IEB
Institut
d'Economia
de Barcelona

IEB Working Paper 2019/10

BUILDING(S AND) CITIES: DELINEATING URBAN AREAS WITH A MACHINE LEARNING ALGORITHM

Daniel Arribas-Bel, Miquel-Àngel Garcia-López, Elisabet Viladecans-Marsal

Cities

IEB Working Paper 2019/10

BUILDING(S AND) CITIES: DELINEATING URBAN AREAS WITH A MACHINE LEARNING ALGORITHM

Daniel Arribas-Bel, Miquel-Àngel Garcia-López, Elisabet Viladecans-Marsal

The Barcelona Institute of Economics (IEB) is a research centre whose goals are to promote and disseminate work in Applied Economics, and to contribute to debate and the decision-making process in Economic Policy.

The **Cities Research Program** has as its primary goal the study of the role of cities as engines of prosperity. The different lines of research currently being developed address such critical questions as the determinants of city growth and the social relations established in them, agglomeration economies as a key element for explaining the productivity of cities and their expectations of growth, the functioning of local labour markets and the design of public policies to give appropriate responses to the current problems cities face. The Research Program has been made possible thanks to support from the **IEB Foundation** and the **UB Chair in Smart Cities** (established in 2015 by the University of Barcelona).

Postal Address:
Institut d'Economia de Barcelona
Facultat d'Economia i Empresa
Universitat de Barcelona
C/ John M. Keynes, 1-11
(08034) Barcelona, Spain
Tel.: + 34 93 403 46 46
ieb@ub.edu
http://www.ieb.ub.edu

BUILDING(S AND) CITIES: DELINEATING URBAN AREAS WITH A MACHINE LEARNING ALGORITHM [*]

Daniel Arribas-Bel, Miquel-Àngel Garcia-López, Elisabet Viladecans-Marsal

ABSTRACT: This paper proposes a novel methodology for delineating urban areas based on a machine learning algorithm that groups buildings within portions of space of sufficient density. To do so, we use the precise geolocation of all 12 million buildings in Spain. We exploit building heights to create a new dimension for urban areas, namely, the vertical land, which provides a more accurate measure of their size. To better understand their internal structure and to illustrate an additional use for our algorithm, we also identify employment centers within the delineated urban areas. We test the robustness of our method and compare our urban areas to other delineations obtained using administrative borders and commuting-based patterns. We show that: 1) our urban areas are more similar to the commuting-based delineations than the administrative boundaries but that they are more precisely measured; 2) when analyzing the urban areas' size distribution, Zipf's law appears to hold for their population, surface and vertical land; and 3) the impact of transportation improvements on the size of the urban areas is not underestimated.

JEL Codes: R12, R14, R2, R4

Keywords: Buildings, urban areas, city size, transportation, machine learning

Daniel Arribas-Bel
University of Liverpool
E-mail: D.Arribas-Bel@liverpool.ac.uk

Miquel-Àngel Garcia-López
Universitat Autònoma de Barcelona and IEB
E-mail: miquelangel.garcia@uab.cat

Elisabet Viladecans-Marsal
Universitat de Barcelona and IEB
E-mail: eviladecans@ub.edu

[*] We are grateful for comments from Gilles Duranton, Laurent Gobillon, Rafael González-Val and Fernando Sanz-Gracia, as well as from participants at the European Meeting of the Urban Economics Association (Amsterdam) and the Economics Catalan Society (Barcelona). Financial support from the Ministerio de Ciencia e Innovación (research projects ECO2013-41310-R and RTI2018-097401-B-I00), Generalitat de Catalunya (research projects 2017SGR796 and 2017SGR1301), and the "Xarxa de Referència d'R+D+I en Economia Aplicada" is gratefully acknowledged..

1. Introduction

Understanding city size and why cities grow are two issues that have attracted the growing interest of researchers in recent decades (Duranton and Puga, 2014). However, in their work one of the main challenges urban economists face, together with the scarcity of data, is just how a city should be defined. Until recently, most available data were provided at the local administrative or local political unit level; yet, using these data has proved problematic: First, because the land size and land use of these units are diverse and the population and economic activity within them are not equally distributed (presenting a mix of both rural and urban land) and, second, because cities can grow beyond their borders spreading into the surrounding area. Given these circumstances, a city definition based on economic characteristics makes little sense. Unfortunately, administrative areas are often used for policy-making purposes but, here again, such areas do not usually reflect any functional reality and may even compromise the effectiveness of resulting policies (Briant et al., 2010). For these motives, an ability to define urban areas more accurately should aid analyses of the heterogeneous nature of policy impacts within the same administrative/political boundaries and of spillovers across functional/economic areas. It should also be helpful in examining the sensitivity of policy evaluation to how economic areas of interest are defined.

We contribute to the literature by developing a new methodology for delineating urban areas. By drawing on a unique database on the precise geolocation of all 12 million buildings in Spain, we design a density-based machine learning algorithm to group buildings within portions of space of sufficient density. In line with Rozenfeld et al. (2011) and Bellefon et al. (2019), our objective is to delineate urban areas following a bottom-up approach. These two papers both define cities through the aggregation of cells based on a density criterion; however, here, we do not rely on micro-aggregations to define the boundaries but use the location of each of the buildings in Spanish territory as our first source of information. One of the improvements provided by our method is that our algorithm uses only 10% of the entire sample at a time (with 1,000 replications) and then extrapolates the structure captured in that subsample to the rest of the dataset. To ensure that the urban areas delineated in this fashion are sufficiently robust, we consider that the buildings belong to an urban area if they are assigned to that urban area in 90% of these replications. We run different tests to provide evidence of the stability of our algorithm and the result of our method is the delineation of 717 urban areas accounting for 75% of the population and occupying less than 5% of the whole territory.

Our dataset provides additional information that we are able to exploit so as to better characterize Spain's urban areas. First, we use information on building heights (measured as the number of floors). Recent papers by Ahlfeldt and McMillen (2018), Brueckner et al. (2017) and Liu et al. (2018) highlight the importance of taking this measure into account to understand the shape of cities and the impact the height of buildings can have on land and housing values. Here, we calculate the vertical land of the delineated urban areas as the footprint of the buildings (horizontal land) multiplied by the number of floors. Our results indicate that the vertical land multiplies by three the amount of developed land. This analysis of the vertical land provides a different perspective on city size (especially in the case of the a country's largest urban areas).

