

Exploring the spatial segmentation of housing markets from online listings

David Abella*,^{†,1} Johann H. Martínez^{†,2}, Mattia Mazzoli,³ Thibault Le Corre,⁴ Julien Migozzi,⁵
Eduard Alonso-Paulí,⁶ Rafel Crespi-Cladera,⁶ Thomas Louail,^{7,8} and José J. Ramasco¹

¹*Instituto de Física Interdisciplinar y Sistemas Complejos IFISC (CSIC-UIB), 07122 Palma de Mallorca, Spain*

²*Instituto de Matemática Interdisciplinar, Departamento de Análisis Matemático y Matemáticas Aplicadas,
and GISC, Universidad Complutense, 28040 Madrid, Spain*

³*ISI Foundation, via Chisola 5, 10126 Turin, Italy*

⁴*Département de Géographie, Université de Montréal, Montréal, Canada*

⁵*School of Geography and the Environment, University of Oxford, Oxford, United Kingdom*

⁶*Departament d'Economia de l'Empresa, Universitat de les Illes Balears, 07122 Palma de Mallorca, Spain*

⁷*UMR 8504 Géographie-cités (CNRS - EHESS - Université Panthéon-Sorbonne,
Université Paris Cité), Campus Condorcet, 93322 Aubervilliers, France*

⁸*UMR 5194 PACTE (CNRS - Sciences Po Grenoble - Université Grenoble Alpes), 38000 Grenoble, France*

The real estate market shows an inherent connection to space. Real estate agencies unevenly operate and specialize across space, price and type of properties, thereby segmenting the market into submarkets. We introduce here a methodology based on multipartite networks to detect the spatial segmentation emerging from data on housing online listings. Considering the spatial information of the listings, we build a bipartite network that connects agencies and spatial units. This bipartite network is projected into a network of spatial units, whose connections account for similarities in the agency ecosystem. We then apply clustering methods to this network to segment markets into spatially-coherent regions, which are found to be robust across different clustering detection algorithms, discretization of space and spatial scales, and across countries with case studies in France and Spain. This methodology addresses the long-standing issue of housing market segmentation, relevant in disciplines such as urban studies and spatial economics, and with implications for policymaking.

[†]Equal contribution.

*Corresponding author: david@ifisc.uib-csic.es

I. INTRODUCTION

The spatial dimension of housing markets is a crucial aspect for urban studies and planning. Understanding the spatial segmentation of the housing market into submarkets [1, 2] has important implications for real estate valuation and investment decisions, which together affect urban development and social equity [2]. Spatial segmentation is the product of many factors such as residential location and the proximity to amenities [2], differences in housing stock [3], price levels [4], and consumer preferences [5].

The spatial division of the real estate market has been studied from different perspectives and with different methods in the literature. Some studies have examined the spatial segmentation of the urban housing market focusing on neighborhood correlations of housing prices [6], the spatial effects of urban public policies on housing values [7], the neighborhood quality and accessibility effects on housing prices [8], while others have determined if a specific property market is spatially segmented into submarkets, and whether accounting for the existence of submarkets improves the accuracy of price modeling [3, 9]. This is especially important for hedonic pricing models that seek to incorporate spatial autocorrelation and heterogeneity [9–12]. Ref. [13] distinguishes two main approaches for spatial segmentation: using pre-defined ge-

ographical boundaries based on *a priori* knowledge, such as local administrative boundaries or expert areas used by market stakeholders, or relying on clustering methods to infer patterns from the structure of the data. For the latter, popular statistical approaches to divide space into submarkets are principal component analysis and hierarchical clustering [2, 4, 14].

The digitization of the housing market [15] provides untapped research opportunities for data-driven studies of market segmentation. With property portals being nowadays the dominant way to create and access market information, online listings constitute a new type of data to study housing markets [16–18]. Scholars studied the spatio-temporal distribution of housing prices [19, 20], revealed the persistence of spatial inequalities in the housing information landscape [21], predicted the social profile of neighborhoods [22], or detected the segmentation of the market from online search patterns [23]. Aside price, pictures or textual descriptions, a listing includes a critical piece of information: the identity of the marketing agency that has posted the listing on the portal. As such, listings constitute digital traces [24] of the work performed by real estate agencies when acquiring, selling or marketing on property portals. It is therefore possible to reconstruct, for each agency, its own portfolio of listings, whose volume and location patterns result from and reflect the heterogeneous practices and market shares of real estate agencies. By informing on *who sells where*, listings offer new ways to examine how real estate agencies unevenly operate and specialize across space, thereby segmenting the market into submarkets [25].

