Consumer Data Research Centre
An ESRC Data Investment
idealista

Improving Housing Submarket Spatial Segmentation

Matthew Law[1], Dani Arribas-Bel[1], Martin Fleischmann[1], Juan Ramón Selva-Royo[2]
[1]Geographic Data Science Lab, University of Liverpool, [2]idealista

Background and Motivation

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 used to represent these submarkets, either individually or through a particular aggregation. When this approach is used to analyse the housing market, for example when calculating regional price indices based on these spatial units, the end result can misrepresent the nature of the underlying property market(s) being studied, an example of the Modifiable Areal Unit Problem. For instance, if a 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.

Data and Methods

The methodology comprises two main sections:

1. Generating spatial segmentations.
2. Assessing these novel segmentations.

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.

Data

The primary data source used is open data from the Spanish Cadastre, which is made available online[1]. This provides detailed information about every building in the country, including its location, footprint, and height. Data for all drivable roads in the study area is obtained from OpenStreetMap via the `OSMnx` Python package.

Base spatial units

Different versions of the segmentations use different base spatial units:

- Morphological tessellation (MT) exhaustively divides the study area into Voronoi polygons ('MT cells') based on the nearest building.
- Enclosed tessellation (ET) first divides the space into 'enclosures'—areas enclosed by drivable streets—using the OSM road network, before following the MT methodology to further subdivide into 'ET cells'.
- 'H3 cells' are regular hexagonal grid cells taken from Uber's open H3 grid indexing system.

Measuring morphometric characters

34 primary characters were used to quantify different elements of urban morphology, calculated using the `momepy` Python package. These include measures like 'building volume' and 'proportion of tessellation cell covered by buildings'.

Incorporating contextual information

'Contextual characters' are generated from a spatial lag of the characters in surrounding cells. A cell is defined as a 'neighbour' if it is a certain

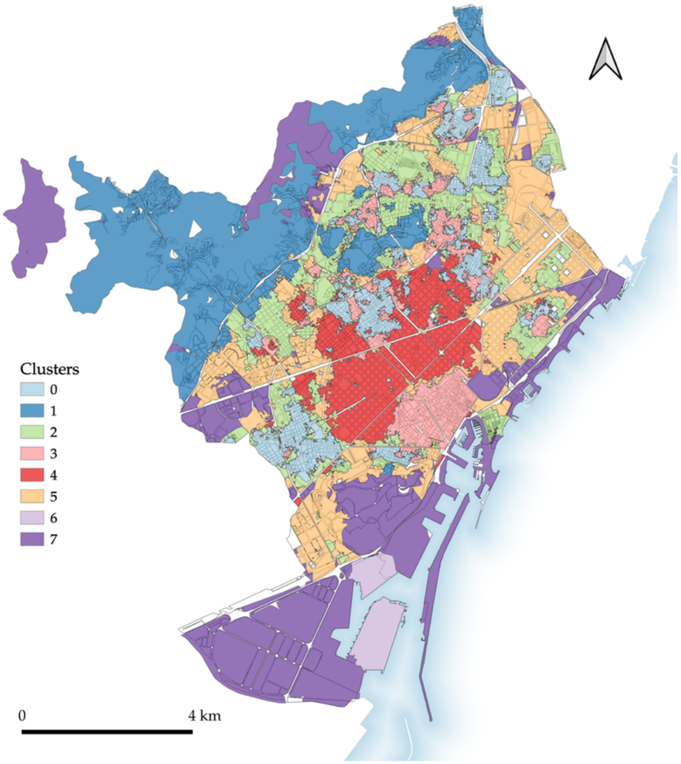
[1] http://www.sedecatastro.gob.es/

number of topological steps away from another cell. The interquartile mean of each cell counted as a neighbour is then computed for each character, generating contextual characters which incorporate information about the surrounding morphology and smooth the subsequent clustering.

Clustering

The main clustering algorithm used models the input data (usually the contextual characters) as a mixture of Gaussians, assigning each base spatial unit to one of these Gaussians (i.e. clusters). An alternative to this Gaussian Mixture Model (GMM) algorithm is Agglomerative Clustering, which imposes a spatial contiguity constraint: all clusters generated must be spatially contiguous. Both algorithms are implemented using the scikit-learn Python package.

Figure 1: an urban morphology based spatial segmentation of Barcelona using ET cells, 5th order spatial weights, and GMM clustering with 8 clusters



Assessing the spatial segmentations

The spatial segmentations generated are assessed for two different purposes in two different ways:

1. How well the segmentation reflects patterns of urban morphology is assessed through a visual assessment of the spatial units. This is achieved by overlaying the segmentation on a basic map of the city, as shown in **Figure 1**, which plots a segmentation using ET cells, 5th order spatial weights, and GMM clustering with 8 clusters.
2. How well the segmentation captures variation in property prices (as a proxy for housing submarkets). This is achieved by measuring the Quartile Coefficient of Dispersion for the sale prices of properties located in each cluster, using proprietary data provided by idealista.

Key Findings

Many of the key findings from the dissertation concern the ways in which varying the choices made over the course of the methodology affects the segmentations consequently generated.

The **spatial unit** used to cluster is key to producing segmentations which accurately reflect the differing urban morphology of a city. When comparing a segmentation of ET cells to an equivalent using H3 cells with identical characters (i.e. characters generated on the ET cells and interpolated to H3 cells), the ET segmentation much better reflects the city's urban morphology.

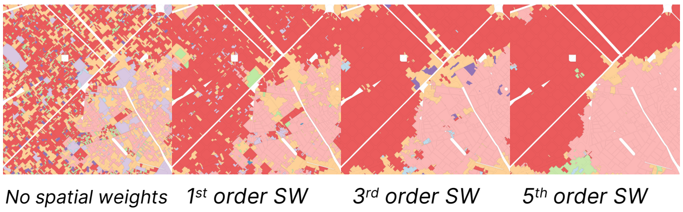
The use of a **spatial lag** of the original morphometric characters proved to be key to producing clear segmentations, particularly when using smaller base spatial units. **Figure 2** shows the effect of an otherwise identical segmentation on contextual characters constructed using different order spatial weights.

The **GMM algorithm** generally worked well, but the **Agglomerative Clustering algorithm** did not produce the desired result: even when a high number of clusters was stipulated, it would generate one cluster which covers the large majority of the study area, and several clusters which cover only very small areas.

Value of the research

Academically, the research adds to a growing literature of quantitative computational urban morphology. **Commercially**, the research builds on ongoing work by idealista to improve the spatial segmentations they use for analysis of geospatial housing data, based on urban morphology.

Figure 2: the effect of an otherwise identical segmentation on contextual characters constructed using different order spatial weights


No spatial weights
1st order SW
3rd order SW
5th order SW