) and spread of small-size prefabricated homes (Buntrock, 2017).

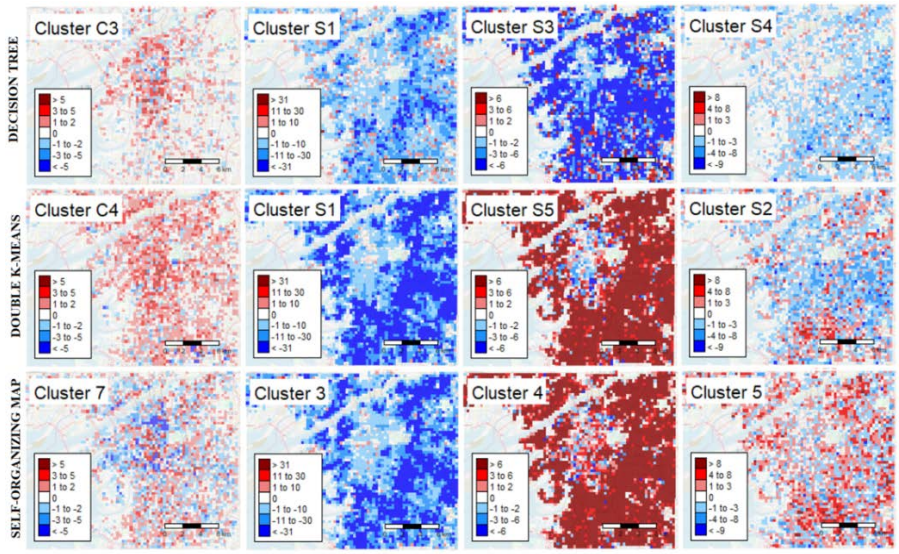

Figure 7. Evolution of building types per cell between 2003-04 and 2013-14 for selected profiles of each model (tree-like classificatory model, double k-means and self-organizing map; class titles available in Table A1)

Finally, large-size detached houses are also gradually disappearing (S4 DTs and S2 k-means), while mid-size detached houses (Cluster 5) show hot and cold spots according to the locations.

6. Conclusion and discussion

This paper puts forward several recommendations and suggestions while processing spatiotemporal data on building stock with machine learning algorithms. Based on the availability of expert knowledge, supervised tree-like classificatory model, constrained clustering and fully unsupervised clustering have been performed. Regardless of the method, we stressed the importance of spatial analysis, which has to be coupled with the usual quality metrics for cluster validation, or with a verification of the class label attributions through random samples and street-based urban imagery for supervised classification. For unsupervised clustering, we also

stressed the importance of using training samples made of inputs drawn from different periods altogether and highlighted the interest of using algorithms that allow obtaining information about the structure of the clustering before deciding the number of clusters.

More specifically, we made use of the Japanese Zenrin® GIS building inventory at two different dates for studying the spatial location and temporal evolution of building types. The first and most important condition is that high quality GIS layers of building footprints with heights (or floors) as attribute data are available for at least two different time points. Another attribute is valuable, but not mandatory: building specialization. Attention should be paid to the time depth between the layers, as it must be sufficient to analyze structural changes in the spatial organization of building types. The temporality behind these structural changes is also dependent on cultural specificities regarding planning, building norms and urban development traditions (deconstruction, dismantlement, renovation, etc.). If the prerequisites are fulfilled, then it is possible to compute a basic set of indicators for each building, such as the one presented in Section 3.1. and runs the aforementioned classification and clustering algorithms with subset merging the different periods altogether. An automated superimposition of a grid then allows counting occurrences of each type, as well as quantify the difference per cell between the different time points. For each category, mappings are performed, with cells that are discretized in the same category, dissolved into a single multipart polygon, thus highlighting localized aggregates and patterns of disappearance and self-regeneration of building types.

