Using Urban Morphology to Improve Housing Submarket Spatial Segmentation

Matthew Law

University of Liverpool

A dissertation submitted in partial fulfilment of the degree of

MSc Geographic Data Science

September 2021

Word count: 12,000

Acknowledgements

This dissertation could not have been written without the help of many people, a few of whom are acknowledged below.

Throughout the project, my supervisor Dani Arribas-Bel has offered expert advice and guidance; throughout the year he has worked to ensure the quality of my remote experience at Liverpool. I am particularly grateful to Martin Fleischmann, not only for his development of `momepy`, an invaluable tool in measuring the morphometric characters of urban spaces, but also for the help he provided with several of the thornier technical issues I encountered while completing this dissertation.

This project was organised through the Consumer Data Research Centre's Master's Dissertation Scheme, and so I thank those involved in the Scheme for facilitating the collaboration with idealista. At idealista, Juan Ramón Selva-Royo was vital in helping determine the direction of the dissertation, and provided me with both necessary data and helpful feedback.

I am forever grateful to my parents and grandparents, who provided me with countless forms of support, not least a pleasant working environment throughout the time I spent on the dissertation. I'm also thankful to all the friends who have in some way helped over the course of this dissertation – particular thanks are due to Ben, who tried his best to stop me leaving all the writing until the last minute; and to Marie-Laure, for foolishly agreeing to proofread.

Matthew Law
15 September 2021

Abstract

Whenever geographic data are aggregated spatially, a decision must be made about the spatial unit into which individual data points are grouped. In analyses of the real estate market, properties are grouped in this way into housing submarkets: sections of the real estate market which share similar characteristics. Typically, existing spatial units (such as administrative neighbourhoods or districts) are used to represent these submarkets, however there is no guarantee that such units align with the housing market dynamics they seek to delineate. This dissertation presents a method to segment an urban area into different spatial units based on its built form – its urban morphology. The spatial segmentations produced are then assessed to determine whether they can be used to represent housing submarkets. Besides the novel segmentations themselves, the dissertation presents several methodological findings. Contextual characters and the transposition of cluster labels onto simpler geometries are shown to be key methods for ensuring spatially coherent segmentations. Segmentations are shown to significantly vary depending on the spatial units clustered to generate the segmentations (with regular grids performing significantly worse than units based on buildings), and on the clustering algorithm employed.

Contents

List of Figures

List of Tables

Abbreviations

API . . . . . Application Programming Interface.

CP . . . . . Cadastral Parcel.

ET . . . . . Enclosed Tessellation.

GMM . . . . . Gaussian Mixture Model.

MAUP . . . . . Modifiable Areal Unit Problem.

MT . . . . . Morphological Tessellation.

OSM . . . . . OpenStreetMap.

OTU . . . . . Operational Taxonomic Unit.

QCoD . . . . . Quartile Coefficient of Dispersion.

UMM . . . . . Urban Morphometrics.

1 Introduction

1.1 Background

Housing submarkets are sections of the real estate market which share similar characteristics. When defined spatially, existing spatial units (such as administrative neighbourhoods) are usually employed to represent these submarkets, either individually or through a grouping of neighbourhoods. When this approach is used to analyse the housing market, for example when producing price indices based on these spatial units, the end result can misrepresent the nature of the underlying property market(s) being studied. For example, if an administrative neighbourhood contains properties of significantly varying prices, the mean price index for the area will be unrepresentative of the properties in the area it seeks to represent.

Urban morphology is the study of the physical form of the built environment. In this dissertation, a methodology is developed to partition a city (using the case study of Barcelona) into novel spatial units based on urban morphology. These are then assessed to determine how well they capture variation in both urban morphology and house prices in the city, and thus their suitability to be used as alternative spatial units to represent housing submarkets.

When carrying out a complex multi-stage process such as the spatial segmentation presented in this dissertation, there are many different parameters which can be varied to produce different configurations of the output (in this case the spatial segmentation generated). This dissertation will examine differing approaches to (housing submarket) spatial segmentation, paying particular attention to the effects of different base spatial units; contextual characters; post hoc methods to simplify the segmentation geometry; and the clustering algorithm used.

This research was undertaken in collaboration with idealista and is based in part on previous work undertaken by Juan Ramón Selva-Royo and David Rey at that company.

1.2 Geographical context

Although intended to be location-agnostic and usable wherever the requisite input data is available, the methods developed over the course of this dissertation were initially applied to the Spanish city of Barcelona, and the segmentations presented in this dissertation are of this city. The city was chosen on the recommendation of idealista, and comprises several distinct urban tissues with differing urban morphology: as such it is a good candidate for the development and assessment of a methodology for spatial segmentation based on urban morphology. Among these are the Ciutat Vella, the medieval old city, and the Eixample, the 19th Century grid of avenues and blocks.

1.3 Aims

The dissertation has three primary aims:

- Develop a methodology to partition an urban area into different spatial units of homogeneous urban morphology ('urban tissues').

- Assess the degree to which these novel spatial units capture variation in house prices, and therefore their suitability to use as spatial housing submarkets.
- Appraise the effects of a range of methodological parameters on the segmentations produced when following the general method presented in the dissertation.

1.4 Structure

This Introduction has introduced the study, outlining the context, motivations, and key aims of the research. The following Literature Review summarises past research in areas relevant to the present study. The Methodology describes the methods used to complete the analyses and to obtain the results of this research. The Results chapter reports the key results of the research presented in this dissertation, which are discussed in greater depth in the Discussion, explaining these results and their implications. Finally, the Conclusion summarises the dissertation's findings and offers potential avenues for further research.

2
Literature Review

Beginning with a discussion of the analytical issues which motivate the study, the chapter goes on to outline several key ways in which housing submarkets have been spatially delineated. The development of urban morphology is then summarised, focussing particularly on the ways in which recent data, technological, and methodological advances have allowed concepts from the field to be operationalised in (computational) quantitative research. Synthesising the two preceding themes, the key ways in which urban morphology has been used to partition space are outlined. The chapter ends with a review of possible gaps in the literature which the present research could seek to fill.

2.1 The pitfalls of partitioning space

A core tool of geographic analysis is the aggregation of observations based on where they happen. In this way, a better understanding can be gained of phenomena that vary over space, be that the distribution of poverty in a city, the spread of a pandemic, or the composition of housing submarkets.

A persistent concern in geographic analysis is that when aggregating observations at certain locations into geographical units, the trends observed (and the results of any subsequent analyses based on these) will change depending on

the spatial unit into which observations are aggregated. This issue was first described as the "Modifiable Areal Unit Problem" (MAUP) by Openshaw and Taylor (1979), and has since been a perennial subject of geographical scrutiny (Openshaw, 1984; Fotheringham and Wong, 1991; Tranmer and Steel, 2001; Duque et al., 2018). While several proposals have been made to minimise its effects (King, 1997; Nakaya, 2000; Holt et al., 1996) it remains the case that "there is no single best method that totally avoids the MAUP as long as data aggregation is involved" (Zhang and Kukadia, 2005, 77). For this reason, it is imperative that when aggregating spatial data, researchers are aware of the MAUP and the possible effects it may have on any conclusions they draw.

One approach to somewhat moderating the impact of the MAUP is to use a method which is not built on the aggregation of spatial units. To give one such example, Calafiore et al. (2021) generate functional neighbourhoods from user origin-destination flow data from Foursquare check-ins, using a spatially weighted community detection algorithm adopted from network science. While consequent analyses employing the resultant spatial units would be as susceptible to the MAUP as any other choice of spatial units, the method avoids the need to choose a base spatial unit from which to build the segmentations.

2.1.1 Spatial housing submarkets

While the term is defined and used in various ways such that "no single definition of a housing submarket exists" (Rae, 2015, 457), a housing submarket can generally be considered to be a set of dwellings sharing similar characteristics (Bourassa et al., 1999), often defined with a spatial contiguity constraint such that there are no multi-part spatial housing submarkets; a line can be drawn on the submarket map linking any dwelling to any other within the same submarket without crossing into a different submarket.

Past research operationalises the concept in a range of ways. Early works use existing divisions: for example Palm (1978) partitions the San Francisco-Oakland region into housing submarkets based on the districts covered by each of the

seventeen Boards of Realtors in the region. Another approach is to make use of convenient existing administrative spatial units, as done by Adair et al. (1996), who determine spatial housing submarkets by amalgamating the existing ward divisions of Belfast into larger groupings with common characteristics.

The turn of the millennium saw the development of a range of quantitative techniques for determining the spatial bounds of housing submarkets, a 'microstructural turn' in housing analysis (Smith and Munro, 2013, 2) resulting from the increased availability of "large micro-datasets that contain geo-coded details of dwellings, their characteristics and values" (Keskin and Watkins, 2017, 1447) and a concomitant advancement in the methods available to analyse these data.

Among these quantitative techniques, Bourassa et al. (1999) produce one of the first examples of cluster-based housing submarket spatial segmentation, defining spatial housing submarkets in Sydney and Melbourne using both *k*-means and Ward's method for agglomerative clustering. Kauko et al. (2002) use two neural network techniques to identify housing submarkets in Helsinki, while Helbich et al. (2013) present "a data-driven spatial regionalization framework for housing market segmentation" (ibid, 885) incorporating several techniques including multiscale geographically weighted regression, principal component analysis, *k*-means, Spatial 'K'luster Analysis by Tree Edge Removal (SKATER), and various checks of predictive accuracy.

2.2 Urban morphology

Urban morphology is the study of the physical form of cities, towns and villages. While the various definitions of urban morphology— Oliveira (2016) cites nine— differ in their details, each considers the elements of which cities are composed, particularly their "urban tissues, streets (and squares), urban plots, [and] buildings" (Oliveira, 2016, 2). By providing a descriptive language to discuss the structure of built environments, urban morphology presents a set of tools to help 'read'

urban forms, and thereby understand the effects of differing urban forms on a wide array of social, economic and environmental processes (Kropf, 2017, 10).

Urban morphology has been associated with (and hence used to understand) many different factors which vary spatially in built environments. A typical application of urban morphology is as a tool to understand the historical development of a town or city, such as Baker's (2009) study of the historical townscape of Hereford. Other applications look at the relationship between urban form and social issues, such as poverty (Vaughan et al., 2005) or public health (Sarkar, 2013); environmental concerns like heat-energy demand and efficiency (Rode et al., 2014); and concerns with both social and environmental components, such as the effect of urban morphology on birdsong loudness and the visibility of green areas (Hao et al., 2015).

2.2.1 Urban morphometrics

Traditional urban morphological research has primarily made use of qualitative methods, using records such as historical and current maps and photographs of the area in question to determine the nature of a settlement's morphology at a given point in time. These methods lend themselves to detailed examinations of the particular historico-geographical context of a given case study settlement, but are labour- and time-intensive, and cannot be easily scaled to larger geographical areas.

In recent years, an increasing availability of both appropriate data and technological tools has made possible a proliferation of quantitative urban morphology, employing an accompanying growing body of methodologies. This mirrors broader trends in geography (Arribas-Bel, 2014; Wolf and Knaap, 2019; Singleton and Arribas-Bel, 2021) and indeed in social science more broadly (Lazer and Radford, 2017), where more and more research is employing methods drawn from data science (and the methodological traditions that precede the term; see Donoho, 2017).

Among the first explicit contributions to the nascent methodology-cum-subfield of urban morphometrics (UMM), Dibble (2016) seeks to establish "a systematic, quantitative and comprehensive process of measuring, defining and classifying urban form" (ibid, vi). Since then, an expanding body of urban morphometric research has been published (Dibble et al., 2019; Araldi and Fusco, 2019; Bobkova et al., 2021), quantitatively examining urban form in a number of contexts and from a number of perspectives.

In this landscape of novel methodological approaches to quantitative urban morphology, Fleischmann et al. (2020b) survey a wide range of such studies and find a catalogue of cases in which the same term is used in multiple quantitative studies of urban morphology, but with different meanings in different studies. In an effort to overcome these terminological inconsistencies, they establish "a systematic and comprehensive framework to classify urban form characters" (ibid, 1). To this end, they introduce the *Index of Elements*, a terminological framework which distinguishes the *Index* of an urban form character (what is being measured, for example 'area' or 'number of neighbours') and the *Element* of urban form being measured (for example 'building' or 'block').

In the interest of minimising this work's contribution to the aforementioned terminological inconsistencies, this dissertation generally adopts this proposed terminological framework, and in all cases seeks to clearly define the key terms used throughout. In addition to the use of 'elements' to describe the element of urban form being measured, henceforth:

- '**character**' is used to refer to a measurable "characteristic … of one kind of urban form that distinguishes it from another kind" (ibid, 2);
- '**tessellation cell**' refers to the spatial unit produced in the process of morphological or enclosed tessellation (described in further detail in the next chapter);
- '**urban tissue**' refers to "a distinct area of a settlement in all three dimensions, characterised by a unique combination of streets, blocks/plot series,

plots, buildings, structures and materials and usually the result of a distinct process of formation at a particular time or period" (Kropf, 2017, 89);

- '**segmentation**' refers a) to the way in which an urban area is divided into spatial units through the clustering methodology described in the next chapter; and b) to the output of such a methodology, as in "this segmentation clearly delineates the Ciutat Vella";
- '**clustering**' refers more narrowly to the part of the methodology in which base spatial units are assigned to a group ('assigned a cluster label') on the basis of the statistical similarity of the characters of each spatial unit.

Past research has laid the groundwork for this study by identifying a wide range of ways in which urban morphology can be measured, and consequently by developing the computational tools to calculate these morphometric characters in a way which is scalable to a large urban area (Fleischmann et al., 2021a).

2.2.2 Defining the spatial unit

Fundamental to a spatial segmentation produced using a clustering methodology (be that according to demography, morphology, or any other character or set thereof) is the smallest spatial unit used. Dibble et al. (2015) define this as the Operational Taxonomic Unit (OTU), borrowing the term used in the biological field of morphometrics to describe the smallest unit used when comparing organisms' characteristics in the process of taxonomic classification. In biology, the OTU is almost always the individual organism, but in urban morphology the choice of OTU is less straightforward.

In his seminal study of the urban morphology of Alnwick, Northumberland, Conzen (1960) refers to the smallest unit of analysis as a 'plot', "a unit of land use … physically defined by boundaries on or above ground" (Conzen, 1960, 5). These plots are then grouped into 'plan-units': morphologically distinct areas of the town, of which Conzen identifies a taxonomy of 13 major and 49 subtypes. The use of plots has been criticised, however, as "more or less ambiguous"

(Kropf, 1997, 1), as the multiple definitions with which the term is used give rise to different—and sometimes contradictory—plot geometries (Kropf, 2018). Mehaffy et al. (2010) study urban morphology through the structure of 'sanctuary areas', defined as "the area between major thoroughfares" (ibid, 23). Other studies have sought to somewhat circumvent the problem of determining the smallest spatial unit on the basis of current use by instead using an arbitrary unit, such as a regular grid (Jochem et al., 2021; Mercadé Aloy et al., 2018; Rode et al., 2014).

Seeking a consistent and universally applicable base spatial unit, Fleischmann et al. (2020a) propose 'morphological tessellation' (MT) as a method for deriving such a unit for use in urban morphometrics. Using only building footprints, the method uses Voronoi tessellation to derive the 'morphological cell', an alternative to the plot as traditionally conceived in studies of urban morphology. Building on this morphological tessellation, Arribas-Bel and Fleischmann (2021) introduce 'enclosed tessellation' (ET) as an alternative spatial unit. In addition to morphological tessellation's use of building footprints, ET incorporates additional barriers to delineate some cell boundaries such as roads, rivers, and railways.

2.2.3 Approaches to urban morphological spatial segmentation

Within urban morphometric research there are multiple approaches to quantitatively partitioning (usually urban) spaces on the basis of their morphology. Both approaches discussed below first divide space into the base spatial units discussed above, before grouping these on the basis of morphology to form novel segmentations.

In methodologies using supervised machine learning techniques, the researcher must first define the classes into which they wish the study area to be categorised. Each base spatial unit is then assigned to one of these existing classes based on its statistical similarity (in the characters of interest) to each class. As an example, Colaninno et al. (2011) present a method to classify Barcelona into seven different morphology-based typologies.

Conversely, methodologies using unsupervised machine learning techniques, the prevailing approach to urban morphological spatial segmentation, do not require this a priori specification of classes into which the base spatial units should be grouped. Examples of this approach include work by Fleischmann et al. (2021b).

2.3 Research gaps

As presented above, urban morphometrics is an actively developing field of research, and as such there is a limited literature presenting quantitative methodologies for producing urban morphology based spatial segmentations, and assessing the methodological parameters which affect these segmentations. In particular, to date no research has incorporated urban morphology into a delineation of housing submarkets.

More generally, there is a need for spatial units which reflect the heterogeneity of variables as they are distributed across urban spaces. Because urban morphology is putatively correlated with a wide range of spatially varying factors (as outlined above), spatial units which reflect a city's urban morphology could be successful at capturing the variation in a wide range of different variables. This may offer one means to reduce the deleterious effects of the MAUP incurred when arbitrary administrative geographies are used to aggregate spatial data.

3
Methodology

This chapter describes the methods used to complete the analyses and to obtain the results of this research. First the general process followed to generate spatial segmentations based on urban morphology is described, before the specific differences between seven different particular segmentations and the methodology used to compare these are each set out.

The aim of the spatial segmentation process is to partition a city into areas with similar urban morphology. To do this, various components of urban morphology ('characters') are measured at a small scale throughout the city. The 'base spatial unit' is the small-scale area at which these characters are (generally) measured and reported. The units with similar values for these morphometric characters are then grouped ('clustered'), thereby aggregating the base spatial units into larger areas with similar urban morphology.

3.1 Base spatial units

In the process of constructing each spatial segmentation, the urban area is first divided into many small spatial units (the OTU as discussed in the previous chapter). These are then used as the building blocks from which the consequent classification of urban space is constructed. When these base spatial units are

changed, the ways in which urban morphology is measured and clustered are changed and thus the resultant spatial segmentations change.

3.1.1 Morphological tessellation

As discussed in the previous chapter, morphological tessellation uses the building footprints to generate a large number of fine-grain spatial units suitable for measuring features of urban morphology, consequently clustering these to create segmentations which reflect different urban tissues.

The spatial units created (MT cells) are produced by generating Voronoi polygons around each building, thereby exhaustively dividing the space in question according to the nearest building to any given point. In this dissertation this process is implemented in Python using the `momepy` package (Fleischmann, 2019).

Because of the minimal data requirement (the only data required are building footprints and the boundary of the area in which to generate the tessellation), MT can be used in a wide range of geographical contexts. For example, MT cells could be generated using building footprints derived purely from satellite or aerial imagery.

3.1.1.1 Data: buildings from the Spanish Cadastre

The buildings used to generate tessellation cells are taken from open data provided by the Spanish Cadastre (*Dirección General del Catastro*), an administrative registry which records the physical and legal characteristics of every property in Spain. Because of this administrative purpose and governmental mandate, the dataset has near-complete coverage at the national level. All unprotected data for each property can be downloaded free of charge, either directly from the Cadastre website or through an API or tool such as the Spanish Inspire Catastral Downloader plugin for QGIS (Soriano, 2021). The data provide detailed information about each building, including the building's exact location, year of construction, height, footprint, and the number of residential units within the building.

These data have been used in previous geographic data science research: Arribas-Bel et al. (2019) used them to delineate improved boundaries for urban areas, while Carpio-Pinedo et al. (2021) used the data to map land use mix as 'walkable trips'. Because the dataset is formulated in accordance with the European INSPIRE (Infrastructure for Spatial Information in the European Community) directive, replicability studies could be relatively straightforward for other European countries with equivalent data.

The data are provided as three different datasets: Cadastral Parcels (CP) describe "the basic unit of ownership" (i.e. the plot) irrespective of whether it has been built upon; Addresses (AD) provides identifiers for the location of each property; and Buildings (BU) gives geospatial information about each building.

3.1.2 Enclosed tessellation

Enclosed tessellation (ET) is an alternative to morphological tessellation which makes use of road network data in addition to building footprints.