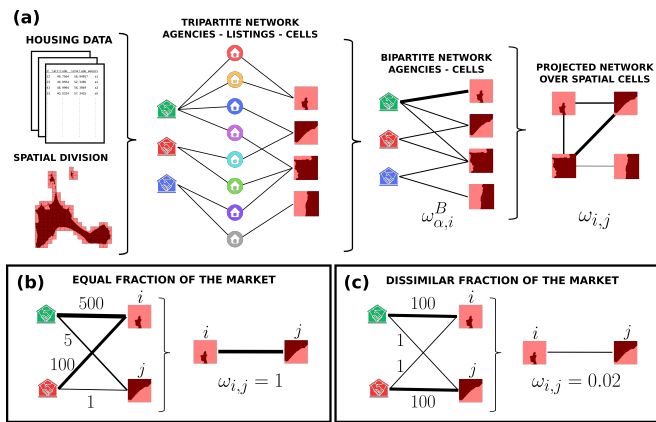


FIG. 1. **Bipartite network construction and projection.** (a) A tripartite network is constructed between real estate agencies, listings, and spatial units obtained from geolocalized housing data and the division of space in regular grid cells. In this network, each listing is connected to its real estate agency and the spatial cell where it is located. This simple tripartite network is contracted into a bipartite network linking agencies and cells, where the link weight $\omega_{\alpha,i}^B$ corresponds to the number of listings the agency α has in the spatial cell i . Finally, the network is then projected over the cells to form a weighed network of spatial units, where the weight $\omega_{i,j}$ of the link between cells i and j quantifies how much they are similar in the market – $\omega_{i,j}$ is properly defined by Equation (2). Two simple examples of the projection process are shown below: with (b) equal and (c) complementary listings distributions for the agencies in the cells.

There is ample evidence underlining how real estate agencies influence market segmentation by determining housing prices, sorting homebuyers into different market channels, and specializing in certain types of neighborhoods and market segments [3, 6, 25–27]. Furthermore, it has been shown that the definition of submarkets based on agencies is far superior to other segmentation techniques [28].

This work introduces a new method to identify the housing market segmentation using geospatial data, complex network analysis techniques, and taking as a basis the local ecosystem of real estate agencies. We build a network structure based on two factors: the *presence* of an agency within a particular area, and the *relative influence* of an agency in this area, determined by the agency’s proportional share of all listings located in the area. Our methodology is applied to the residential property market in 3 Spanish provinces and 3 French urban areas, for which we have a rich, high resolution dataset sourced from property portals. We find that the market in those regions is divided into a hierarchy of subregions. We test the robustness of our results against different community detection algorithms, scales, and administrative boundaries in different countries.

II. MATERIALS AND METHODS

A. Data description

For Spain, we analyze listings published on the portal *Idealista.com* [29]. The dataset covers a 2-year time period, from January 2017 to December 2018 and it comprises a comprehensive collection of online listings georeferenced with their (lat, long) coordinates in the Spanish provinces of Balearic Islands, Barcelona, and Madrid. These listings were posted by more than 50,000 real estate agencies, each identified with its unique id. There are about one million listings for sales, and over 800,000 for rentals.

French listings were obtained from the portal *SeLoger.com* [30]. The dataset includes all listings posted in the country over a 6-month period from July to December 2019 - representing over 2 million sale listings. Geographical information is only available at the administrative and census levels, such as ZIP codes (“*code postal*”), municipalities (“*communes*”), and census tracts (“*IRIS*”), the finest and basic scale for sub-municipal information in France. We focus on three major urban areas: Paris, Marseilles and Toulouse.

For both datasets, we focus on houses and apartments, and do not consider farms or rural parcels.

B. Building a network

We begin by discretizing the space into spatial units (square grid cells, municipalities, districts, postal codes, census-tracts, etc). This allows us to label each listing according to the spatial unit it falls into, along with the agency that posted this listing. By doing so, we build a tripartite relation between agencies, listings, and spatial units. Based on this structure, we can build a weighted bipartite network that connects agencies and spatial units, where the link weight $\omega_{\alpha,i}^B$ accounts for the number of listings posted by agency α that are located in the spatial unit i . The resulting network contains all the information about the spatial characteristics of the housing market.

Bipartite networks can be projected to create networks with a single type of nodes [31–33]. In our case, we project it to build a new weighted network connecting spatial units (see Fig. 1(a) for schematic representation, taking as an example the discretization of space with square grid cells). Let us assume that we have N spatial units and N^a real estate agencies. The set of all agencies operating in the entire area is $\{\alpha\}$, while the subset operating in the spatial unit i is denoted by $\{\alpha\}_i$. The fraction of listings in i that belong to a certain agency α is

$$f_{\alpha,i} = \frac{\omega_{\alpha,i}^B}{\sum_{\gamma \in \{\alpha\}_i} \omega_{\gamma,i}^B}, \quad (1)$$

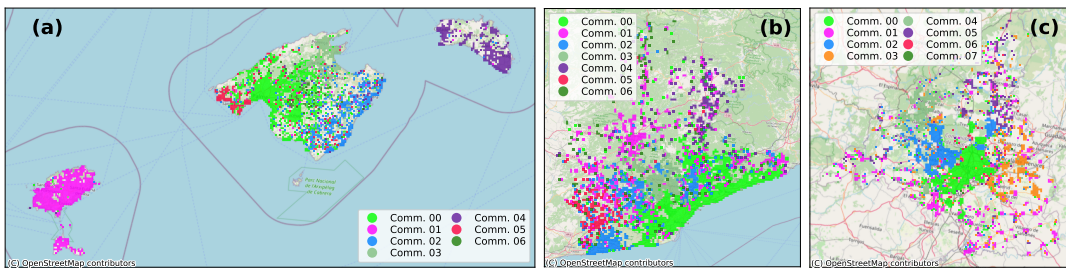


FIG. 2. **Market segmentation for 1 km square cells.** Communities from the projected network for the three Spanish provinces studied: Balearic Islands (a), Barcelona (b), and Madrid (c). The spatial cells are 1 km square cells. The communities shown are detected using the Louvain algorithm with a consensus clustering of 1000 realizations. The underground map data is rendered from OpenStreetMap under ODbL.

