Comparing and modeling land use organization in cities

Maxime Lenormand, Miguel Picornell, Oliva Garcia Cantú, Thomas Louail,
Ricardo Herranz, Marc Barthelemy, Enrique Frías-Martínez, Maxi San
Miguel, José Javier Ramasco

Comparing and modeling land use organization in cities

Maxime Lenormand,[1] Miguel Picornell,[2] Oliva G. Cantú-Ros,[2] Thomas Louail,[3, 4] Ricardo Herranz,[2] Marc Barthelemy,[3, 5] Enrique Frías-Martínez,[6] Maxi San Miguel,[1] and José J. Ramasco[1]

[1]*Instituto de Física Interdisciplinar y Sistemas Complejos IFISC (CSIC-UIB),*
Campus UIB, 07122 Palma de Mallorca, Spain
[2]*Nommon Solutions and Technologies, calle Cañas 8, 28043 Madrid, Spain*
[3]*Institut de Physique Théorique, CEA-CNRS (URA 2306), F-91191, Gif-sur-Yvette, France*
[4]*Géographie-Cités, CNRS-Paris 1-Paris 7 (UMR 8504), 13 rue du four, FR-75006 Paris, France*
[5]*Centre d'Analyse et de Mathématique Sociales, EHESS-CNRS (UMR 8557),*
190-198 avenue de France, FR-75013 Paris, France
[6]*Telefónica Research, 28050 Madrid, Spain*

The advent of geolocated ICT technologies opens the possibility of exploring how people use space in cities, bringing an important new tool for urban scientists and planners, especially for regions where data is scarce or not available. Here we apply a functional network approach to determine land use patterns from mobile phone records. The versatility of the method allows us to run a systematic comparison between Spanish cities of various sizes. The method detects four major land use types that correspond to different temporal patterns. The proportion of these types, their spatial organization and scaling show a strong similarity between all cities that breaks down at a very local scale, where land use mixing is specific to each urban area. Finally, we introduce a model inspired by Schelling's segregation, able to explain and reproduce these results with simple interaction rules between different land uses.

arXiv:1503.06152v2 [physics.soc-ph] 3 Dec 2015

INTRODUCTION

Land use patterns appear as a natural result of citizens' and planners' interaction with the urban space. However, in a feedback loop, they also play a major role in the experience that residents and visitors have of a city [1]. Land use patterns have an effect on the liveability of neighborhoods and even on the health of the local residents [2]. On the other hand, land use and transportation display a well-established relation [3–6]. Transport demand depends on the location of residence and business areas, while the presence of new transport lines or facilities such as metro stations can substantially modify the land use mixing in a given area of the city. These ideas lie behind the development of the so-called Land Use Transport Interaction (LUTI) models [7, 8], which are commonly employed in transport planning around the globe [9].

An important issue regarding land use refers to the methods employed to estimate it. City Hall registers, surveys or satellite images have been used in the past to this end [10–16]. The emergence of geolocated ICT technologies introduces extra capabilities to directly measure the use that citizens make of each urban space. The information is exhaustive in terms of spatial and temporal resolution, allowing for the detection of concentrations of people second by second along days, weeks and months. Information from mobile phone call records [17–30], geolocated tweets [15, 31–34], credit card use [35] or FourSquare [22] has been considered in the literature. Different data sources have been compared, finding a consistent agreement among the estimations on human concentrations and mobility obtained from different ICT data [27], as well as between ICT data and more traditional techniques [21–24, 27, 29, 36]. Such wealth of information together with the ability to process massive data brought by the Internet era

allows the systematic comparison of features across cities. This analysis can lead to the discovery and confirmation of properties that have been hypothesized to be common to all cities, and also to laws providing insights into the way a property scales with city size. Some examples of these properties include number of patents filed, unemployment rates, GDP per capita, business diversity, consumption of resources, length of road networks, or even crime density [37–43]. The finding of these laws raises the hope of the existence of a coherent framework for city science [38, 39, 41, 42, 44–46].

In this work, we explore land use patterns in the five most populous urban areas of Spain. Land use information is obtained from mobile phone records using a new framework based on network theory and systematic comparisons of land use distribution across the five cities are performed at different scales. Our results reveal common features in the land use types' spatial distributions, which can be understood with a model introduced also here. The similarities break down when the land use type mixing is studied at very short spatial scales, exposing patterns characteristic to each city.

MATERIALS AND METHODS

A network approach to detect land use

Our database is composed of aggregated and anonymized call records during 55 days between September and November 2009 in Spain. Every time a user receives or makes a call, the event is registered together with the tower (BTS) providing the service. The positions of the BTSs are geo-referenced and so the activity levels of each spatial area can be tracked in time. For this work, we select the five most populated metropolitan areas of Spain: Madrid (with a popula-
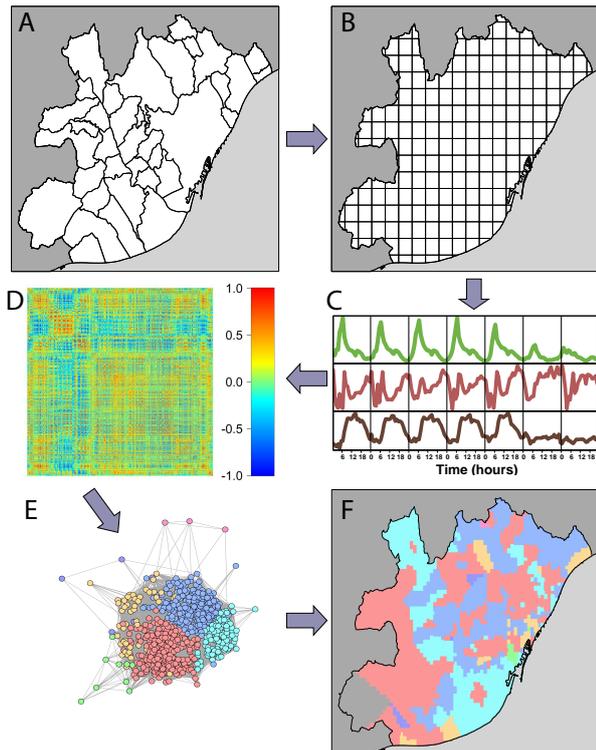
Figure 1. **Steps of the method to detect land use.** (A-B) The urban area is divided in cells of equal area. (C) For each cell, we calculate an activity profile in terms of phone calls along time during the days of the week. (D) A Pearson correlation matrix between cell activities is computed. Then the matrix formed by correlations over a threshold value δ is used to define an undirected weighted network (E), which is clusterized using community detection techniques and the results plotted again on the city map (F).

tion over 5.5 million people), Barcelona (3.2 million), Valencia (1.5 million), Seville (980,000) and Bilbao (900,000). There is no unique definition for the border of an urban area. It may refer, for instance, to official, census or economic delimitation of the cities. Since the focus here is on urban land use, we are interested in identifying the inner zones of each city and, therefore, we use the definition of the metropolitan transportation offices: only areas served by metro or urban buses are considered. This is, nevertheless, an important question because the selection of borders may influence the scaling analysis when comparing across cities [42, 47].

The space of the urban areas is divided following a Voronoi tessellation with the BTS location as centers. The extension of the areas served by each BTS is very different, since it depends on the expected peaks of demand. To ensure a common geographical framework, the five urban areas are divided in a grid with square cells of $500 \times 500 \, m^2$ to which the activity is mapped. This should prevent spurious effects due to the Voronoi areas heterogeneity (see the Appendix for a detailed description of the cities and the division process).

The activity (number of users) in each cell is monitored in time and then processed as illustrated in Figure 1. Average activity profiles are estimated over each day of the week hour by hour in every cell. These profiles are normalized by the total hourly activity to subtract the trends introduced by the circadian rhythms. A Pearson correlation coefficient is then calculated between the activities of every pair of cells, obtaining a correlation matrix describing the level of similarity between activity profiles. The correlations can take positive and negative values. Distributions of these values are shown in the Figure S4 of the Appendix. In order to remove non-significant and negative correlations we only consider Pearson correlation coefficients higher than a threshold δ. As a result, we obtain one weighted network per urban area. We first note that variations of the threshold do not produce significant changes in the properties of the resulting network. The results in the main text refer to a value of δ equal to the correlation distribution dispersion.

Once the networks are built, their mesoscopic structure is analyzed using clustering techniques. The main advantage of community detection algorithms in networks compared to more classical clustering techniques based on dissimilarity matrix is that the number of clusters do not need to be fixed *a priori*. However, it is important to note that different clustering methods can lead to distinct partitions of the networks. We report next results obtained with Infomap [48], while a systematic comparison with results obtained with other clustering tools is provided in the Appendix. As mentioned previously, Infomap does not require the input of a predetermined number of clusters. Therefore, it is interesting to find that in the five cities, between 98 and 100% of the cells are covered with only 4 groups. Figure 2 shows how the activity looks like for each of these four clusters in Madrid (similar plots for Barcelona, Valencia, Seville and Bilbao are included as Figure S9 and S10).

