

FEATURE RECOGNITION AND CLUSTERING FOR URBAN MODELLING

Exploration and analysis in GIS and CAD

André CHASZAR

O-Design Consulting and Research, New York, NY, USA, and Delft University of Technology, Delft, Netherlands
AC@0-design.com

and

José Nuno BEIRÃO

Bquadrado and Technical University Lisbon, Lisbon, Portugal
jnb@fa.utl.pt

Abstract. In urban planning exploration and analysis assist the generation, measurement, interpretation and management of the modelled urban environments. This frequently involves categorisation of model elements and identification of element types. Such designation of elements can be achieved through attribution (e.g. ‘tagging’ or ‘layering’) or direct selection by model users. However, for large, complex models the number and arrangement of elements makes these approaches impractical in terms of time/effort and accuracy. This is particularly true of models which include substantial numbers of elements representing existing urban fabric, rather than only newly generated elements (which might be automatically attributed during the generation process). We present methods for identification and categorisation of model elements in models of existing and proposed urban agglomerations. We also suggest how these methods can enable exploration of models, discovery of identities and relationships not otherwise obvious, and acquisition of insights to the models’ structure and contents which are not captured, and may even be obscured, by manual selection or automated pre-attribution.

Keywords. City information modelling; data mining; feature recognition; geometric-content-based-search; urban typologies.

1. Introduction

A fundamental assumption of urban design (more so than -planning) is that urban forms influence the working of urban structures. Thus, not all urban activities and

processes can be accommodated by identical (or nearly homogeneous) structures, but instead variegated forms are needed. The scale(s) at which such variety are needed and would be detectable also vary, and it is not always at an 'urban' scale that important differences occur which influence the life of the city. For example, a wide range of residential, commercial, educational, cultural and other activities can be accommodated within apparently identical block structures of Barcelona's Eixample. Yet in most cities the blocks and buildings, street patterns, open spaces and other basic urban elements do vary in form, by accident or intention. This is generally viewed as a positive characteristic for urban agglomerations to have, though at some threshold variety may come to seem excessive (Lynch, 1960). Thus, legibility of the city is achieved by some balance of predictability and surprise, of uniformity and variation, neither homogeneous nor random. The recognition of the need in the city for variation with order fits to the concept of urban typologies, not just at the level of entire agglomerations (city types) but also at the level of their elements (block types, street pattern types, etc.) The effort to read and project cities in terms of typologies also aims to support planning and control, for example in terms of predictability based on the assumption that the workings of particular types and the interactions among them are sufficiently understood (Branch, 1985). Yet no general consensus exists regarding exactly which types would adequately capture the characteristics of cities, which suggests a need for flexibility in defining and discovering types, achieved through ongoing processes of categorization.

That processes of categorization, or of 'typifying', are ongoing rather than fixed has been well demonstrated elsewhere (Bowker and Star, 1999), and we take it as given that in urbanism types are largely cultural constructs rather than 'truths' (Habraken, 2000). In either case, however, designers and planners working with large models of urban agglomerations can benefit from tools that help them make sense of the models and the modelled entities by categorizing (and re-categorizing) through analyses of the data available. Furthermore, new methods may induce the emergence of types otherwise not identifiable. Such a possibility is especially relevant in a field where ambiguity is a constant factor, as in the case of the urban environment.

In this paper we present and compare some methods for analyzing and categorizing urban forms which can help urban designers both to make sense of the structure and workings of existing urban agglomerations (towns, cities, formal or informal) and to find potential problems and opportunities in their future development. The methods on which we focus are 'clustering' (Agrawal et al., 1998; Witten and Frank, 2005) and 'user-guided feature recognition' (Chaszar, 2011a, 2011b). These methods both proceed from data contained in digital urban models - whether GIS or CAD - and process these data to find similarities and differences of possible significance. Finding such categories, patterns and other relationships among elements of the urban fabric enables designers to draw conclusions about existing conditions and also to take action when needed by proposing changes to

those conditions – especially, in the present case, physical conditions, in contrast with regulatory, policy and other possibly relevant changes.

The abilities conferred by methods such as those described here are important for various reasons. First, while models of existing urban conditions are often annotated or associated with data describing the functions (actual or legislated) and activities of particular buildings, blocks and/or parcels, such data may be unavailable, missing or incorrect. Analysis of forms can then be supplementary and may enable the finding of inconsistencies in the data. Second, projected urban designs may also include data denoting the intended uses of urban elements, but the reading of potentials by the designers themselves may under-represent the range of possible readings. Analysis of forms can then be exploratory, amplifying the models' roles as 'learning devices' (Hatchuel, 2002). The identification of possible correspondences between shapes and existing or future uses can to a certain extent be made without recourse to such techniques, of course; however, (semi)automation of the analysis as described here can be very valuable as a practical matter when models are large – consisting of hundreds, thousands or more elements – and it can also enable detection of more subtle similarities and differences than would a more conventional, 'manual' approach (Chaszar, 2011a, 2011b).