The first step in creating the ET is to generate enclosures from the street network and city boundary. These areas are similar to the Sanctuary Area used in previous studies of urban form (Mehaffy et al., 2010; Dibble et al., 2019), albeit they are constructed of all drivable roads and not solely of main roads. Within each of these enclosures, Voronoi polygons are generated around each building, partitioning the space within each enclosure according to the nearest building to create enclosed tessellation cells. Like MT, the ET cells were generated in Python using `momepy`.

Because this is ultimately a segmentation of housing submarkets, ET cells containing no buildings are dropped at this stage. This means that unlike MT, the segmentation is not spatially exhaustive: certain areas without buildings are excluded from the final classification (commonly roundabouts and the central reservations of dual carriageways). This difference is discussed at greater length in the Discussion. Figure 3.1 shows the process of generating enclosed tessellation cells.

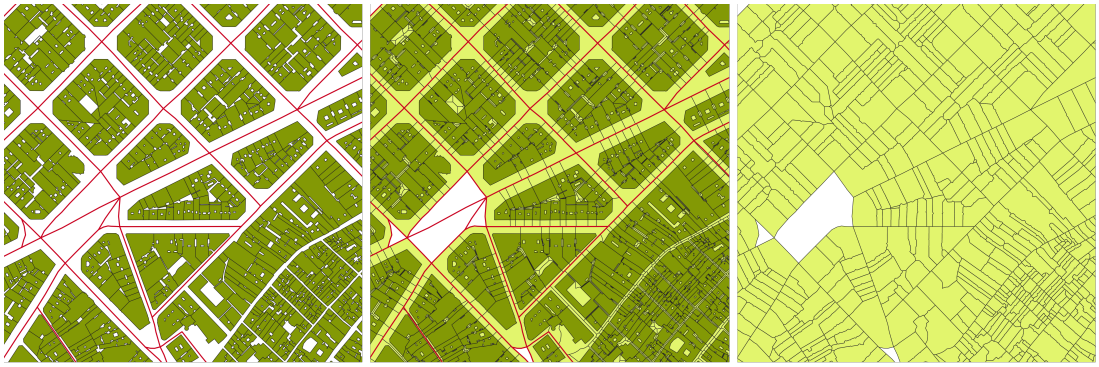

Figure 3.1: Generating enclosed tessellation: input buildings (dark green) and roads (red) data; plotting Voronoi polygons around each building within the enclosures formed by the roads; final tessellation.

3.1.2.1 Data: roads from OpenStreetMap

The road network is taken from the collaborative online OpenStreetMap (OSM) project. OSM is a relatively novel source of volunteered geographic information, but since its 2004 inception has already been used in a wide range of geographical research (Jokar Arsanjani et al., 2015). Although there have been questions raised about the quality and completeness of crowdsourced geospatial data (Mooney and Minghini, 2017), studies of OSM's road network have found a high level of completeness (Barrington-Leigh and Millard-Ball, 2017) and so this is unlikely to be a problem, particularly in prominent urban areas in an economically developed country such as Spain.

The `OSMnx` Python package (Boeing, 2017) is used to query the Overpass OSM application programming interface (API) and return all roads marked as drivable by general vehicular traffic[1] within a 100 metre buffer around the boundary of the city.

[1]Specifically, the function used to return 'drivable' roads removes all routes tagged as `abandoned`, `bridleway`, `bus_guideway`, `construction`, `corridor`, `cycleway`, `elevator`, `escalator`, `footway`, `path`, `pedestrian`, `planned`, `platform`, `proposed`, `raceway`, `service`, `steps`, `track`, `alley`, `driveway`, `emergency_access`, `parking`, `parking_aisle`, or `private`; `motor_vehicle = no`, and `motorcar = no`. See the source code for further detail.

3.1.3 Blocks

The block is a similar unit to the enclosures used in the process of generating ET cells, being delineated by the street network provided. They differ from enclosures in that blocks are built back up from ET cells, and so do not include those spaces surrounded by streets but not containing any building. Blocks are again generated using `momepy`.

3.1.4 H3

H3 (Brodsky, 2018) is a hierarchical, hexagonal, global grid system initially developed by Uber Technologies to process and visualise the large amounts of spatial data they collect. It completely partitions the surface of the Earth into hexagons derived from a projection of the globe as a spherical icosahedron. For this reason each hexagonal H3 cell has approximately the same area as any other at the same resolution; the H3 grid does not suffer from the same problems of latitude-dependent size distortion a square grid on a Mercator projection would have. H3 was implemented using the `h3` Python bindings. A cell resolution of 10 is used, at which level each cell has an edge length of 65.9 metres and an area of 15,047.5 m^2.

Figure 3.3 compares each of the four base spatial units employed in the dissertation.

3.2 Measuring morphometric characters

The characters used as inputs when carrying out spatial segmentation are the data which determine the ultimate segmentations, and so it is imperative that these successfully capture the desired features: in this study, the morphology of an area.

In clusterings which use the full set of characters (see Table 3.1) there are three different elements on which characters are measured: buildings, tessellation cells, and blocks. Some characters—such as the area of the element—can be measured on all three elements, while others—such as the coverage area ratio,

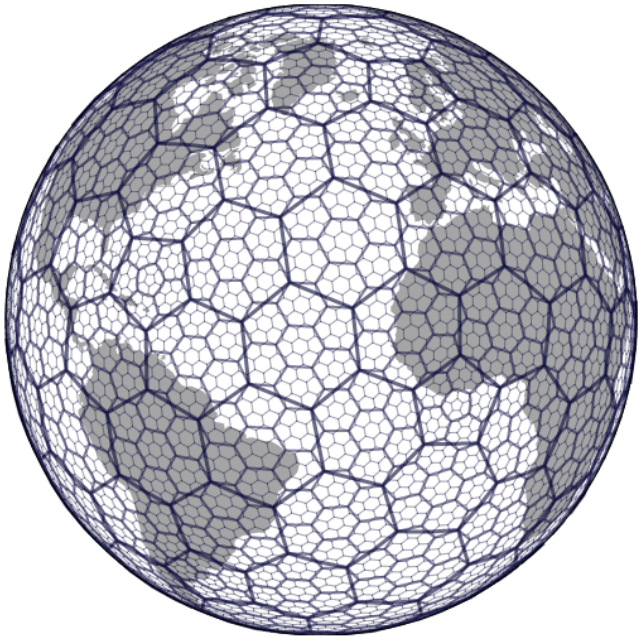

Figure 3.2: H3 partitions the globe into (mostly) hexagonal cells. Source: Brodsky (2018).

which reports the proportion of a tessellation cell covered by a building—can only be measured on (a) certain element(s).

3.2.0.1 Data: building information from the Spanish Cadastre; roads from OpenStreetMap

The two data sources used to generate morphometric characters are those already introduced: the Spanish cadastral dataset, and the drivable roads from OSM.

The cadastral dataset provides multiple forms of geospatial information about each building: `Building` describes the geometry and attributes of each building, while the more granular `BuildingPart` gives separate information about the constituent parts of buildings, including the height of each part. This height (used to calculate several morphometric characters) is derived from the `numberOfFloorsAboveGround` attribute, which is multiplied by three to approximate a height value in metres (assuming that each floor is approximately three metres tall). As the attribute is only provided for `BuildingParts`, a spatially weighted average of the heights of

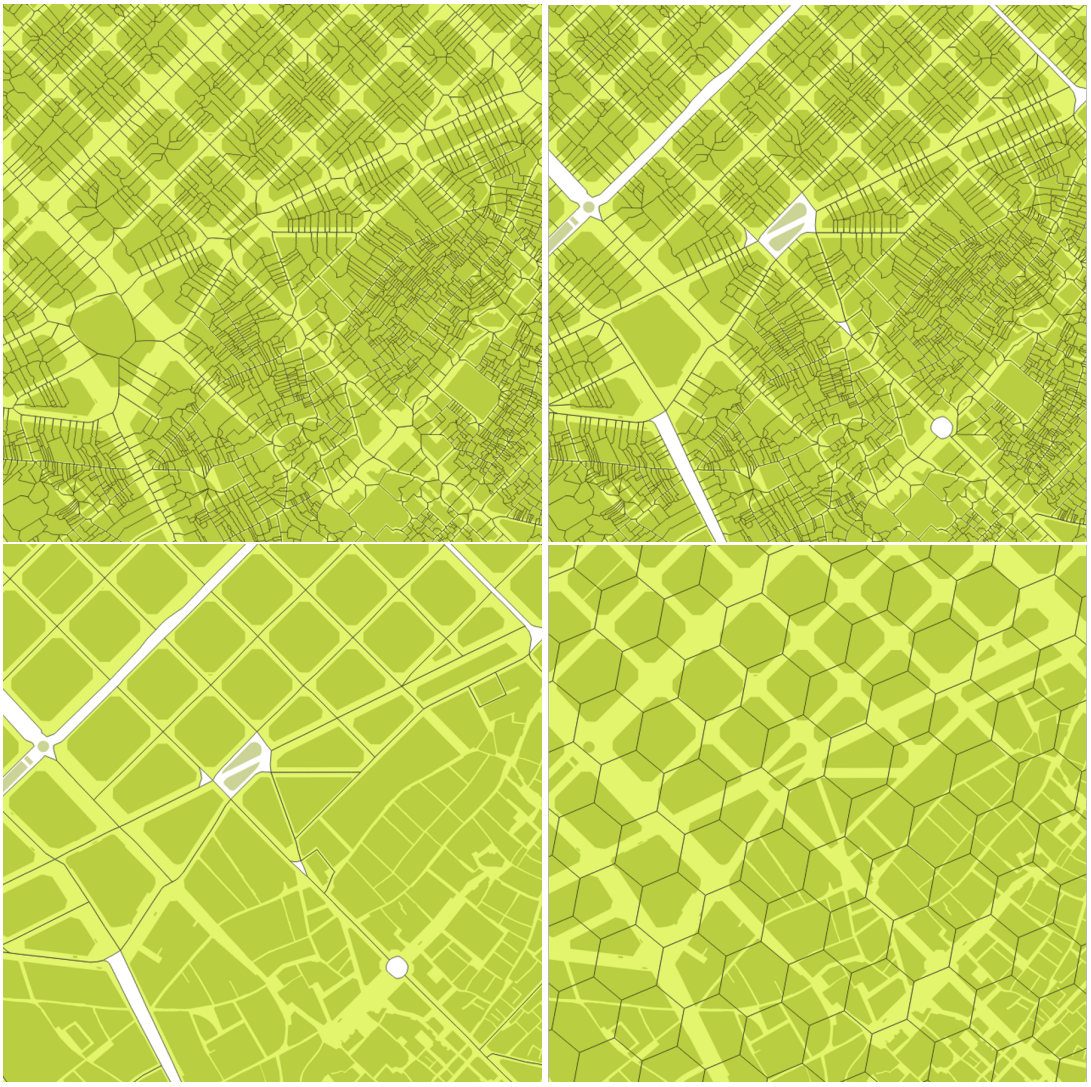

Figure 3.3: Different base spatial units: morphological tessellation, enclosed tessellation, blocks, and H3.

each building's constituent parts is calculated to produce height information for each building. Before this was done, the `numberOfFloorsAboveGround` attribute was manually inspected and corrected in certain cases: for example some building parts reported a value of 106, despite the tallest building in Barcelona only having 44 floors.

3.2.1 Primary characters

Each primary character quantifies an aspect of urban morphology with respect to a particular element (as described above). As wide a range of characters as

reasonably possible is initially calculated in order to capture as many different aspects of urban form as possible. The primary characters are calculated using `momepy`. The initial set of characters is detailed in Table 3.1. The character descriptions are primarily adapted from Fleischmann (2021) Fleischmann (2021) and `momepy` documentation.

3.2.2 Selecting characters

The information provided by each character must vary spatially, that is to say that if the values of a character are randomly distributed throughout a city or do not vary at all, the character is not useful in a *spatial* segmentation, which is premised on dividing the city based on the ways in which given input values vary over space. For this reason, the spatial autocorrelation of each character throughout the study area is assessed by calculating the Global Moran's I. If this is not statistically significant, the character is not spatially autocorrelated and therefore is discarded. All characters used in the segmentations were found to exhibit significant spatial autocorrelation (positive Moran's I value and $p < 0.005$).

A degree of correlation is expected between certain characters, but if multiple characters are highly correlated this may be indicative of two or more characters representing essentially the same concept, thus not providing a significant amount of additional information and skewing the results of the clustering. For this reason, the degree of correlation between each of the characters is inspected. As the distributions of values in each character is not necessarily normal, Spearman's rank correlation coefficient is computed to quantify the degree to which each character is correlated to each other character. Figure 3.4 provides an example of this process for the characters generated on ET cells.

On the basis of Figure 3.4, four characters were removed on the basis of exhibiting too high a degree of collinearity: `blg_CentroidCornersMean`, `blg_Perimeter`, `blk_Perimeter`, and `blk_CompactnessWeightedAxis`.

Table 3.1: Initial set of urban morphometric characters. Those included in the H3 clustering are indicated with '*'.

Variable name	Index	Element	Description	H3
blg_Height	Height	Building	Height of building	*
blg_FloorArea	Floor area	Building	Floor area of building	*
blg_Area	Area	Building	Area of building	*
blg_Volume	Volume	Building	Volume of building	*
blg_Perimeter	Perimeter	Building	Perimeter of building	*
blg_CourtyardArea	Courtyard area	Building	Area of holes in buildings (aka courtyards)	*
blg_FormFactor	Form factor	Building	Compactness of building	*
blg_VolumeFacadeRatio	Volume to façade ratio	Building	Ratio of building volume to area of the façade, a proxy for volumetric compactness	*
blg_CircularCompactness	Circular compactness	Building	Ratio of area of building's enclosing circle to its area	*
blg_Corners	Corners	Building	Number of corners the building has	*
blg_Squareness	Squareness	Building	Squareness of building	*
blg_EquivalentRectangularIndex	Equivalent rectangular Index	Building	Deviation of building from an equivalent rectangle	*
blg_Elongation	Elongation	Building	Ratio of shorter to longer dimension of the minimum bounding rectangle, a proxy for the deviation of the shape from a square	*
blg_CentroidCornersMean	Centroid to corner mean	Building	Mean distance from building centroid to its corners	*
blg_CentroidCornersSD	Centroid to corner standard deviation	Building	Standard deviation of distance from building centroid to its corners	*
blg_Orientation	Orientation	Building	Orientation of the building	*
blg_SharedWallsRatio	Shared walls ratio	Building	Ratio of shared walls to total perimeter	*
blg_CellAlignment	Cell alignment	Building	Calculate the difference between cell orientation and orientation of building	
tess_Orientation	Orientation	Tessellation cell	Orientation of the tessellation cell (calculated via the bounding rectangle)	
tess_LongestAxisLength	Longest axis length	Tessellation cell	Length of the longest axis of tessellation cell	
tess_Area	Area	Tessellation cell	Area of tessellation cell	
tess_CircularCompactness	Circular compactness	Tessellation cell	Ratio of area of tessellation cell's enclosing circle to its area	
tess_EquivalentRectangularIndex	Equivalent rectangular Index	Tessellation cell	Deviation of tessellation cell from an equivalent rectangle	
tess_WeightedNeighbours	Weighted neighbours	Tessellation cell	Number of tessellation cell neighbours divided by the cell perimeter	
tess_CoverageAreaRatio	Coverage area ratio	Tessellation cell	Proportion of tessellation cell covered by a building	
tess_FloorAreaRatio	Floor area ratio	Tessellation cell	Ratio of floor area to tessellation cell area	
blk_Area	Area	Block	Area of block	
blk_Perimeter	Perimeter	Block	Perimeter of block	
blk_CircularCompactness	Circular compactness	Block	Ratio of area of block's enclosing circle to area of block	
blk_EquivalentRectangularIndex	Equivalent rectangular Index	Block	Deviation of block from an equivalent rectangle	
blk_CompactnessWeightedAxis	Compactness-weighted axis	Block	Compactness-weighted axis of block, a proxy of permeability of an area	
blk_Orientation	Orientation	Block	Orientation of the block (calculated via its bounding rectangle)	
blk_WeightedNeighbours	Weighted neighbours	Block	Number of block neighbours divided by the block perimeter	
blk_WeightedBuildings	Weighted buildings	Block	Number of buildings within the block divided by the block area	

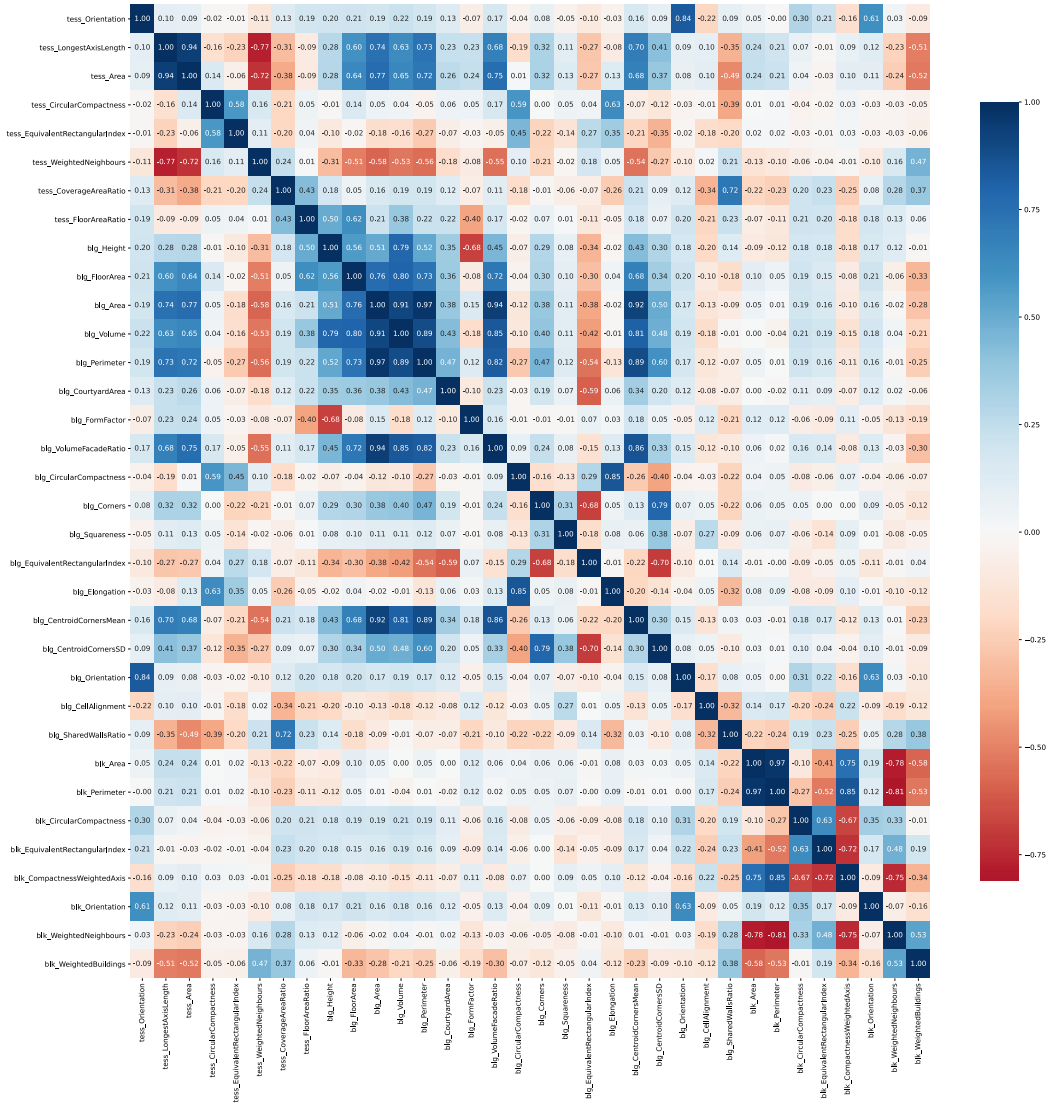

Figure 3.4: Correlation matrix of morphometric characters.