Each of the clusters can be associated with a main land use:

1. **Residential (red)**, which is characterized by low activities from 8am to $5 - 6$pm. For the cells composing this group, the activity peaks around $7 - 8$am and during the evening. In the weekend, the activity is almost constant except for the night hours;

2. **Business (blue)**, where the activity is significantly higher during the weekdays than during the weekends. Furthermore, it concentrates from 9am to $6 - 7$pm. This land use designation can be related to a wide range of commercial, retail, service and office uses;

3. **Logistics/Industry (cyan)**, where, as for Business, the activity is higher during the weekdays. We observe a large peak between 5am and 7am followed by a smaller peak around 3pm. This cluster can be related to transport and distribution of goods: for example, "Mercamadrid"

(the largest distribution area of Madrid) belong to this cluster;

4. **Nightlife (orange)**, which is characterized by high activity during the night hours (1am-4am), especially during the weekends. During the weekdays, these areas show higher activity between 9am and 6pm, as for the Business cluster, which may be hinting a certain level of mixing in the land use. Some examples of this category are the "Gran Via" in Madrid and the "Ramblas" of Barcelona where abound theatres, restaurants and pubs mixed with offices and shops. This is typically the smallest cluster of the four in number of cells.
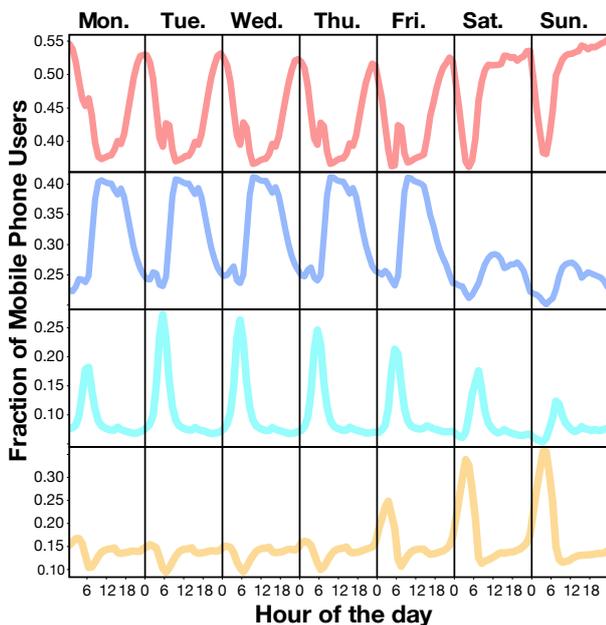

Figure 2. **Temporal patterns associated with the four clusters for the metropolitan area of Madrid.** In red: Residential cluster; In blue: Business; In cyan: Logistics/Industry; And in orange: Nightlife.

More systematically, we compared the land use patterns obtained with our algorithm with cadastral data. The dataset contains information about land use for each cadastral parcel of the metropolitan area of Madrid and Barcelona (about 650,000 parcels). In particular, we have for each cadastral parcel the net internal area devoted to Residential, Business and Industrial uses. This information can be used to identify the dominant cadastral land use in each grid cell classified as Residential, Business and Industrial uses by the community detection algorithm. To do so we need to define a rule to determine what is the dominant land use in a cell. Intuitively, one would tend to identify the dominant land use in a cell as the land use class with the largest area. However, the nature of both land use assignations is very different: the cadastral data is based on the net internal area officially devoted to each activity and not necessarily on

the number of people performing it. Therefore, Residential use is the land use class with the largest area in most of the cell leading to an over-representation of Residential cells in the metropolitan area. To circumvent this limitation we introduce two thresholds to identify Business and Logistics cells with cadastral data in order to obtain a distribution of the fraction of cells according to the land use type similar to the one obtained with our algorithm (see the Appendix for details). The overall agreement is high, we find a percentage of correct predictions equal to 65% for Madrid and 60% for Barcelona which is consistent with values obtained in other studies, 54% in [21] and 58% in [23]. Furthermore, for both case studies, almost all land use types have a percentage of correct predictions higher than 50% (see the Appendix for details).

RESULTS

Comparison of cities

Once defined the clusters, we can compare the proportion of land use type over the five case studies. Figure 3 shows the fraction of cells and the fraction of mobile phone activity averaged over the time period according to the land use patterns for each metropolitan area. We find similar results for the five case studies, the Residential land use type represents in average about 40% of the cells and the mobile phone activity against 30% for the Business category and less than 15% for the Logistics and Nightlife clusters.
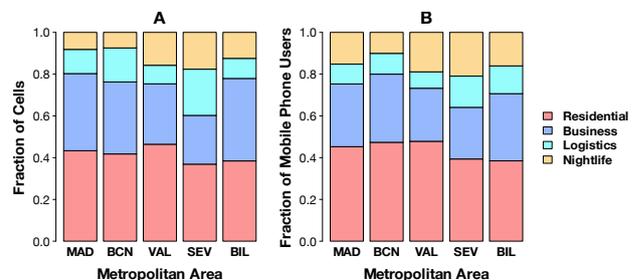

Figure 3. **Fraction of cells (A) and mobile phone users (B) according to the type of land use for each case study.** The fraction of mobile phone users is averaged over the 168 values of the time period.

We can also study how the cells in each cluster are organized in the city's space. For the sake of comparison, we arbitrarily consider as city center the location of the City Hall and build a histogram with the number of cells at a certain distance from it. Since each city has a different spatial extension, distances are normalized by dividing by the maximal distance in each city so as to produce comparable results. The distributions are shown in Figure 4, where average curves over all cities have been superimposed. It is interesting to note certain similarity in the distribution of cells for all urban areas. City size acts as a
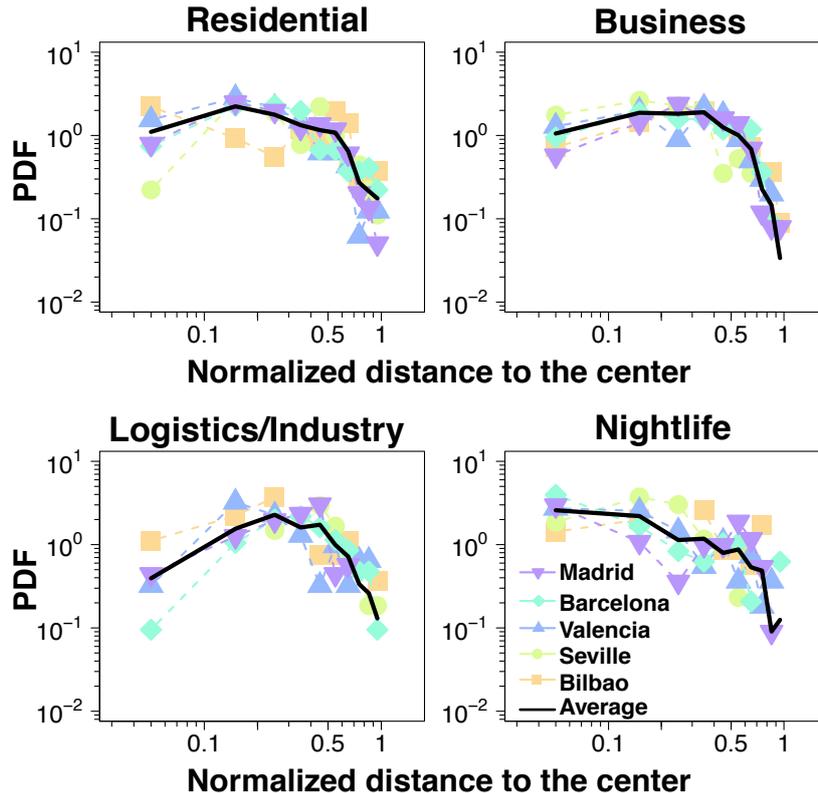
Figure 4. **Distribution of the distance between the cells and the City Hall according to the type of land use.** The distance has been normalized by the maximum distance in each city.

natural cutoff in the distributions, although no simple functional shape is found in any of the clusters. Residential cells are well distributed across the cities but with a maximum not very far from the center. Business cells appear at a similar distance as Residential but peaking a little further. Logistics and Industry are preferentially located in the periphery, while the Nightlife cells are well distributed along the urban areas but slightly more concentrated in the central areas.

In order to quantify land use distribution patterns, we use the Ripley's K [49] defined as

$$K(r) = \frac{A}{n^2} \sum_i^n N_i(r), \qquad (1)$$

where A is the city area, r the search radius (a geographical scale), the index i runs over the cells in the urban area and n is the total number of cells. $N_i(r)$ stands for the number of cells of a given type within a distance r from the cell i. This indicator measures the spatial heterogeneity of a given type of cells. The baseline for homogeneous random systems is a growth $K(r) = \pi r^2$ until reaching A. If the value of $K(r)$ is over the random curve for a certain r it implies that the system is clusterized at that scale. Since cities have different sizes, both $K(r)$ and the radius must be normalized by their maximum values (A for $K(r)$ and the maximum distance for r). Curves for the normalized Ripley's K for each city and land use type are

displayed in Figure 5A as a function of the normalized radius. The $K(r)/A$ for each city are always above the green curve corresponding to a random distribution of land use types, indicating coarsening of land use. We find a scaling-like curve for all the land use types with most of the cities following well the general trend with some small deviations for Nightlife in Seville.

Deepening the analysis, we can also define an entropy index to characterize the land use spatial organization. Let us consider a frame containing the full urban area, which is, in turn, sub-divided in a certain number D^2 of equal divisions. Each of these subdivisions, B_i, intersects the elementary cells so a certain fraction of area falls in each of the land use types: f_i^R in the Residential cluster, f_i^B in Business, f_i^L in Logistics, and f_i^N in Nightlife. An entropy index, E_i, can be defined for B_i as