where the index γ runs over all the agencies operating in i . In the projected network, we define the influence weight between two spatial units i and j as

$$\omega_{i,j} = \frac{\sum_{\gamma \in \{\alpha\}_{ij}} f_{\gamma,i} f_{\gamma,j}}{\frac{1}{2} \left[\sum_{\gamma \in \{\alpha\}_i} f_{\gamma,i}^2 + \sum_{\beta \in \{\alpha\}_j} f_{\beta,j}^2 \right]}, \quad (2)$$

where $\{\alpha\}_{ij} \equiv \{\alpha\}_i \cap \{\alpha\}_j$ is the subset of agencies operating in i and j . The weight $\omega_{i,j} = 1$ if the agencies operating in i and j are the same, and cover an equal fraction of the market in both spatial units. If the market distribution is similar, but not equal, the weight will deviate from 1. Reciprocally, if no common agency is found across the two spatial units, the weight is zero and there is no link between them. Fig. 1(b) and 1(c) show examples of the influence weights between two spatial units with equal distribution of the listings in (b), for which $\omega_{i,j} = 1$, and a complementary distribution in (c) with a value of $\omega_{i,j} = 0.02$. Note that our influence weight is related to the participation ratio introduced by Derrida *et al.* in [34].

The projected network is thus built with the spatial units as nodes, which are connected with links weighted according to Equation 2. A group of spatial units strongly connected between them implies that they share a common ecosystem of agencies, that operate with a similar market share in these units. Searching for clusters in this weighted spatial network should therefore inform us on the spatial segmentation of the housing market, the clusters corresponding to submarkets. In the network literature, such clusters are commonly referred to as communities, with numerous methods proposed to detect them [35]. We use several classic community detection algorithms [36–40] that account for network weights, including Louvain [39], Infomap [37], and OSLOM [40]. These algorithms enable us to classify the spatial units into communities. Since these algorithms are stochastic, we perform several realizations of each method, and perform consensus clustering [41] for higher stability.

III. RESULTS

A. Segmenting the market according to agencies' operations

We start by analyzing the spatial segmentation that arises from the data geolocated in the Balearic Islands, Barcelona, and Madrid using 1 km-sided square cells. Fig. 2 presents the communities listed according to their size, from larger to smaller. Even though our methodology does not consider spatial proximity, we observe spatial segmentation in adjacent regions with few exceptions. For the Balearic Islands, we observe that spatial constraints, such as insular nature of the environment, affect the segmentation of the housing market: while the same submarket covers Minorca or Ibiza-Formentera, Majorca is divided into four different ones. It is noteworthy that the submarkets that emerge in all these three provinces are slightly larger than municipalities.

To study the robustness of identified submarkets in each of the three provinces, we run several community detection algorithms, and compare the communities obtained across realizations of different algorithms. We define as a network partition the classification of the cells in communities, $X = \{x_0, x_1, \dots, x_{|X|-1}\}$, where each community x_i is a set of cells. The partition X has $|X|$ communities in this notation. Every cell must be in at least one community, but in some clustering methods a cell may belong to several. In order to compare two partitions X and Y , we compute a confusion matrix C^{XY} in which each element is defined as

$$C_{ij}^{XY} = |x_i \cap y_j|, \quad (3)$$

where x_i and y_j are communities in the partitions X and Y , respectively, and $|\cdot|$ stands for the cardinal (number of elements) of a set. An element C_{ij}^{XY} can be zero if there is no overlap between the communities, and it can be large if the two communities coincide across the partitions. We reorder then the elements of the matrix C^{XY} to have the largest values in the pseudo-diagonal. Note that C^{XY} is not necessarily a squared matrix because the

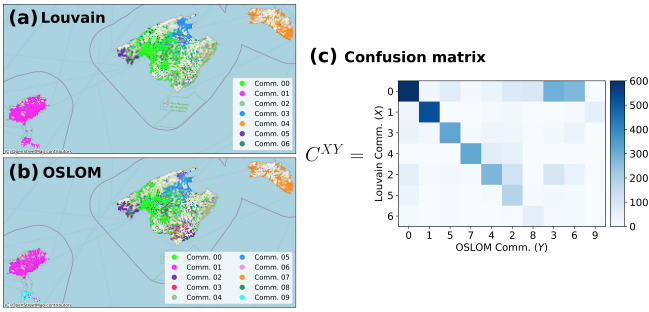


FIG. 3. **Agreement between different partitions.** Partition result of the community detection methods at the Balearic Islands using Louvain algorithm (a) and OSLOM method (b) 1km square cells. The confusion matrix C^{XY} of the two partitions (c), is ordered according to the maximum overlap. The underground map data is rendered from OpenStreetMap, under ODbL.

number of communities in each partition may differ. This process is essentially the identification of the communities in one partition that correspond to the communities in the other. This is a statistical match, given that the cells of a community in X may be distributed in several communities in Y . As shown in Fig. 3, if the partitions between the two methods are similar, we must observe a strong pseudo-diagonal in the confusion matrix. The sum of the elements of this pseudo-diagonal is the number of cells clustered in the same way in the two partitions. To compute a measure of the agreement between two partitions, we use the fraction $H(X, Y)$ [42, 43] defined as

$$H(X, Y) = \sum_{i=0}^{\min(|X|, |Y|)-1} \frac{C_{ii}^{XY}}{N}, \quad (4)$$

where the matrix C^{XY} is ordered to maximize the pseudo-diagonal, and N is the total number of cells. $H(X, Y)$ is a metric commonly used in the literature to compute the accuracy between community detection algorithms [44–51], its value is bounded in the interval $(0, 1]$, but it has the downside that $H(X, Y)$ depends on the size of the communities. To determine if the value of $H(X, Y)$ is significant, it is necessary to compare it with a randomized version of the partitions, $H(X_r, Y_r)$, in which the cells are reshuffled at random across the communities of each partition respecting the community sizes.