3.2.3 Contextual characters

As there is no explicitly spatial input to the clustering algorithm used (for all but one of the segmentations), there is no guarantee that any clusters generated from purely the primary characters will form spatially contiguous areas (that is, the eventual spatial segmentations). Furthermore, since the base units are small (particularly in the case of ET/MT cells, which only delineate the area around a single building) they cannot uniquely capture the morphology of the urban tis-

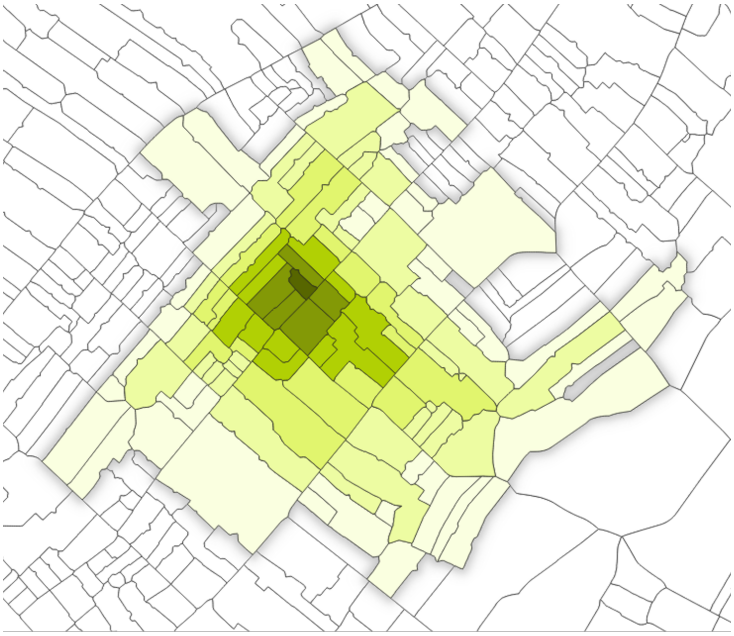

Figure 3.5: An example of the topological neighbours used to create contextual characters for a tessellation cell: the darkest green is the original cell and each shade lighter represents one extra degree of topological distance from this cell.

sue to which they belong. The use of spatial weights to incorporate information about the neighbouring spatial units therefore expands the geographic extent of morphometric information supplied while preserving the spatial granularity provided by the base units in a way which would not be achieved by simply aggregating the characters to less granular spatial units.

When used with ET/MT cells, the contextual characters also reflect how humans perceive (urban) space. While the measured geographic distance may be equal, a person's mental conception of the distance between two points will be different if the area between these is a built environment with a high density of streets and buildings, compared to an open landscape with fewer roads or features. Using topological steps based on ET/MT cells can therefore be seen as a more theoretically sound reflection of the relationship between different elements than a simple distance buffer or k nearest neighbours, and a better way to describe the morphology of the vicinity of a given cell (Fleischmann et al., 2020a).

To generate contextual characters, a spatial weights matrix is created for the base spatial unit: this records which cells are neighbours of a given cell. The

definition of 'neighbour' in this context depends on the number of topological steps stipulated when computing the spatial weights (the order of contiguity, k). If $k = 1$, queen contiguity-based spatial weights are used such that any cell sharing a common edge or vertex is counted as a neighbour. If $k \leq 2$, the spatial weight will define as a 'neighbour' both the cells included as neighbours in the $k = 1$ criterion and *their neighbours* according to the same criterion, and so on such that as k increases so does the maximum number of topological steps between a cell and its 'neighbour' (see Figure 3.5).

For each character and each cell, the interquartile mean (IQM) is then calculated from that character's values in all cells defined as a neighbour in the spatial weights matrix provided. The IQM is calculated by taking the mean average of all values between the lower and upper quartile when ordered, and as such is less affected by outliers than a simple mean of all values.

3.3 Clustering

In all but one segmentations presented in this dissertation, the clustering of cells into urban tissues was carried out using a Gaussian Mixture Model (GMM), a probabilistic derivative of k-means. GMM models the distribution of each dimension (in our case each (contextual) character) of each cluster as a Gaussian distribution, rather than using k-means' simpler distance-based model. This allows clusters to have different shapes (in the hyperspace in which they are clustered), for instance ellipses, rather than classifying each point according to its nearest cluster centroid (as measured by Euclidean distance in n-dimensional hyperspace, where n = the number of characters). The GMM is implemented with the `GaussianMixture` algorithm from the scikit-learn Python package (Pedregosa et al., 2011).

Agglomerative Clustering is an alternative clustering algorithm, employed because it allows the incorporation of a spatial constraint; it can be stipulated that

when mapped, the clusters generated must all be spatially contiguous[2]. Unlike GMM, Agglomerative Clustering requires a spatial weights matrix: this allows it to operationalise the spatial constraint by providing a definition of which cells are and are not counted as neighbours.

3.3.1 Determining n

GMM requires the number of clusters (n) to be specified a priori, a decision which must satisfy both statistical and theoretical considerations.

An elbow plot is a typical method used to determine the appropriate number of clusters in an unsupervised clustering of this nature. The x-axis plots the number of clusters, while the y-axis charts a metric quantifying the goodness of fit of a clustering with n clusters. The Bayesian Information Criterion (BIC) is one such measure, providing a quantification of the clustering fit, penalising a higher number of clusters in order to deter overfitting. Figure 3.6 shows an example elbow plot charting the number of clusters against the BIC for a segmentation using ET cells with contextual characters from 5th order spatial weights. The plot serves to illustrate the diminishing returns of improvements in the clustering fit as n increases, and therefore acts as a guide to choosing the 'optimal' n.

It is important that diagnostic statistics and methods such as the BIC not be blindly followed and seen to unambiguously indicate the optimal n. While there may be many statistically distinct clusters may be distinguished, these may not reflect real/clear differences in urban morphology.

3.4 Comparing different segmentations

Many different segmentations were carried out varying different parameters. Seven of these resultant spatial segmentations are reported below: these have been se-

[2]Except, that is, in cases where the boundary of the urban area itself contains 'islands' (literal or figurative). As an example, the spatial units which comprise the island at the port of Barcelona (a cruise terminal connected by bridge to the mainland) are assigned to the same cluster as much of the mainland from which they are separated by water (thus not fulfilling the spatial contiguity criteria).

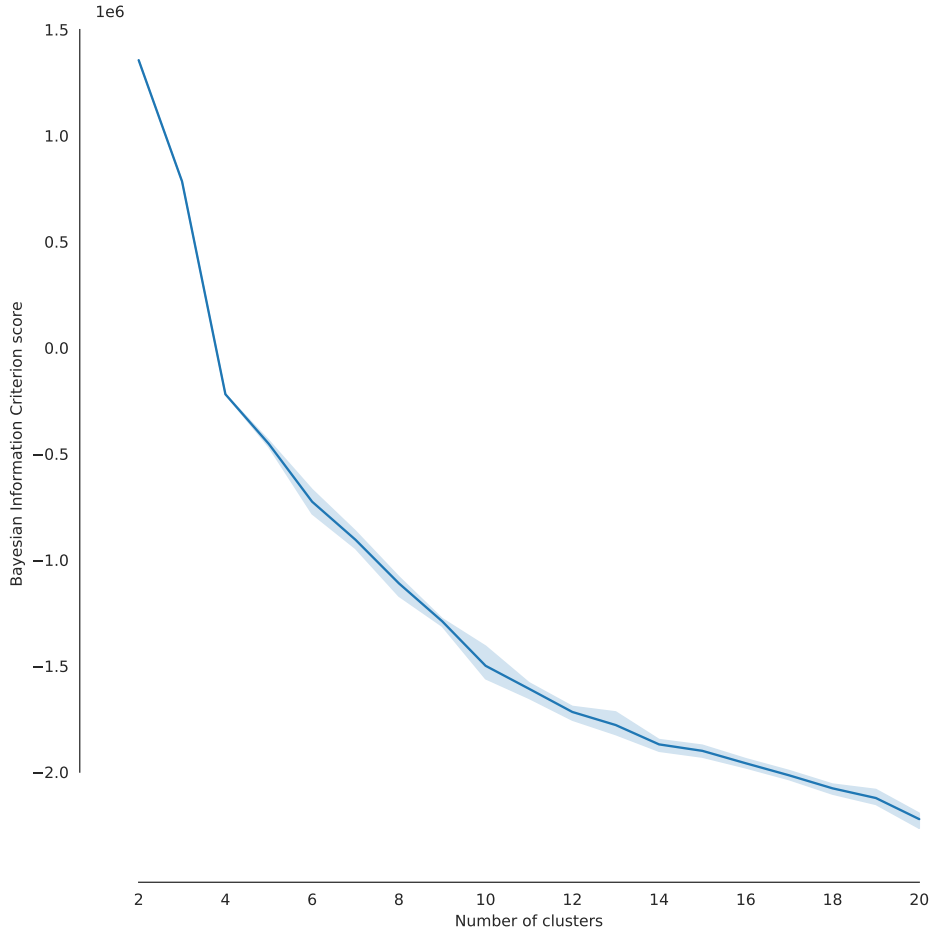

Figure 3.6: An elbow plot for a segmentation with ET cells and contextual characters from 5th order contiguity weights. The shaded area represents the 95% confidence interval.

lected to showcase a range of differing possible approaches to certain elements of the methodology, and the effects these methodological changes have on the segmentations generated. Figure 3.7 presents a graphical comparison of the methodologies used to generate each of these segmentations.

3.4.0.1 Morphological tessellation

This segmentation generates all characters in Table 3.1 using MT cells. 5th-order contiguity spatial weights are then used to compute the IQM for each

character in the neighbourhood of each MT cell. These values are used as inputs to a GMM clustering to generate the segmentation.

3.4.0.2 Enclosed tessellation

This segmentation is identical to the morphological tessellation segmentation, save for its use of enclosed tessellation cells in place of morphological tessellation cells. While seemingly minor, carrying out a segmentation with this as the sole difference allows any differences between the two spatial units when carrying out an UMM-based clustering to be methodically examined.

ET transposed to block geometry As the ET cells are small geometries which may for example separate buildings sharing a wall, any segmentation using this geometry may have 'messy' borders between clusters, even when using high-order contiguity weights. To reduce this issue, this segmentation takes the output from the above enclosed tessellation segmentation and classifies each block according to the cluster label which takes up the greatest proportion of the block's area. This is done using the `area_join` function in the `tobler` package in Python (Knaap et al., 2021).

ET transposed to H3 geometry As with the above transposition to block geometry, but with the level 10 H3 cells.

3.4.0.3 H3 'basic'

The 'basic' H3 clustering entirely eschews the use of ET or MT cells, instead only calculating the characters marked as '*' in Table 3.1, all of which have the building as their element. The value of each character in each H3 cell is then calculated from a spatially weighted average of the values of the buildings within the cell, computed using `tobler`'s `area_interpolate` function.

3.4.0.4 H3 with ET characters

This segmentation uses the full set of characters generated using ET cells, transposed onto the H3 geometry. A first order contiguity weight (using the H3 cell as the unit) is then used to create contextual characters, which are used as the input to a GMM clustering to generate the final segmentation.

3.4.0.5 Spatially constrained MT

This segmentation uses the full set of characters generated using MT cells, but substitutes the GMM algorithm for scikit-learn's `AgglomerativeClustering` regionalisation algorithm. This stipulates that all clusters must be spatially contiguous.

3.5 Assessing segmentations

For each segmentation, two distinct—albeit related—factors should be assessed: how well the segmentation captures differences in urban form types; and how well the segmentation captures the variation in other variables of potential interest, such as property prices.

3.5.1 Relation to urban morphology

The former of these assessments is significantly more challenging, not lending itself to any obvious quantitative form of validation and hence requiring a greater degree of subjective judgement. Past studies which segment cities on the basis of their urban morphology use "visual observation and personal knowledge of the city" (Fleischmann et al., 2021b, 20) to assess the validity of their segmentations, but this is challenging in cases when the author does not have this personal knowledge of the city in question (as is the case in this dissertation). In lieu of this, each segmentation is mapped over the outlines of the city's blocks (as given by the cadastral data), allowing a visual comparison of the segmentation to underlying urban structure. Note that these outlines should not be

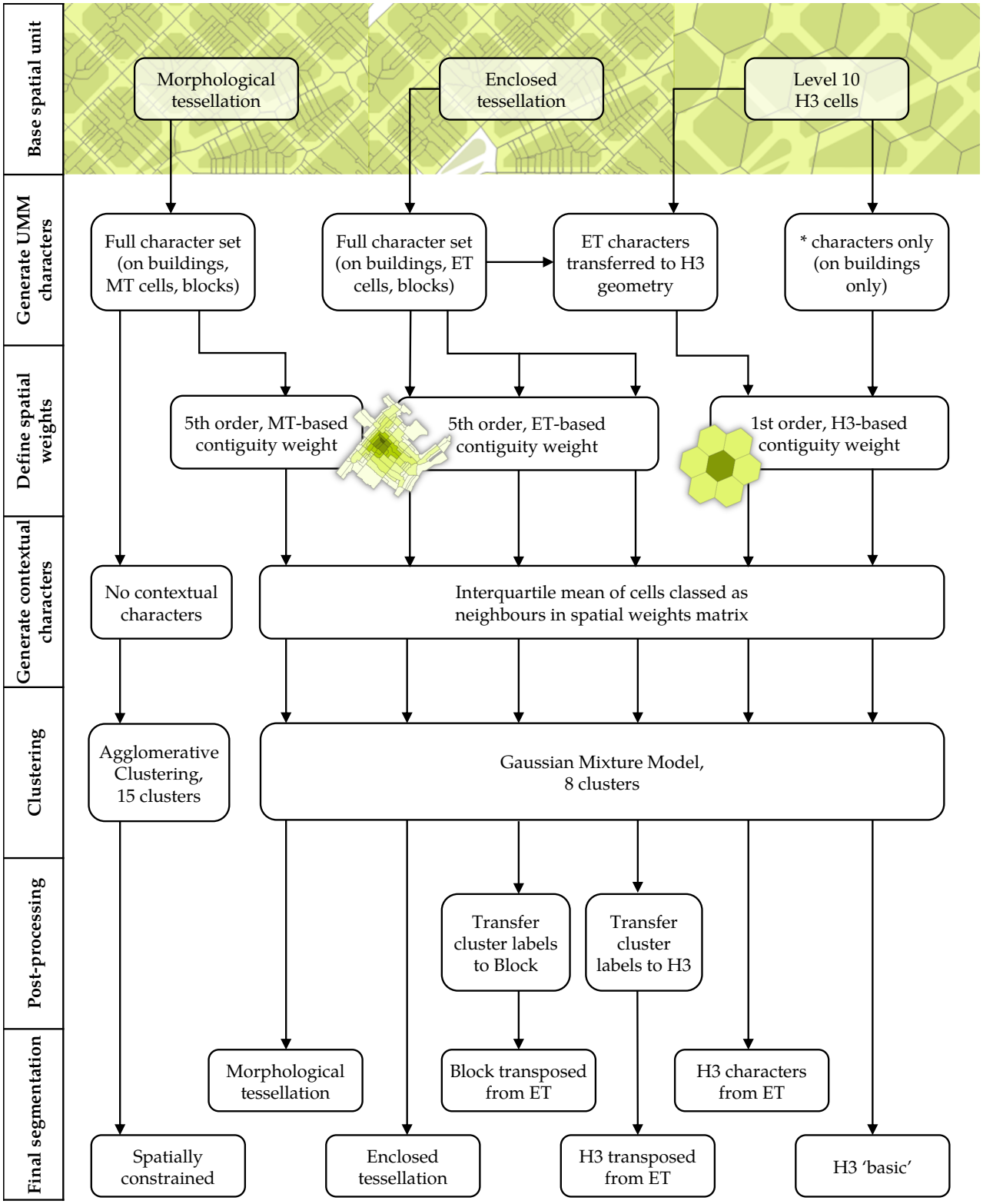

Figure 3.7: A flowchart showing the methodological differences in the construction of each segmentation.

mistaken for those of the base spatial units that have been clustered to produce the segmentations.

3.5.2 Relation to property prices

Fleischmann et al. (2021b) also validate their clustering by measuring the correlation with other urban dynamics expected to correlate (the age of buildings, land use, and qualitative classification of urban form in official planning documents). This principle of validation by comparison with a correlate is echoed in the following section, which assesses the degree to which the segmentations produced capture variation in property prices across a city.

3.5.2.1 Data: house price indices

In addition to the cadastral dataset, idealista provide their own proprietary real estate data, which is used to assess the utility of the segmentations generated as proxies for housing submarkets. The data provide the average sale price of all properties within a cadastral parcel, and are sourced from idealista's online real estate marketplace listings. Figure 3.8 shows the coverage of the house price indices data used: those cadastral parcels with values for the average sale property price are blue, while those without this information are red.

When making this evaluation, clarity about the spatial unit actually being assessed is essential. In all of the novel segmentations reported, there is a distinction between the *types* identified, and the distinct *polygons* into which the city is partitioned. In the following analysis, the *types* describe each of the numbered clusters (8 of them in most segmentations) as integrated units of comparison, irrespective of their geographical location. This means that two areas on opposite sides of a city—both classified as having the same morphological type and thus assigned to the same cluster—will be counted as belonging to the same group.

Conversely, *polygons* treat each individual polygon in the segmentation as a separate group, such that two non-contiguous areas with the same initial clus-

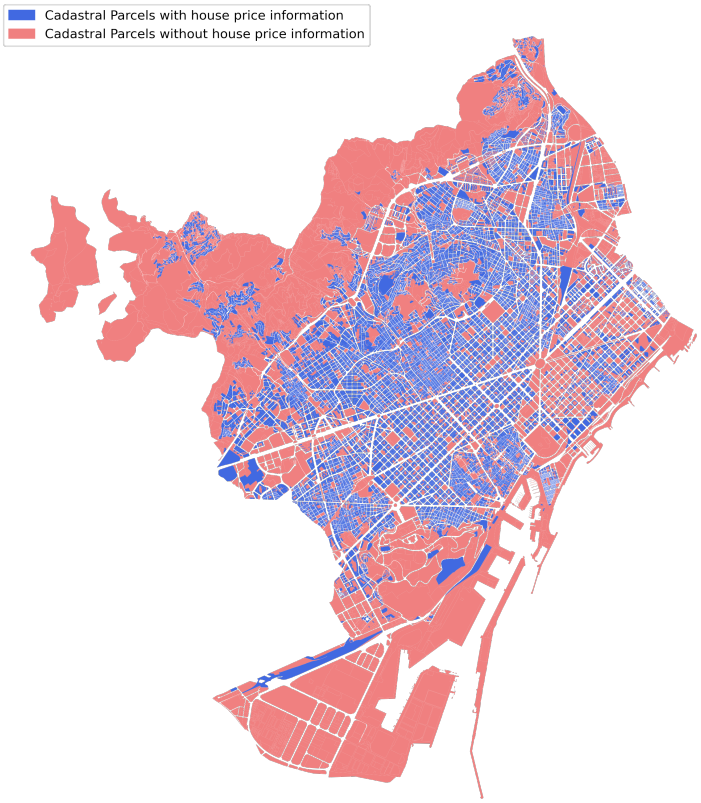

Figure 3.8: A map of the spatial coverage of the house price indices data in Barcelona.

ter label will be counted as belonging to separate groups, and hence be treated separately when measuring within-group variation[3].

For both spatial units, (types and polygons), the relationship between each segmentation and variation in property prices is quantified using the quartile coefficient of dispersion [QCoD; Zwillinger and Kokoska (1999), 17]. As the name suggests, this metric measures the dispersion of values within each unit, specifically that of the CP-level idealista house sale price indices described above.