Second, we dispose of information on the use of the buildings (residential vs non-residential) and the methodology we adopt also allows us to define the employment centers within our delineated urban areas. Thus, we can identify 2,056 employment centers, representing 63% of the total vertical land. However, only 70 centers house more than 10,000 jobs and just seven are home to more than 50,000 jobs. These results are in line with the well-documented evidence that economic activity within urban areas is markedly more concentrated and presents different patterns of location to those presented by residential areas. This exercise highlights an additional use of our algorithm, namely, it provides a better understanding of a city's internal structure.

Various methodologies have been developed that define urban areas as a collection of smaller units. A common approach in this regard relies on commuting patterns. Here, as long as population mobility plays a key role both in an economic system's performance and in the daily life of individuals, the journey-to-work relationship between two areas allows researchers to determine whether they belong to the same local labor market and, hence, if they can be considered to form part of the same urban area (see Duranton, 2015, for a review). However, because of the lack of commuting data for some developing countries, an increasing number of papers in recent years have opted to use information on the distance between lights in nighttime satellite images to delineate urban areas. Henderson et al. (2018) provide an overview of the applications of night light data in economics and Dingel et al. (2019) use such data to define metropolitan areas. Similarly, instead of using night light data a number of studies employ land cover data. This information, provided by NASA and, more recently, by the European Space Agency (ESA), among others is also available at the global scale. Examples of this approach include Chowdhury et al. (2018), who use such data to estimate urban areas, and Baragwanath et al. (2019), who use them to define urban markets. Finally, recent developments in communication technologies have facilitated studies of how people use space in cities, providing an important new tool for urban research, especially for areas where data are scarce or simply not available. The work of Louail et al. (2014) and Büchel and von Ehrlich (2019) are good examples of how cell phone data records can be used to understand the spatial structure of cities.

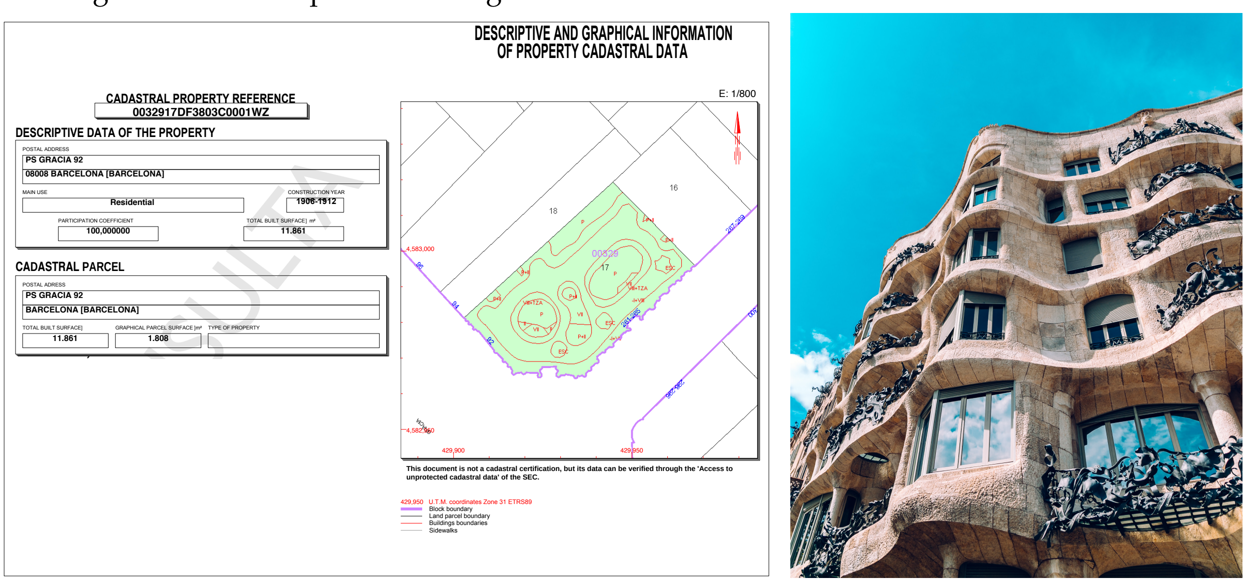
The delineation method proposed here offers several advantages. First, it does not attempt to aggregate administrative units. Second, it only takes into consideration areas that have been developed (i.e. buildings), ignoring undeveloped regions of the territory. Third, detailed information about the buildings allows us to characterize more accurately the structure of the city in terms of its verticality and the location of residential and non-residential activities. Fourth, our approach is more robust than other methodologies and allows to explore the stability of a boundary because it relies on the computation of several candidate solutions that we then combine to arrive at our preferred solution. And, finally, if the appropriate information is available, our algorithm can be replicated for other countries for two reasons: a) it is computationally scalable to large datasets and b) buildings are homogenous units across different countries.

The rest of this paper is organized in six sections and three appendices. In Section 2 we describe the data. Section 3 explains the methodology employed to delineate the urban areas. Section 4 presents the results. In Section 5 we compare our delineated urban areas with other delineations. Finally, in Section 6, we highlight the most important findings and draw our final conclusions. Appendix A includes the technical details of our algorithm; Appendix B reports some robustness checks of the algorithm; and, Appendix C shows summary statistics.

2. Data

Our dataset, provided by the Spanish Cadaster (Dirección General del Cadastro), comprises a unique three-dimensional description of all buildings in Spain[1], geolocated with metric precision for the year 2017. The Cadaster is an administrative registry, supervised by the Ministry of Finance, that contains a description of all real estate data (urban and rural). In other words, the Cadaster constitutes a record of the physical, legal and economic characteristics of all the properties in the country, with one of its main uses being to provide accurate information for the tax system. For example, one of the main local taxes in Spain, the property tax, is dependent on the information contained in this register. In this regard, it should be noted that the law holds that the registration of every property is mandatory and free of charge. This rule guarantees that the data cover the universe of buildings.