Figure 4(a-c) compares the three community detection algorithms (Louvain, OSLOM, and Infomap) used for different provinces. In all cases, the agreement between the communities detected from the real partition is higher than that of the randomized communities. The OSLOM-Louvain comparison exhibits the highest agreement, which is significant in all provinces. In the Balearic Islands, a robust and statistically significant agreement is evident among all methods. However, when examining Barcelona and Madrid, Infomap detects a large commu-

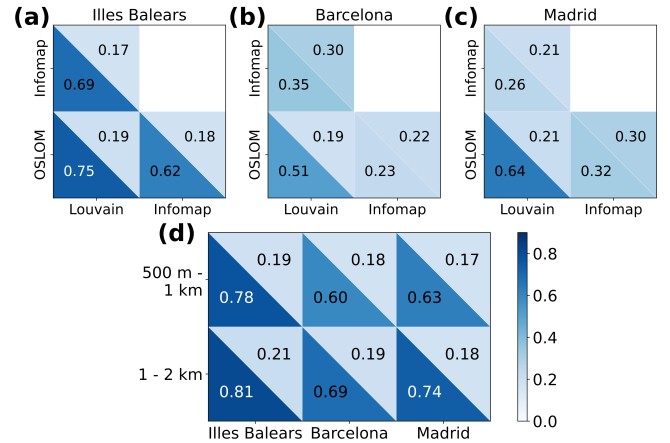


FIG. 4. **Agreement across three methods and cell sizes.** Agreement across the different community detection methods for the network in Balearic Islands (a), in the province of Barcelona (b) and of Madrid (c). The metric used to compute the agreement between method partitions is $H(X, Y)$, shown in the lower triangles for each pair of methods, denoted by X and Y . The upper triangles display the value $H(X_r, Y_r)$, being X_r and Y_r the partitions randomized (preserving the communities size). In (d), comparison of partitions obtained with the Louvain method for networks generated with different cell sizes: 500 m-sided vs 1 km-sided cells (top row), and 1 km-sided vs 2 km-sided cells (bottom row).

nity probably due to the high density of the network, and this does not compare well with the other methods which detect more communities. In fact, the value of $H(X, Y)$ approaches the one of the randomized model. This issue is absent in the Balearic Islands, where the network has a stronger intrinsic spatial division into different islands.

So far, we have focused on the results for the networks built with 1 km-sided square cells. It is, nevertheless, important to check whether the results may vary depending on the scale of the unit cells. We thus recalculate the networks taking as basis square cells of side 500 m and 2 km and compute the communities using the Louvain method with consensus clustering. The cells of the different scales have been delimited to keep spatial coherence: four 500 m cells form one of the 1 km cells used in the previous figures, and four 1 km cells aggregate to form a 2 km cell. This hierarchical structure allows us to compare communities at various levels because we can identify the cells across scales. For example, if a 2 km cell belongs to a community, then the four 1 km cells composing it share the same community label. In parallel, we also run the community detection algorithm in the network composed of 1 km cells, and then we can use the confusion matrix and $H(X, Y)$ to compare the partitions at these two scales using 1 km cells. Note that the calculation of $H(X, Y)$ requires the same number of basic units in the two partitions. Figure 4(d) shows the results of this analysis, where we use 1 km-sided cells as a reference for comparison with the other scales. In all

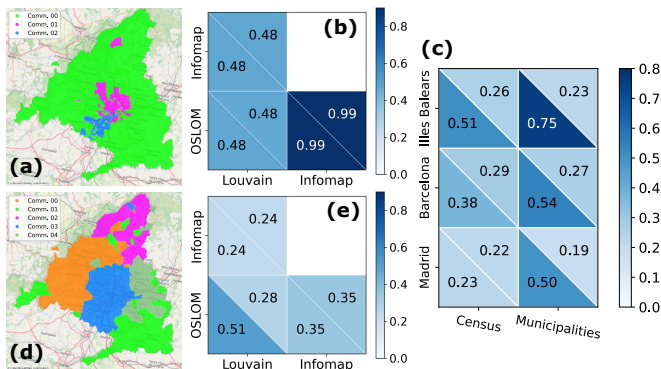


FIG. 5. **Community detection from networks using administrative spatial units.** Communities detected using census areas (a) and municipalities (d) as spatial units to build the network in Madrid. The clustering method employed is the Louvain algorithm. The agreement across the different methods for the census (b) and municipalities (e). (f) shows the communities' agreement between 1 km cells and administrative boundaries networks for all Spanish provinces. The agreement in (b)-(c)-(e) is computed using $H(X, Y)$ (lower triangles) compared with the value randomizing the communities (upper triangles). The underground map data is rendered by OpenStreetMap, under ODbL.

cases, we notice a consistently high and statistically significant level of agreement. This demonstrates that our methodology generates communities that remain robust across the three spatial scales.