Regarding the clustering applications, the methodological propositions could be improved, especially by focusing on maintaining output coherence over different settings or initializations. A matrix of similarity indexes can for example be calculated among the clustering results to evaluate the robustness of the protocols to pseudorandom number generation (seed), and cross-validation methods, such a k-fold or Jackknife resampling, can be used to evaluate the robustness of the outputs to the training samples.Finally, algorithms such as k-means or self-organizing maps are efficient for the detectionof spherical and well-separated clusters (minimization of intra-cluster distances and maximization of inter-cluster distances). Other algorithms could be better suited for the detection of other structures (axial, annular, etc.) such as density-based methods, or for the detection of clusters following a similar behavior for a subset of variables only, such as Bayesian networks.

Yet, interesting preliminary results already stand out regarding specific locations and evolution trajectories of certain building types in Osaka, Japan. It shows that the ongoing transformations that characterize urban areas are linked, at least to some extent and to different degrees, to the inherited urban forms. The supervised tree-like classificatory model is the model that provides the greatest contributions in term of thematic knowledge. Indeed, provided the thresholds are established by an expert, there is little room for mistakes with the exception of data accuracy. By contrast, many unknowns remain in cluster analysis, for which the usual quality metrics are barely able to provide answers. Eventually, expert knowledge is not required to perform unsupervised clustering, but it becomes mandatory to evaluate the quality of

the partitions. This research opens interesting perspectives, such as the study of the disappearance of some building types, or the gradual replacement of some types by other. Several methods, from standard correlation matrices to geographically weighted regressions, as well has the addition of other indicators, could allow exploring these dynamics in conjunction with other factors, such as population evolution, socio-demographic characteristics, planning policies, etc. We will also remark that building substitution through destruction-reconstruction is particularly relevant in Japanese cities. In other contexts, and more particularly in European cities, building renovation should also be included when studying sociodemographic dynamics in conjunction with building types. For the interested reader, detailed thematic results for the case study of Osaka arepresented and discussed in Perez *et al.* (2023).

Acknowledgment

This study is supported by Joint Research Program no. 774 at CSIS, UTokyo (Zmap TOWN II 2013/14 Shapefile Osaka prefecture).

Bibliography

Araldi A. and Fusco G. (2019). From the street to the metropolitan region: Pedestrian perspective in urban fabric analysis. *Environment and Planning B: Urban Analytics and City Science*, vol. 46, n° 7, p. 1243-1263.

Araldi A., Emsellem D., Fusco G., Tettamanzi A. and Overal D. (2023). Ordinary building types of France and their geographic distribution. Computer-aided quantitative analysis using official national-scale spatial data. *Revue Internationale de Géomatique* (RIG), in press.

Asami Y. and Niwa Y. (2008). Typical lots for detached houses in residential blocks and lot shape analysis. *Regional Science and Urban Economics*, vol. 38, n° 5, p. 424-437.

Behnisch M., Hecht R., Herold H. and Jiang B. (2019). Urban big data analytics and morphology. Special issue editorial. *Environment and Planning B: Urban Analytics and City Science*, vol. 46, n° 7, p. 1203-1205.

Berghauser Pont M. *et al.* (2019). The spatial distribution and frequency of street, plot and building types across five European cities. *Environment and Planning B: Urban Analytics and City Science*, vol. 46, n° 7, p. 1226-1242.

Bonnin P., Nishida M. and Inaga S. (2014). *Vocabulaire de la spatialité japonaise.*CNRS Editions, 560 p.

Buhnik S. (2010). From Shrinking Cities to *Toshi no Shukushō*: Identifying Patterns of Urban Shrinkage in the Osaka Metropolitan Area, *Berkeley Planning Journal*, vol. 23, n° 1, p. 132-155.

Buntrock D. (2017). Prefabricated housing in Japan. *Offsite Architecture Constructing the Future*, Smith E. and Quale J. (Eds.), Routledge: New York, NY, USA, p. 190-213.

Case Scheer B. (2017). Urban Morphology as a Research Method. *Planning Knowledge and Research*, Sanchez T.W. (ed.), p. 167-181.