$$E_i = -\sum_\alpha f_i^\alpha ln(f_i^\alpha), \qquad (2)$$

where α runs over the four clusters. The entropy E_i is then averaged over all the divisions to obtain a global metric for the city at a given scale $E(D)$. $E(D)$ tends to zero if the land use within the divisions becomes unique, as occurs for instance at large D (small spatial scales). On the other extreme, when $D \to 1$, $E(D)$ converges to a fixed value describing the full city. Figure 5B shows how the average entropy be-
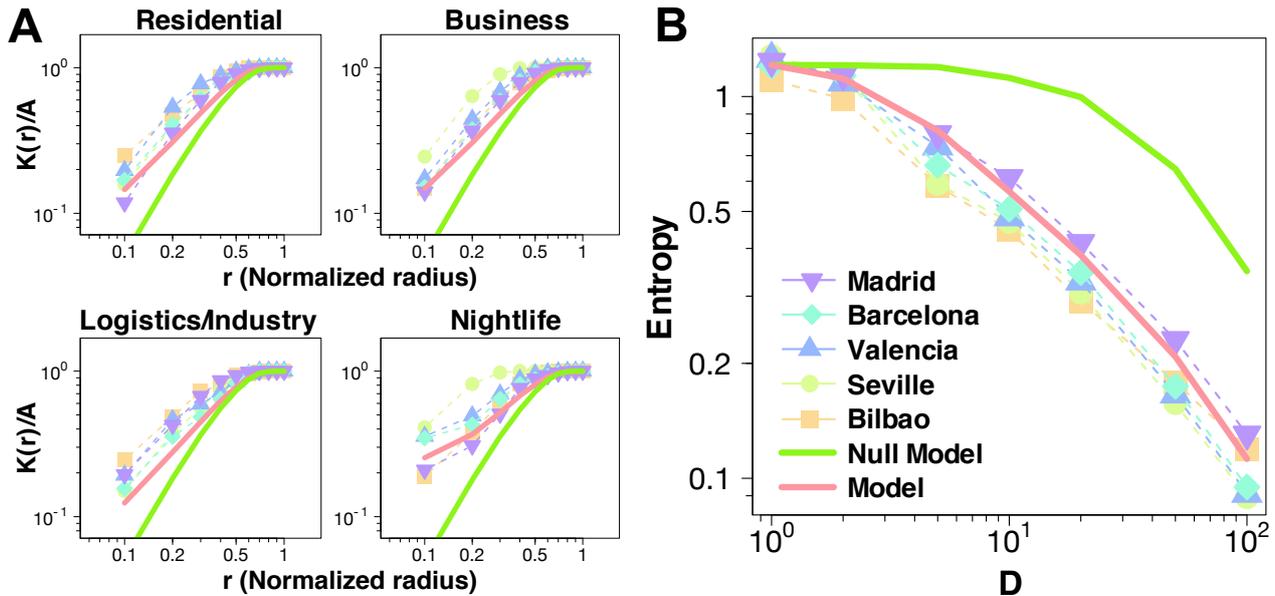
Figure 5. **Comparison of the observed and the simulated Ripley's K and average entropy index.** (A) Ripley's K divided by the city area as a function of the search radius. The radius has been normalized with the maximum value in each urban area. (B) Average entropy index as function of the lateral number of divisions (inverse scale) D. The color and symbols of the curves represent different cities. The red curve corresponds to our model results and the green curve is the outcome of a random null model. Results for our model were obtained with a calibrated value of $\gamma = 0.8$. The red and green curves display the average over 100 realizations.

haves with D. The curves are similar across cities, recalling the shape of scaling functions. This is not surprising if the concept of a fractal-like distribution of the city activity applies as has been previously discussed in the literature [37–42].

Modeling land use

Urban land use models in the literature are typically built with relative elaborated mechanisms [50, 51]. If basic in the rules, the models typically refer to characteristics of cities such as the population or activity distributions [44–46]. The shape of $E(D)$ can be explained, however, by a simple model inspired by Schelling's segregation [52]. It is important to stress that this model is not intended to reproduce all the processes leading to the land use formation, but to explain the scaling of its spatial distribution patterns.

The basic framework is a lattice in 2D representing the urban space. Initially, a variable t_i with a land use type is assigned to every cell i at random (Residential R, Business B, Logistics L or Nightlife N). The global fraction of cells of each type respects the proportions found in the empirical data in such a way that $E(D = 1)$ coincides with the observations. A *satisfaction* index, S_i, is then defined per cell taking into account its type and those of its neighbors. Similarly to Schelling's model, we assume that the satisfaction increases when a cell is surrounded by cells of its own type. Otherwise, S_i depends on the particular com-

binations of types. Some land uses *attract* each other as, for instance, Residential and Business, while others repel as Residential and Logistics. The existence of rules of attraction and repulsion between different types of land use has already been considered in the past like for example in [53, 54]. To be specific if p_t^i is the fraction of neighbors of i of type t, then S_i is calculated as

$$\text{if } t_i = L, \qquad S_i = \delta_{p_L^i, 1},$$

$$\text{if } t_i = N, \qquad S_i = p_N^i \, \delta_{p_L^i, 0},$$

$$\text{if } t_i = R, B, \quad S_i = \begin{cases} \delta_{p_L^i, 0} & \text{with probability } \gamma, \\ \\ p_{R,B}^i \, \delta_{p_L^i, 0} & \text{with probability } 1 - \gamma, \end{cases}$$

where $\delta_{p,x}$ is the Kronecker delta (equal to one if $p = x$, zero otherwise) and γ is the only model parameter. Note that the first condition implies that for Logistic cells $S_i = 1$ only if they are surrounded by cells of the same type, and that cells of other types have zero satisfaction if surrounded by any Logistic one. With this rule, we introduce a tendency to locate Industry and Logistics out of the core areas of the cities. Residential and Business cells have a certain tolerance to the R, B and N types with γ acting as a mixing control parameter: if $\gamma = 0$, mixing is not favored. A global satisfaction measure is defined as the sum over all the cell satisfaction indices, $S = \sum_i S_i$.
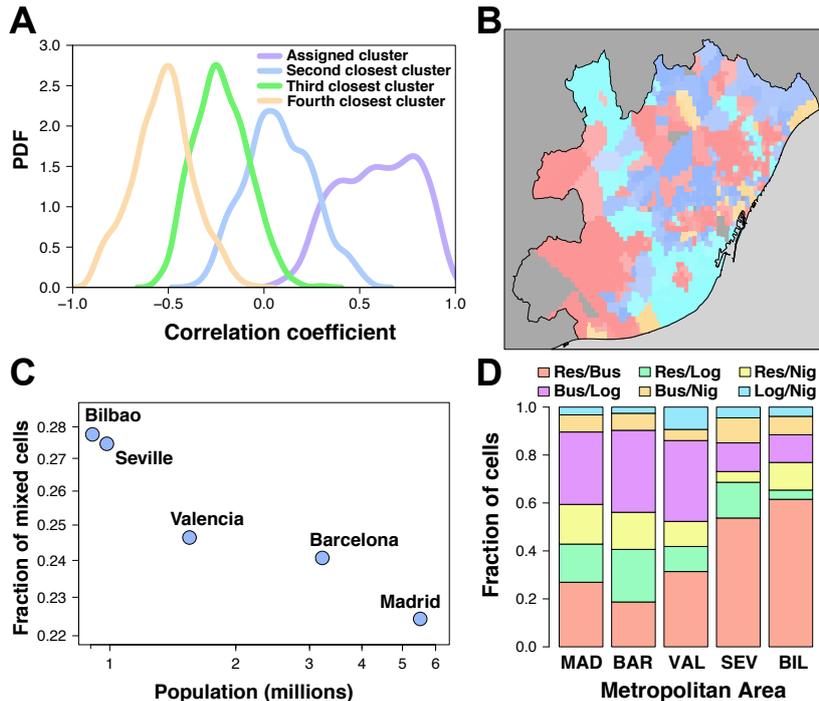
Figure 6. **Land use mixing.** (A) Distribution of the Pearson correlations between cells activity and the average cluster profiles. (B) Map of Barcelona displaying the four clusters with the colors varying from white to the baseline according to the intensity of the relation with the assigned cluster of each cell. The color code is red for Residential, blue for Business, cyan for Logistics/Industry and orange for Nightlife. (C) Fraction of mixed cells as a function of the city population. (D) Fraction of cells classified by the type of land use mixing among those with two types of land use.

The model is updated by choosing random pairs of cells and interchanging their land use if the exchange increases S. This process is repeated until the satisfaction reaches a stationary state.

Calibrating the single parameter γ, we can reproduce the observed $K(r)/A$ and $E(D)$ scaling in the real urban areas (see red curves in Figure 5). The value of the mixing parameter at which the best average results are obtained is $\gamma = 0.08$ (Figure S8). For comparison sake, we have included a null model in which the land use types are distributed at random, keeping the real proportions (green curves in Figure 5). Unsurprisingly, the null model fails at reproducing the curves obtained with the data, mainly because, generally, areas of a given land use type tend to cluster together to form land use zones. More interestingly, assuming that the satisfaction of a cell increases when it is surrounded by cells of its own type and that Logistic cells and cells repel each other (i.e. $\gamma = 0$) is also insufficient to reproduce the properties observed in the data (for more details see Figure S8 in Appendix). Therefore, taking into account the attraction between Residential and Business areas seems to be crucial for the reproduction of land use spatial organization in cities.

Mixing of land use types

So far, we have considered that each elementary cell has a unique land use type associated. This condition can be easily relaxed. If an average activity profile is defined for each of the four clusters, a Pearson correlation coefficient between the activity profile in each cell and the clusters' averages can be calculated. The distribution of correlation values is shown in Figure 6A. The highest correlation value corresponds typically to the cluster at which the cell is assigned. Still, in some cases, positive correlation values are found for other or even two other clusters. For every cell, we can quantify the intensity of its relation with each cluster by summing over these positive correlations and normalizing by the total. A map of the Barcelona metropolitan area with the intensity of each cell relation with its assigned cluster is shown in Figure 6B. The colors represent the four main type of cells and the color saturation is related to the correlation: darker if the correlation is high, paler otherwise. Most cells match well with their original assigned cluster, keeping darker colors, while some are brighter, implying a higher level of land use mixing.

We arbitrarily define a cell as mixed when the normalized correlations fall within the interval $0.3 - 0.7$ for other clusters besides the assigned one. The fraction of mixed cells as a function of the city popu-

lation is displayed in Figure 6C. Larger cities show lower mixing and contain areas devoted to more specific purposes. Figure 6D illustrates how the land use types combine in each cell. Business and Residential integrate well together as in our model, increasingly so for smaller cities. The mixing proportions are city-dependent and act as a fingerprint to characterize each urban area. This feature may be used to classify cities with similar land use mixing patterns. Besides population size, causal links between level of spatial mixing and city shape, area, age of the city, function, etc. will be explored in future investigations. The mixing proportion can be either related to the organization of cities as monocentric or polycentric [28, 55]. Smaller cities display a more monocentric structure, which can be associated with the mixing of land use types given the most restricted space.

DISCUSSION

In summary, we introduce a method to automatically detect land use from electronic records and apply it to the five largest urban areas of Spain in order to perform a systematic comparison across them on the land use distribution. The urban space is divided in a regular grid to prevent geographic heterogeneity and to maintain the spatial scale under control. The user activity profiles are monitored in each unit cell along time, and then a correlation matrix is established between the profiles of every pair of cells. This correlation matrix encodes the functional network of each city. We analyze them by using network clustering techniques, which ensures that cells showing similar use profiles are grouped together. This method has been applied to the five most populated Spanish cities: Madrid, Barcelona, Valencia, Seville and Bilbao. Since the delimitation of urban areas could affect the results, the definition of the municipal transport offices is employed in each case. Interestingly given that the method is unsupervised, four groups consistently appear as dominant in all cities. They correspond to activity profiles compatible with main land uses in Residence, Business, Logistics/Industry and Nightlife. Not only the types are the same across cities, but also the proportions of cells and area devoted to each type are similar.