B. Comparison with networks obtained from administrative boundaries

In this section, we examine how incorporating administrative spatial boundaries to build networks impacts the detection of communities. In many cases, the geographical information for listings is only available at the level of existing administrative boundaries and statistical units, which are by design more heterogeneous than square cells.

We aggregate listings into administrative and statistical spatial units to determine if the emergent submarkets are stable and consistent when comparing with the ones observed with the networks built with square cells. In this case, we consider municipalities and census tracts as they are the most common administrative divisions applied to spatial statistics.

Fig. 5 shows the communities found in the province of Madrid. We observe clear differences between the results obtained using census tracts (Fig. 5(a)-(b)) and using municipalities (Fig. 5(d)-(e)). The results for census tracts are characterized by a large community that covers almost all the territory and the agreement between methods is not significant. In contrast, the results using municipalities have a good and significant OSLOM-Louvain

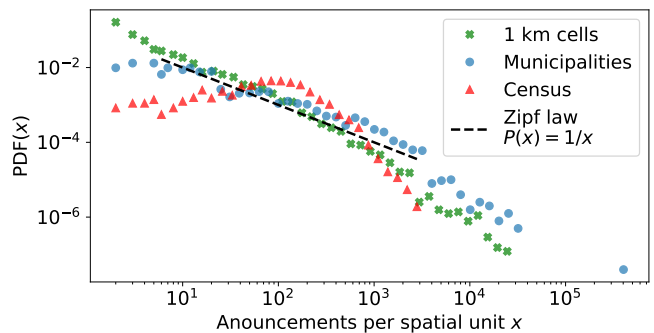


FIG. 6. **Distribution of listings for different spatial units.** Each spatial unit is shown by a different color and marker: green crosses (1 km-sided cells), blue circles (municipalities), and red triangles (census). The dashed black line shows the slope of a Zipf law distribution.

agreement. Keeping Louvain as the reference method, we compare the partitions of the networks originated from 1 km, census tracts, and municipalities in Fig. 5(c). The communities in the networks using cells and municipalities show significant agreement, while those based on census tracts show non-significant values in Barcelona and Madrid.

While the distribution of listings per spatial unit in the other cases follows a heterogeneous distribution, well-described by a Zipf law, the one for census tracts follows a more homogeneous distribution (see Fig. 6). This effect is a consequence of how the census tracts are built, forcing the population in each unit to be similar by a heterogeneous selection of the space included in each unit. This distribution is directly translated into the network weights and thus impacts the spatial segmentation method.

C. Recovering the submarkets from census level data

Multiple datasets, such as our French data, are available at census level. To maintain the broad applicability of our spatial segmentation methodology, we have devised a data aggregative method to recover the results obtained at the cell and municipality levels. This technique enables us to restore the Zipf law pattern using data gathered at the census level and to find similar segmentation results regardless of the basic spatial units.

We start with listings at a census scale, such that each listing is associated to an agency and a census tract. The first step is to divide the space into square cells, as we did in Section II. The cells intersect with the census tracts. We then associate each listing to a cell with a probability proportional to the overlapping area between the listing census tract and the cell. This process is repeated for all the listings to reconstruct a tripartite network of agencies-listings-cells, from which we can follow the

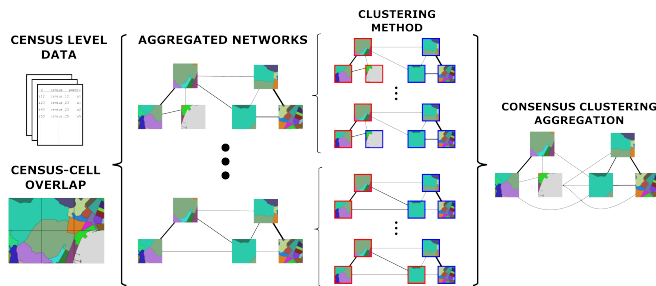


FIG. 7. **Stochastic aggregative method using census level data.** From a census-level listing and the spatial division of the census in square cells, we generate an ensemble of networks. In this ensemble, each listing within a census tract is associated to a cell with a probability based on the overlapping area between the census and the cell. For each of these cell networks, we run a community detection algorithm multiple times. The next step involves combining the results from these partitioned networks through consensus clustering, resulting in an aggregated network.

methodology explained to reach a cell-cell network and a segmentation in submarkets (communities). We observe that in the final networks the Zipf law distribution of listings per cell is recovered.

Since the assignation of listings to cells is stochastic, the projected network is different each time the process is repeated. To avoid uncertainty, we construct an ensemble of these networks. For each network, we run the community detection algorithm multiple times. Once our cells are labeled with a community, we perform consensus clustering to aggregate all partitions from all aggregated networks of our ensemble into a single consensus aggregated network. We represented this process in detail on Fig. 7.