Figure 1: An example of building information contained in the Cadaster: La Pedrera



Source: http://www.sedecatastro.gob.es (left image). Photo by Florencia Potter on unsplash.com (right image).

All unprotected data referring to each property, identified by its cadastral reference, can be downloaded from the Cadaster at http://www.sedecatastro.gob.es. These data include all information about the building except that referring to its ownership and value. For each building, this url provides access to an online form that provides basic information and which can be downloaded in PDF format. It also gives access to a detailed map of the building that can be downloaded in GIS format and detailed information about such characteristics as: 1) the building's exact location and total built surface, 2) the year of construction, 3) its use (residential or non-residential), 4) height (number of floors above ground), 5) its footprint (m^2),

[1]Except for the Basque Country and Navarra.

and 6) the number of total units that are contained in each building and, specifically, the number of residential units (dwellings). By way of illustration, Figure 1 shows the online form and footprint map of Antoni Gaudí's well-known building, 'La Pedrera', in Barcelona. The online form indicates that it was built between 1906 and 1912, and records its postal address, footprint (1.808 km^2) and main use (residential).

Table 1 reports the main figures to be drawn from the database. Thus, in Spain there are more than 12 million buildings (that is, 0.25 buildings per capita; 75% of them with a residential use) made up of just more than 37 million units (63% of which are classified as dwellings). As discussed in the Introduction, it is especially useful to exploit the information available about building heights and footprint to obtain what we denote as the 'horizontal' area (the building's footprint) and the 'vertical' area (which is obtained by multiplying the building's footprint by the number of floors). The horizontal area of the buildings in Spain covers 3,099 km^2 (less than 1% of the country's total surface area), while the vertical area is nearly three times that of its horizontal area (8,468 km^2), which corresponds roughly to the average height of three floors per building.

Table 1: Buildings in Spain

	Counts		Areas
Buildings	12,069,635	Horizontal area = $\sum$(Building footprint)	3,099 km^2
Residential	75.7 %	Percentage of Spain's land area	0.6 %
Non-residential	24.3 %		
Units within buildings	37,011,784	Vertical area = $\sum$(Footprint × floors)	8,468 km^2
Residential	63.3 %	Residential	65.2 %
Non-residential	36.7 %	Non-residential	34.8 %
Average number of floors	2.9		

Figure 2: Buildings in Spain

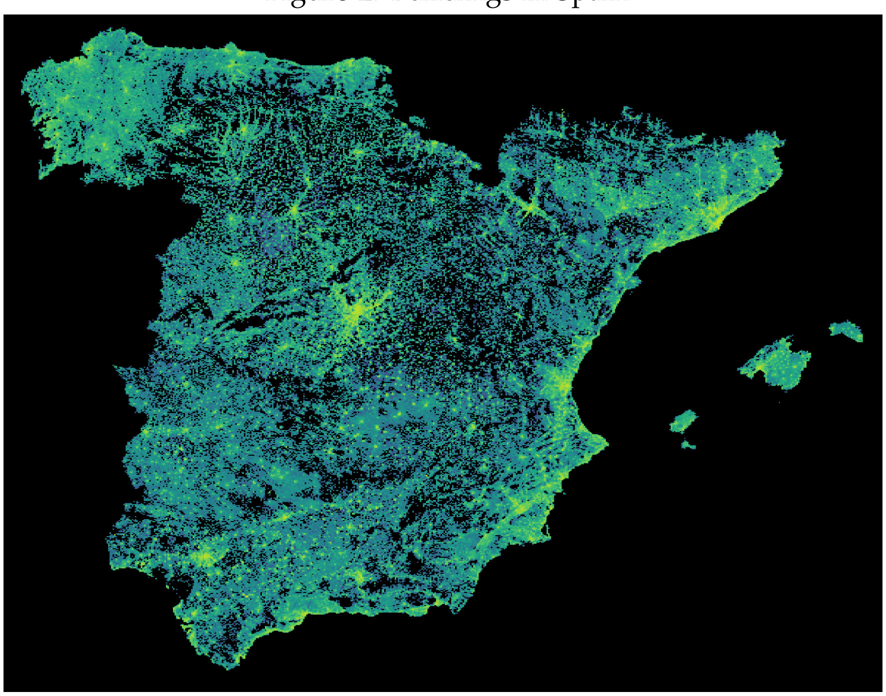
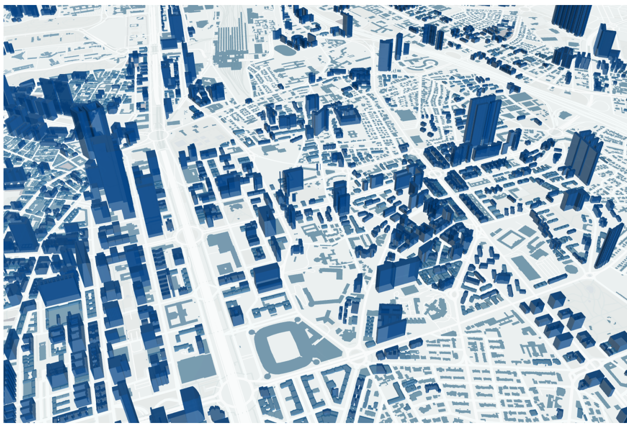


Figure 2 shows the distribution of buildings across Spanish territory (the colored dots reflecting the density of buildings): thus, black identifies areas without any buildings, while the blue, green and yellow dots indicate areas with an increasing concentration of buildings (with yellow showing the highest concentrations). The areas with most yellow dots appear along the Spanish coast and in the center of the country, with the Madrid area being the brightest.

Figure 3 presents 3-D illustrations of the location of buildings in central areas of the present-day municipalities of Madrid, Barcelona, Sevilla, Zaragoza, Valencia and A Coruña. Interestingly, the concentration of buildings and the verticality of these areas is quite distinct.