We also study the distribution of the four land use types at different spatial scales. We define the Ripley's K and the entropy index for each land use type and the behavior of both metrics is explored as the spatial scale varies from the full city (macroscopic scale)

to a single cell (microscopic). The five cities show similar scaling curves for the metrics, implying comparable structures regarding how the four types amalgamate at the urban level. The shape of the scaling curves can be explained by a simple model that has been proposed in this work. The model is based on a Schelling-like segregation in which the different land use types interact to generate a spatial distribution in the city. Cells in a given land use type tend to maximize the number of neighbors undergoing equivalent uses. This rule induces a tendency to coarsening in land use types. The different land uses interact by attracting each other, such as services and residential areas, or by repelling like industry and almost any other type. The calibration of a single parameter regulating the intensity of the attraction between services, residential uses and nightlife is enough to reproduce the scaling curves observed in the real cities. Moreover, we also demonstrate that a model without land use type interactions cannot recreate the empirical scaling.

Similarities across cities break down when one focuses on how the land use types mix microscopically within each unit cell. A characteristic mixing profile is detected for every urban area, providing an individual city fingerprint. Further data on other cities could help to elucidate whether different typologies exist at this microscopic mixing level. In conclusion, despite further data from other countries and sources could be important to confirm our results, we find that a coherent picture emerges in the land use organization of major urban areas and that its origin can be explained with a basic model.

ACKNOWLEDGEMENTS

[1] C. Humphries. The science of cities: Life in the concrete jungle. *Nature*, 491:514515, 2012.

[2] J. F Sallis, B. E. Saelens, L. D. Frank, T. L. Conway, D. J. Slymen, K. L. Cain, J. E. Chapman, and J. Kerr. Neighborhood built environment and income: examining multiple health outcomes. *Social science & medicine*, 68(7):1285–1293, 2009.

[3] D.A. Badoe and E.J. Miller. Transportationland-use interactions: Empirical findings and implications for modeling. *Transportation Research D: Transport &*

Environment, 5D(4):235–263, 2000.

[4] R. Cervero. Built environments and mode choice: toward a normative framework. *Transportation Research Part D: Transport and Environment*, 7(4):265 – 284, 2002.

[5] S. Krygsman, M. Dijst, and T. Arentze. Multimodal public transport: an analysis of travel time elements and the interconnectivity ratio. *Transport Policy*, 11(3):265–275, 2004.

[6] Y.-H. Tsai. Quantifying urban form: Compactness versus 'sprawl'. *Urban Studies*, 42:141–161, 2005.

[7] R. Cervero and K. Kockelman. Travel demand and the 3ds: Density, diversity, and design. *Transportation Research Part D: Transport and Environment*, 2(3):199 – 219, 1997.

[8] P. Waddell, G. F. Ulfarsson, J. P. Franklin, and J. Lobb. Incorporating land use in metropolitan transportation planning. *Transportation Research Part A: Policy and Practice*, 41(5):382 – 410, 2007.

[9] K. Bartholomew. Land use-transportation scenario planning: promise and reality. *Transportation*, 34(4):397–412, 2007.

[10] For surveys see in spain, for instance, the central office for cadastral information (sede electrónica de la dirección general del catastro (sec) in spanish). accessible via the url: **http://www.sedecatastro.gob.es**.

[11] J.-P. Donnay and D. Unwin. Modelling geographical distributions in urban areas. In Jean-Paul Donnay, M.J. Barnsley, and P.A. Longley, editors, *Remote Sensing and Urban Analysis*, pages 205–224. Taylor & Francis, 2001.

[12] X. Yang and C. P. Lo. Using a time series of satellite imagery to detect land use and land cover changes in the atlanta, georgia metropolitan area. *International Journal of Remote Sensing*, 23(9):1775–1798, 2002.

[13] K. T. Geurs and B. van Wee. Accessibility evaluation of land-use and transport strategies: review and research directions. *Journal of Transport Geography*, 12(2):127–140, June 2004.

[14] S. Wu, B. Xu, and L. Wang. Urban land use classification using variogram-based analysis with aerial photographs. *Photogrammetric Engineering and Remote Sensing*, 72:812–822, 2006.

[15] V. Frías-Martínez, V. Soto, H. Hohwald, and E. Frías-Martínez. Characterizing urban landscapes using geolocated tweets. In *SocialCom/PASSAT*, pages 239–248. IEEE, 2012.

[16] S. Hu and L. Wang. Automated urban land-use classification with remote sensing. *International Journal of Remote Sensing*, 34:790–803, 2013.

[17] J. Reades, F. Calabrese, A. Sevtsuk, and C. Ratti. Cellular census: Explorations in urban data collection. *Pervasive Computing, IEEE*, 6(3):30–38, 2007.

[18] M. C. Gonzalez, C. A. Hidalgo, and A.-L. Barabasi. Understanding individual human mobility patterns. *Nature*, 453(7196):779–782, June 2008.

[19] J. Reades, F. Calabrese, and C. Ratti. Eigenplaces: analysing cities using the space time structure of the mobile phone network. *Environment and Planning B: Planning and Design*, 36(5):824–836, 2009.

[20] V. Soto and E. Frías-Martínez. Automated land use identification using cell-phone records. In *Proceedings of the 3rd ACM international workshop on MobiArch*, HotPlanet '11, pages 17–22, New York, NY, USA, 2011. ACM.

[21] J. L. Toole, M. Ulm, M. C. González, and D. Bauer. Inferring land use from mobile phone activity. In *Proceedings of the ACM SIGKDD International Workshop on Urban Computing*, UrbComp '12, pages 1–8, 2012.

[22] A. Noulas, C. Mascolo, and E. Frías-Martínez. Exploiting foursquare and cellular data to infer user activity in urban environments. In *Proceedings of the 2013 IEEE 14th International Conference on Mobile Data Management - Volume 01*, MDM '13, pages 167–176, 2013.

[23] T. Pei, S. Sobolevsky, C. Ratti, S. L. Shaw, and C. Zhou. A new insight into land use classification based on aggregated mobile phone data. *International Journal of Geographical Information Science*, 28:1988–2007, 2014.

[24] S. Grauwin, S. Sobolevsky, S. Moritz, I. Gódor, and C. Ratti. Towards a comparative science of cities: using mobile traffic records in new york, london and hong kong. *ArXiv e-print*, arXiv:1406.4400, 2014.

[25] V. Frías-Martínez, V. Soto, A. Sánchez, and E. Frías-Martínez. Consensus clustering for urban land use analysis using cell phone network data. *Int. J. Ad Hoc Ubiquitous Comput.*, 17(1):39–58, 2014.

[26] V. Frías-Martínez and E. Frías-Martínez. Spectral clustering for sensing urban land use using twitter activity. *Eng. Appl. Artif. Intell.*, 35:237–245, 2014.

[27] M. Lenormand, M. Picornell, O. Garcia Cantú, A. Tugores, T. Louail, R. Herranz, M. Barthélemy, E. Frías-Martínez, and J. J. Ramasco. Cross-checking different source of mobility information. *PLoS ONE*, 9(8):e105184, 2014.

[28] T. Louail, M. Lenormand, O. Garcia Cantú, M. Picornell, R. Herranz, E. Frías-Martínez, J. J. Ramasco, and M. Barthelemy. From mobile phone data to the spatial structure of cities. *Scientific Reports*, 4(5276), 2014.

[29] P. Deville, C. Linard, S. Martin, M. Gilbert, F. R. Stevens, A. E. Gaughan, V. D. Blondel, and A. J. Tatem. Dynamic population mapping using mobile phone data. *Proceedings of the National Academy of Sciences*, 111:15888–15893, 2014.

[30] T. Louail, M. Lenormand, M. Picornell, O. Garcia Cantú, R. Herranz, E. Frías-Martínez, J. J. Ramasco, and M. Barthelemy. Uncovering the spatial structure of mobility networks. *Nature Communications*, 6:6007, 2015.

[31] B. Gonalves and D. Sánchez. Crowdsourcing dialect characterization through twitter. *PLoS ONE*, 9:e112074, 2014.

[32] B. Hawelka, I. Sitko, E. Beinat, S. Sobolevsky, P. Kazakopoulos, and C. Ratti. Geo-located twitter as a proxy for global mobility patterns. *Cartography and Geographic Information Science*, 41:260–271, 2014.

[33] M. Lenormand, A. Tugores, P. Colet, and J. J. Ramasco. Tweets on the road. *PLoS ONE*, 9(8):e105407, 2014.

[34] M. Lenormand, B. Goncalves, A. Tugores, and J. J. Ramasco. Human diffusion and city influence. *arXiv preprint arXiv:1501.07788*, 2015.

[35] M. Lenormand, T. Louail, O. Garcia Cantú, M. Picornell, R. Herranz, M. Barthelemy, M. San Miguel, and J. J. Ramasco. Influence of sociodemographic characteristics on human mobility. *Scientific Reports*, 2015.

[36] M. Tizzoni, P. Bajardi, A. Decuyper, G. Kon Kam King, C.M. Schneider, V. Blondel, Z. Smoreda, M.C. Gonzalez, and V. Colizza. On the use of human mobility proxies for modeling epidemics. *PLoS Comput Biol*, 10:e1003716, 2014.

[37] L.M.A. Bettencourt, J. Lobo, D. Helbing, C. Kuhnert, and G.B. West. Growth, innovation, scaling, and the pace of life in cities. *Proceedings of the National Academy of Sciences*, 104(17):7301–7306, 2007.

[38] M. Batty. The Size, Scale, and Shape of Cities. *Science*, 319(5864):769–771, February 2008.

[39] L. Bettencourt and G. West. A unified theory of urban living. *Nature*, 467(7318):912–913, 2010.

[40] L. M. A. Bettencourt. The Origins of Scaling in Cities. *Science*, 340(6139):1438–1441, June 2013.

[41] M. Batty. *The New Science of Cities.* The MIT Press, 2013.