To verify the results of the aggregative method, we perform a comparison of the submarkets obtained out of different networks. Starting with our Spanish data, where the listings are geolocated using exact coordinates, we build networks at the level of 1 km cells, census tracts and municipalities. We then apply the method to aggregate the census tracts to the cells. This gives us a fourth family of networks, which we call aggregated cells network. We then run community detection methods and compare them across the networks, taking as a basis the partition obtained from the network of aggregated cells (see Fig. 8). For all cases, the agreement exhibited by partitions of the aggregated cells network and the original cells or the municipalities is very high (and significant compared to the randomized communities). Therefore, by reconstructing the network with the aggregative method, we recover the original communities at the cell and municipality levels and avoid the issues caused by the natural spatial heterogeneity of census tracts.

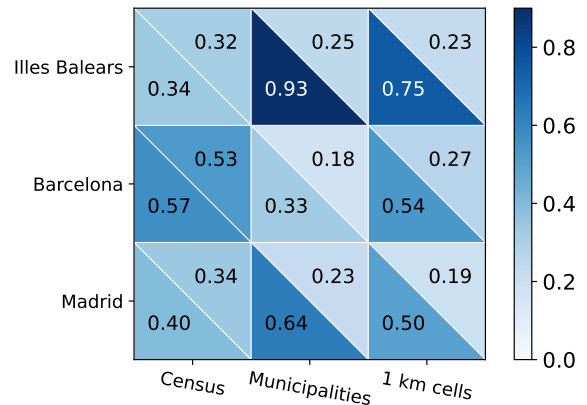


FIG. 8. **Comparison between the communities from aggregated cells network and other spatial units.** Each column shows the agreement between the communities of the 1 km aggregated cells networks (from census data) and the networks obtained from the other spatial units: Census, Municipalities, and 1 km cells from the original latitude longitude coordinate data. Each row shows the results for each province: Balearic Islands, Barcelona and Madrid. The agreement is computed via the fraction of correctly detected cells $H(X, Y)$ (lower triangles) compared with the value randomizing the communities (upper triangles).

D. Comparison across countries

In this section, we investigate whether the emergent spatial segmentation revealed by our method is a unique feature of the Spanish market, or can be understood as a more general phenomenon across geographical contexts. To this end, we use listing data for three major French urban areas, namely, Marseilles, Paris, and Toulouse. Since we do not have exact coordinates for the listings, which are only located at a census tract level, we have to employ the stochastic aggregative technique described in the previous section to obtain the networks at the cell level or to aggregate the data at the municipality (*commune*) level (since the census tracts can be grouped within each *commune*).

Communities emerge in these French urban areas at aggregated cell level as well (see Fig. 9). The communities are contiguous in space, similar to the ones observed in Spain, suggesting that listings (as a source of information on listed properties and agencies) allow us to study the spatial segmentation of the housing market through a data-driven, bottom-up method that foregrounds the practices of key market intermediaries.

We repeat the exercise of comparing networks built from different spatial divisions. If France exhibits the same structures found in the Spanish dataset, we would expect the communities found from the aggregated cells and municipality networks to coincide, being the ones from the network of IRIS level very different. Fig. 10 displays the agreement between the communities using

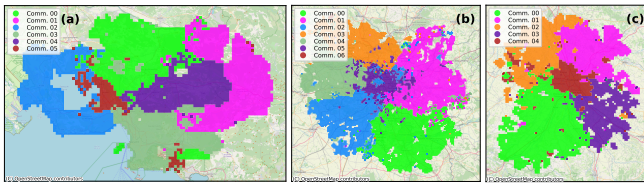


FIG. 9. **Spatial segmentation for 1 km aggregated cells constructed from IRIS level data for France.** Communities detected at the stochastic projected network for the 3 French FUA studied: Marseilles-Aix en Provence (a), Paris (b), and Toulouse (c). The communities shown are detected using the Louvain algorithm with a consensus clustering of 200 clustering method realizations for each of the 100 stochastic networks generated in the IRIS to cell aggregative process. The underground map data is rendered by OpenStreetMap, under ODbL.

aggregated cells and administrative divisions (IRIS and communes). All values of the agreement are significant when compared with the randomized communities, but the largest agreement is found between aggregated cells and communes in all places, echoing results with the Spanish data. This indicates that our aggregative method is a general tool to compute a robust spatial segmentation of the housing market.

IV. CONCLUSIONS

In this study, we present a new method for analyzing the spatial segmentation of housing markets through the activity of real estate agencies, using online listings to extract information on the location of both the property and the marketing agency. We apply this method to analyze comprehensive datasets of geolocated listings in two different countries: Spain and France.

Our methodology is based on dividing space into spatial units, to construct a tripartite network between listings, real estate agencies, and spatial units. We project the network, taking into account the presence and influence of real estate agencies. To divide our projected networks, we use different classic community detection algorithms that account for network weights, such as Louvain, Infomap, and OSLOM. Our methodology generates a spatial segmentation into regions that happen to be spatially connected and larger than municipalities. This segmentation into submarkets remains robust across different community detection algorithms, scales, and administrative boundaries across different countries.