A lower QCoD for a type/polygon, and a lower average QCoD for all the types/polygons within a segmentation is therefore indicative of units which contain more similar properties, and therefore can be seen to better represent housing submarkets. The QCoD is calculated as:

[3]In GIS terms, types can be thought of as MultiPolygons or a multipart feature: they may consist of several distinct geometries; while polygons can be thought of as the Polygons resulting from an 'explode' operation on a multipart feature: each polygon is one spatially contiguous geometry.

$$QCoD = \frac{Q_3 - Q_1}{Q_3 + Q_1}$$

Where Q_1 and Q_3 denote the lower and upper quartile values, respectively. Its use of the quartile values ensures that the final statistic is not skewed by extreme outliers, such as one very expensive property in an otherwise inexpensive neighbourhood, while still communicating the spread of values within a given type or polygon. As some of the individual polygons will comprise only very small areas (as discussed in the next chapter), polygons containing fewer than ten cadastral parcels with attached house price data are excluded from analyses which calculate and report the polygon-level QCoD.

In addition to the types and polygons from each of the seven segmentations, the QCoD is also computed for existing spatial segmentations which may be used to represent spatial housing submarkets, allowing a comparison of the novel segmentations to existing spatial units. These existing spatial units are:

1. The **neighbourhoods** are Barcelona's 73 *barris* (Catalan) / *barrios* (Spanish).
2. The **districts** are the 10 *districtes municipals* (Catalan) / *distritos municipales* (Spanish) into which Barcelona is divided.
3. The '**idealista polygons**' are taken from the units currently used internally within idealista as a bespoke spatial unit for use in analyses involving geographical aggregations of data: Barcelona contains 69 of these units.

The boundaries used for the neighbourhoods and districts are those provided by the Ajuntament de Barcelona on the Open Data BCN website (Institut Municipal d'Informàtica, 2017), while the idealista polygons have been provided by the company.

Because different segmentations will produce units of different areas, which will in turn affect the potential dispersion within these units, for each segmentation the mean of the areas of types/polygons in the segmentation is reported, along with a boxplot showing the distribution of these type/polygon areas.

4 Results

This chapter reports the key results of the research presented in this dissertation. First, the different segmentations introduced in the previous chapter are reported, and the relationship of these to patterns of urban morphology is qualitatively examined. Second, the results of the quantitative evaluation of the relationship between each segmentation and house prices are reported.

Further supplementary results are reported in the Appendix.

4.1 Relationship to urban morphology

Perhaps the most intuitive way to compare the seven segmentations presented in this dissertation is graphically. Figures 4.1 through 4.7 map the way each segmentation partitions the city of Barcelona. While the scale of these city-level maps limits the amount of detail which each can show, sites of particular interest in certain segmentations are highlighted throughout the Discussion.

It should be noted that clusters have been coloured with the same colours where a clear equivalence is perceptible (for example when identified, the cluster most closely corresponding to the Ciutat Vella is pink), but these are so coloured for ease of comparison only and should not be taken to indicate any formal connection between the clusters produced by different segmentations.

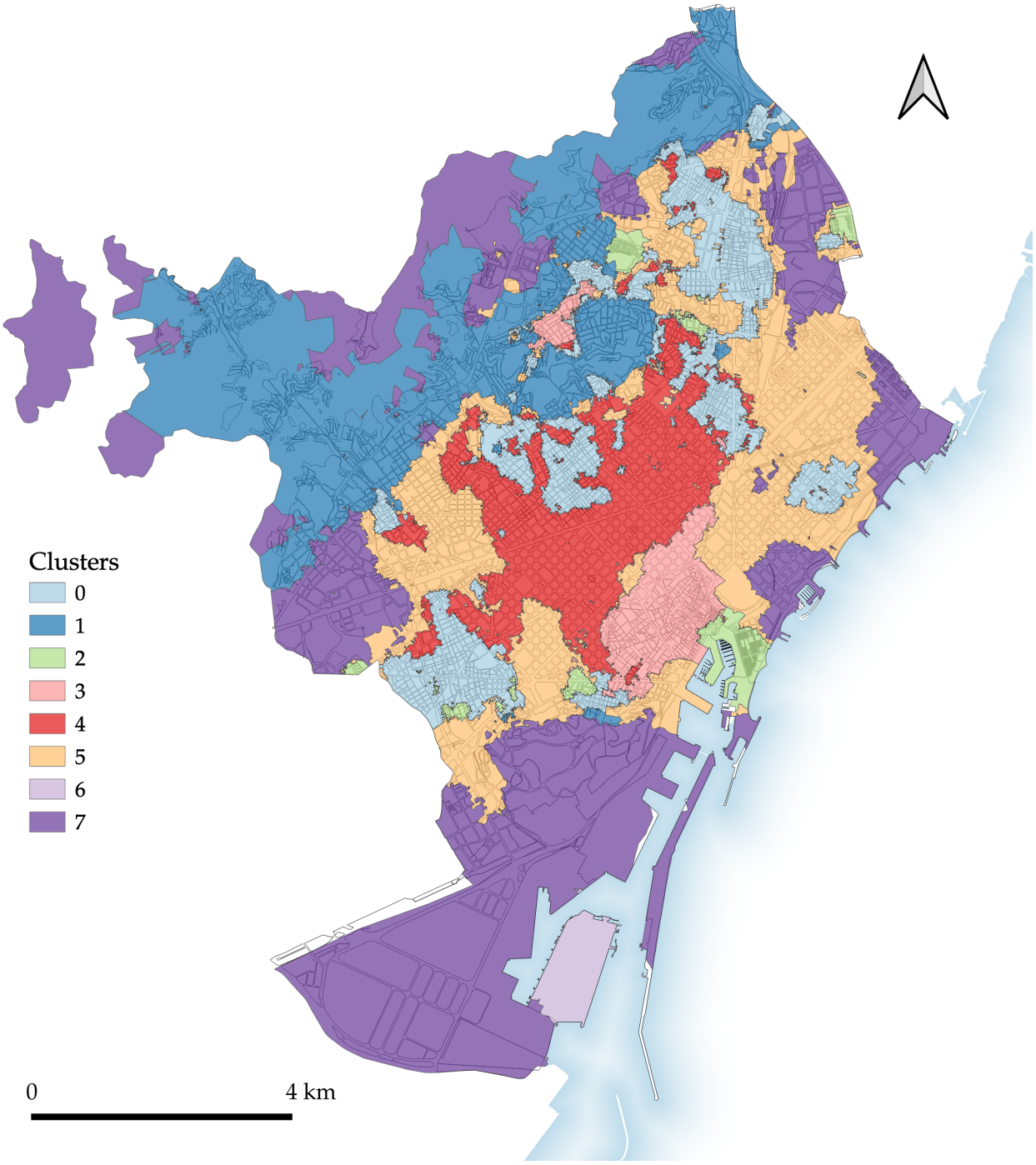

Figure 4.1: Segmentation 1: Morphological tessellation.

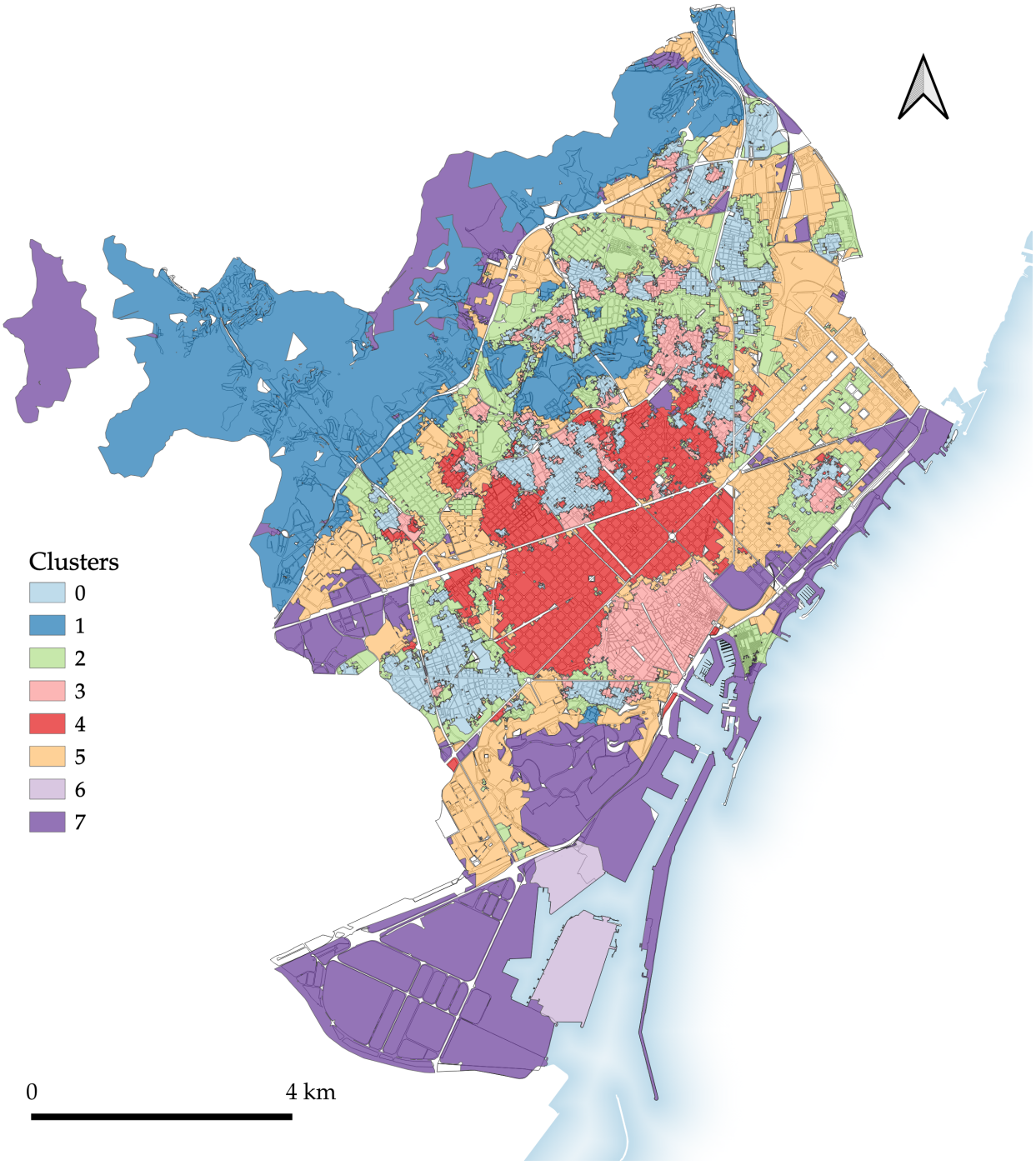

Figure 4.2: Segmentation 2: Enclosed tessellation.

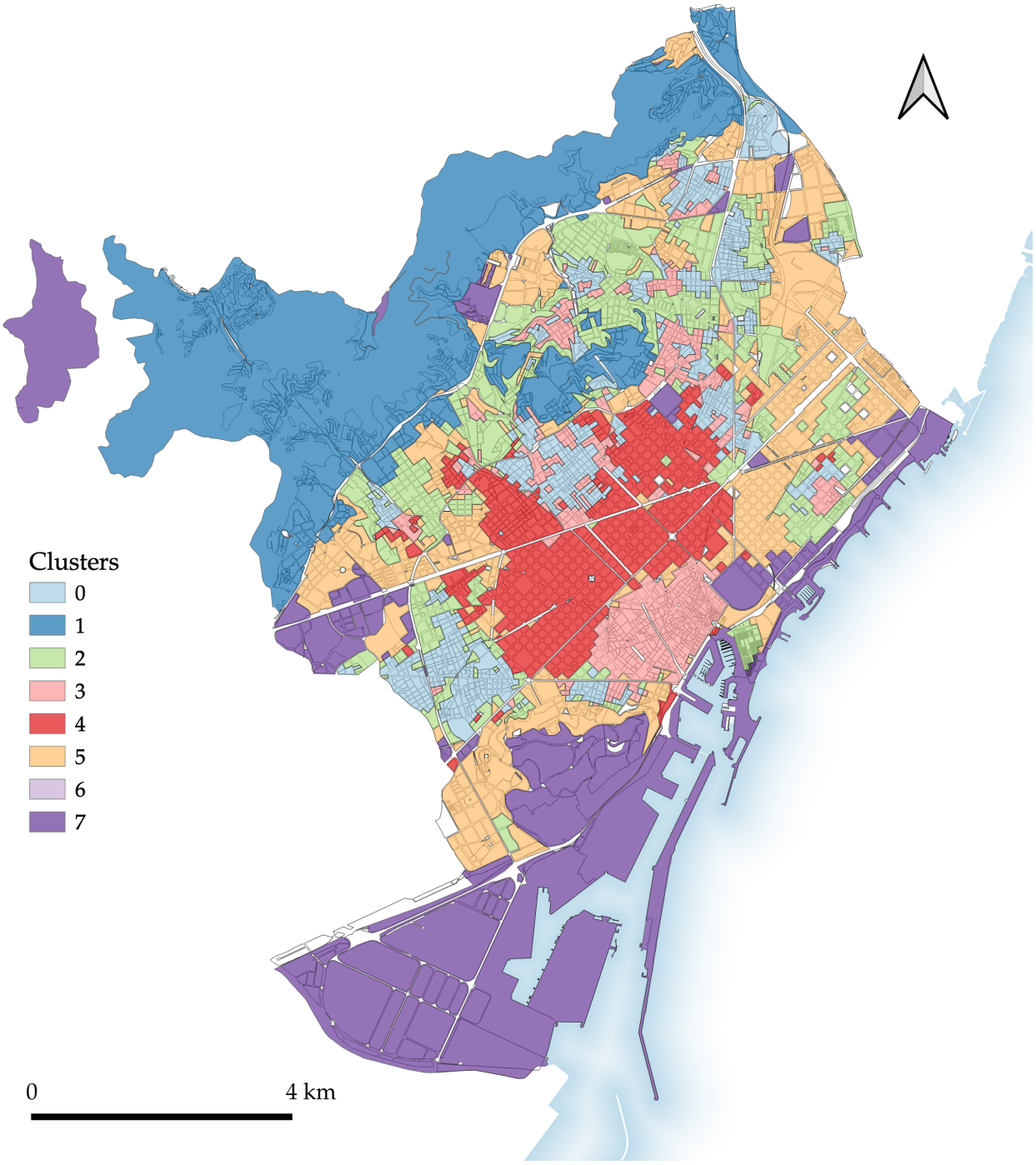

Figure 4.3: Segmentation 3: Enclosed tessellation transposed to block geometry.

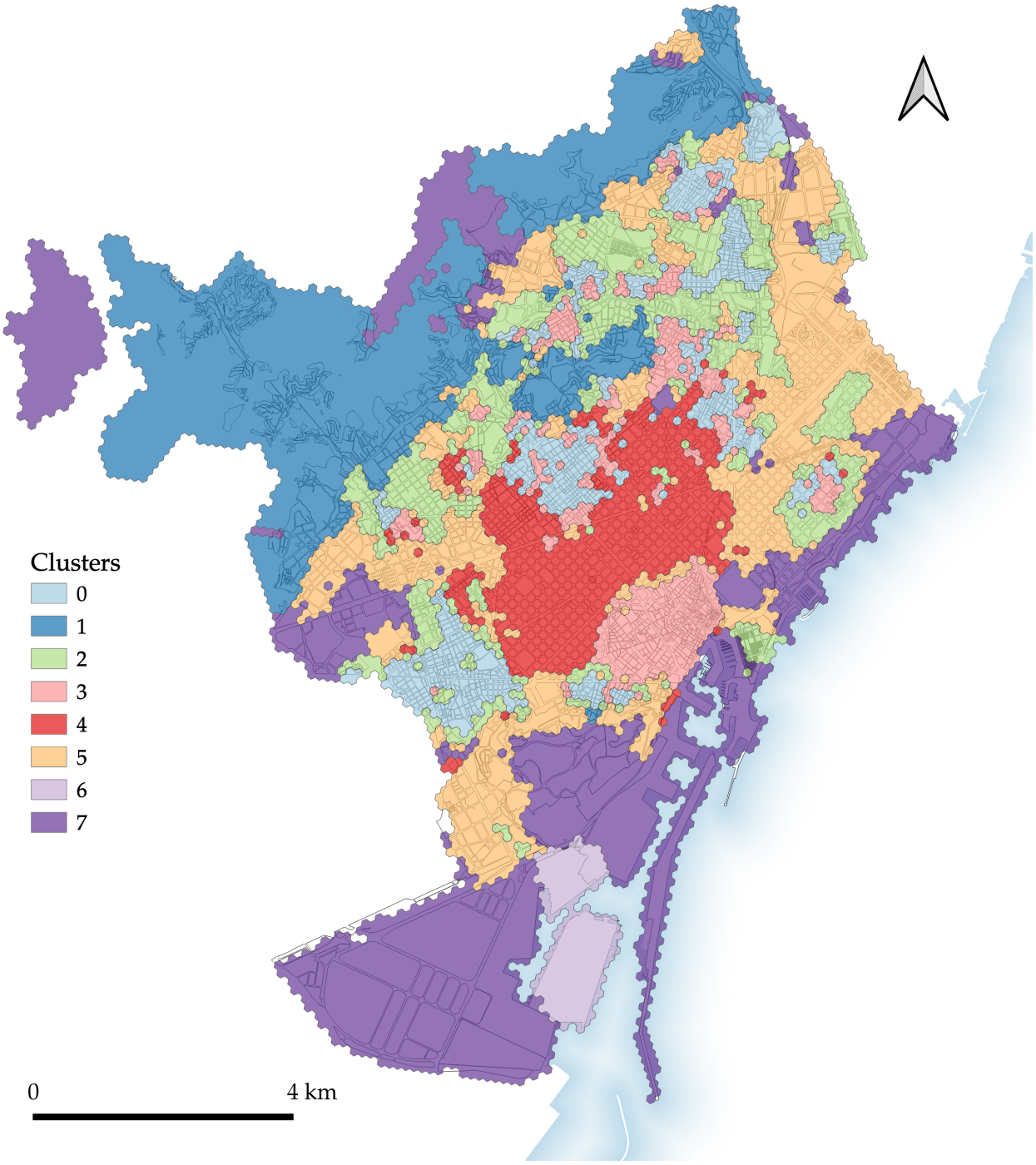

Figure 4.4: Segmentation 4: Enclosed tessellation transposed to H3 geometry.