[42] E. Arcaute, E. Hatna, P. Ferguson, H. Youn, A. Johansson, and M. Batty. Constructing cities, deconstructing scaling laws. *Journal of The Royal Society Interface*, 12:20140745, 2015.

[43] L. G. A. Alves, H. V. Ribeiro, E. K. Lenzi, and R. S. Mendes. Distance to the scaling law: A useful approach for unveiling relationships between crime and urban metrics. *PLoS ONE*, 8(8):e69580, 2013.

[44] H. A. Makse, J. S. Andrade, M.l Batty, S. Havlin, and H. E. Stanley. Modeling urban growth patterns with correlated percolation. *Phys. Rev. E*, 58:7054–7062, 1998.

[45] R. Louf and M. Barthelemy. Modeling the polycentric transition of cities. *Phys. Rev. Lett.*, 111:198702, 2013.

[46] R. Louf and M. Barthelemy. How congestion shapes cities: from mobility patterns to scaling. *Scientific Reports*, 4:5561, 2014.

[47] R. Louf and M. Barthelemy. Scaling: lost in the smog. *Environment and Planning B: Planning and Design*, 41(5):767–769, 2014.

[48] M. Rosvall and C. T. Bergstrom. Maps of random walks on complex networks reveal community structure. *Proceedings of the National Academy of Sciences*, 105(4):1118–1123, 2008.

[49] B.D. Ripley. The second-order analysis of stationary point processes. *Journal of Applied Probability*, 13:255–266, 1976.

[50] A. G. Wilson. The Use of Entropy Maximising Models, in the Theory of Trip Distribution, Mode Split and Route Split. *Journal of Transport Economics and Policy*, 3(1):108–126, 1969.

[51] J. Decraene, C. Monterola, G. K. K. Lee, T. G. G. Hung, and M. Batty. The emergence of urban land use patterns driven by dispersion and aggregation mechanisms. *PLoS ONE*, 8(12):e80309, 12 2013.

[52] T. Schelling. Dynamic models of segregation. *Journal of Mathematical Sociology*, 1, 1971.

[53] P. Krugman. *The Self-Organizing Economy.* Blackwell, 1996.

[54] X. Feng and D. Levinson. *Evolving transportation networks.* Springer Science & Business Media, 2011.

[55] C. Roth, S. M. Kang, M. Batty, and M. Barthelemy. Structure of urban movements: Polycentric activity and entangled hierarchical flows. *PLoS ONE*, 6(1):e15923, 2011.

[56] S. Maslov and K. Sneppen. Specificity and stability in topology of protein networks. *Science*, 296(5569):910–913, 2002.

[57] A. Lancichinetti, F. Radicchi, and S. Fortunato. Community detection algorithms: a comparative analysis. *Physical Review E*, 80(5):056117, 2009.

[58] A. Lancichinetti, F. Radicchi, and J. J. Ramasco. Statistical significance of communities in networks . *Physical Review E*, 81(4):046110, 2010.

[59] A. Lancichinetti, F. Radicchi, J. J. Ramasco, and S. Fortunato. Finding Statistically Significant Communities in Networks. *PLoS ONE*, 6(4):e18961+, April 2011.

[60] V.D. Blondel, J.L. Guillaume, R. Lambiotte, and E.L.J.S. Mech. Fast unfolding of communities in large networks. *J. Stat. Mech*, page P10008, 2008.

[61] http://www.sedecatastro.gob.es/OVCInicio.aspx.

APPENDIX

Case studies

In this study, we focused on the five biggest metropolitan areas of Spain, Madrid, Barcelona, Valencia, Seville and Bilbao (Figure S1). These metropolitan areas are very different in terms of sizes and populations (Table S1). For all cities we have selected as urban area the one served by public transportation (bus and metro) instead of the official definition that in the case of Seville includes a much larger extension relatively depopulated.

Data pre-processing

Mobile phone records of anonimyzed users during 55 days (hereafter noted T) within the period of September-November 2009 were aggregated in two different ways. The aggregated data corresponds to the number of users per hour and per base transceiver stations (BTSs) identified with UTM (WSG84) coordinates. A user may appear connected to more than one BTS within a period of one hour. To avoid over counting people the following criteria was used when

Figure S1. **Map of the metropolitan areas.**

Table S1. **Summary statistics on the metropolitan areas.**

Metropolitan area	Number of municipalities	Number of inhabitants	Area (km^2)
Madrid	27	5,512,495	1,935.97
Barcelona	36	3,218,071	634
Valencia	43	1,549,855	628.81
Sevilla	8	983,852	352
Bilbao	34	908,916	500.2

aggregating the data: each person shall count only once per hour. If a user is detected in k different positions within a certain 1-hour time period, each registered position will count as $(1/k)$ "units of activity". From this aggregated data activity per BTS and per hour is calculated for each day. In order to compute the number of mobile phone users $P_{g,d}(h)$ in a grid cell g (dimension 0.5×0.5 km^2) for a day $d \in T$ between h and $h+1$, where $h \in |[0,23]|$, we first computed the Voronoi cells associated with each BTS.

Voronoi cells

First we remove the BTSs with zero mobile phone users and we compute the Voronoi cells associated with each BTSs of the metropolitan area (hereafter called MA). We remark in Figure S2A that there are four types of Voronoi cells:

1. The Voronoi cells contained in MA.

2. The Voronoi cells between MA and the territory outside the metropolitan area.

3. The Voronoi cells between MA and the sea (noted S).

4. The Voronoi cells between MA, the territory outside the metropolitan area and the sea.

To compute the number of users associated with the intersections between the Voronoi cells and MA we have to take into account these different types of Voronoi cells. Let m be the number of Voronoi cells (ie BTSs), $N_{v,d}(h)$ the number of users in a Voronoi cell v (on day d at time h) and A_v the area of v, $v \in |[1,m]|$. The number of users $N_{v \cap MA,d}(h)$ in the intersection between v and MA is given by the following equation:

$$N_{v \cap MA,d}(h) = N_{v,d}(h)\left(\frac{A_{v \cap MA}}{A_v - A_{v \cap S}}\right) \qquad (1)$$

We note in Equation S1 that we have removed the intersection of the Voronoi area with the sea, indeed, we assume that the number of users calling from the sea are negligible. Now we consider the number of mobile phone users $N_{v,d}(h)$ and the associated area A_v of the Voronoi cells intersecting MA (Figure S2B).

Grid cells

Let n be the number of grid cells, the number of mobile phone users $N_{g,d}(h)$ (on day d at time h) is given by the following equation, $\forall$ g $\in |[1,n]|$:

$$N_{g,d}(h) = \sum_{v=1}^{m} N_{v,d}(h)\frac{A_{v \cap g}}{A_v}. \qquad (2)$$

Then the set of days T is divided into subsets $T_w \subset T$ and the average number of mobile phone users is computed for each day of the week w (Equation S3).

$$N_{g,w}(h) = \frac{\sum_{d \in T_w} N_{g,d}(h)}{|T_w|} \qquad (3)$$

The average number of mobile phone users for the metropolitan areas according to the time and the day of the week are plotted on Figure S3. The profile curve shows two peaks, one peak around 12AM and another one around 7PM. It also shows that the number of mobile phone users is higher during weekdays than during weekend.

N is normalized such that the total number of users at a given time on a given day is equal to 1, Equation S4,

$$\hat{N}_{g_0,w}(h) = \frac{N_{g_0,w}(h)}{\sum_{g=1}^{n} N_{g,w}(h)} \qquad (4)$$

This normalization allows for a direct comparison between sources with different absolute user's activity. For a given grid cell $g = g_0$ we defined the temporal distribution of users $\hat{N}_{g_0}$ as the concatenation of the temporal distribution of users associated with each day of the week. For each grid cell we obtained a temporal distribution of users (also called signal) represented by a vector of length 24×7. It is possible that some grid cells have exactly the same signal because some Voronoi cells may contain several cells, in this case the grid cells have been aggregated (Figure S2C).
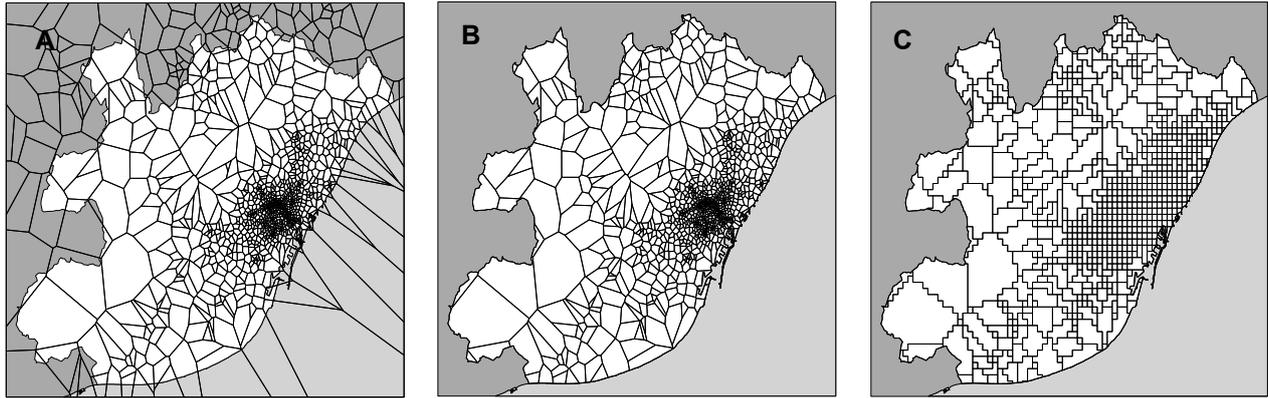
Figure S2. **Map of the metropolitan area of Barcelona.** The white area represents the metropolitan area, the dark gray area represents territory surrounding the metropolitan area and the light grey area represents the sea. (A) Voronoi cells of the mobile phone antennas point pattern. (B) Intersection between the Voronoi cells and the metropolitan area. (C) Recording sites composed of grid cells of dimension 0.5×0.5 km^2.
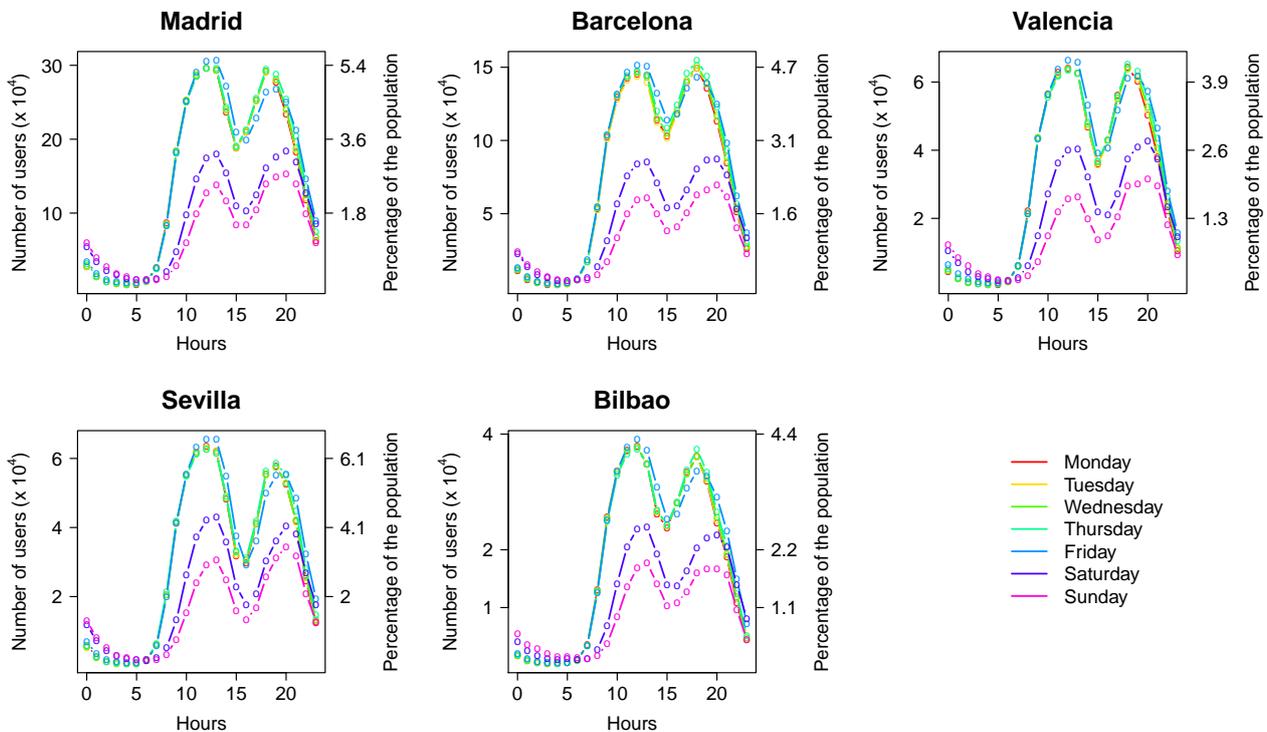