Figure 3: Zooming in on Spain

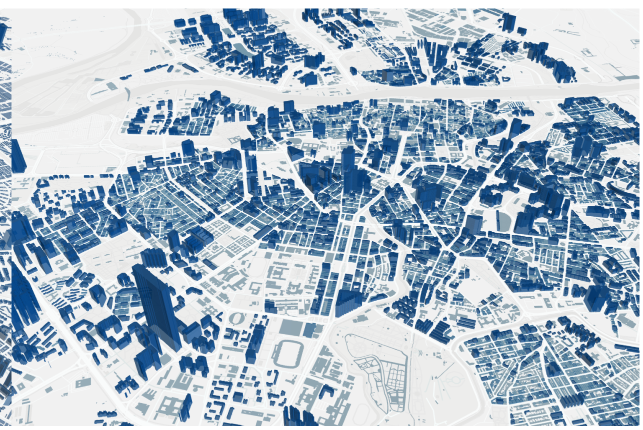
(a) Madrid



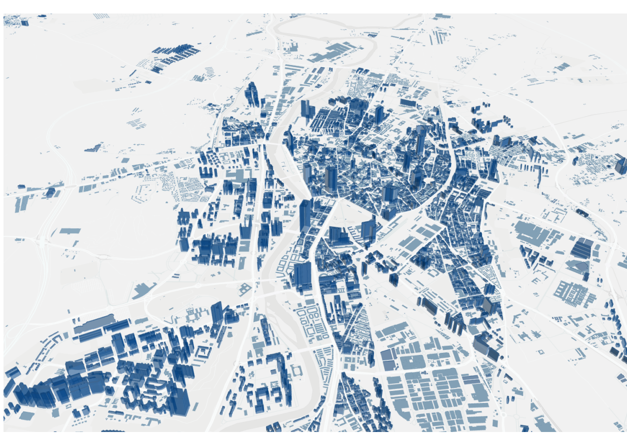
(b) Barcelona

(c) Sevilla

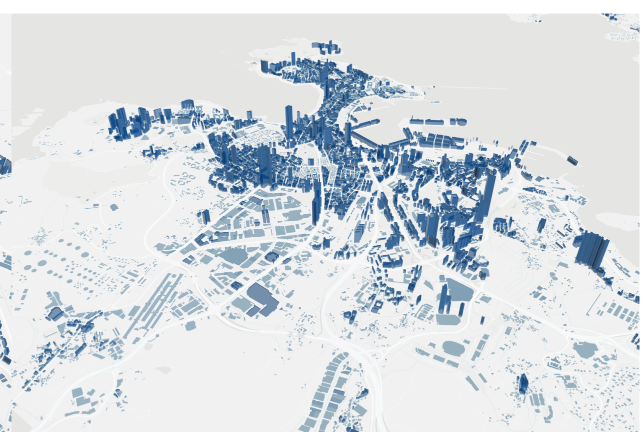
(d) Zaragoza



(e) Valladolid



(f) A Coruña



3. Delineating urban areas with buildings and a machine learning algorithm

We delineate urban areas as portions of land with a minimum, uninterrupted level of building density. To do so, we develop a novel approach as an extension of a well-understood machine learning algorithm (DBSCAN; Ester et al., 1996) that we name 'Approximate DBSCAN' (A-DBSCAN)[2]. Its purpose is to detect robust clusters of buildings that reach a minimum density threshold. To achieve this, our algorithm requires two input parameters: first, the minimum number of buildings that each urban area (cluster) needs to include to be considered so; and, second, a maximum search distance in which to count surrounding buildings to check whether the first criterion is satisfied. Once a set of buildings is identified as a cluster, our method draws its surrounding boundary using the α-shape algorithm Edelsbrunner et al. (1983), a widely used approach to delineate tight bounding boxes.

Figure 4 illustrates how DBSCAN works for a random group of buildings (Figure 4a) when the minimum number of buildings is set at four. The algorithm first chooses a building (in red), draws a circle with a radius equal to the chosen distance threshold (the second parameter) and evaluates the minimum number criterion (Figure 4b). In this case, the criterion is satisfied and this building is labelled as a 'core' point. The algorithm continues to run by drawing circles around the other points and evaluating the minimum number criterion. Figure 4c shows all the buildings that satisfy the minimum number criterion and which are core points. All these buildings/points are reachable, that is, there is a direct connection from one building to another or an indirect link via paths that cross through other core buildings.

Figure 4d shows other buildings (in blue) that do not satisfy the minimum number criterion but which are reachable from some core buildings (i.e. they are within the core building circles). These are the so-called 'border' points and they also belong to the delineated urban area. Finally, Figure 4e shows a building (in green) which, after drawing the circle with a radius equal to the distance threshold, does not satisfy the minimum number criterion. This type of building/point is the so-called 'noise' point and does not belong to the delineated urban area.