Charrad M., Ghazzali N., Boiteau, V. and Niknafs A. (2014). NbClust: An R Package for Determining the Relevant Number of Clusters in a Data Set. *Journal of Statistical Software*, vol. 61, n° 6, p. 1-36.

Chorley R. and Hagget P. (1967). Models, Paradigms and the New Geography. *Socio-Economic Models in Geography*, Routledge, pp. 1-24,.

Coutard N., Lévy J-P., Barles S. et Blanc N. (2013). Écologies urbaines. *Le développement durable à découvert*, Euzen A., Eymard L et Gaill F. (eds.), CNRS editions.

D'Acci L. (ed.) (2019). *The Mathematics of Urban Morphology*, Birkhäuser.

Dempster A., Laird N. and Rubin D (1977). Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society. Series B*, vol. 39, n° 1, p. 1-38.

Fujii Y. 2008. Shrinkage in Japan. *Shrinking Cities Volume 3: Japan*, edited by P. Oswalt, 9-12. Berlin: Project Office Philipp Oswalt.

Fleischmann M., Feliciotti A., Romice O. and Porta S. (2021). Methodological Foundation of a Numerical Taxonomy of Urban Form. *Environment and Planning B: Urban Analytics and City Science*. vol. 49, n° 4, p. 1283-1299.

Fusco G. and Perez J. (2019). Bayesian Network Clustering and Self-Organizing Maps under the Test of Indian Districts. A comparison. *Cybergeo: European Journal of Geography*, paper 887.

Hecht R., Meinel G. and Buchroithner M. (2015). Automatic identification of building types based on topographic databases – A comparison of different data sources. *International Journal of Cartography*, vol. 1, n° 1, p. 18-31.

Haldiki M., Batistakis Y. and Vazirgiannis M. (2001). On Clustering Validation Techniques. *Journal of Intelligent Information Systems*, vol. 17, n° 2/3, p. 107-145.

Hartmann A., Meinel G., Hecht R. and 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*, vol. 5, n° 8, 142.

Islam H., Jollands M. and Setunge S. (2015). Life cycle assessment and life cycle cost implication of residential buildings - A review. *Renewable and Sustainable Energy Reviews*, vol. 42, p 129-140.

ISUF (2021). *Annual Conference Proceedings of the XXVIII International Seminar on Urban Form: Urban Form and the Sustainable and Prosperous City*. Ed. by Feliciotti, A. and Fleischmann, M., 29TH June - 03rd 2021, Glasgow.

Jaccard P. (1901). Étude comparative de la distribution florale dans une portion des Alpes et des Jura. *Bulletin de la Société Vaudoise des Sciences Naturelles*, 37, p. 547-579.

Larkham P.J. (2019). Extending Urban Morphology: Drawing Together Quantitative and Qualitative Approaches. *The Mathematics of Urban Morphology*, D'Acci L. (ed.) Birkhäuser, p. 503-515.

Lüdecke D., Ben-Shachar M., Patil I. and Makowski D (2020). Extracting, Computing and Exploring the Parameters of Statistical Models using R. *Journal of Open Source Software*, vol. 5, n° 53, 2445.

Lynch K. A. (1981). *A Theory of Good City Form*. MIT Press.

Marsh R. (2016). Building lifespan: Effect on the environmental impact of building components in a Danish perspective. *Architectural Engineering and Design Management*, vol. 13, n° 2, p. 80-100.

Jacobs J. (1961). *The Death and Life of Great American Cities*. Random House, New York.

Kohonen T. (1982). Self-organized formation of topologically correct feature maps. *Biological Cybernetics*, vol. 43, p. 59-69.

Kollwitz M., Luotonen E. and Huuhka T. (2022). How changes in urban morphology translate into urban metabolisms of building stocks: A framework for spatiotemporal material flow analysis and a case study. *EPB: Urban Analytics and City Science*, Online First.