In the next section we review the context of the problem, briefly describing the practice of City Information Modelling (CIM) and related practices as well as some relevant challenges encountered within these practices. We then present a set of alternative methods for exploring and navigating the contents of such models, focusing on 'clustering' and 'user-guided feature recognition' utilising mainly geometric characteristics of model elements, though adaptable also to include topological and non-geometric data. Subsequently we describe the urban model on which tests were performed, present test results, and discuss how the application of these methods improves model users' abilities to analyse and interpret City Information Models and other models of urban environments.

2. Context – Practices and Problems

CIM here is understood as: an urban design environment integrating design and analysis tools allowing for the manipulation of designs not just through geometric parameters but also by manipulating the design attributes including some qualitative properties (Beirao et al., 2009). GIS tools are normally used in the analysis of urban environments by querying geographic databases, usually resorting to structured query languages (SQL) (Egenhofer, 1994). In the practice of City Information Modelling we stress the importance of integrating the analytical platform with the design platform maintaining the ability to perform the analysis tasks; in this case being able to analyse proposals integrated in a context and easily changing them. We also see that the GIS environment has only very basic tools to enable queries

based on geometrical characteristics of the urban environment. As discussed above, a typical analytical task would be finding particular urban types within an urban context. Typology identification can be particularly useful in contexts with incomplete or sparse databases, still a recurrent problem in many countries. Furthermore, in the context of developing countries it is often impossible to keep an updated database due to fast informal developments as well as to the difficulties in accessing the emerging areas for data survey. In these latter situations, many times the only updated easy-to-access element is an aerial photo from which only shape and a few other properties may be extracted. Typically, in such situations the designer is forced to rely on some kind of process to further fill the gaps in information.

In what follows we first describe the two methods of concern and describe their application to a particular urban model, contrasting the results obtained. We then discuss these findings in terms of their significance for future work with urban models and for future development of the methods.

3. Methods – Clustering and UGFR

The ‘clustering’ approach comprises a number of largely automated, statistically based methods of detecting correlations among groups of items via their attributes. The attributes examined can be data of any kind (e.g. economic, demographic, etc.), including geometric data like length, perimeter, area, etc. ‘Clusters’ are formed on the basis of the degrees to which the analysed items’ attributes are identical or similar. While clustering analyses can be performed using all available attributes of the analysed items, more commonly some subset of attributes is chosen for analysis, which gives a degree of user-control, if not actual interaction. Subsets of attributes are defined considering meaningful correlations which may be conclusive for the identification of types or for classification in general (Gil et al., 2012). Clustering methods require also in certain cases that users outwit redundancy in the data to avoid classifying as a type what may be several expressions of the same attribute and not a real cluster of different attributes. Such statistical methods may be used to identify and classify different block and street types. It should be stressed that the attributes used in this example are all numerical and extracted from the geometry in a GIS environment (from a shapefile). Some of the attributes are directly obtained from the geometry and others calculated from basic meaningful relations among the primary ones (e.g.: FAR – floor area ratio, GSI – ground space index, OSR – open space ratio). All these attributes are then stored in a database where the statistical analyses and clustering are subsequently performed.

The ‘user-guided feature recognition’ (UGFR) method detects similarities and differences among analysed items by comparing each item to a ‘feature vector’ which describes the attributes of interest. Note that whereas ‘attributes’ tend to be single data items attached to a model element (or sub-element), ‘features’ tend to be emergent

characteristics of the attributes, elements and/or sub-elements. While this is a common approach in most object recognition methods, most such methods aim to define a comprehensive feature vector – or even a ‘vector of vectors’ – which can capture all features (or attributes) of interest needed to identify a particular object type and distinguish it from other types (Tangelder and Velkamp, 2008). This normally aims for and leads to a fully automated analysis process with no user control. In contrast, the ‘user-guided feature recognition’ method applied here aims to focus on salient features as chosen by the user, and allows for user-control via selection and/or editing of the features considered salient for discriminating types or otherwise significant in terms of their urban nature. While ‘user-guided feature recognition’ was developed specifically to deal with geometric data, it could also be extended to include any other sorts of attributed data, like ‘clustering’: use, physical conditions (ground type, water and other resources), number of entrances, demographics, etc. In any case, the method gives a visual, easy to assess, view of specific geometry-based properties of the components of the urban environment (Chaszar, 2011a, 2011b).