We discovered a limitation of our method when the spatial units exhibit a highly heterogeneous area distribution, and the Zipf law of the distribution of listings per spatial unit is not fulfilled, as in the case of census tracts. To overcome this limitation and extend our methodology to heterogeneous-level data, we developed a method to create an aggregated network via stochastic reconstruc-

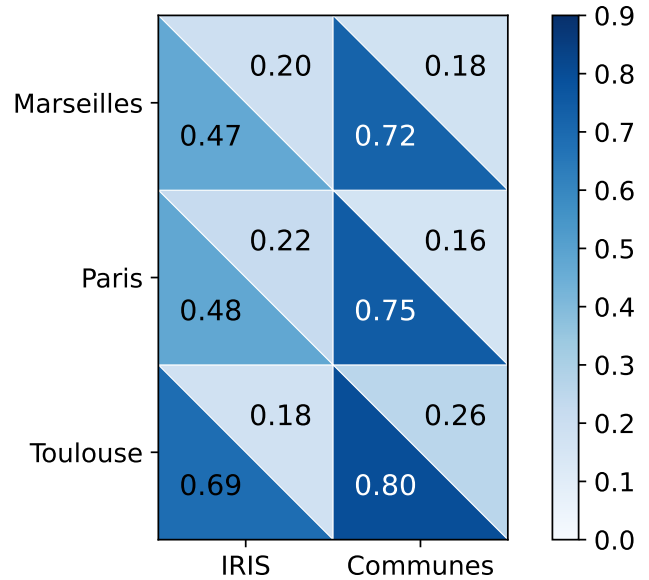


FIG. 10. **Comparison between the 1 km aggregated cells communities and the political units communities in France.** Each column shows the agreement between the 1 km aggregated cells and the French political spatial units: IRIS and Communes. Each row shows the results for each FUA: Marseilles-Aix en Provence, Paris, and Toulouse. The agreement is computed via the fraction of correctly detected cells (lower triangles) compared with the value randomizing the communities (upper triangles).

tion and consensus clustering aggregation. This methodology exhibits good accuracy when compared with the communities from the original high-precision data.

To summarize, we have developed a new methodology that uses listings data to evaluate the spatial segmentation of housing markets into spatially-coherent submarkets. This methodology is generally applicable to different datasets of geolocated listings to infer the submarkets that emerge from the uneven presence and influence of real estate agencies across space. The market-based supra-municipal communities that emerge from the data are found to be robust. Future research should investigate how identifying the submarkets created by market intermediaries can inform policymaking and improve price modeling.

AUTHOR'S CONTRIBUTIONS

Conceptualization: DA, JHM, MM, TLC, JM, TL, JJR; Methodology and analysis: DA, JHM, MM, JJR; Data acquisition and curation: DA, JHM, MM, TLC, JM, EA-P, RC-C, TL, JJR; Writing (original draft preparation): DA, JJR; Writing (review and editing): DA, JHM, MM, TLC, JM, EA-P, RC-C, TL, JJR; Visualization: DA, JHM; Funding acquisition: RC-C, TL, JJR.

All authors read and approved the final manuscript.

ACKNOWLEDGEMENTS

DA, JHM, MM, EA-P, RC-C and JJR acknowledge funding from the CAIB (Government of the Balearic Islands) through the project NouLloguer (PRD2018/43). DA and JJR received partial funding from the Agencia Estatal de Investigación (AEI, MCI, Spain) MCIN/AEI/10.13039/501100011033 and Fondo Europeo de Desarrollo Regional (FEDER,UE) under project APA-SOS (PID2021-122256NB-C22) and the Maria de Maeztu Program for units of Excellence in R&D, grant CEX2021-001164-M. JM, TL and TLC thank the *Groupe SeLogger*

for their precious collaboration and for making the data available through a partnership (Ref. CNRS N. 238072). The *Groupe SeLogger* cannot be held responsible for the completeness, reliability and veracity of the results of this study.

AVAILABILITY OF DATA AND MATERIALS

The projected networks at the spatial resolution of 1 km cell, census-tract, and municipality are available at Zenodo [52] and Github [53]. These links also include the code and additional code to perform the stochastic aggregative method from generic census data.