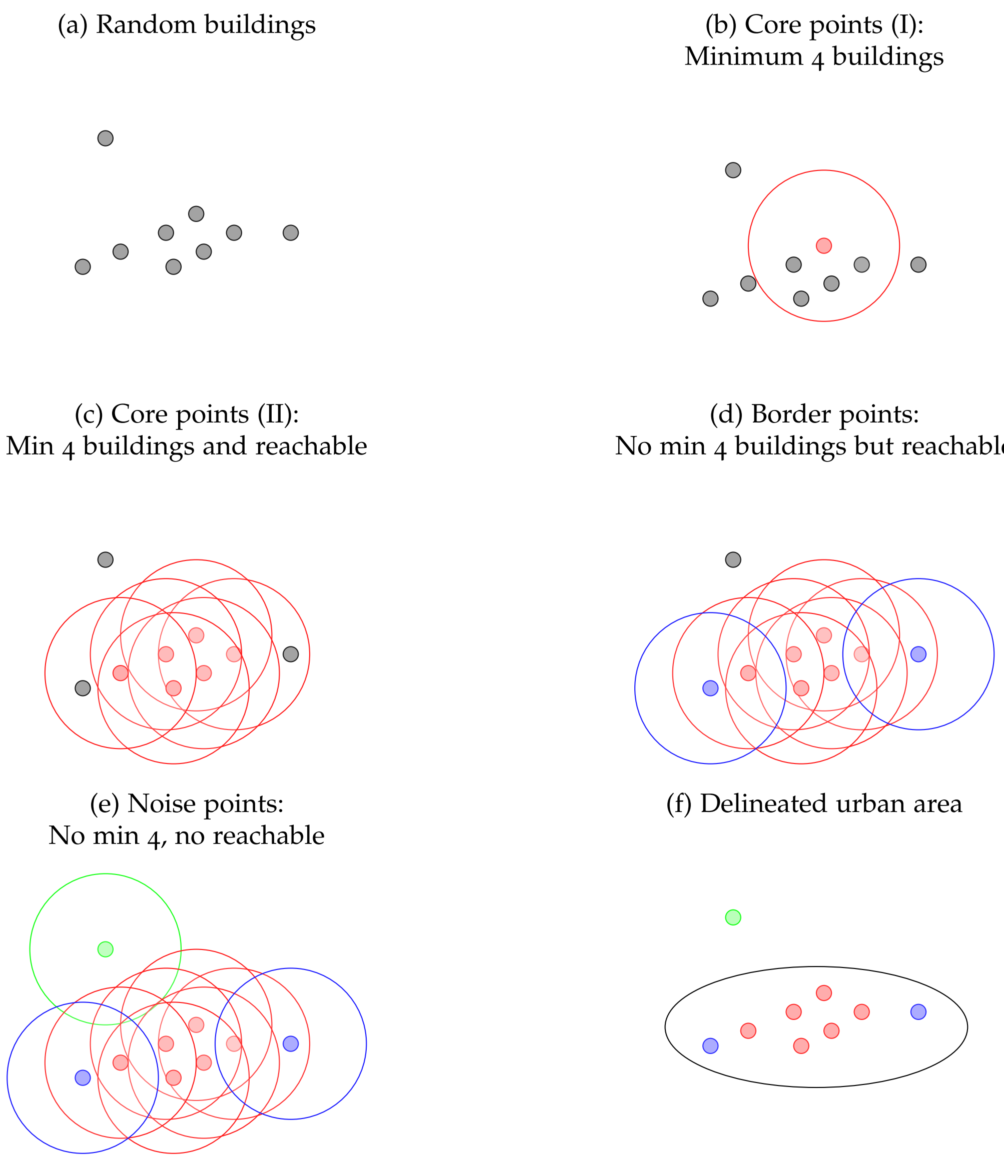
The final delineated urban area (Figure 4f) is made up of core and border buildings (red and blue points) but does not include any noise buildings (green points). By definition, the core of the delineated urban area has the highest density of buildings and the border area the lowest. As a result, our definition of the urban area is in line with the traditional idea of a city, that is, a place with high levels of building density and with an urban spatial structure in which building density decreases towards its the boundaries.

When applying the algorithm to our building dataset, we set the two parameters – that is, the minimum number of buildings and the distance threshold – based on knowledge and evidence from the Spanish urban system. First, we set the minimum number to 2,000 buildings in order to ensure the urban areas we delineate house, at least, 5,000 people. On average, the Spanish household comprises 2.5 members (Instituto Nacional de Estadística, 2018). This minimum threshold is set assuming that the average building is a single family house, which is not exactly

[2]An open-source implementation of A-DBSCAN, written in Python following the `scikit-lean` API, is available at https://github.com/darribas/adbscan_buildings.

the case of some areas of Spain. However, we do not want to underestimate, or rule out altogether, the newer settlements built largely in accordance with that model of urban development. For the maximum distance threshold, our preferred results use 2,000 m. This parameter is chosen based on information about Spanish commuting patterns. The average daily distance commuted by a person in Spain's biggest cities is approximately 4 km (Cascajo et al., 2018) which, divided by two, yields 2 km per trip.

Figure 4: An example of (Approximate-)DBSCAN algorithm



As mentioned, our approach uses a machine learning algorithm, based on the original DBSCAN developed by Ester et al. (1996). DBSCAN facilitates cluster identification based on measures of density without the need for auxiliary geographies. However, from a computational point of view, it does not scale well and, more importantly, it does not include any mechanism to ensure the robustness of the clusters (or, in our case, urban areas). To address this shortcoming, we specifically developed the A-DBSCAN extension. Thus, we propose turning the original algorithm into an ensemble that combines a number of exact DBSCAN runs (1,000 replications) on random subsamples (10% of the original dataset) that are expanded to the rest of the sample through a nearest-neighbor algorithm. These solutions are then summarized in one final set of urban areas (clusters) in which buildings are classified based on their most common occurrence. That is, to ensure robustness, the buildings belonging to an urban area are those assigned to that urban area in at least 90% of the replications. Otherwise, these buildings are classified as noise points and do not belong to any urban area. A more detailed and technical explanation of A-DBSCAN is provided in Appendix A.

We perform several experiments to explore the degree of agreement between our algorithm and the original DBSCAN. An ideal test in this context would be to compare the results of the two algorithms when applied to the entire dataset of Spanish buildings. However, this is not computationally feasible (indeed, one of the reasons for the development of A-DBSCAN is precisely to overcome this computational hurdle). Instead, we consider different parts of Spain characterized by varying numbers of buildings and population, urban areas of different size, and by different geographical features. We are then able to run both algorithms on these subsets and to compare their delineated urban areas. To do so, we use the 'adjusted Rand index' (Hubert and Arabie, 1985), a measure of similarity between two groups or classifications that is widely used in machine learning. In our case, we compare the set of delineated urban areas and the buildings that make up each area when using (1) our algorithm (with the 1,000 replications) and (2) the original DBSCAN. Mathematically,

$$\text{Rand index} = \frac{a+b}{a+b+c+d}$$

where a is the number of buildings that are assigned to the same urban areas in (1) and (2); b is the number of buildings that are assigned to different urban areas in (1) and (2); c is the number of buildings that are assigned to the same urban areas in (1) and to different urban areas in (2); d is the number of buildings that are in different urban areas in (1) and in the same urban areas in (2). Intuitively, $a+b$ can be considered as the number of agreements (i.e., buildings assigned to the same urban areas in (1) and (2)) and $c+d$ as the number of disagreements (i.e., buildings assigned to different urban areas in (1) and (2)). As a result, the Rand index measures the ratio of agreements between the two methods over the total number of buildings, and its values range between 0, dissimilarity, and 1, maximum similarity[3].