These geometric-content-based search and classification methods are complementary to others such as data queries and clustering by data mining the geographic database (Gil et al., 2012), based on traditional GIS platforms and geographic databases. The complementarity of these methods can be further highlighted by pointing out that the results obtained from UGFR can also be subject to clustering for finding other kinds of correlations among properties of an urban environment which may be used to fine-tune a typology search procedure. The main difference in the proposed method is that complementary results may be achieved by searching particular geometric features. Whereas conventional data queries typically only retrieve model elements explicitly attributed with the specified properties, the user-guided methods enable selection via derived properties as well. And while data mining identifies clusters of elements with cohering sets of properties, it tends to ignore properties found in multiple clusters, which may nevertheless be important in characterising the types generated.

The methods proposed are also of interest when the geographic information is limited but the geometric information is relatively complete, and in particular in the preparation of the working base in an area lacking good reliable geographic information. In many countries, geographic information is many times not much more than flattened geometric information not correctly classified nor attributed (e.g. layered). It may be particularly useful in the case of emergent settlements where the information is usually insufficient and constantly growing and mutating. In such circumstances the first steps of the procedures aim at classifying all geometric representations in order to correct, update and complete the database. These procedures typically resort to visual recognition of geometrical features in the representations. The proposed methods support this activity through semi-automated, user-guided feature recognition. Moreover, because UGFR runs directly on the geometric model

(within the CAD environment) the method becomes quite interactive with the design process easily allowing analysis at different stages of design development. The results of our testing, presented below, indicate that such methods can contribute to the interactive flow of information during the design decision process, whether this process involves comparing different scenarios of urban design/planning or refining designs/plans to meet particular goals, such as the development of adaptable urban systems capable of accommodating rapid transformations of contemporary societies.

4. Models and Results

We performed tests for the research described here using models based on data from an existing urban unit, a cluster of small villages linked by a collection of industrial and commercial facilities focused on stone quarrying and processing. The present urban structure results from different occupations and transformations through time with different morphological characteristics related with the dominant activity of each period. The urban fabric thus comprises rural structures later occupied by industrial facilities (storehouses and factories), structures dependent on existing quarries, and new suburban housing developments. (The settlements are 30km from Lisbon).

The models include data pertaining to physical structures and features (e.g. buildings, streets, waterways) as well as geometrically describable non-physical features (e.g. property lines, administrative zones) and also other non-physical data attributed to the geometric items (e.g. population, income, land use). These data were originally contained only in a GIS model, but the geometrically describable items were exported to a CAD environment for more advanced visualization and to enable use of the UGFR tools. The model comprised approximately 3000 geometric elements, of which approximately 2000 were buildings, a modest size for urban agglomerations but already too large for effective analysis by ‘manual’ means – i.e. mainly visual inspection, with subsequent selection and colouring, assignment to CAD file ‘layers’ and/or other means of annotation.

Application of the ‘clustering’ method results in categorizations that capture more detailed and in-depth information than current manual procedures. Gil et al. (2012) clearly showed the capacity of the statistical methods to capture, for instance, several types of peripheral blocks that would probably be classified as a single category if traditional morphological analysis had been used. Additionally, clustering allows identifying the variability range of the parameters in each type and to pinpoint the characteristics of the local archetype – i.e. centroid – for each cluster (type). However, as an ‘unsupervised learning’ technique, it allows only minimal scope for user interaction.

Applying the UGFR method resulted in various user-selected categorizations of buildings according to various criteria, including simpler size-based and more complex shape-based ones. For more detailed discussion of the UGFR interface and

evaluation logic, readers are referred to Chaszar (2011a, 2011b). Size-based metrics include area, volume and volume-area ratio, as well as maximum and/or minimum dimensions. Shape-based criteria also may depend in the case of UGFR on some dimensional information (in contrast with many other object/feature recognition methods which are scale-free), but they can also involve more complex calculations using the basic geometric data. Some simple example categorizations by shape are ‘tower-like’ and ‘slab-like’ buildings. Note that reporting of search/classification analysis results can be either binary – i.e. a model item/object either meets or does not meet the stipulated criteria – or else give a more nuanced feedback, in which degrees of criterion satisfaction are also indicated (similarly to relevance ranking in common search engines). Although not implemented here, more complex feature recognition queries can also be composed, evaluating objects for general characteristics such as their ‘roughness’, ‘porosity’, etc., as well as for more specifically urbanistically motivated shape properties which can further influence the objects’ roles within the urban ensemble.