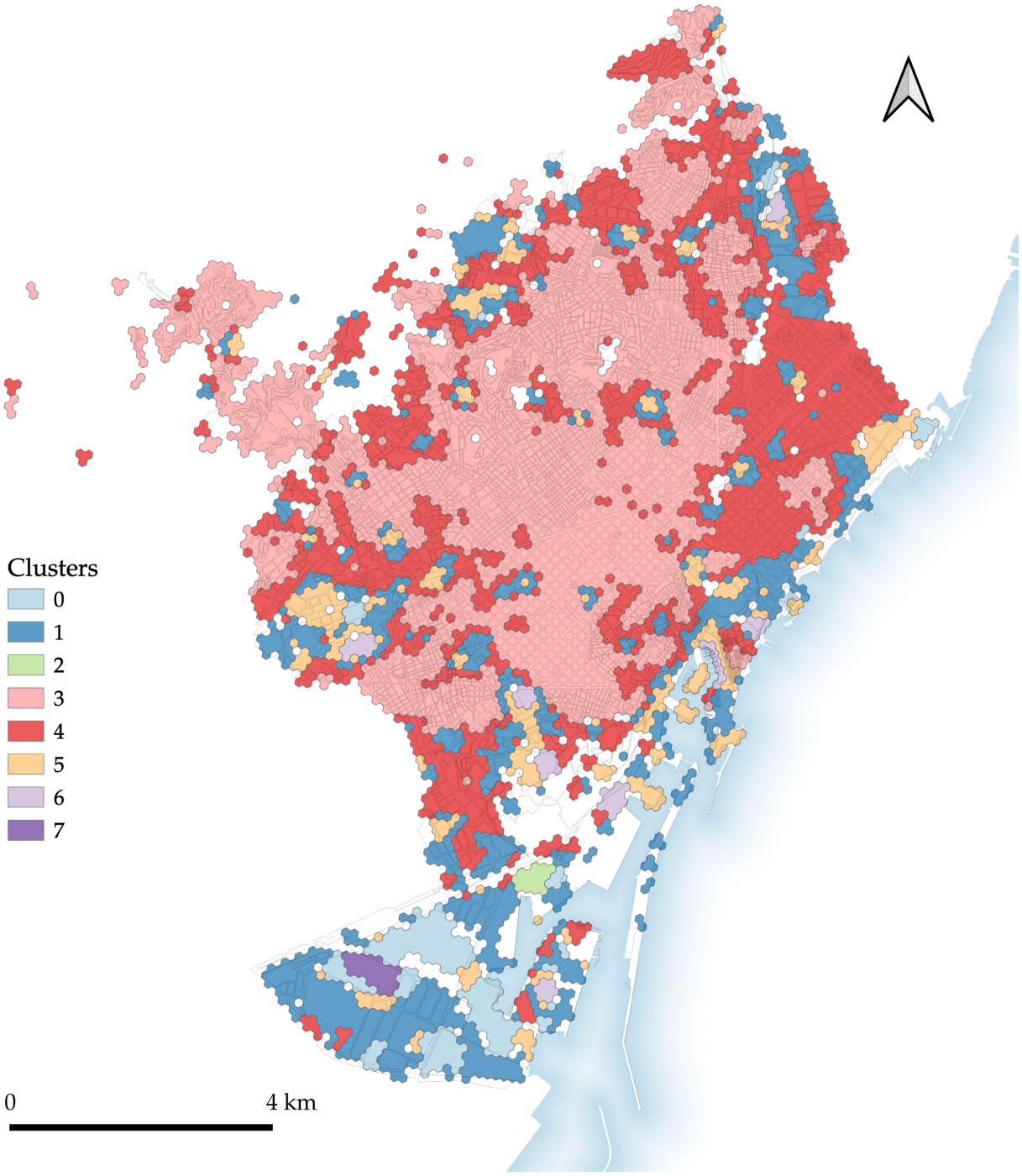

Figure 4.5: Segmentation 5: H3 'basic'.

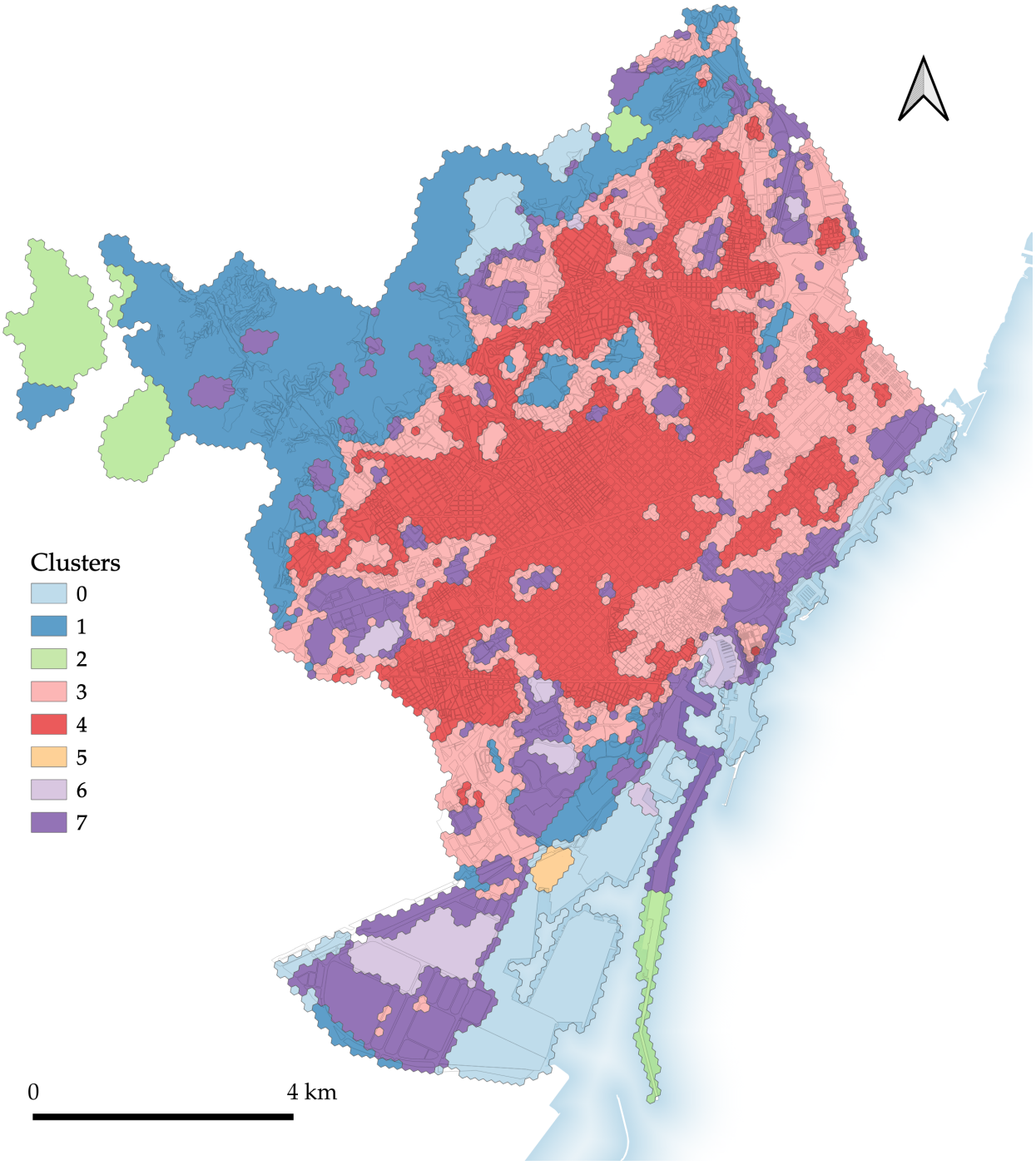

Figure 4.6: Segmentation 6: H3 clustering using characters from enclosed tessellation.

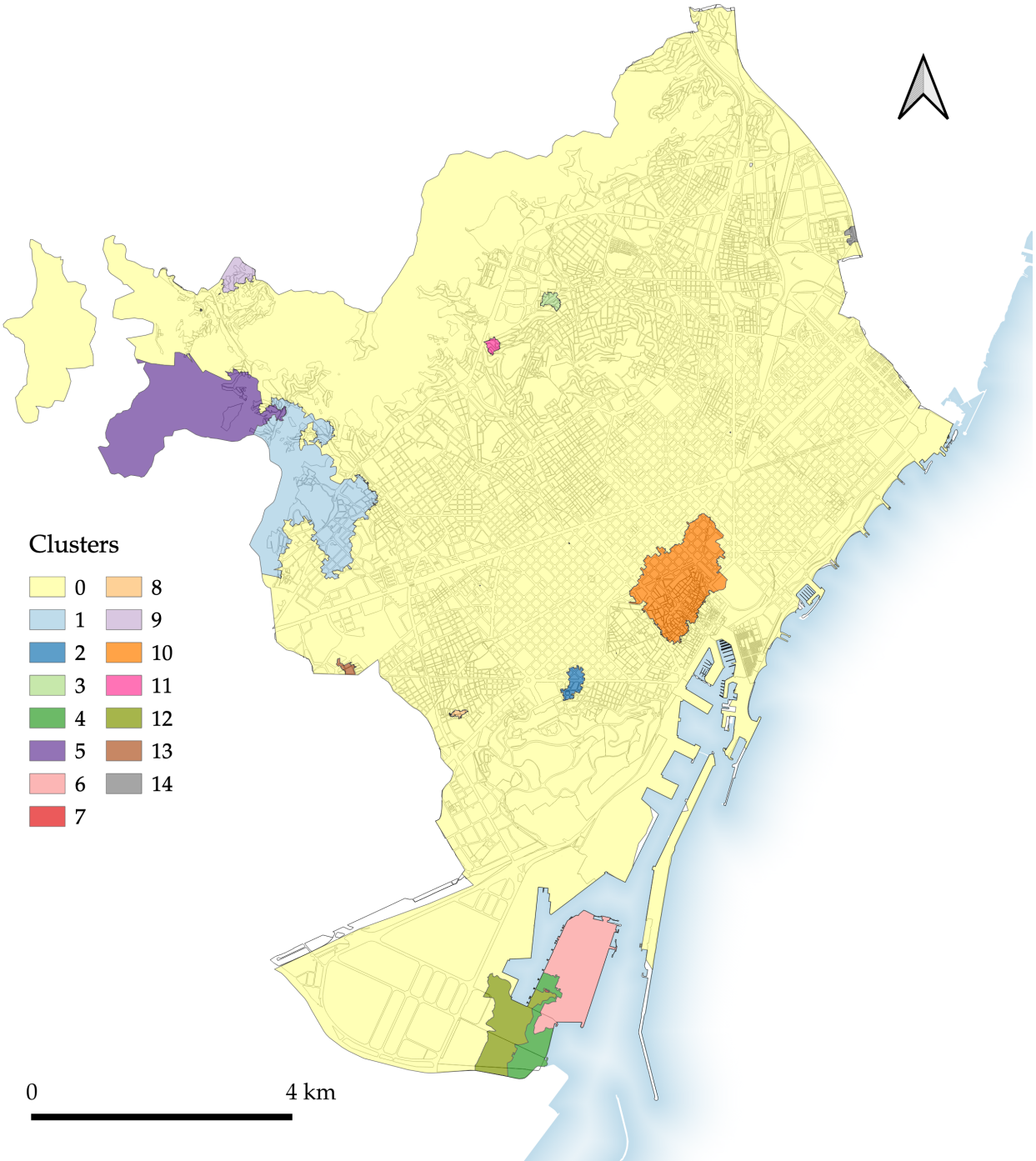

Figure 4.7: Segmentation 7: Spatially constrained clustering on morphological tessellation cells.

Although discussed at greater length in the next chapter, a simple initial description of each of the segmentation maps makes clear certain key results. The **morphological tessellation segmentation** (Figure 4.1) relatively successfully picks out certain key urban tissues, including those of the Ciutat Vella and of the Eixample, although both have clear room for improvement. For example, the Ciutat Vella type also includes a section of the Eixample immediately to the north of the old city, while the Eixample type excludes certain parts of the Eixample ((mis)classified as part of the pink and orange types) while also incorporating other areas beyond the Eixample grid.

The **enclosed tessellation segmentation** (Figure 4.2) identifies similar overall morphological trends to the preceding MT segmentation, but also comprises key differences. Immediately conspicuous are the areas omitted from classification – those enclosures not containing any buildings. The segmentation is also visibly more 'fragmented' than its MT counterpart, particularly in the northern part of the city.

This fragmented nature is a motivating factor in the development of the next two segmentations, which transpose the ET segmentation to **block** (Figure 4.3) and **H3** (Figure 4.4) geometries.

The **H3 'basic' segmentation** (Figure 4.5) tests the performance of a segmentation produced without using ET or MT cells at any point. Its results leave a lot to be desired: despite the use of contextual characters (defined with neighbouring H3 cells), few if any of the segmentation's types or polygons could be said to clearly constitute distinct urban tissues.

The consequent **H3 clustering using ET characters** (Figure 4.6) somewhat improves on the preceding 'basic' segmentation. While morphologically distinct areas are delineated to some extent, the differentiation of different types is inferior to that found in the ET or MT segmentations.

Finally, a visual inspection of the **spatially constrained segmentation** (Figure 4.7) immediately makes obvious the imbalance in the size of its clusters: 90.3% of the total area is assigned to one cluster.

4.2 Relationship to house price indices

For each segmentation, multiple metrics are calculated to measure the dispersion within/between the clusters generated. As discussed above, both type and polygon[1] level metrics are reported below.

In order to allow a comparison between the novel spatial segmentations mapped above and the existing spatial segmentations which may be used to represent spatial housing submarkets, three of the latter are included within the comparison below. The neighbourhoods, districts, and idealista polygons are mapped in Figure 4.8. This makes clear the ways in which the idealista polygons are largely coterminous with the city's administrative neighbourhoods, but sometimes merge these neighbourhoods, as with those which form the two orange idealista polygons in Nou Barris[2]; and sometimes separate these neighbourhoods in two, as with la Marina del Prat Vermell at the very South of the city. In other cases the geometry of the original neighbourhoods has been simplified or otherwise altered to create the idealista polygons.

4.2.1 Type metrics

Table 4.1 reports the average Quartile Coefficient of Dispersion of the average residential property sale price for all types, the mean of the areas of types in each segmentation, and the number of types included in each segmentation. Also included for each segmentation is a boxplot showing the distribution of these type areas.

Figure 4.9 plots each type from each segmentation, comparing its area (on the x-axis) with the QCoD of house prices within the type (on the y-axis). It shows a definite relationship between these two variables, and for this reason the average QCoD alone (provided in Table 4.1) is insufficient to make a fair

[1] As set out in the previous chapter, *type* describes all cells in the city with a certain cluster label, whereas *polygons* refer to the separate geographies of contiguous cells with a certain cluster label.

[2] Ciutat Meridiana, Vallbona, and Torre Baró in the North of the district; and Can Peguera and el Turó de la Peira in the South.

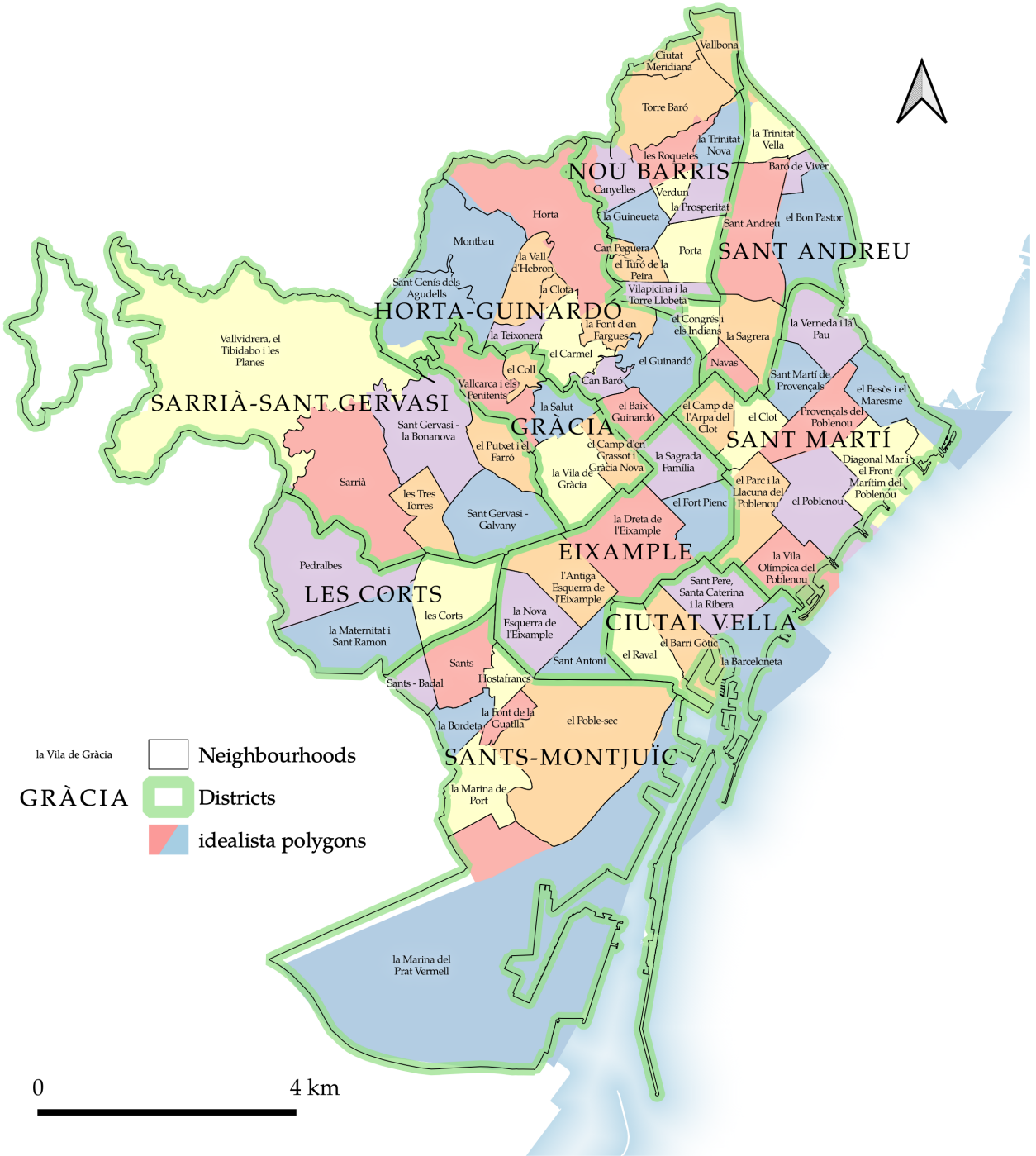

Figure 4.8: Existing spatial units: neighbourhoods, districts, and idealista polygons in Barcelona.

Table 4.1: Average type values for each segmentation.

		Type area (km^2)		
Segmentation	Mean QCoD	Mean	Distribution	Number of units
Morphological tessellation	0.206	12.62		8
Enclosed tessellation	0.201	11.92		8
ET transposed to block	0.202	14.53		6
ET transposed to H3	0.202	13.47		8
H3 basic	0.140	10.13		8
H3 with ET characters	0.222	13.47		8
Spatially constrained MT	0.080	6.73		15
Existing neighbourhoods	0.071	1.39		73
Existing districts	0.115	10.17		10
idealista polygons	0.075	1.42		69

comparison of how well segmentations capture variations in property prices: all else being equal, larger areas will tend to have larger dispersions. By plotting both the QCoD and the area of the spatial unit whose internal house price dispersion the QCoD records, a better judgement can be made about how well a given segmentation captures variation in house prices *given its area*. Note that in Figure 4.9, the dot for one type (the largest type in the spatially constrained segmentation) is not displayed, as its area of 91.2km^2 makes it an extreme outlier.

Because Figure 4.9 plots every type separately, it is difficult to make a clear judgement about how well a segmentation performs overall. This is evident by the spread of dots of the same colour in different areas of the chart: while one type within a segmentation may have a low QCoD, this may be offset by other types within the same segmentation having much greater dispersion. In order to allow better judgements of the overall performance of segmentations, Figure 4.10 therefore plots the average areas of the types in each segmentation against the corresponding average house price QCoDs, allowing the direct comparison of each segmentation's average values for these two metrics. The grey line plots a simple linear regression model fit to the points on the graph. Although clearly a poor predictor of the average QCoD of a given segmentation, the line serves as a visual aid in understanding how well a segmentation can be seen to capture house price dispersion, given the size of the units into which it partitions space

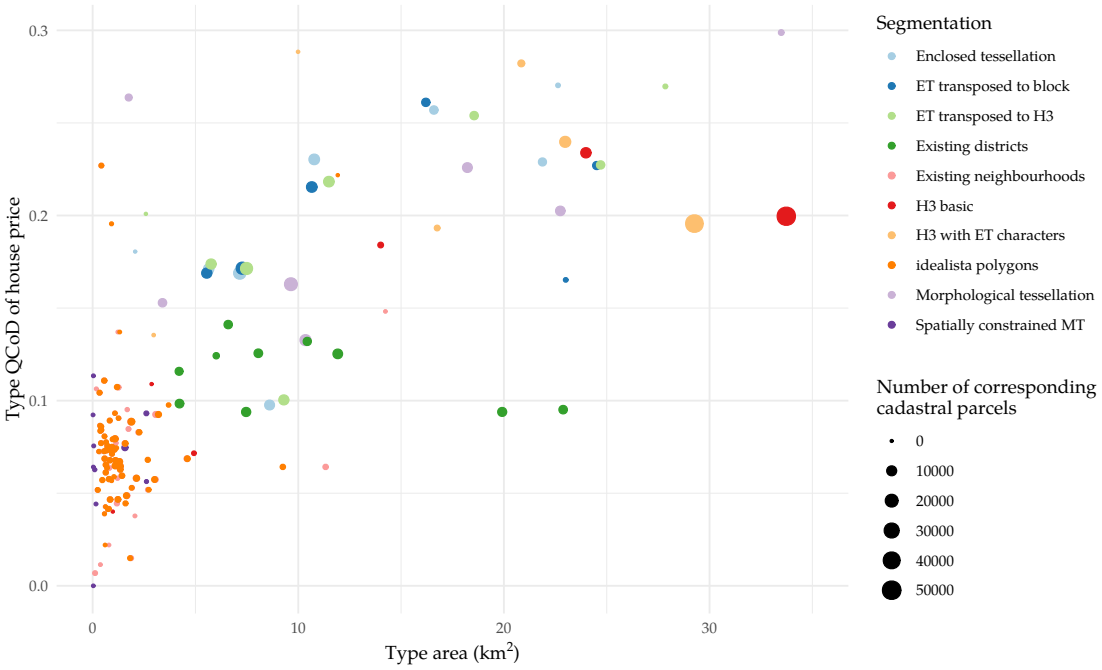

Figure 4.9: Every type plot by house price QCoD and area.