-
- [1] P. Morawakage, G. Earl, B. Liu, E. Roca, and A. Omura, The Journal of Real Estate Finance and Economics pp. 1–40 (2022).
- [2] S. C. Bourassa, M. Hoesli, and V. S. Peng, Journal of Housing Economics **12**, 12 (2003).
- [3] B. Keskin and C. Watkins, Urban Studies **54**, 1446 (2017).
- [4] A. C. Goodman and T. G. Thibodeau, Journal of housing economics **7**, 121 (1998).
- [5] C. Leishman, G. Costello, S. Rowley, and C. Watkins, Urban studies **50**, 1201 (2013).
- [6] R. Palm, Economic Geography **54**, 210 (1978).
- [7] C. Baumont, Papers in Regional Science **88**, 301 (2009).
- [8] R. A. Dubin, Regional science and urban economics **22**, 433 (1992).
- [9] H. Usman, M. Lizam, and B. Burhan, Real Estate Management and Valuation **29**, 16 (2021).
- [10] A. Páez, Journal of Geographical systems **11**, 311 (2009).
- [11] C. Bitter, G. F. Mulligan, and S. Dall’erba, Journal of geographical systems **9**, 7 (2007).
- [12] B. Case, J. Clapp, R. Dubin, and M. Rodriguez, The Journal of Real Estate Finance and Economics **29**, 167 (2004).
- [13] L. Hu, S. He, and S. Su, Computers, Environment and Urban Systems **94**, 101775 (2022).
- [14] S. C. Bourassa, F. Hamelink, M. Hoesli, and B. D. MacGregor, Journal of Housing Economics **8**, 160 (1999).
- [15] A. Rae, in *The Routledge Handbook of Housing Economics* (Routledge, 2024), ISBN 978-0-429-32733-9, num Pages: 10.
- [16] S. Sawyer and K. Crowston, AMCIS 1999 Proceedings p. 5 (1999).
- [17] G. Boeing and P. Waddell, Journal of Planning Education and Research **37**, 457 (2017).
- [18] G. Boulay, D. Blanke, L. Casanova Enault, and A. Granié, The Professional Geographer **73**, 115 (2021), ISSN 0033-0124, publisher: Routledge .eprint: <https://doi.org/10.1080/00330124.2020.1824678>, URL <https://doi.org/10.1080/00330124.2020.1824678>.
- [19] Y. Yao, J. Zhang, Y. Hong, H. Liang, and J. He, Transactions in GIS **22**, 561 (2018).
- [20] J. F. Adolfsen, B. M. Monsted, A. M. B. Schmith, A. T.-A. Martinello, S. Gudiksen, and K. F. Sonberg, Working Paper Danmarks National Bank p. 28 (2022).
- [21] G. Boeing, Environment and Planning A: Economy and Space **52**, 449 (2020).
- [22] E. C. Delmelle and I. Nilsson, Computers, Environment and Urban Systems **88**, 101658 (2021), ISSN 0198-9715, URL <https://www.sciencedirect.com/science/article/pii/S019897152100065X>.
- [23] A. Rae, Housing Studies **30**, 453 (2015).
- [24] M. J. Salganik, *Bit by Bit: Social Research in the Digital Age* (Princeton University Press, Princeton, 2017), ISBN 978-0-691-15864-8.
- [25] R. Palm, **66**, 266 (????), ISSN 0016-7428, URL www.jstor.org/stable/213885.
- [26] L. Bonneval, *Les agents immobiliers: pour une sociologie des acteurs des marchés du logement* (ENS Éditions, 2017).
- [27] M. Besbris and J. W. Faber, in *Sociological Forum* (Wiley Online Library, 2017), vol. 32, pp. 850–873.
- [28] C. Leishman, G. Costello, S. Rowley, and C. Watkins, Urban studies **50**, 1201 (2013).
- [29] *idealista — Casas y pisos, alquiler y venta. Anuncios gratis*, URL <https://www.idealista.com/>.
- [30] *SeLogger*, URL <https://www.seloger.com/>.
- [31] M. E. Newman, Proceedings of the national academy of sciences **98**, 404 (2001).
- [32] M. E. Newman, Physical review E **64**, 016131 (2001).
- [33] T. Zhou, J. Ren, M. Medo, and Y.-C. Zhang, Physical review E **76**, 046115 (2007).
- [34] B. Derrida and H. Flyvbjerg, Journal of Physics A: Mathematical and General **20**, 5273 (1987), URL <https://dx.doi.org/10.1088/0305-4470/20/15/039>.
- [35] S. Fortunato, Physics Reports **486**, 75 (2010).
- [36] M. E. Newman and M. Girvan, Physical review E **69**, 026113 (2004).
- [37] M. Rosvall and C. T. Bergstrom, Proceedings of the national academy of sciences **105**, 1118 (2008).
- [38] V. D. Blondel, J.-L. Guillaume, R. Lambiotte, and E. Lefebvre, Journal of statistical mechanics: theory and experiment **2008**, P10008 (2008).
- [39] V. A. Traag, L. Waltman, and N. J. Van Eck, Scientific reports **9**, 5233 (2019).
- [40] A. Lancichinetti, F. Radicchi, J. J. Ramasco, and S. Fortunato, PLOS ONE **6**, 1 (2011), URL <https://doi.org/>

- 10.1371/journal.pone.0018961.
- [41] A. Lancichinetti and S. Fortunato, Scientific reports **2**, 336 (2012).
- [42] M. Girvan and M. E. Newman, Proceedings of the national academy of sciences **99**, 7821 (2002).
- [43] D. Hric, R. K. Darst, and S. Fortunato, Physical Review E **90**, 062805 (2014).
- [44] L. Danon, A. Diaz-Guilera, J. Duch, and A. Arenas, Journal of statistical mechanics: Theory and experiment **2005**, P09008 (2005).
- [45] J. Duch and A. Arenas, Physical review E **72**, 027104 (2005).
- [46] Z. Li, S. Zhang, R.-S. Wang, X.-S. Zhang, and L. Chen, Physical review E **77**, 036109 (2008).
- [47] R. K. Darst, Z. Nussinov, and S. Fortunato, Physical Review E **89**, 032809 (2014).
- [48] P.-Y. Chen and A. O. Hero, IEEE Transactions on Signal Processing **63**, 5706 (2015).
- [49] B. Saoud and A. Moussaoui, Physica A: Statistical Mechanics and its Applications **460**, 230 (2016).
- [50] S. Wang, J. Liu, and X. Wang, Journal of Statistical Mechanics: Theory and Experiment **2017**, 043405 (2017).
- [51] S. Fortunato and D. Hric, Physics reports **659**, 1 (2016).
- [52] Zenodo (2024), URL <https://doi.org/10.5281/zenodo.11093099>.
- [53] D. Abella, *Data and materials for: "Exploring the spatial segmentation of housing markets from online listings"*, URL <https://github.com/davidabbu/>.