5. Discussion

On the basis of the investigations described, we can make some observations at varying degrees of generality regarding the use of ‘clustering’ and UGFR analysis methods as well as regarding the use of GIS and CAD for such analyses. Specifically within the research reported here, it was evident that even relatively simple size-based metrics can be effective in giving a quick overview of building/block type distributions. The more complex shape-based categorizations, however, can capture other characteristics which may be relevant to the existing or future potential roles of urban elements. UGFR clearly identifies some of the recurrent building types in the study area. For instance, industrial buildings pointedly stand out in a search for predominantly flat buildings (‘slab-like buildings’). When combined with large area building search the type is captured in almost every instance. House types imply more refined analyses. Contemporary single houses tend to be a bit larger (and also more ‘square like’) than the older traditional houses. The latter tend to have simpler polygons as footprint albeit being predominantly irregular while the former, although more regular, tend to have more complex polygons as footprint. Size matters, of course, certainly in terms of the character of an urban area as well as in other more technically defined performance criteria. Designers generally get a quick and perhaps unconscious sense of variations of granularity in the urban fabric without recourse to analysis tools or even to software (as from a simple map), but adding abilities to process the relevant data reliably, to define regions based on categories of element sizes, to visually record and easily communicate such results, can all assist the tacit and explicit reasoning processes involved in design and planning. Shape-based analyses, on the other hand, underscore the fact that categories are malleable, and that any

particular item may be grouped differently, depending on the criteria chosen. While a 'clustering' approach tends to assume that 'more data is better' for forming durable categories based on correlations among all available attributes (all presumably weighted equally), a UGFR approach emphasizes the importance of judgement, exercised through selection of relevant attributes and of their relative importance.

More generally, although the two methods as explained here are not that different in terms of the parameters directly extracted from the model (length, width, orientation, area, perimeter and so forth), they differ greatly in terms of their working environment: UGFR allows for more interactivity with the design process as it works within the CAD environment. All of the analyses performed could in principle be programmed and executed within a GIS environment, provided that this includes sufficiently rich geometric data as well as functions capable of manipulating those data in geometrically sophisticated ways. In practice, such capabilities are much more characteristic of CAD environments, and practically absent from GIS ones, so the continued need for transferring data seems likely. Another option may be to implement the clustering and/or UGFR algorithms within a more sophisticated database environment such as PostgreSQL, which would allow formulation of more complex queries and allow also the execution of functions to derive object (or even inter-object) properties from the given attribute data – as does UGFR within CAD – and/or otherwise process the data – as for clustering or other statistical and data-mining procedures. However this again would separate the analysis functionality from the design environment, tending to hinder the formulation and testing of insights by the model users.

As such systems accommodate varying degrees of user-interaction, the more interactive versions are more conducive to model/data exploration and thus more supportive of the role of models as 'learning devices' (Hatchuel, 2002). This in turn we would expect to lead to more and better insights gained, and thus to better comprehension of existing conditions as well as better design and planning (which, if competently carried on, may even result in better cities). Still, the level of interactivity offered poses challenges of usability and comprehensibility. More specifically, a system with very many settings and controls for customization of analyses by users may become too cumbersome for general use, instead appealing only to relatively few enthusiasts willing to devote considerable effort to mastering them. On the other hand, inadequate controllability (via user interaction) can result in misleading analyses due to inappropriate assumptions built into the system – e.g. regarding salience of data, processing of data, the number of types/categories, etc. – and/or due to inappropriately detailed (or approximate) analyses being the default, perhaps only, option.

All in all, it is apparent that both 'clustering' and 'user-guided feature recognition' methods are potentially helpful in analysis of urban models, especially but not exclusively those including geometric data. The exact ways in which such

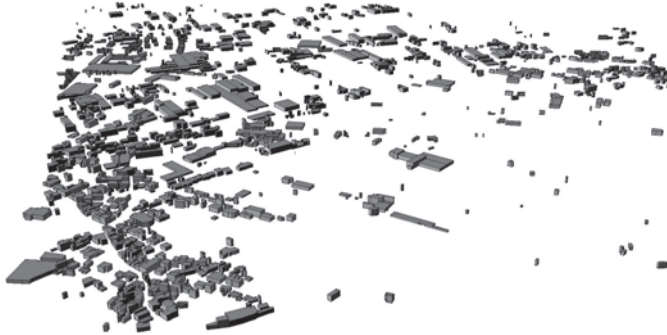


Figure 1.