(in this case, the different types). Segmentations plot below the line can be seen to have lower levels of dispersion given the average areas of their types, and therefore better capture variation in house prices. Conversely, segmentations plot above the line can be seen to have higher QCoD values than might be expected given the average areas of their types, and therefore perform worse as a delineation of housing submarkets.

It should be noted that this system of interpretation is not derived from any particular empirical base, but rather the plots provide a useful heuristic for comparing the performance of segmentations which divide space into different numbers of units of different sizes.

4.2.2 Polygon metrics

Treating each type—each colour on the segmentation map—as the ultimate spatial unit generated by the segmentation process is one way of assessing the segmentations. Alternatively, each *polygon* produced can be seen as a separate spatial unit: in this conceptualisation, areas which are assigned the same cluster

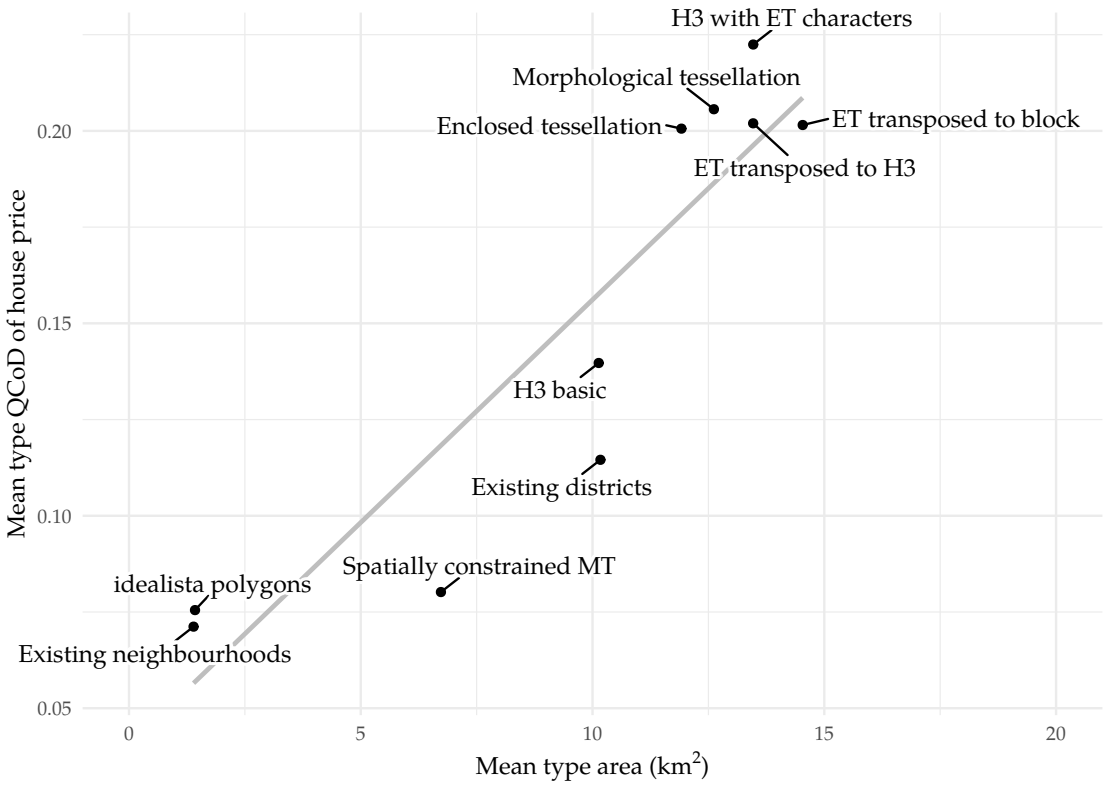

Figure 4.10: Segmentation averages of typologies plot by house price QCoD and area.

label—the same colour on the map—but are located in different parts of the city will be counted as separate units.

Table 4.2 shows the average results of the same statistics as Table 4.1 for the same segmentations, but calculated at the polygon level rather than the type level.

As a generality, the smaller a unit is, the fewer house price data points it is likely to contain: larger units are therefore generally more robust when calculating the QCoD, and types more robust than polygons. Because many of the polygons have very small areas (notably those composed of a single or very few MT or ET cells), many contain few cadastral parcels with attached house price information: of the 3,721 polygons into which the ten segmentations examined here are divided[3], 1,732 (46.5%) correspond to fewer than ten house price data points, including 632 (17%) with no corresponding cadastral parcels. Figure 3.8

[3]Seven novel and three existing.

demonstrates that there is also a geographical pattern to this validation data, meaning that certain kinds of types and polygons are more likely to have few or no relevant data from which to calculate the QCoD.

This makes a calculation of the QCoD for house prices within these polygons at best not robust and at worst impossible. For this reason, the averages shown in Table 4.2 and plot in Figure 4.12 exclude polygons containing fewer than ten house price data points. As these polygons only contain a few cadastral parcels which are geographically proximate and therefore likely also numerically proximate in terms of average house price, their QCoD is likely to be low: there is likely to be minimal dispersion in an area containing only a few properties. Including these polygons would therefore have weighted the average QCoD values to suggest lower levels of dispersion, but only on the grounds of including many small polygons, which would likely not be seen as constituting distinct housing submarkets. A version of Figure 4.12 which does not exclude these polygons in its calculations is provided in Appendix A as Figure A.1.

The polygon area distribution boxplots are again shown alongside the average area value, but the range for these has been artificially truncated: values (all outliers) larger than 18km^2 are beyond the range of the plot. Because the types and polygons for the spatially constrained segmentation are identical (since spatial contiguity is a condition of the clustering process, so the types it generates cannot be multipart geometries[4]), the 91.2km^2 type discussed previously is also counted as a polygon[5]. If plotted with a range inclusive of this polygon, the majority of the other boxplots would be rendered illegible, so for this reason the range is limited.

The different units represented in the two tables is evident when comparing the 'Number of units' column in each table: in Table 4.1 this reports the number of types (eight for the enclosed tessellation segmentation; the same when it is transposed to the H3 geometry), while in Table 4.2 this reports the number of polygons

[4]Except in the few cases where the city boundary itself contains separate geometries, such as islands.

[5]Or to be more precise, the large majority of the type is counted as a polygon: as can be seen in Figure 4.7, it also encompasses two islands (one literal and one figurative), which are counted as separate polygons.

Table 4.2: Average polygon values for each segmentation.

		Type area (km^2)		
Segmentation	Mean QCoD	Mean	Distribution	Number of units
Morphological tessellation	0.076	0.79		119
Enclosed tessellation	0.067	0.24		311
ET transposed to block	0.066	0.26		322
ET transposed to H3	0.074	0.45		206
H3 basic	0.071	0.53		116
H3 with ET characters	0.075	0.97		81
Spatially constrained MT	0.080	8.71		11
Existing neighbourhoods	0.071	1.36		73
Existing districts	0.115	9.90		10
idealista polygons	0.073	1.26		68

(1,666 for the enclosed tessellation segmentation; reducing to 249 when this is transposed to the H3 geometry). This difference is also reflected in the areas of the units: polygons tend to be much smaller than the types to which they belong.

Because the types and polygons of the existing spatial units (neighbourhoods, districts, and idealista polygons) are spatially coterminous (their types include few or no multipart geometries), they show the fewest changes when comparing types with polygons.

Figure 4.11 replicates Figure 4.9, but reporting the values for polygons and not types. In order to more clearly show the spread of values, the area of polygons is mapped to the x-axis logarithmically: as shown in the axis label, each axis tick multiplies by a factor of ten (0.1 km^2, 1 km^2, 10 km^2, etc). This is made necessary by the distribution of areas among the polygons being examined: while there are a few notable polygons with large areas[6], there are many with very small areas. 53.4% of polygons are smaller than 0.01 km^2 (10,000 m^2, or about the size of Liverpool's Abercromby Square) and 34.2% of polygons are smaller than 0.001 km^2 (1,000 m^2, or about a twelfth the size of a block in the Eixample).

[6]Of the 3,721 polygons into which the ten segmentations examined here are divided, 185 (5%) have an area of more than 1 km^2; 32 (0.86%) have an area of more than 5 km^2; and 16 (0.43%) have an area of more than 10 km^2.

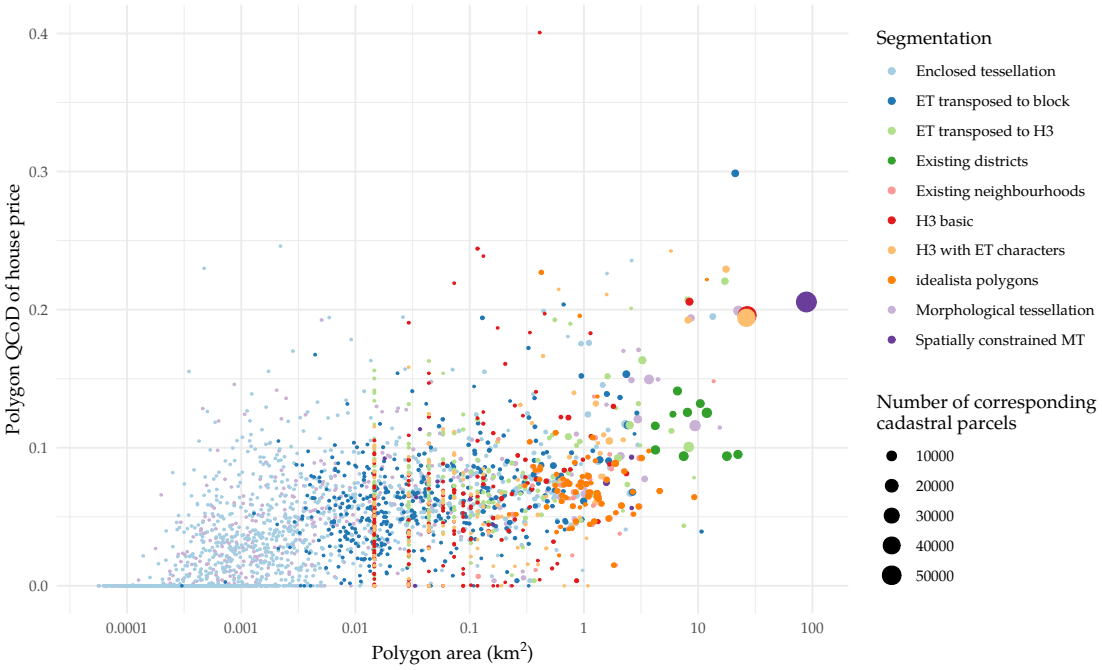

Figure 4.11: Every polygon plot by house price QCoD and area.

Figure 4.12 replicates Figure 4.10, but again reporting the values for polygons and not types.

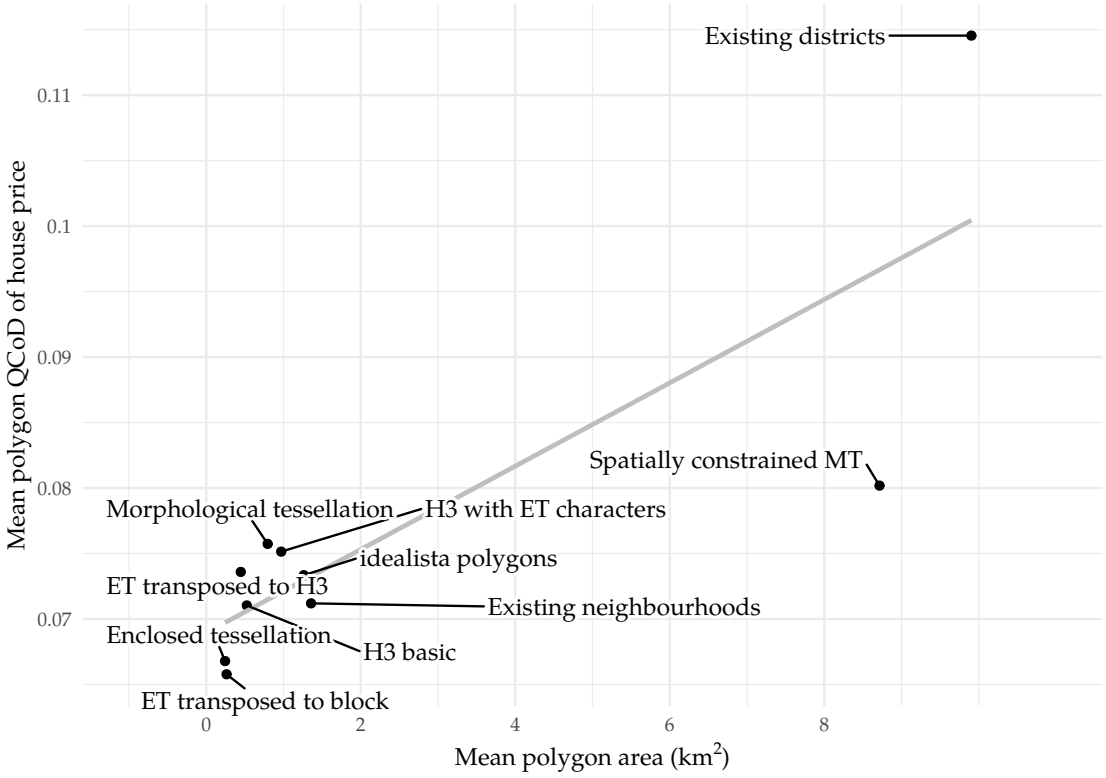

Figure 4.12: Segmentation averages of polgyons plot by house price QCoD and area.

5 Discussion

This chapter discusses in greater depth the results presented in the previous chapter, explaining these results and their implications.

Broadly, the discussion focuses on the two key concerns of the dissertation: the creation of spatial segmentations which reflect the urban morphology of a city, and the creation of spatial segmentations which can be used as spatial housing submarkets. Certain measures and evaluations of the segmentations focus on one or the other of these concerns—for instance the QCoD assesses solely the segmentations' suitability to represent housing submarkets—while other assessments of the segmentation are germane to both concerns. Many of the key findings are of a methodological nature, demonstrating the ways in which altering certain elements of the methodology affects the segmentation produced. These methodological findings often correspond to limitations in the research, and more broadly in the methodology employed.

5.1 Creating spatial segmentations to reflect housing submarkets

As reported in the previous chapter, the QCoD allows the degree to which segmentations capture variation in house prices to be quantified, allowing compar-

isons between different segmentations to be made on this criterion.

The similar type-level average QCoD values for the MT, ET, ET transposed to H3, and ET transposed to block segmentations (see Figure 4.10) suggest that transposing to the H3 or block geometry has minimal effect on the degree of house price variation captured by the segmentation. This kind of transposition may therefore be a good way to reduce the 'fragmentary' nature of the segmentation and 'clean up' its geometry to a certain extent, without damaging its performance with regards to delineating housing submarkets.

Figure 4.10 also highlights the difficulties of making direct comparisons between segmentations which divide an area into a different number of units and/or units with different areas. The existing neighbourhoods and idealista polygons have lower QCoDs than any of the novel segmentations, but they also have significantly lower average type areas and a higher number of types, and so to some extent this should be expected. Interestingly, judging solely on the metrics shown in Figure 4.10 could lead to the conclusion that the spatially constrained segmentation performs very well; however the geographical distribution of its clusters (see Figure 4.7) means that the segmentation would likely be of little use in practice.

A comparison of Figure 4.10 and Figure 4.12 highlights the different conclusions reached when different spatial units (the type and polygon) are used as the basis for analysis. While the positions of segmentations on each chart varies, certain key trends are true of both. Both demonstrate the general relationship between average unit area and average unit QCoD, and in both most of the novel segmentations form a cluster of relatively similar values, suggesting relatively small differences between the performances of these segmentations on the metrics plotted on these charts[1].

[1]The notable exceptions are the spatially constrained segmentation, an exception in both type and polygon analyses; and H3 'basic', an exception in the type analysis due primarily to its relatively low average type area.

5.2 Creating coherent spatial segmentations

5.2.1 Fragmentary segmentations

Throughout the process of producing spatial segmentations, one issue encountered was that the spatial units generated were fragmentary and had limited spatial contiguity. When using the GMM algorithm to cluster base spatial units, the clustering algorithm has no knowledge of the topological relationships between the cells it is clustering, and as such there is no guarantee that any segmentations produced will be contiguous. The fragmentary kinds of segmentations produced by such clusterings cannot be seen to accurately represent either urban morphology or housing submarkets: although the specific details may differ dependent on the conceptualisation and scale of either of these areas, both are certainly larger than the tessellation cells used in ET/MT-based segmentations. Hence, a segmentation which assigns an individual ET/MT cell different cluster labels to all of its neighbours can be seen to have failed to adequately delineate either urban tissues or housing submarkets: in this resultant segmentation, the single tessellation cell will be classed as a distinct polygon, but will not by any reasonable measure be representative of the concepts it seeks to represent.

An argument could be made that, while using the 'polygons' of the segmentation makes more conceptual sense when using the segmentations to represent housing submarkets, different types of urban morphology may be better conceptually aligned with the 'types' of the segmentation, and therefore fragmentary segmentations are less problematic when classifying urban morphology.

5.2.2 Limiting fragmentary segmentations with contextual characters

As explained in the Methodology, an imperfect solution to the problem of fragmented segmentations is the use of contextual characters constructed by averaging the values of nearby cells. Indeed, these are incorporated into all segmentations presented in the previous chapter (save that which stipulated a spatial

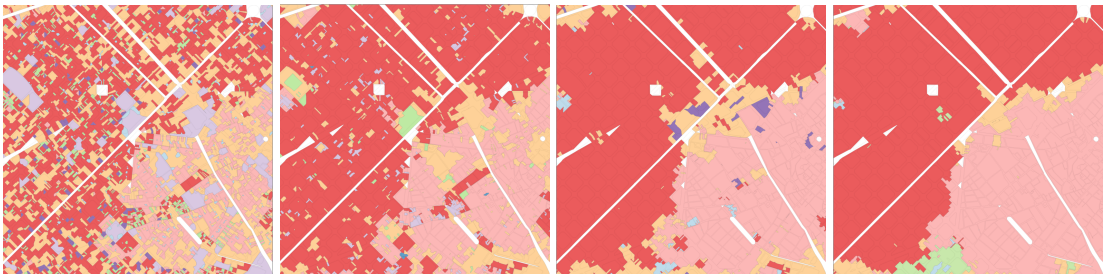

Figure 5.1: The effect of an otherwise identical segmentation on contextual characters constructed using different order spatial weights. From left to right: no contextual characters; 1st order spatial weight, up to 3rd order spatial weight, up to 5th order spatial weight.

constraint). Figure 5.1 demonstrates the difference that changing the neighbour criteria when generating contextual characters makes to the final segmentation. While vague spatial patterns can be discerned, the clustering using only the primary characters of each cell (on the far left) is far too fragmented to offer a useful segmentation of either urban morphology or housing submarkets. The subsequent maps plot segmentations identical in all factors save the neighbour criteria used to generate the contextual characters, showing clusterings based on contextual characters built using neighbours up to a topological distance of one, three, and five.