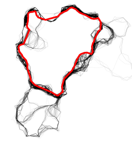
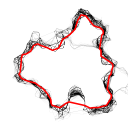
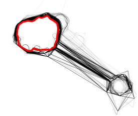
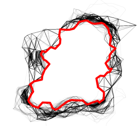
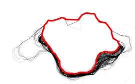
The first three columns in Table 2 present the results of these comparisons for six different parts of Spain. Column 1 shows their geographical location; column 2 reports the size of their population, and column 3 presents the corresponding Rand index. In general, the degree of similarity between the delineated urban areas when using the two methods is quite high, with the maximum value being recorded for Sevilla (with a 97% degree of similarity)[4]. This is remarkable because our algorithm uses only 10% of the entire sample to calculate the exact DBSCAN (with

[3] We use the adjusted version of this index that corrects for the probability of buildings being assigned to the same urban areas by chance. To do so, we use the implementation in the Python library `scikit-learn` (Pedregosa et al., 2011).

[4] Additional computations for other randomly selected parts of the country show values that are always above 80%.

1,000 replications), and then extrapolates the structure captured in that subset to the rest of the dataset. Altogether, these results provide evidence of the efficiency and effectiveness of our algorithm.

Table 2: A-DBSCAN

Section	Population	Rand index	Largest urban area	Boundaries	Stability index
	7,136,015	0.795	Madrid		15.043
	5,871,695	0.987	Barcelona		15.074
	1,313,335	0.985	Sevilla		8.516
	849,035	0.947	Zaragoza		52.098
	493,895	0.980	Valladolid		29.082
	439,810	0.962	A Coruña		10.602

Notes: Gridded population data from National Institute of Statistics (INE). http://ine.es/censos2011_datos/cen11_datos_resultados_rejillas.htm

An additional advantage of our algorithm is its sampling approach, since it allows us to explore the stability of the delineations. In other words, given that each of the final delineated urban areas are based on 1,000 replications, it is possible to quantify the degree of agreement between the 1,000 delineations for each urban area. The last three columns in Table 2 explore this dimension for the largest urban areas (column 4) found in each of the aforementioned parts of Spain. In column 5 we draw the final boundaries of each urban area (the thicker red line) on top of the delineations of each replication (thinner black lines). The figures allow us not only to compare the overall stability between delineations, but also to identify areas within a given urban area of greater and lesser stability. For instance, while A Coruña's northern side displays a high degree of agreement, its southern border is more variable across replications, suggesting a more nuanced boundary. This approach also identifies borderline cases associated with more disperse developments that meet the requirements imposed by the algorithm in a large number of replications but not in enough to grant final assignation into the urban area. Zaragoza is a good example of this situation.

To summarize these visual displays, we compute a Stability index, which is based on the average difference between the delineated area in each single replication and that of the final delineated urban area. We express its value as a percentage of the final surface to correct for city size effects:

$$\text{Stability index} = \frac{\sum_r |A_r - \mathring{A}|}{R} \frac{100}{\mathring{A}}$$

where A_r is the surface of the boundary obtained in replication r, $\mathring{A}$ is the surface of the final delineation, and R is the total number of replications. This measure captures the extent to which individual boundaries drawn as part of our method spatially overlap with the final boundary chosen. Each individual drawing of the boundary might be larger than the final one (as illustrated in the visualizations in Table 2) and include buildings that do not form part of the final delineated urban area. However, the difference between the two offers a measure of the stability of the final delineation and of the extent to which urban development is clearly delimited in the periphery of an urban area or, on the contrary, the degree to which it fades away progressively. The index has a lower bound of zero for the case of complete stability, when all replications agree exactly, and is not upper bounded (the difference between the final delineation and each replication can be arbitrarily large). Column 6 in Table 2 reports the Stability indexes for the selected urban areas. In general, the values show high degree of stability in the delineations. Once again, the case of Sevilla stands out as it shows the closest value to zero and, as a result, the highest degree of stability between the delineations. In contrast, and as mentioned when discussing the boundaries (column 5), less stability is found in Zaragoza because of a disperse development that is not assigned to the final delineation of its urban area.

In summary, our algorithm for delineating urban areas has certain advantages over other methods. First, it is density-based and, in combination with our building dataset, identifies urban areas that only contain continuous parcels of space where the building density exceeds a minimum threshold. As discussed, this feature is in line with the traditional idea of a city that has come into existence because of the agglomeration economies created by the high concentration of population and firms. On the other hand, delineations based on, for example, commuting and/or administrative boundaries include large areas of undeveloped land, which reduces the overall city density. Second, our delineated urban areas are spatially continuous collections of buildings rather than exogenous aggregations, such as grid cells or administrative boundaries. Such ex-ante groupings may be necessary when there is no information available about individual locations or even justified in some specific cases but, generally, they imply a loss of granularity. Furthermore, they can potentially distort the final conclusions of analyses based on them: the so-called 'Modifiable Areal Unit Problem' (MAUP) (Openshaw, 1984, Briant et al., 2010). Third, our algorithm is robust to marginal changes in the data and has a "built-in" approach to explore solution stability. Furthermore, similar to bootstrap estimates, the proposed sampling approach ensures the results are computationally scalable and, thus, feasible in large datasets like ours (with more than 12 million buildings).

4. Urban areas in Spain

4.1 Main results

As can be seen in Table 3, our method delineates 717 urban areas that account for approximately 75% of the Spanish population. These areas contain 5.7 million buildings (roughly half of Spain's total), made up of 26 million units (72% of the total). The sum of building footprints (the horizontal area) is 1,596 km^2. When we also take into account the buildings' floor area (the vertical area), this figure is multiplied by three (4,869 km^2). Interestingly, the average number of floors in the buildings in these urban areas is three. When considering the buildings' footprint together with the land lying between the buildings (streets, roads, parks, etc.), the total surface of the delineated urban areas reaches 22,469 km^2, which represents nearly 5% of the surface area of the whole of Spain. The use of the buildings of these urban areas is mainly residential (83.1% of the cases). As for the population size of the delineated urban areas (Table 3 Panel B), ten of them have more than 500,000 inhabitants and represent one third of the Spanish population. Together with the next 47 biggest urban areas (those with a population between 100,000 and 500,000 inhabitants), they represent 52% of the whole Spanish population. The data for these ten urban areas seem to indicate that the biggest ones in terms of population have larger surfaces and vertical lands. For the delineated urban areas, the correlation between their population and surface is 0.89, while that between their population and vertical land is even bigger (0.98).