Figure S3. **Average number of mobile phone users per hour according to the day of the week for the five metropolitan areas.**

Functional network

Choice of δ

In the method used to extract the functional network from the mobile phone data presented in the main text we apply a threshold δ to the correlation matrix in order to remove the noise and negative correlations from the correlation matrix. Hence, we have to choose a value of δ high enough to remove the noise but not too high in order to preserve the structure and the properties of the network. Figure S4 displays the distribution of the weights (i.e. correlation coefficient) for the five case studies. One can observe that these distributions can be approximated by a Gaussian distribution. Therefore, we have decided to keep only edges with a weight higher than the weight distribution's standard deviation. In Figure S4 we note
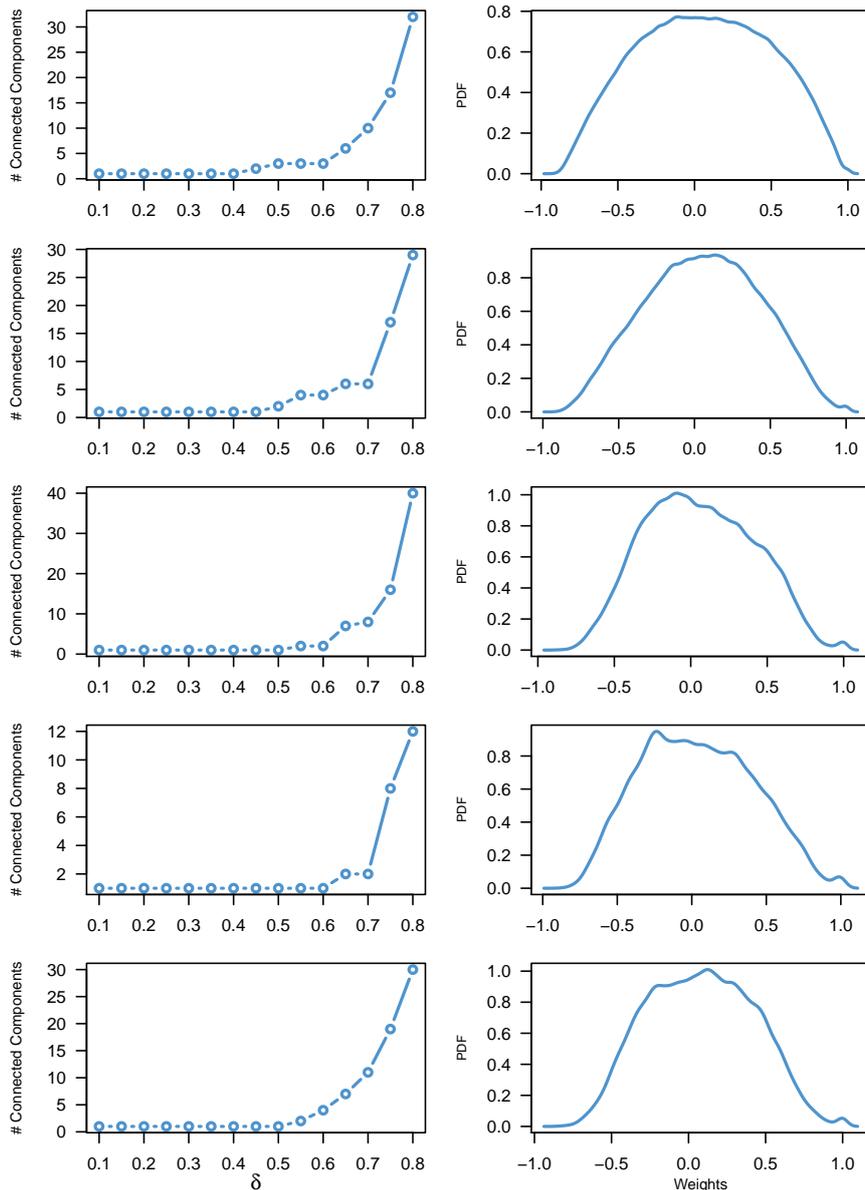
Figure S4. **Number of connected components as a function of** δ **(Left) and weight distribution (Right) for the five case studies.** From top to bottom, Madrid, Barcelona, Valencia, Sevilla and Bilbao.

Table S2. **Statistical properties of the functional networks.**

City	SD	N	E	$<k>$	$<k>/N$	C	L	C_r	L_r
Madrid	0.42	1,381	222,227	321.8	0.233	0.69	2.04	0.31	1.77
Barcelona	0.38	652	46,573	142.9	0.219	0.62	2.02	0.29	1.79
Valencia	0.35	351	13,847	78.9	0.225	0.66	2.06	0.31	1.84
Sevilla	0.38	188	3,700	39.2	0.209	0.62	2.15	0.26	1.81
Bilbao	0.35	267	8,915	66.8	0.25	0.67	2.03	0.39	1.76

ies. In these tables we can observe the threshold (SD), the number of nodes (i.e number of cells) (N), the number of edges (E), the average degree ($<k>$), the average clustering coefficient (C) and the average shortest path length (L). The average clustering coefficient C_r and the average shortest path length L_r have been obtained with a randomly rewired network preserving the degree of the original network by permuting links (4 x (number of edges) times) [56]. We observe that the five networks are very similar, characterized by a high clustering coefficient and low average shortest path.
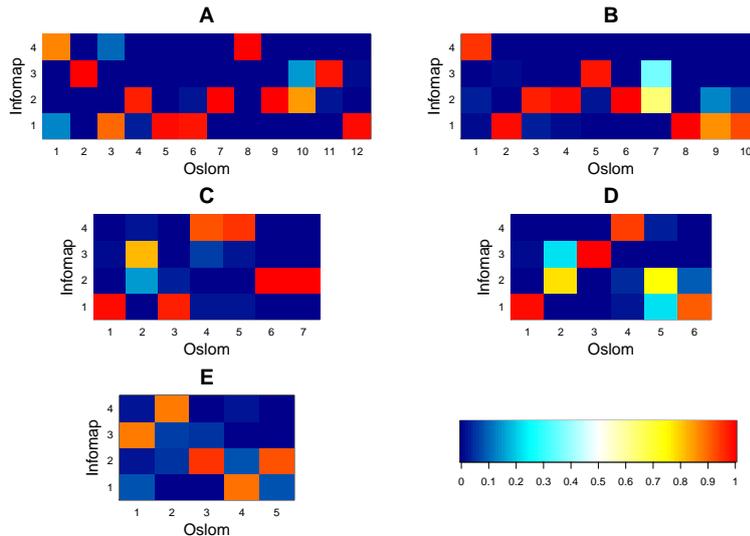
that for δ lower than the weight distribution's standard deviation (around 0.4, see details in Table S2) the number of connected components is equal to 1.

Table S2 summarizes the statistical properties of the functional networks obtained for the five case stud-

Figure S5. Contingency tables between the partitions obtained with Infomap and OSLOM for each case study. (A) Madrid. (B) Barcelona. (C) Valencia. (D) Sevilla. (E) Bilbao. Each row represents a cluster obtained with Infomap and each column represents a cluster obtained with Oslom. The matrices have been normalized so that the sum of each column is equal to one.

Community detection

Community detection in complex networks has recently been the subject of an abundant literature and a large number of algorithms has been proposed the last few years. The purpose of these algorithms is to identify closely connected groups of nodes within a network. To do so, several techniques are used such as maximizing the modularity, measuring probability flows of random walks or optimizing the local statistical significance of communities.