Besides acting simply as a methodological tool to create cleaner segmentations, contextual characters also serve a conceptual role. The central reason that the segmentation using only primary characters is too fragmented to be of any practical use is that the tessellation cells clustered (aspatially, with GMM) hold no information about their 'surroundings'. The urban morphology of a given tissue cannot, however, be defined based only on one building (or tessellation cell). Accordingly, the contextual characters serve to partially determine the scale of the areas demarcated in the resultant segmentation: the wider the range of neighbours included in the contextual characters, the larger (in general) the polygons in the resultant segmentation will be.

The choice of topological distance used to determine the contextual characters should therefore also reflect the expected scale of an urban tissue. To give an

intentionally extreme example, it would not make sense to have contextual characters which averaged the character values of all cells within three kilometres, since the morphometric characters of a location three kilometres away should not be thought of as having any influence on the urban morphology of the primary cell, or the urban tissue to which it belongs. Thus, when employing contextual characters in this way there is a trade-off to be made between the breadth of contextual information included (and the concomitant degree of smoothing this produces) and the amount of granularity lost in the final segmentation. For each additional order of topological distance included as a neighbour in the contextual characters, the segmentation loses the ability to discern smaller urban tissues.

5.2.3 Limiting fragmentary segmentations by transposing onto different geometries

An alternate approach to the issue of fragmentary segmentations is to carry out the established methodology to generate a spatial segmentation using the chosen tessellation with small base spatial units, and subsequently transpose the segmentation generated onto a coarser geometry. This is demonstrated in this dissertation by transposing the clusters produced in the ET segmentation onto the block and H3 geometries (see Figures 4.3 and 4.4).

By aggregating the cluster labels to larger geometries it is ensured that none of the resulting segmentation's distinct spatial units (i.e. 'polygons') are smaller than the cell in these larger geometries (i.e. the block or H3 cell). This is reflected in plots such as Figure 4.11, which shows that very small polygons (those with areas of less than 0.01 km^2) are almost entirely the preserve of the MT and ET segmentations (plotted in lilac and light blue respectively). As expected, the transposed segmentations are somewhat 'smoothed': because the transposition process (as described in the Methodology) assigns each block/H3 cell the cluster label which covers the largest proportion of its area, smaller 'outlier' geometries (such as solitary ET/MT cells) are usually erased from the segmentation, 'cleaning' the output.

5.3 Creating spatial segmentations from different base spatial units

5.3.1 H3

The effects of using different base spatial units are not limited to the geometries of the segmentations produced. When different spatial units are used as an *input* to the clustering, substantially different segmentations are produced, even when using the same set of characters. This can be most clearly seen by comparing the segmentation which transposes the *labels* from the ET segmentation to the H3 geometry *after* clustering (shown in Figure 4.4) with that which transposes the *characters* generated with ET segmentation to the H3 cells *before* clustering (shown in Figure 4.6). The two differ only subtly in their methodology—the former uses the enclosed tessellation cells as data points when clustering, while the latter uses H3 cells—but the resultant segmentations are markedly different. While both distinguish between certain broad morphological variations (classifying the dense Eixample as distinct from the more rural Northwest or the industrial area at the South of the city), the clustering performed using ET cells much better captures morphologically homogeneous areas. For instance, the Eixample and Ciutat Vella are both more tightly and accurately delineated in the ET clustering than in the H3 equivalent.

An explanation for this may be that the use of H3 changes the density of cells (i.e. rows input to the GMM clustering). Whereas the enclosed tessellation contained more cells per km^2 in the denser old city than in the sparser rural areas in the Northwest, a key design feature of the regular H3 grid is that it weights each area equally in this respect: the number of cells per km^2 remains consistent irrespective of whether the square kilometre is in the densest area of the city or the sparsest.

To some extent, using different base spatial units as inputs to the clustering stage of the methodology can be thought of as changing the 'weighting' given to different areas. When a clustering is performed on ET cells, the algorithm

has more information[2] about areas of the city which are by their nature more information-dense. Conversely when a clustering uses H3 cells (or any other regular grid), the 'information density' is uniform over space, meaning that as many rows are dedicated to a square kilometre of undeveloped countryside as are to a square kilometre of bustling city centre. The segmentations produced using H3 cells (such as that shown in Figure 4.6) suggest that this misaligned information density may act as a limiting factor on the degree to which the segmentation is able to discern urban tissues. This echoes Arribas-Bel and Fleischmann's (2021) assertion that "choosing a spatial unit that does not closely match [the] distribution [of urban fabric] will subsume interesting variation and will hide features" (ibid, 7).

5.3.2 Morphological and enclosed tessellation

The segmentations produced using MT and ET cells present a good opportunity for a comparison of the effects of these different base spatial units. It would be expected, ex ante, that there is little difference between the two, as the units differ in only small ways and are largely similar in their construction.

A key distinction between the two tessellations is that enclosed tessellation[3] is not spatially exhaustive, discarding any enclosure (that is, a space enclosed by drivable roads) which does not have a building within it. As the spatial segmentation is intended to delineate *housing* markets, there is a clear theoretical basis for discarding those spaces which do not contain any housing. Whilst this was an intentional methodological choice (indeed, following prior non-exhaustive segmentations produced by idealista), it may have had some undesirable effects.

A common issue is that when roads with two carriageways recorded in OSM as separate lines are used to create enclosures, the middle strip (i.e. the central reservation of the road in question) is discarded. This can then cause an artificial

[2]Literally more rows, and also a greater proportion of the total rows in the input data.

[3]In this dissertation that is: Arribas-Bel and Fleischmann's original description of enclosed tessellation does not remove any cells and notes that the spatial-exhaustiveness of the tessellation is a key feature.

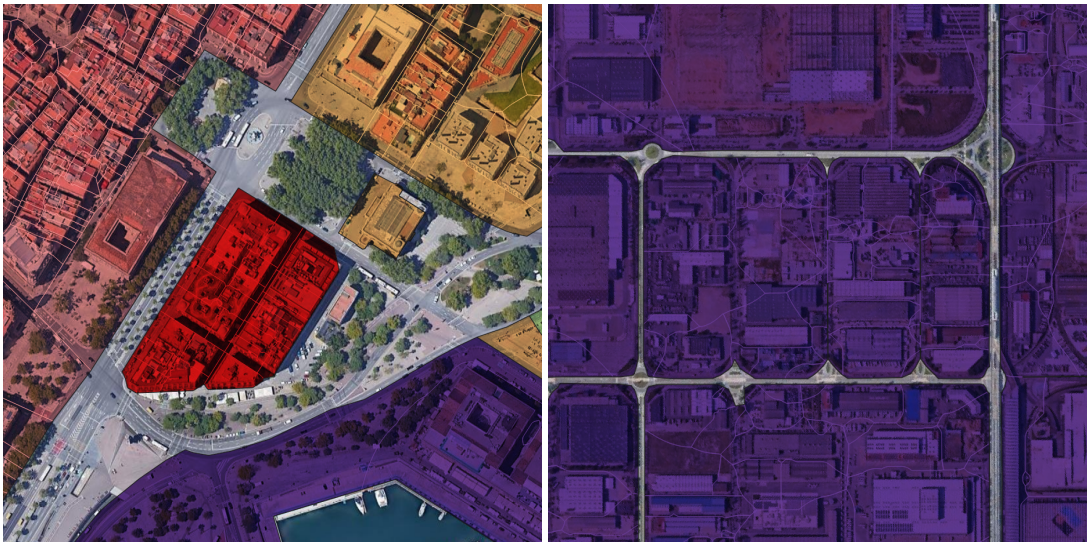

Figure 5.2: Islands produced by enclosed tessellation.

barrier between cells when calculating the contextual characters, either making the topological relationship with would-be neighbouring cells more indirect, or isolating the cell(s) entirely. Figure 5.2 shows two examples of instances where the use of enclosed tessellation has rendered certain groups of cells islands, topologically disconnected from any surrounding units and therefore unable to incorporate the local morphology into contextual characters.

Arguably, there is some justification for having two-lane highways present a greater conceptual barrier than the wall between two buildings, but the absolute barrier effect found in the current implementation of ET is on balance too great. A possible alternative would be to keep the 'empty' ET cells generated, ignoring them when calculating morphometric characters based on information about buildings within the cells, but allowing their use at the contextual character stage. In this way the 'empty' ET cells would be assigned the average values of their neighbours, and not act as an absolute barrier when calculating contiguity-based spatial weights for contextual characters. Alternatively, an additional step could be added, using the ET cells as the input to a further Voronoi tessellation. In this way, any 'empty' spaces (those enclosures without buildings) would be divided according to their nearest (non-empty) ET cell and appended to these cells, creating a spatially exhaustive tessellation based on enclosures and in which every

cell contains a building.

5.4 Creating spatial segmentations with different clustering algorithms

The results show that, while the segmentations generated are contingent on a number of different parameters, the Gaussian Mixture Model clustering algorithm used to produce the majority of segmentations is capable of building clusterings which accurately discern key urban tissues. The results of the segmentation produced using Agglomerative Clustering in place of GMM highlight the imperative role played by the choice of algorithm. Even when using identical input data, different algorithms (even when aiming to demarcate the same thing) can produce vastly different clusterings.

6 Conclusion

6.1 Summary of findings

In conclusion, this dissertation has developed a methodology to segment an urban area into different spatial units based on its urban morphology, and assessed the degree to which these novel segmentations reflected both urban morphology and housing market dynamics. A quantitative assessment was used to investigate the degree to which the segmentations produced captured variation in house prices as a proxy for housing market dynamics: this highlighted the differences between segmentations produced using different methodologies. This form of assessment should only serve as an approximate guide to interpretation, and unqualified conclusions should not be drawn on the basis of these results.

The primary contribution of this dissertation has been methodological. Throughout the study, the effects of alternate methodologies has been examined and discussed, leading to a number of key results. The dissertation found the use of contextual characters (effectively spatial lags) to be essential to developing spatially coherent segmentations. It was also shown that segmentations can be made more 'clean' by 'transposing' the cluster labels from segmentations produced by clustering smaller base spatial units onto larger, simpler geometries, with each unit (e.g. block or H3 cell) in these geometries labelled according

to the cluster label which takes up the largest proportion of the cell's area in the original segmentation. The selection of base spatial unit has been found to greatly affect the resultant segmentation, corroborating previous assertions that the spatial unit should match the distribution of the variable of interest when producing spatial clusterings of this nature. This finding also suggests that the use of other base spatial units—such as regular grids—may hamper the accurate spatial segmentation of a study area based on a (set of) variable(s). The choice of clustering algorithm was also found to be crucial to the generation of satisfactory segmentations: the substitution of one clustering algorithm for another engendered extensive changes to the segmentations produced.

6.2 Further research

The number of parameters involved in a research methodology of this scope is such that it is not possible to examine in detail the effect of changing every possible variable; this dissertation has only focused on what may be some of the most salient of these.

Future research could build upon these findings by expanding the study area, testing the method using data from a range of cities, or indeed countries.

While this dissertation has focused on producing segmentations which accurately delineate a city based on urban morphology, there is potential to further refine these areas to ensure their suitability for use as delineating housing submarkets. This could be achieved by following the methodological improvements proposed in the previous chapter, or by incorporating a further smoothing process into the methodology.

There is potential for significant further investigation into the algorithm used in the clustering process. For example, further exploring algorithms which incorporate spatial constraints, or making use of GMM's ability to assign each datapoint (i.e. tessellation cell) a probability of cluster membership. Additionally,

the goodness-of-fit of the clusters these algorithms generate could be examined with measures such as geosilhouettes (Wolf et al., 2021).

Perhaps one overriding issue in seeking to delineate clear and defined boundaries between different urban tissues and/or housing submarkets is that the definitions of both of these concepts are contested. This is by no means to suggest that the approach/endeavour should be abandoned or that any attempt to delineate such areas is futile, but rather to recognise that to do so it to create a 'wrong'—but potentially useful—simplification of these complex concepts. As is often the case, the map's clean lines belie the complexity of the areas they seek to represent.

Appendices

A
Additional Figures

The following are additional figures not essential to the primary narrative of the dissertation, but which may be of interest as supplementary reference materials.

Table A.1 recreates Table 4.2, but does not exclude polygons containing fewer than ten house price data points.

Figure A.1 recreates Figure 4.12, but again does not exclude polygons containing fewer than ten house price data points. By comparing the two figures, it is clear which segmentations included the most polygons with few corresponding house price data points. For instance, the morphological tessellation segmentation moves from far below the straight line to above it when these values are excluded.

Figure A.2 shows a segmentation produced using enclosed tessellation cells and clustering primary characters (i.e each tessellation cell has no information about its neighbours), while Figure A.3 shows a segmentation produced using ET cells and clustering contextual characters incorporating information from cells up to 3rd order topological distance away.

Table A.1: Average polygon values for each segmentation, all polygons included.

		Type area (kmš)		
Segmentation	Mean QCoD	Mean	Distribution	Number of units
Morphological tessellation	0.031	0.15		657
Enclosed tessellation	0.030	0.06		1666
ET transposed to block	0.061	0.23		412
ET transposed to H3	0.070	0.43		249
H3 basic	0.058	0.22		370
H3 with ET characters	0.066	0.57		188
Spatially constrained MT	0.080	5.61		18
Existing neighbourhoods	0.071	1.36		75
Existing districts	0.115	8.48		12
idealista polygons	0.075	1.33		74

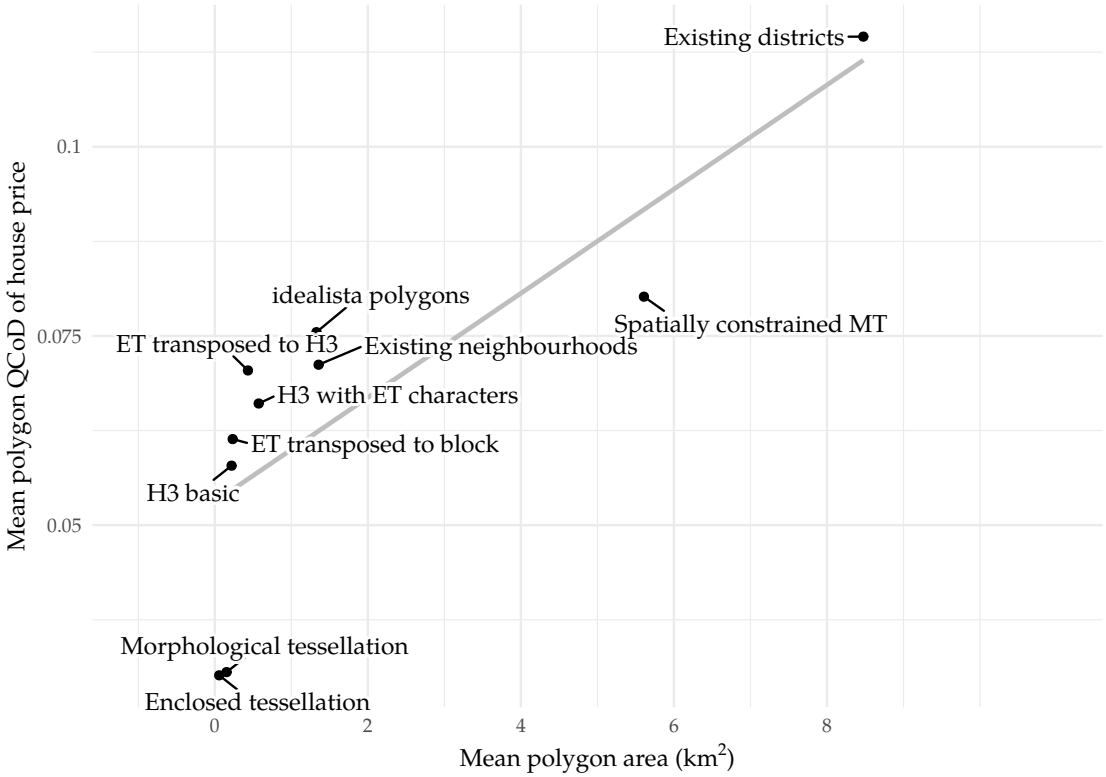

Figure A.1: Segmentation averages of polgyons plot by house price QCoD and area.

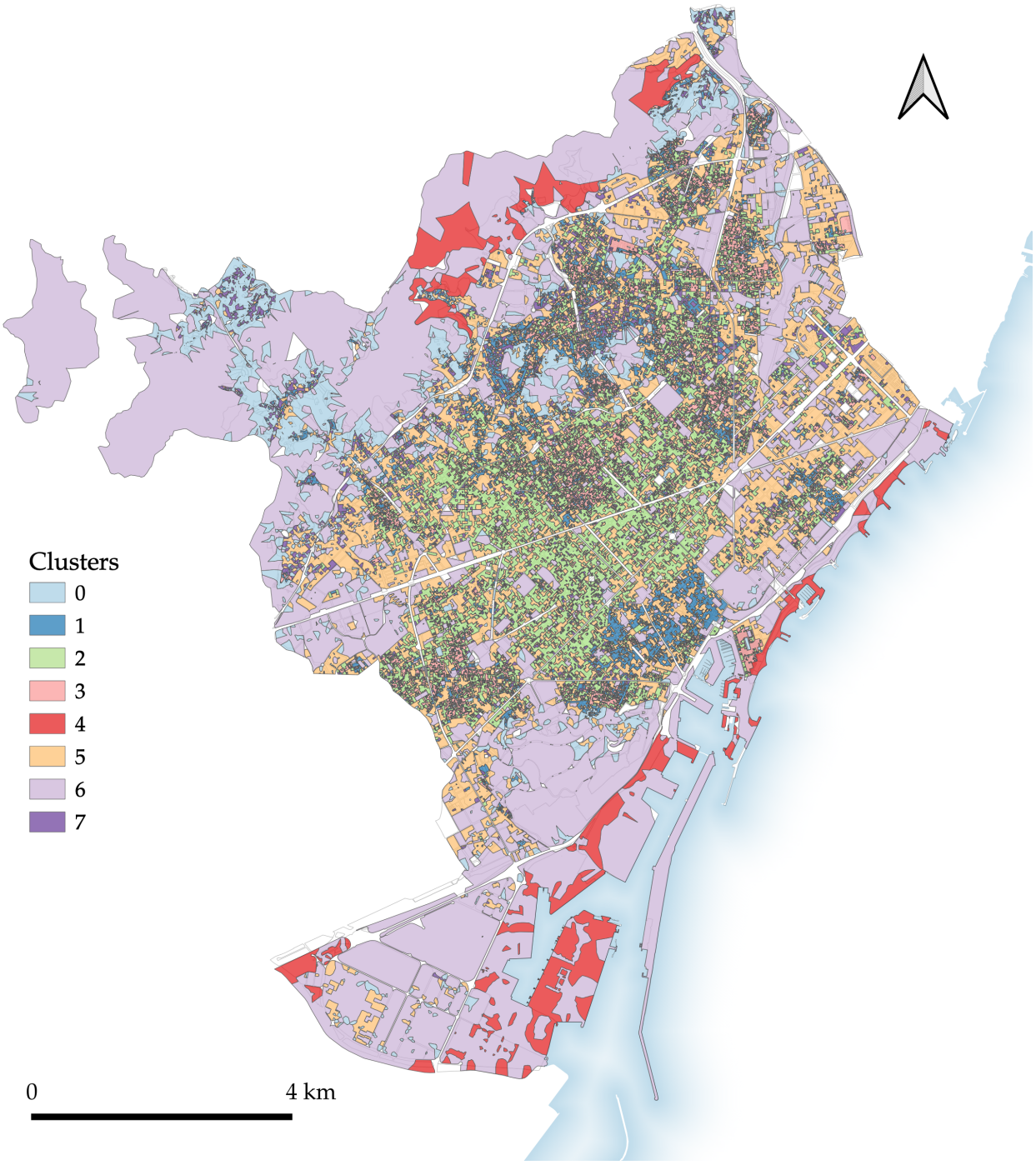

Figure A.2: Enclosed tessellation with primary characters.