Table 3: Buildings and delineated urban areas

Panel A: Counts and areas for all delineated cities

	Counts		Areas
Buildings	5,719,503	Horizontal area = $\sum$(Building footprint)	1,596 km^2
Percentage of total buildings	47.4 %	Percentage of total horizontal area	51.5 %
Residential	83.1 %		
Non-residential	16.9 %		
Units within buildings	26,606,814	Vertical area = $\sum$(Footprint $\times$ floors)	4,869 km^2
Percentage of total units	71.9 %	Percentage of total vertical area	57.5 %
Residential	63.9 %	Residential	71.5 %
Non-residential	36.1 %	Non-residential	28.5 %
Average number of floors	3.1	Delineated surface	22,469 km^2
		Percentage of Spain's land area	4.4 %

Panel B: Number of urban areas by population size

	Urban areas	Population	Percentage of Spain's pop
All	717	35,015,936	74.8 %
Population $\leq$ 5,000	131	472,080	1.0 %
5,000 $<$ Population $\leq$ 10,000	220	1,566,350	3.4 %
10,000 $<$ Population $\leq$ 25,000	189	2,971,230	6.4 %
25,000 $<$ Population $\leq$ 100,000	120	5,719,630	12.2 %
100,000 $<$ Population $\leq$ 500,000	47	8,742,245	18.7 %
Population $>$ 500,000	10	15,544,400	33.2 %

Notes: In 2011, 46,815,916 inhabitants lived in Spain. Population is computed using population grid data (1×1 km cells within the boundaries of our urban areas) from the 2011 Population Census.

These results are obtained when the algorithm considers the number of buildings to be found within a 2,000-meter threshold. However, we also calculated the algorithm modifying this distance to see how it affects our results. Table B.1 in Appendix B presents the results of the delineated urban areas with different distance thresholds (1,500 m, 1,600 m, 1,800 m, 2,200 m, 2,400 m and 2,500 m). As expected, as the threshold distance falls, the number of delineated urban areas increases and the percentage of population contained in the new delineated urban areas diminishes. Thus, for the smallest distance threshold (i.e. 1,500 m), we obtain 773 urban areas (with 70% of the population). In contrast, as the distance increases, we obtain fewer urban areas containing more population. For example, when the distance was greatest (i.e. 2,500 m), 699 urban areas are delineated containing 79% of the total Spanish population.

Table 4 describes the characteristics of the ten largest urban areas in Spain in terms of population, number of buildings and units, surface, horizontal and vertical area and number of floors. It is interesting to see the different structures presented by these urban areas. For example, Madrid and Barcelona contain a similar number of inhabitants (4.52M and 4.37M, respectively) but the surface area and the number of buildings they contain is quite different. Barcelona has nearly twice the surface area and twice the number of buildings as Madrid. The other eight urban areas are much smaller.

Table 4: The largest delineated urban areas

	Population	Buildings	Units	Surface (km^2)	Horizontal area (km^2)	Vertical area (km^2)	Floors
Madrid	4,515,845	198,517	2,901,415	691	116	412	4.3
Barcelona	4,375,970	381,730	2,891,685	1,191	133	512	3.9
Valencia	1,654,565	174,281	1,290,016	628	66	218	3.7
Sevilla	1,214,080	203,270	767,174	635	53	149	3.0
Málaga	840,150	81,916	660,119	309	33	100	3.2
Zaragoza	646,785	23,612	475,756	88	15	60	4.9
Murcia	619,740	120,767	484,300	427	37	99	2.9
Santa Cruz	584,520	120,724	384,778	356	27	79	3.0
Las Palmas	578,735	93,690	372,973	319	20	65	3.2
Granada	514,010	112,904	434,175	342	24	75	3.0

Notes: Population is computed using population grid data (1×1 km cells within the boundaries of our urban areas) from the 2011 Population Census. Surface is total land of the delineated urban area. Horizontal area refers to the sum of building footprints. Vertical area is the sum of floor footprints. Floors refers to the average number of floors.

Figure 5 shows the geographical location of the 717 urban areas (with colors ranging from green to yellow with increasing density of buildings within the urban area). As can be seen, most of the urban areas are concentrated along the Mediterranean coast, and in the center and south of Spain. The smaller scale illustration in Figure 6 shows the delineated urban areas in the region of Barcelona.

It is interesting to compare our results with those reported by Bellefon et al. (2019) who also delineate the French urban areas using building density but the authors apply a new dartboard methodology. Although France and Spain are similar countries (in terms of economic development, location, etc.), some aspects of their respective urban structures differ considerably. France has 0.5 buildings per capita (covering 0.9% of the land), while in Spain there are 0.25 buildings per capita (covering just 0.6% of the territory). Likewise, the average number of floors in French buildings is two, while, as we have seen, in Spain it is three. Thus, initially it would appear that France presents a less dense urban system. When applying the delineation method to French buildings, 7,223 urban areas are obtained of which 695 have a core. These areas concentrate 75% of the French population. This last figure is quite similar to the one reported here for the Spanish urban areas.