In this paper, we have decided to use the Infomap method proposed in [48]. Infomap finds communities by using the probability of flow of random walks on the network as a proxy for information flow in the real system and then decompose the graph into groups of nodes among which information flows easily. As shown in [57], this method gives good results, however, to evaluate the robustness of the results, the analysis has also been performed with two other clustering methods, Oslom [58, 59] and Louvain [60]. Oslom is a method based on a topological approach to detect statistically significant cluster whereas Louvain is based on modularity optimization which means finding the optimal partition maximizing the density of links within clusters and minimizing the density of links between clusters.

In order to compare the partition obtained with the different method we have plotted in Figure S5 and S6, respectively, the contingency tables between the partitions obtained with Infomap and Oslom and Infomap and Louvain for each case study. In these figures, each plot represents a contingency table C in which each element C_{ij} is the number of nodes which belong to the cluster i detected with Infomap and to the cluster j detected with Oslom or Louvain. The matrices have been normalized so that the sum of each column is equal to one. This normalization allows us to study how the nodes belonging to the groups obtained with Oslom or Louvain are distributed among the clusters found with Infomap. First, we can observe that the number of communities detected with Louvain or Oslom is always greater or equal to the ones obtained with Infomap. Indeed, Louvain has detected a similar number of clusters whereas the number of communities detected with Oslom increases with the size of the metropolitan area, from 5 clusters for Bilbao to 12 for Madrid. However, it is worth noting that in most of the cases, more than 80% of the nodes belonging to the Oslom and Louvain's clusters are gathered in one Infomap cluster. This means that even if the size of the partitions are different, we observe that clusters obtained with Louvain and Oslom are sub-clusters of clusters identified with Infomap.

Comparison with cadastral data

In order to validate the results we compared the land use patterns obtained with our algorithm with cadastral data available on the Spanish Cadastral Electronic Site [61]. The dataset contains information about land use for each cadastral parcel of the metropolitan area of Madrid and Barcelona (about 650,000 parcels). In particular, we have for each cadastral parcel the net internal area devoted to Residential, Business and Industrial uses. We can use these data to identify the dominant cadastral land use in each grid cell classified as Residential, Business and Industrial uses by the community detection algorithm. To do so we need to define a rule to determine what is
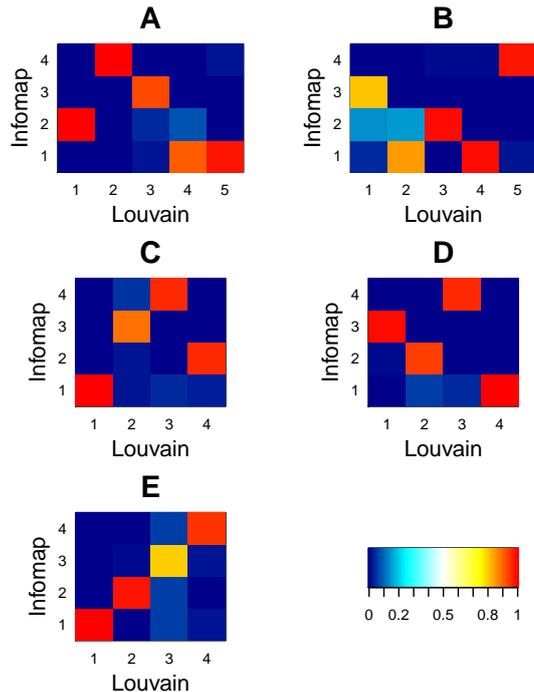
Figure S6. **Contingency tables between the partitions obtained with Infomap and Louvain for each case study.** (A) Madrid. (B) Barcelona. (C) Valencia. (D) Sevilla. (E) Bilbao. Each row represents a cluster obtained with Infomap and each column represents a cluster obtained with Louvain. The matrices have been normalized so that the sum of each column is equal to one.

the dominant land use in a cell. Intuitively, one would tend to identify the dominant land use in a cell as the land use class with the largest area. However, Residential use is the land use class with the largest area in most of the cell leading to an over-representation of Residential cells in the metropolitan area. To circumvent this limitation we introduce two thresholds δ_{Bus} and δ_{Log} to identify Business and Logistics cells with cadastral data. If the fraction of area devoted to Business in a grid cell is higher than δ_{Bus} then the grid cell is classified as Business. Otherwise, if the fraction of area devoted to Logistics is higher than δ_{Log} then the grid cell is classified as Industry. Finally, if the fraction of area devoted to Business and Logistics is, respectively, lower than δ_{Bus} and δ_{Log} then the grid cell is classified as Residential.

Hence, we can adjust the values of these two thresholds in order to obtain a distribution of the fraction of cells according to the land use type similar to the one obtained with our algorithm. To this end we have calibrated these parameters by minimizing the L^2 distance between the distribution of the fraction of cells according to the land use type obtained with the cadastral data and the one obtained with our algorithm for the municipality of Barcelona which represents 20% of the metropolitan area of Barcelona. In Figure S7, we can observe that the minimum is reached for $\delta_{Bus} = 0.2$ and $\delta_{Log} = 0.2$. Now we can use these values to identify the dominant cadastral land use in each grid cell of the metropolitan area of
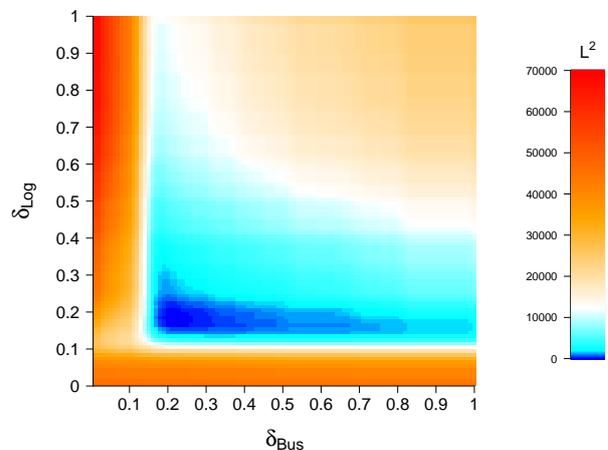

Figure S7. L^2 distance between the distribution of the fraction of cells according to the land use type (Residential, Business and Logistics) obtained with our algorithm and the cadastral data as a function of δ_{Bus} and δ_{Log} for the municipality of Barcelona.

Barcelona and Madrid.

We find a percentage of correct predictions equal to 65% for Madrid and 60% for Barcelona which is consistent with values obtained in other studies, 54% in [21] and 58% in [23]. Furthermore, for both case studies, almost all land use types have a percentage

Table S3. **onfusion matrix of the classification for Madrid and Barcelona.** For the Residential, Business and Logistics rows and columns, the value in the ith row and the jth column gives the percentage of grid cells classified as use i by the cadastral classification which are classified as belonging to the class j by the algorithm. The Total is the distribution of the percentage of cells according to the land use type obtained with our algorithm (row) and the cadastral data (column) with the threshold values $\delta_{Bus} = 0.2$ and $\delta_{Log} = 0.2$.

Madrid

Algorithm Cadastral	Residential	Business	Logistics	Total
Residential	**71.23**	21.67	7.1	49.04
Business	30.84	**62.33**	6.83	39.55
Logistics	22.9	32.06	**45.04**	11.41
Total	49.74	40.42	9.84	100

Barcelona

Algorithm Cadastral	Residential	Business	Logistics	Total
Residential	**68**	26.55	5.45	47.5
Business	28.99	**52.17**	18.84	35.75
Logistics	13.4	32.99	**53.61**	16.75
Total	45.77	37.65	16.58	100

of correct predictions higher than 50% (Table S3).