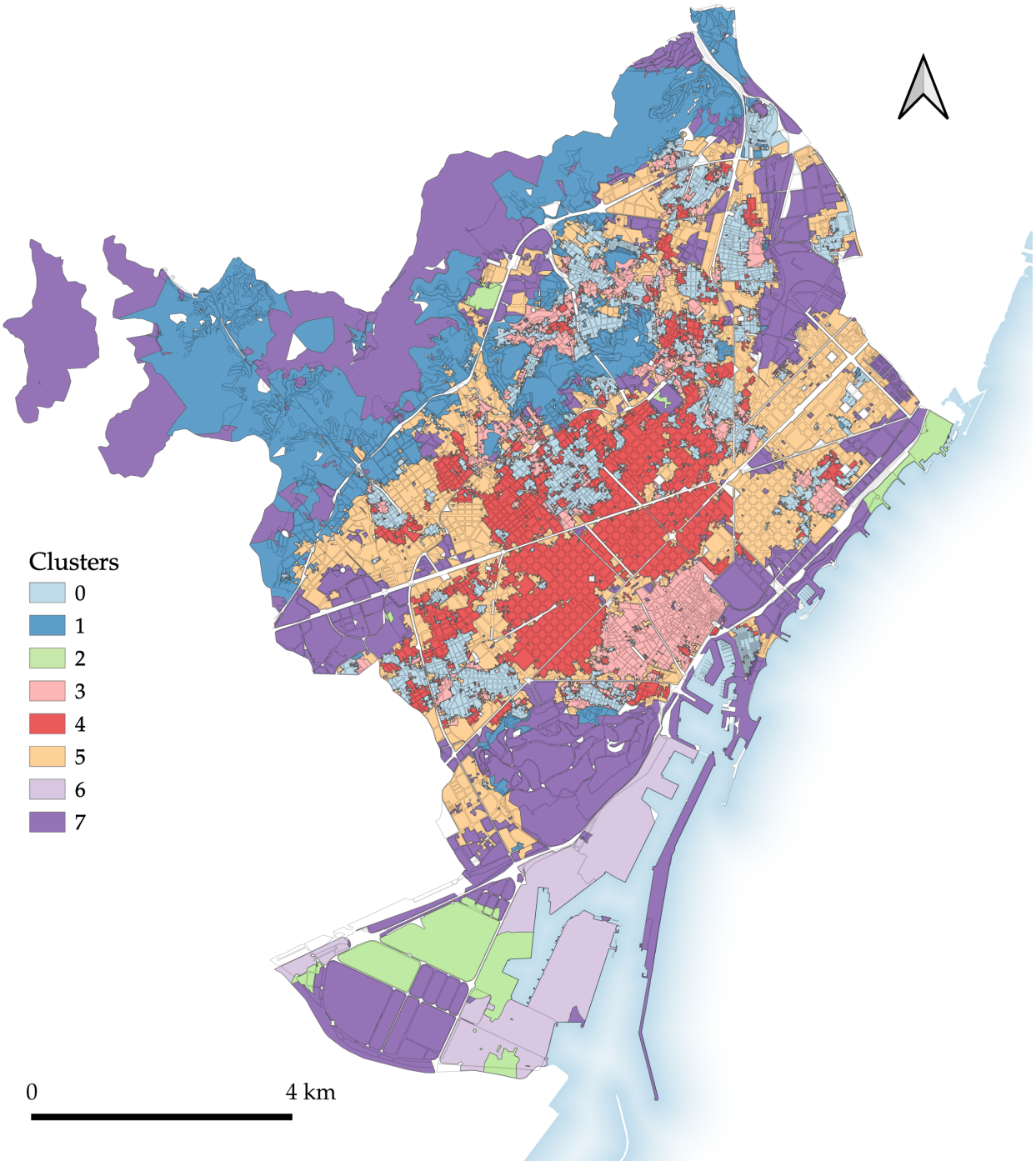

Figure A.3: Enclosed tessellation with contextual characters generated using 3rd order spatial weights.

Works Cited

A. S. Adair, J. N. Berry, and W. S. McGreal. 1996. Hedonic modelling, housing submarkets and residential valuation. *Journal of Property Research* 13, 1 (01 1996), 67–83. https://doi.org/10.1080/095999196368899 Publisher: Routledge.

Alessandro Araldi and Giovanni Fusco. 2019. From the street to the metropolitan region: Pedestrian perspective in urban fabric analysis. *Environment and Planning B: Urban Analytics and City Science* 46, 7 (09 2019), 1243–1263. https://doi.org/10.1177/2399808319832612 Publisher: SAGE Publications Ltd STM.

Daniel Arribas-Bel. 2014. Accidental, open and everywhere: Emerging data sources for the understanding of cities. *Applied Geography* 49 (05 2014), 45–53. https://doi.org/10.1016/j.apgeog.2013.09.012

Dani Arribas-Bel and Martin Fleischmann. 2021. Spatial Signatures - Understanding (urban) spaces through form and function. (2021).

Daniel Arribas-Bel, Miquel-Àngel Garcia-López, and Elisabet Viladecans-Marsal. 2019. Building(s and) cities: Delineating urban areas with a machine learning algorithm. *Journal of Urban Economics* (11 2019), 103217. https://doi.org/10.1016/j.jue.2019.103217

Nigel Baker. 2009. *A characterisation of the Historic Townscape of Central Hereford*. Technical Report. Hereford. https://www.herefordshire.gov.uk/downloads/file/1438/historic-townscape-of-central-herefordpdf

Christopher Barrington-Leigh and Adam Millard-Ball. 2017. The world's user-generated road map is more than 80% complete. *PLOS ONE* 12, 8 (08 2017), e0180698. https://doi.org/10.1371/journal.pone.0180698

Evgeniya Bobkova, Meta Berghauser Pont, and Lars Marcus. 2021. Towards analytical typologies of plot systems: Quantitative profile of five European cities. *Environment and Planning B: Urban Analytics and City Science* 48, 4 (05 2021), 604–620. https://doi.org/10.1177/2399808319880902 Publisher: SAGE Publications Ltd STM.

Geoff Boeing. 2017. OSMnx: New methods for acquiring, constructing, analyzing, and visualizing complex street networks. *Computers, Environment and Urban Systems* 65 (09 2017), 126–139. https://doi.org/10.1016/j.compenvurbsys.2017.05.004

Steven C. Bourassa, Foort Hamelink, Martin Hoesli, and Bryan D. MacGregor. 1999. Defining Housing Submarkets. *Journal of Housing Economics* 8, 2 (06 1999), 160–183. https://doi.org/10.1006/jhec.1999.0246

Isaac Brodsky. 2018. H3: Uber's Hexagonal Hierarchical Spatial Index. https://eng.uber.com/h3/

Alessia Calafiore, Gregory Palmer, Sam Comber, Daniel Arribas-Bel, and Alex Singleton. 2021. A geographic data science framework for the functional and contextual analysis of human dynamics within global cities. *Computers, Environment and Urban Systems* 85 (01 2021), 101539. https://doi.org/10.1016/j.compenvurbsys.2020.101539

José Carpio-Pinedo, Manuel Benito-Moreno, and Patxi J. Lamíquiz-Daudén. 2021. Beyond land use mix, walkable trips. An approach based on parcel-level land use data and network analysis. *Journal of Maps* 17, 1 (01 2021), 23–30. https://doi.org/10.1080/17445647.2021.1875063 Publisher: Taylor & Francis.

Nicola Colaninno, Josep Roca, and Karin Pfeffer. 2011. Urban form and compactness of morphological homogeneous districts in Barcelona: towards an automatic classification of similar built-up structures in the city. (01 2011).

M. R. G. Conzen. 1960. Alnwick, Northumberland: A Study in Town-Plan Analysis. *Transactions and Papers* (*Institute of British Geographers*) 27 (1960). https://doi.org/10.2307/621094 Publisher: Royal Geographical Society (with the Institute of British Geographers), Wiley.

Jacob Dibble. 2016. *Urban Morphometrics: Towards A Quantitative Science of Urban Form*. Ph. D. Dissertation. Glasgow. http://digitool.lib.strath.ac.uk/R/?func=dbin-jump-full&object_id=27955

Jacob Dibble, Alexios Prelorendjos, Ombretta Romice, Mattia Zanella, Emanuele Strano, Mark Pagel, and Sergio Porta. 2015. Urban Morphometrics: Towards a Science of Urban Evolution. *arXiv:1506.04875* [*physics*] (06 2015). http://arxiv.org/abs/1506.04875 arXiv: 1506.04875.

Jacob Dibble, Alexios Prelorendjos, Ombretta Romice, Mattia Zanella, Emanuele Strano, Mark Pagel, and Sergio Porta. 2019. On the origin of spaces: Morphometric foundations of urban form evolution. *Environment and Planning B: Urban Analytics and City Science* 46, 4 (05 2019), 707–730. https://doi.org/10.1177/2399808317725075 Publisher: SAGE Publications Ltd STM.

David Donoho. 2017. 50 Years of Data Science. *Journal of Computational and Graphical Statistics* 26, 4 (10 2017), 745–766. https://doi.org/10.1080/10618600.2017.1384734 Publisher: Taylor & Francis.

Juan C. Duque, Henry Laniado, and Adriano Polo. 2018. S-maup: Statistical test to measure the sensitivity to the modifiable areal unit problem. *PLOS ONE* 13, 11 (11 2018), e0207377. https://doi.org/10.1371/journal.pone.0207377 Publisher: Public Library of Science.

Martin Fleischmann. 2019. momepy: Urban Morphology Measuring Toolkit. *Journal of Open Source Software* 4, 43 (11 2019), 1807. https://doi.org/10.21105/joss.01807

Martin Fleischmann. 2021. *The Urban Atlas: Methodological Foundation of a Morphometric Taxonomy of Urban Form*. Ph. D. Dissertation. University of Strathclyde, Glasgow.

Martin Fleischmann, Alessandra Feliciotti, and William Kerr. 2021a. Evolution of Urban Patterns: Urban Morphology as an Open Reproducible Data Science. *Geographical Analysis* n/a, n/a (2021). https://doi.org/10.1111/gean.12302

Martin Fleischmann, Alessandra Feliciotti, Ombretta Romice, and Sergio Porta. 2020a. Morphological tessellation as a way of partitioning space: Improving consistency in urban morphology at the plot scale. *Computers, Environment and Urban Systems* 80 (03 2020), 101441. https://doi.org/10.1016/j.compenvurbsys.2019.101441

Martin Fleischmann, Alessandra Feliciotti, Ombretta Romice, and Sergio Porta. 2021b. Methodological Foundation of a Numerical Taxonomy of Urban Form. *arXiv:2104.14956* [*physics*] (04 2021). http://arxiv.org/abs/2104.14956 arXiv: 2104.14956.

Martin Fleischmann, Ombretta Romice, and Sergio Porta. 2020b. Measuring urban form: Overcoming terminological inconsistencies for a quantitative and comprehensive morphologic analysis of cities. *Environment and Planning B: Urban Analytics and City Science* (03 2020), 2399808320910444. https://doi.org/10.1177/2399808320910444 Publisher: SAGE Publications Ltd STM.

A S Fotheringham and D W S Wong. 1991. The Modifiable Areal Unit Problem in Multivariate Statistical Analysis. *Environment and Planning A: Economy and Space* 23, 7 (07 1991), 1025–1044. https://doi.org/10.1068/a231025 Publisher: SAGE Publications Ltd.

Yiying Hao, Jian Kang, and Johannes D. Krijnders. 2015. Integrated effects of urban morphology on birdsong loudness and visibility of green areas. *Landscape and Urban Planning* 137 (05 2015), 149–162. https://doi.org/10.1016/j.landurbplan.2015.01.006

Marco Helbich, Wolfgang Brunauer, Julian Hagenauer, and Michael Leitner. 2013. Data-Driven Regionalization of Housing Markets. *Annals of the Association of American Geographers* 103, 4 (07 2013), 871–889. https://doi.org/10.1080/00045608.2012.707587 Publisher: Routledge.

D. Holt, D. G. Steel, M. Tranmer, and N. Wrigley. 1996. Aggregation and Ecological Effects in Geographically Based Data. *Geographical Analysis* 28, 3 (1996), 244–261. https://doi.org/10.1111/j.1538-4632.1996.tb00933.x

Institut Municipal d'Informàtica. 2017. Administrative units of the city of Barcelona - Open Data Barcelona. (06 2017). https://opendata-ajuntament.barcelona.cat/data/en/dataset/20170706-districtes-barris Type: dataset.

Warren C Jochem, Douglas R Leasure, Oliver Pannell, Heather R Chamberlain, Patricia Jones, and Andrew J Tatem. 2021. Classifying settlement types from multi-scale spatial patterns of building footprints. *Environment and Planning B: Urban Analytics and City Science* 48, 5 (06 2021), 1161–1179.

https://doi.org/10.1177/2399808320921208 Publisher: SAGE Publications Ltd STM.

Jamal Jokar Arsanjani, Alexander Zipf, Peter Mooney, and Marco Helbich (Eds.). 2015. *OpenStreetMap in GIScience*. Springer International Publishing. https://doi.org/10.1007/978-3-319-14280-7

Tom Kauko, Pieter Hooimeijer, and Jacco Hakfoort. 2002. Capturing Housing Market Segmentation: An Alternative Approach based on Neural Network Modelling. *Housing Studies* 17, 6 (11 2002), 875–894. https://doi.org/10.1080/02673030215999 Publisher: Routledge.

Berna Keskin and Craig Watkins. 2017. Defining spatial housing submarkets: Exploring the case for expert delineated boundaries. *Urban Studies* 54, 6 (05 2017), 1446–1462. https://doi.org/10.1177/0042098015620351 Publisher: SAGE Publications Ltd.

Gary King. 1997. *A Solution to the Ecological Inference Problem: Reconstructing Individual Behavior from Aggregate Data*. Princeton University Press, Princeton.

Eli Knaap, Renan Xavier Cortes, Sergio Rey, Dani Arribas-Bel, James Gaboardi, Martin Fleischmann, and Patty Frontiera. 2021. *pysal/tobler: Release v0.8.2*. Zenodo. https://doi.org/10.5281/zenodo.5047613

Karl Kropf. 1997. When is a plot not a plot: problems in representation and interpretation. Citation Key: kropf1997plot.

Karl Kropf. 2017. *The Handbook Of Urban Morphology*. John Wiley & Sons, Ltd, Chichester, UK. https://doi.org/10.1002/9781118747711 DOI: 10.1002/9781118747711.

Karl Kropf. 2018. Plots, property and behaviour. *Urban Morphology* 22 (04 2018), 5–14.

David Lazer and Jason Radford. 2017. Data ex Machina: Introduction to Big Data. *Annual Review of Sociology* 43, 1 (2017), 19–39. https://doi.org/10.1146/annurev-soc-060116-053457

Michael Mehaffy, Sergio Porta, Yodan Rofè, and Nikos Salingaros. 2010. Urban nuclei and the geometry of streets: The 'emergent neighborhoods' model. *URBAN DESIGN International* 15, 1 (03 2010), 22–46. https://doi.org/10.1057/udi.2009.26

Josep Mercadé Aloy, Francesc Magrinyà Torner, and Marina Cervera Alonso de Medina. 2018. Descifrando la forma urbana: un análisis de patrones de agrupamiento basado en SIG. *GeoFocus Revista Internacional de Ciencia y Tecnología de la Información Geográfica* 22 (12 2018), 3–19. https://doi.org/10.21138/GF.612

Peter Mooney and Marco Minghini. 2017. *A Review of OpenStreetMap Data*. Ubiquity Press, 37–59. https://doi.org/10.5334/bbf.c

Tomoki Nakaya. 2000. An Information Statistical Approach to the Modifiable Areal Unit Problem in Incidence Rate Maps. *Environment and Planning A: Economy and Space* 32, 1 (01 2000), 91–109. https://doi.org/10.1068/a31145 Publisher: SAGE Publications Ltd.

Vítor Oliveira. 2016. *Urban Morphology: An Introduction to the Study of the Physical Form of Cities*. Springer International Publishing. https://doi.org/10.1007/978-3-319-32083-0

S Openshaw. 1984. Ecological Fallacies and the Analysis of Areal Census Data. *Environment and Planning A: Economy and Space* 16, 1 (01 1984), 17–31. https://doi.org/10.1068/a160017 Publisher: SAGE Publications Ltd.

Stan Openshaw and P. J Taylor. 1979. *A Million or so Correlation Coefficients: Three Experiments on the Modifiable Areal Unit Problem*. Pion, London, 127–144. https://liverpool.idm.oclc.org/login?url=https://search.ebscohost.com/login.aspx?direct=true&db=cat00003a&AN=lvp.b1216796&site=eds-live&scope=site

Risa Palm. 1978. Spatial Segmentation of the Urban Housing Market. *Economic Geography* 54, 3 (07 1978), 210–221. https://doi.org/10.2307/142835 Publisher: Routledge _eprint: https://www.tandfonline.com/doi/pdf/10.2307/142835.

Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, Jake Vanderplas, Alexandre Passos, David Cournapeau, Matthieu Brucher, Matthieu Perrot, and Édouard Duchesnay. 2011. Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research* 12, 85 (2011), 2825–2830. http://jmlr.org/papers/v12/pedregosa11a.html

Alasdair Rae. 2015. Online Housing Search and the Geography of Submarkets. *Housing Studies* 30, 3 (04 2015), 453–472. https://doi.org/10.1080/02673037.2014.974142 Publisher: Routledge.

Philipp Rode, Christian Keim, Guido Robazza, Pablo Viejo, and James Schofield. 2014. Cities and Energy: Urban Morphology and Residential Heat-Energy Demand. *Environment and Planning B: Planning and Design* 41, 1 (02 2014), 138–162. https://doi.org/10.1068/b39065 Publisher: SAGE Publications Ltd STM.

Chinmoy Sarkar. 2013. *The Science of Healthy Cities: Deciphering the associations between urban morphometrics and health outcomes*. Ph. D. Dissertation. http://orca.cf.ac.uk/47613/

Alex Singleton and Daniel Arribas-Bel. 2021. Geographic Data Science. *Geographical Analysis* 53, 1 (2021), 61–75. https://doi.org/10.1111/gean.12194

Susan J. Smith and Moira Munro. 2013. *The Microstructures of Housing Markets*. Routledge. Google-Books-ID: Bc3hAQAAQBAJ.

Patricio Soriano. 2021. *Spanish Inspire Catastral Downloader (V1.1)*. https://github.com/sigdeletras/Spanish_Inspire_Catastral_Downloader

Mark Tranmer and David G Steel. 2001. Ignoring a Level in a Multilevel Model: Evidence from UK Census Data. *Environment and Planning A: Economy and Space* 33, 5 (05 2001), 941–948. https://doi.org/10.1068/a3317 Publisher: SAGE Publications Ltd.

Laura Vaughan, David L. Chatford Clark, Ozlem Sahbaz, and Mordechai (Muki) Haklay. 2005. Space and exclusion: does urban morphology play a part in social deprivation? *Area* 37, 4 (2005), 402–412. https://doi.org/10.1111/j.1475-4762.2005.00651.x

Levi Wolf and Elijah Knaap. 2019. Learning Geographical Manifolds: A Kernel Trick for Geographical Machine Learning. (05 2019). https://doi.org/10.31235/osf.io/75s8v

Levi J Wolf, Elijah Knaap, and Sergio Rey. 2021. Geosilhouettes: Geographical measures of cluster fit. *Environment and Planning B: Urban Analytics and City Science* 48, 3 (03 2021), 521–539. https://doi.org/10.1177/2399808319875752 Publisher: SAGE Publications Ltd STM.

Ming Zhang and Nishant Kukadia. 2005. Metrics of Urban Form and the Modifiable Areal Unit Problem. *Transportation Research Record* 1902, 1 (01 2005), 71–79. https://doi.org/10.1177/0361198105190200109 Publisher: SAGE Publications Inc.

Daniel Zwillinger and Stephen Kokoska. 1999. *CRC Standard Probability and Statistics Tables and Formulae*. CRC Press.