Lee J. and Newman G. (2017). Forecasting Urban Vacancy Dynamics in a Shrinking City: A Land Transformation Model. *International Journal of Geo-Information*, vol. 6, n° 4, 124.

MLIT (2007). White Paper on Land, Infrastructure, Transport and Tourism in Japan. *Ministry of Land, Infrastructure, Transport and Tourism Publication*, Ministry of Land: Tokyo, Japan.

Newman P. and Kenworthy J. (1999). *Sustainability and Cities: Overcoming Automobile Dependence*. Washington. Island Press.

Oliveira V. and Madeiros V. (2016). 'Morpho: Combining morphological measures'. *Environment and Planning B: Planning and Design*, vol. 43, n° 5, p. 805-825.

Paquot T. (2019). *Désastres urbains*, La Découverte, Paris, 264 p.

Perez J., Fusco G., Araldi A. and Fuse T. (2019a). Identifying building typologies and their spatial patterns in the metropolitan areas of Marseille and Osaka. *Asia-Pacific Journal of Regional Science*, 4, p. 193-217.

Perez J., Araldi A., Fusco G. and Fuse T. (2019b). The Character of Urban Japan: Overview of Osaka-Kobe's Cityscapes. *Urban Science*, vol. 3, n° 4, 105.

Perez J., Ornon A. et Usui H. (2020). Classification of residential buildings into spatial patterns of urban growth: A morpho-structural approach. *Environment and Planning B: Urban Analytics and City Science*, vol. 48, n° 8, p. 2402-2417.

Perez J., Fusco G. and Sadahiro Y. (2023). Shrinkage and Morphological Change: A Study of Building Type Evolution in the Osaka-Kobe City-region in Japan, Submitted.

Rand W M. (1971). Objective criteria for the evaluation of clustering methods. *Journal of the American Statistical Association*, 66, p. 846-850.

Rossi A. (1982). *The Architecture of the City*. MIT Press, Cambridge, MA.

Salingaros N. (2005). *Principles of Urban Structure*. Techne Press, Delft.

Salingaros N. (2006). *A Theory of Architecture*. Umbau Verlag.

Sennet R. (2018). *Building and Dwelling: Ethics for the City*. London: Penguin Books.

Sakamoto K., Iida A. and Yokohari M. (2017). Spatial Emerging Patterns of Vacant Land in a Japanese City Experiencing Urban Shrinkage. A Case Study of Tottori City. *Urban and Regional Planning Review*, vol. 4, p. 111-128.

Shelton B. (2012). *Learning from the Japanese City: West Meets East in Urban Design*. Taylor & Francis.

Usui H. et Perez J. (2020). Are patterns of vacant lots random? Evidence from empirical spatiotemporal analysis in Chiba Prefecture, east of Tokyo. *Environment and Planning B: Urban Analytics and City Science*, vol. 49, n° 3, p. 777-793.

Appendices

R session: information and package versions

```
1: R version 4.2.0 (2022-04-22 ucrt)
2: Platform: x86_64-w64-mingw32/x64 (64-bit)
3: Running under: Windows 10 x64 (build 19044)
4: kohonen_3.0.11      exactextractr_0.8.2 tmap_3.3-3         raster_3.5-15       sp_1.5-0
5: forcats_0.5.1       stringr_1.4.0       dplyr_1.0.9        purrr_0.3.4         readr_2.1.2
6: tidyr_1.2.0         tibble_3.1.7        ggplot2_3.3.6      tidyverse_1.3.1     lwgeom_0.2-8
7: sf_1.0-7            fmsb_0.7.3          factoextra_1.0.7          FNN_1.1.3.1         tmaptools_3.1-1
8: tmap_3.3-3          OpenStreetMap_0.3.4
```