Calibration of γ

The value of γ was calibrated in order to reproduce the evolution of the entropy index as a function of the number of divisions by side obtained with the data (red line in Figure 4 in the main text). We chose the value of γ minimizing the Euclidean distance between the observed values and the average values obtained with the model with 100 replications. The best results have been obtained with the value $\gamma = 0.8$ (Figure S8).
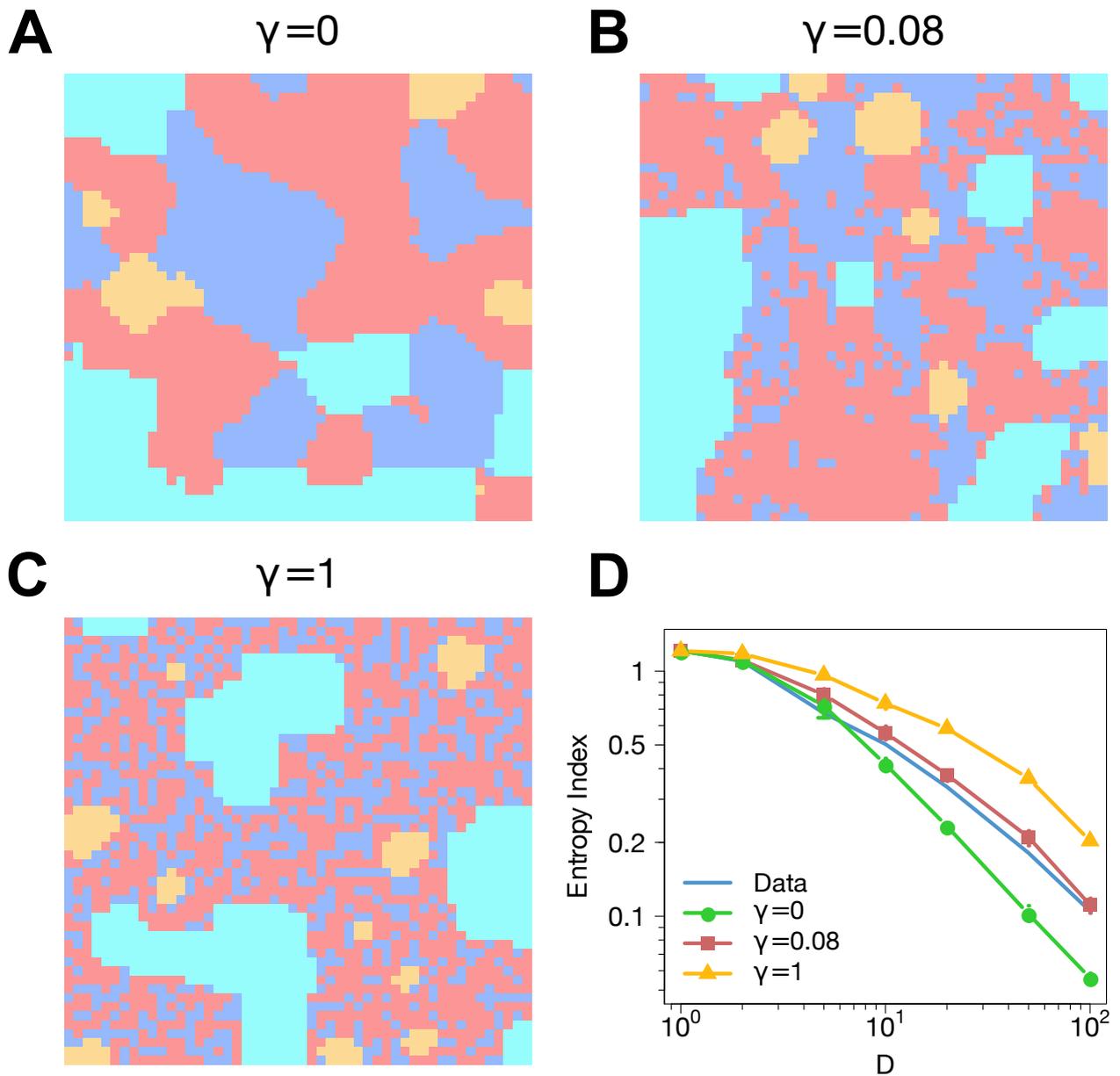
Figure S8. Results obtained with different values of γ ($\gamma = 0$, $\gamma = 0.8$ and $\gamma = 1$) and $T = 500,000$. The 2D lattice used to represent the urban space is composed of $50 \times 50 = 2,500$ cells. The model seems to converge after $300,000$ iterations but to ensure the convergence all the results shown in the paper were obtained with $500,000$ iterations.
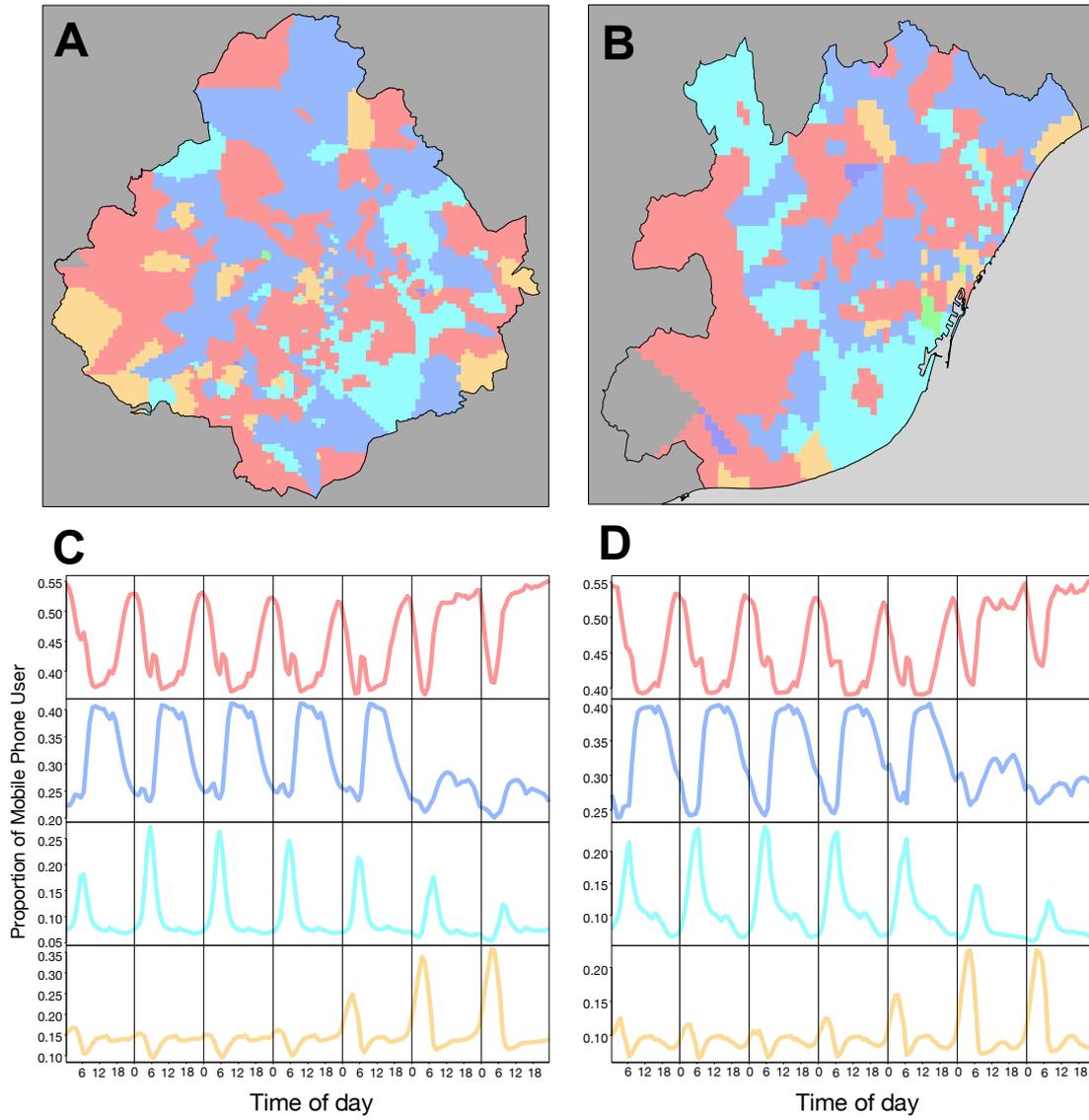
Figure S9. (A-B) Geographical location of the clusters for Madrid and Barcelona. (C-D) Temporal patterns associated with the four clusters for both metropolitan areas. In red, Residential cluster; In blue, Business; In cyan, Logistics; And in orange, Nightlife.
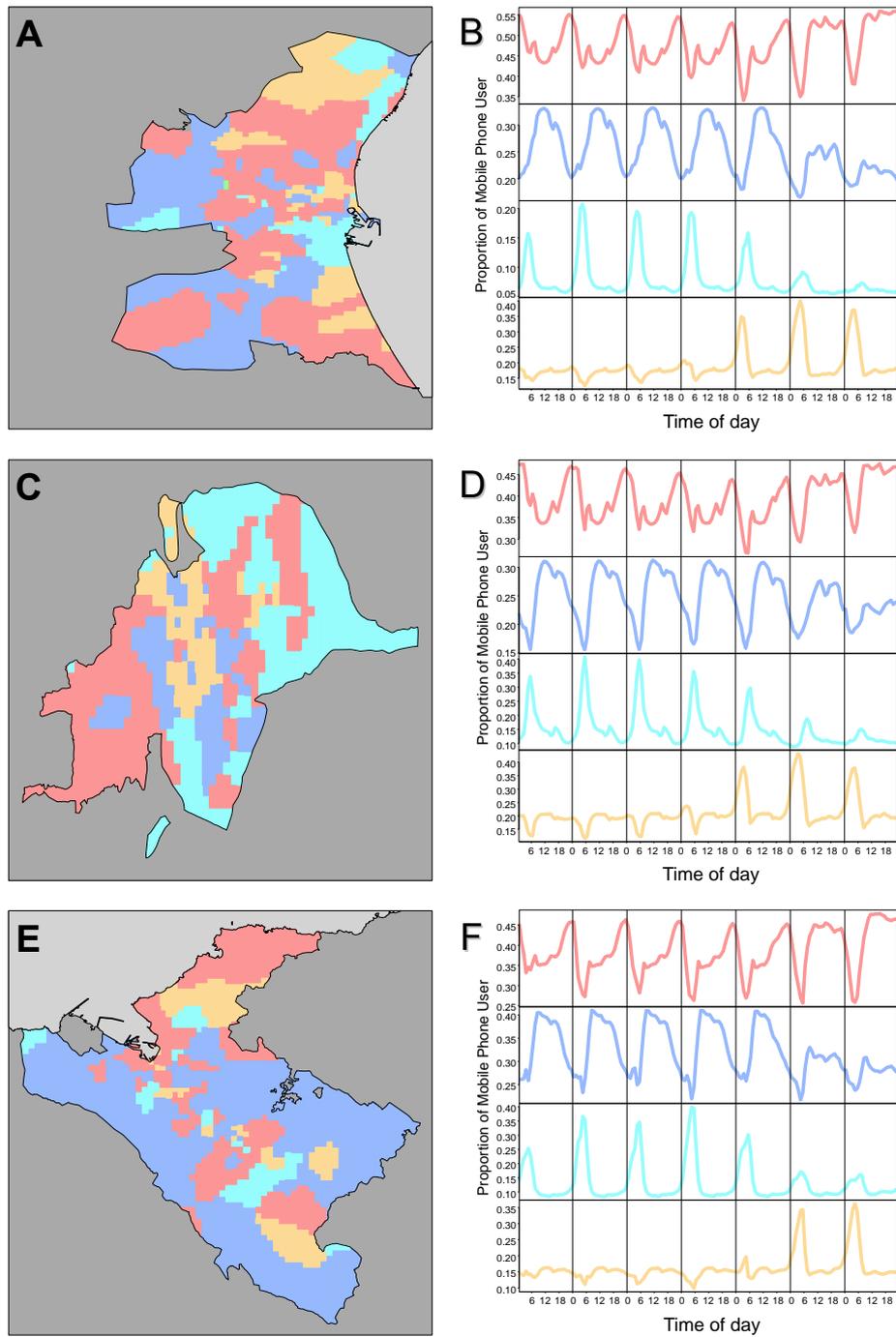
Figure S10. (A), (C) and (E) Geographical representation of the communities for Valencia (A), Sevilla (C) and Bilbao (E). (B), (D) and (F) Temporal patterns associated with the communities for the metropolitan area of Valencia (B), Sevilla (D) and Bilbao (F). In red, the Residential community; In blue, the Business community; In cyan, the Logistics/Industry community; In orange, the Nightlife community.