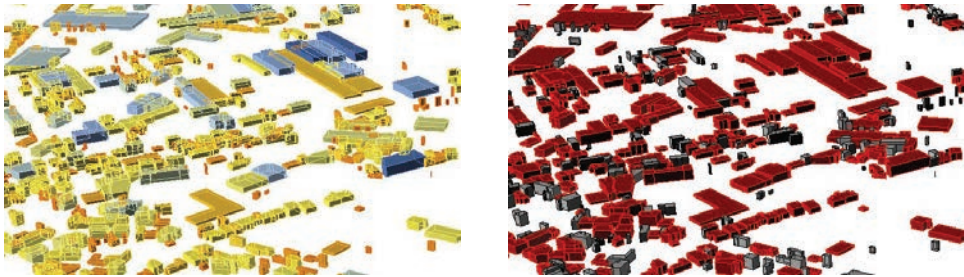


Figure 2 (a) & (b).

methods are applied, and the ends in which they result depend of course on the aims and abilities of the designers/planners using them, as on the overall processes of conceptualization and execution in which they are embedded. The combination of both methods may in future work enable the possibility of capturing even more refined patterns or types, identifying their attributes, the variability range of their parameters and even their archetypes.

6. Conclusion

We have presented methods for analyzing the forms of urban structures – in this case particularly blocks and buildings – which can help designers to comprehend the current conditions of cities and to project their futures. Different approaches to finding possibly significant similarities and differences in urban forms have been described, utilizing mainly geometric data contained in GIS and CAD models of urban agglomerations. These approaches have varying types and degrees of user-interactivity, which may make one or the other more or less suitable for particular urban analysis and design tasks. Such systems enable users to define and/or discover types of urban

elements rather than confining them to searching for pre-defined, ‘known’ types. This includes the ability to define/discover categorizations at varying degrees of discrimination, from very rough (e.g. all buildings over, and all under or equal to X sq.m. in footprint) to very nuanced (e.g. buildings with various degrees of self-shading on one or more relevant facades). Simple size-based metrics tend to have less discriminatory capability of course – not distinguishing between tall, long, or more compact forms of urban structures, for instance – but even within such a criterion as many categorizations may be made as desired. By working within the CAD environment UGFR becomes an interactive design support analysis tool, with advantages including: (1) providing a more intuitive reflection on the meaning of shapes and (2) bringing analysis processes closer to common design practices. Both clustering and User-Guided Feature Recognition can include also non-geometric data which may give further insights to the existing and potential future workings of the modelled cities. Further work is needed to extend such analysis methods to include other physical urban elements such as open (land and water) spaces, transportation networks, etc. Future development might also include analyses of spatial relationships among urban objects, rather than only of the characteristics of objects themselves, providing more accurate insight on, for example, the rather obscure area of public space morphology.

References

- Agrawal, R.; Gehrke, J.; Gunopulos, D.; Raghavan, P.: 1998, Automatic subspace clustering of high dimensional data for data mining applications, *Proc. ACM SIGMOD*, 94–105.
- Beirão, J., Duarte, J., Stouffs, R.: 2009, An urban grammar for Praia, *Proc. eCAADe*, 575–584.
- Bowker, G.C. and Star, S.L.: 1999, *Sorting things out: Classification and its consequences*, MIT Press, Cambridge, MA.
- Branch, M.: 1985, *Comprehensive city planning: introduction & explanation*, Planners Press, Chicago.
- Chaszar, A.: 2011a, Navigating complex models in Collaborative work for integrated (and sustainable) design, *Proceedings CAADFutures*, Liege.
- Chaszar, A.: 2011b, Spatial query and object recognition to support 3D digital building model use, *Proceedings EG-ICE*, Twente.
- Egenhofer, M.J.: 1994, Spatial SQL: a query and presentation language, *IEEE Transactions on knowledge and data engineering*, **6**(1), 86–95.
- Gil, J., Beirão, J.N., Montenegro, N., and Duarte, J.P.: 2012, On the discovery of urban typologies: Data mining the multi-dimensional morphology of urban areas, in *Urban Morphology*, **16**(1), 27–40.
- Habraken, N.J.: 2000, *The Structure of the Ordinary*, MIT Press, Cambridge, MA.
- Hatchuel, A.: 2002, Towards design theory and expandable rationality, *Journal of Management and Governance*, **5**(3-4), 260–273.
- Lynch, K.: 1960, *The image of the city*, MIT Press, Cambridge, MA.
- Tangelder, J. and Veltkamp, R.: 2008, A survey of content based 3D shape retrieval methods, *Multimedia Tools and Applications*, **39**(3), 441–471.
- Witten, I. H. and E. Frank.: 2005, *Data Mining: Practical Machine Learning Tools and Techniques*, Morgan Kaufmann.