Algorithm IC: Indicators computation on a GIS inventory named “DATA”

```
1:    DATA_BUFFER <- st_buffer(DATA, 0.1) # neighbors 1/2
2:    DATA <- cbind(DATA, lengths(st_intersects(DATA _BUFFER)) - 1) # neighbors 2/2
3:    DATA$area<- st_area(DATA) # area 1/1
4:    DATA$floor_area<- DATA$floor* DATA$area# floor area 1/1
5:    DATA$perimeter<- st_perimeter(DATA) # elongation 1/2
6:    DATA$elongation<- (pi*(2*(sqrt(DATA$area/pi))))/DATA$perimeter# elongation  2/2
7:    DATA_HULL <- st_convex_hull(BU0304) # convexity 1/2
8:    DATA$convexity<- DATA$area/st_area(DATA_HULL) # convexity 2/2
```

Algorithm CEK: Asupervised decision tree function with 3 classes applied on a GIS inventory named “DATA”

```
1: DTs <- function(row) {
2:   X1 <- row[['var1']]
3:   X2 <- row[['var2']]
4:   X3 <- row[['var3']]
5:   if(X1 == 1363& X2 >1000 & X2 == 0) {   # condition examples
6:     class <- 'C1'# label attribution
7:   }
8:   else if(conditions) {
9:     class <- 'C2'
10:  }
11:   else if(conditions) {
12:     class <- 'C3'
13:  }
14:   return(class)
15: }
16: DATA$class<- as.factor(apply(DATA, 1, DTs))
```

Algorithm CLEK: k-means with multi-temporal subset made of DATA 1 and 2, collective and single-family homes and test data mapping

```
1: sample <- rbind(subset(DATA1, var == condition)[sample(1:100, round(nrow(subset(DATA1, var ==
condition))*0.6), replace=TRUE),], subset(DATA2, var == condition)[sample(1:100,
round(nrow(subset(DATA2, var == condition))*0.6), replace=TRUE),])# condition shall be collective OR single-family
homes
2: kmeans.sample<- kmeans(as.matrix(sample[,c(5:8)]), 4) # 5:8 variables used for clustering; 4: number of clusters
3: kmeans.mapping<- cbind(subset(DATA1, var == condition), cl.kmeans = get.knnx(kmeans.sample$center,
subset(DATA1, var == condition)[,c(5:8)], 1)$nn.index[,1]) # mapping example for DATA 1
```

Algorithm CNEK: SOM with multi-temporal subset made of DATA 1 and 2 and test data mapping

```
1: sample <- data.matrix(rbind(subset(DATA1, var == condition)[sample(1:100, round(nrow(subset(BUILDING1,
```

```
var == condition))*0.6), replace=TRUE),], subset(DATA2, var == condition)[sample(1:100,
round(nrow(subset(DATA2, var == condition))*0.6), replace=TRUE),])# condition shall be collective AND single-family homes
2: som.grid<- somgrid(xdim = 15, ydim = 15, topo = 'hexagonal', toroidal = F) # setting the map to 225 nodes
3: som.model<- som(sample, maxNA.fraction = 1, grid = som.grid, keep.data = T, rlen = 500) # SOM clustering
4: som.hc.cluster<- as.data.frame(cbind("nodes_IDS" = 1:225, "hc_clusters" =
cutree(hclust(dist(som.model$codes[[1]])), 9))) # hierarchical clustering with 9 clusters
5: map.DATA1 <- kohonen::map(som.model, DATA1) # test data mapping example for DATA1
6: DATA1.clsom <- cbind(DATA1,"nodes_IDS" = map.DATA1[[1]]) # add SOM nodes ID to original data
7: DATA1.clhcsom <- DATA1.clsom %>%
left_join(som.hc.cluster, by = "nodes_IDS") # get correspondences between SOM nodes ID and hierarchical clustering clusters ID
```

Algorithm GSC: Grid superimposition and count of inputs DATA1 and 2 per cell

```
1:    mesh <- st_sf(st_make_grid(DATA1, cellsize = 250)) # 250 meters grid creation and superimposition over the test region DATA1
2:    mesh<-mesh[lengths(st_intersects(mesh, DATA1)) >0,] # filter cells with no overlap with buildings
3:    DATA1_C <- tapply(st_geometry(DATA1), DATA1$class, FUN = function(x) lengths(st_intersects(mesh, x))) # count inputs per cell for DATA1
4:    res.temp<- cbind(mesh, data.frame(sapply(DATA1_C, function(x) x[1:max(lengths(DATA1_C))]))) # df with DATA1
5:    DATA2_C <- tapply(st_geometry(DATA2), DATA2$class, FUN = function(x)       lengths(st_intersects(mesh, x))) # count inputs per cell for DATA2
6:    MESH.DAT <- cbind(res.temp, data.frame(sapply(DATA2_C, function(x) x[1:max(lengths(DATA2_C))]))) # df with DATA1 and DATA2
```

Algorithm LSM: Spatial mapping, example for a class label aggregated with MESH.DAT

```
1:      osm_test_region<- read_osm(MESH.DAT, ext=1) # import basemap
2:      tm_shape(osm_test_region) + tm_rgb(alpha = 0.6) + # plot basemap with transparency
tm_shape(st_geometry(st_union(subset(MESH.DAT, class == discretization values)))) + tm_fill("palegoldenrod", alpha = 0.6) +
tm_shape(st_geometry(st_union(subset(MESH.DAT, class >= discretization values)))) +
tm_fill("coral", alpha = 0.7) +
tm_shape(st_geometry(st_union(subset(MESH.DAT, class >= discretization values)))) +
tm_fill("violetred", alpha = 0.5) +
tm_shape(st_geometry(st_union(subset(MESH.DAT, class >= discretization values)))) +          tm_fill("darkorchid4", alpha = 0.5) +
tm_shape(st_geometry(st_union(subset(MESH.DAT, class >= discretization values)))) +              tm_fill("gray1", alpha = 0.5) +
      tm_layout("title", frame = F, legend.frame = T, title.bg.color = T) +
      tm_scale_bar(position = c("right", "bottom")) +
      tm_add_legend('fill', labels = c("discretization values"),
            col = c("palegoldenrod","coral","violetred","darkorchid4","gray1"), alpha = 0.8, title = ' ') +
      tm_legend(position=c("left", "bottom"))
```

Table A1. *Cluster counts, evolution and characteristics*

Tree	**C1**	**C2**	**C3**	**C4**	**S1**	**S2**	**S3**	**S4**	**S5**
Count 2013-20114	9.908	6.855	6.139	42.366	132.691	38.422	115.827	5.742	70.925
Evolution (%)	37.24	13.28	11.17	18.09	-18.54	-19.48	-16.95	-32.26	26.03
Common characteristics	Collective				Single family home				
					Adjoining			Detached	
Individual characteristics	Large	Elongated	Adjoining & narrow	Small	Verys mall & elongated	Intricate	Elongated &/or compact	Large	Small to mid-size
K-MEANS	**C1**	**C2**	**C3**	**C4**	**S1**	**S2**	**S3**	**S4**	**S5**
Count 2013-2014	7.959	32.458	14.590	10.261	65.929	17.987	64.572	97.164	117.955
Evolution (%)	-15.04	29.19	23.45	22.78	-58.16	-11.15	22.02	-15.41	70.34
Common characteristics	Collective				single family home				
Individual characteristics	Very small & compact	Large & compact	Adjoining, large & compact	Tall, intricate & /or elongated	Adjoining & elongated	Large	Intricate & detached	Compact, mostly detached & elongated	Small (mostly compact/adjoining)
SOM	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**
Count 2013-2014	120.465	8.559	83.727	131.032	22.074	42.261	5.712	9.086	5.959
Evolution	-13.53	-57.46	-47.17	69.03	14.04	9.54%	-10.09	102.18	12.63
Individual characteristics	Detached & compact	Adjoining& compact	Elongated, adjoining & small	Adjoining, compact &small	Detached & mid-size	Average	Tall, narrow & elongated	Adjoining, large & compact	Tall, narrow &/or elongated