Figure 5: Delineated urban areas in Spain

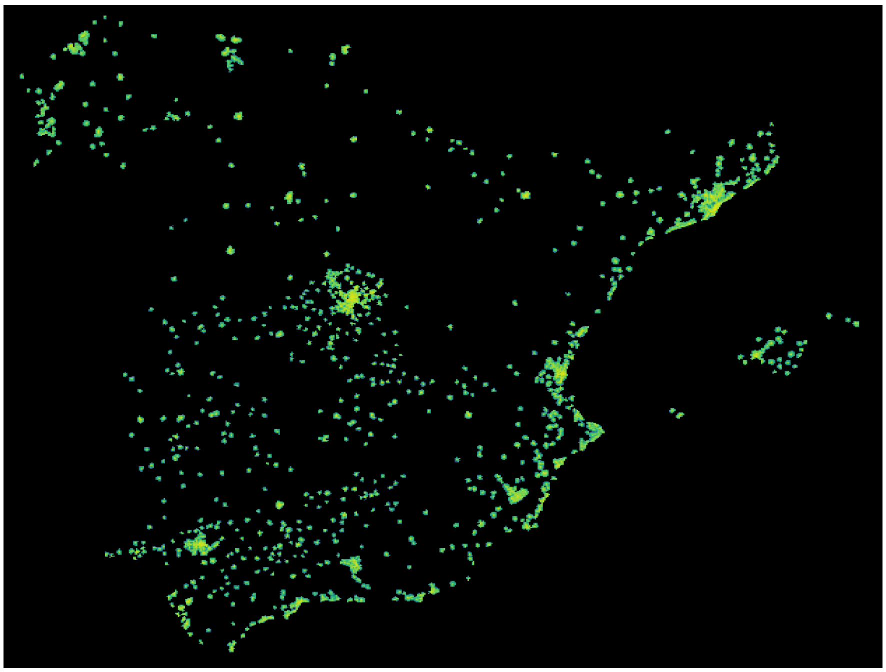
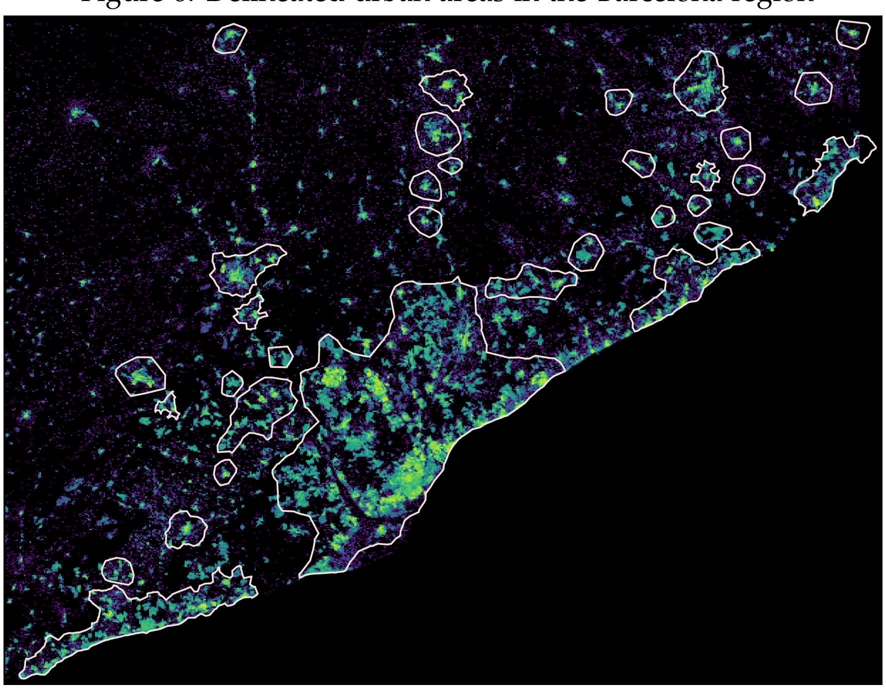


Figure 6: Delineated urban areas in the Barcelona region



4.2 Identifying employment centers within the urban areas

An interesting exercise to illustrate a further application of our method is the identification of employment centers within each of the urban areas. The goal of this exercise is to focus specifically on the concentration of economic activity. With this purpose in mind, we adopt a similar approach to that described in Section 3, albeit with some differences. Most importantly, we focus on units rather than on buildings, given that a single building may include several units (especially if it contains more than one floor). This is a prevalent feature of CBDs and other forms of employment concentration where firms and workers cluster to benefit from density. Using units instead of buildings also allows us to account for the difference between vertically dense areas and those clustered only horizontally. This implies, in the case of this exercise, working only with the non-residential units.

Although the identification mechanism is similar to that used when delineating city boundaries, a few changes have to be introduced. First, instead of running a single instance of the algorithm for the entire dataset, we apply the method to each urban area delineated in the previous stage. Second, for each dataset of urban area buildings, we run A-DBSCAN using 50% of the sample. We retain 90% as the stability threshold and 1,000 replications to generate the delineations. Third, while each point still represents a single building, we now weight them based on the number of non-residential units that the building houses. Finally, we adapt our two algorithm parameters (minimum number of buildings per urban area and distance threshold) to identify employment centers in accordance with methods proposed in the literature. The most frequently used are those based on density thresholds (Giuliano and Small, 1991, McMillen and Smith, 2003, Giuliano et al., 2007, Muñiz et al., 2008) and density peaks (McMillen, 2001, Redfearn, 2007, Garcia-López et al., 2017b,a), where an employment center is a place whose primary feature is a high density of workers (and certainly one with a density higher than that of nearby locations). Giuliano and Small (1991) and McMillen and Smith (2003) define this density as 2,500 employees per km^2. Assuming ten employees per non-residential unit (Fariñas and Huergo, 2016), the threshold we need to impose is 250 non-residential units per km^2. By considering a distance threshold of 250 m, the minimum number of non-residential units is therefore 49[5].

Table 5 presents the results of the delineation of the employment centers within each of the urban areas. Panel A shows that the footprint of the employment centers amounts to 886 km^2 (that is, 55% of the horizontal land of all the urban areas). Interestingly, the economic activity inside the urban areas is clearly more concentrated and presents a distinct pattern of location to that of residential use. When we analyze the vertical land associated with these employment centers, the surface increases to 3,060 km^2 (that is, 63% of the vertical land of all the urban areas). Thus, unsurprisingly, insofar as the average number of floors is 3.5 in the employment centers, the density of buildings in these areas is higher.

Panel B shows that, with no restriction on the size of the employment centers, the 717 urban areas contain 2,056 employment centers. However, when we establish a minimum number of jobs per center (so as to take the largest economic agglomerations into consideration), the number of

[5]Since the circle around each unit has an area of 0.196 km^2 (=$(0.250)^2 \times \pi$), the minimum number of units can be obtained by multiplying this area by the minimum density, that is, $49 = 0.196$ km^2 $\times$ 250 units/km^2.

employment centers falls. Thus, there are only 70 employment centers with more than 10,000 jobs and just seven centers with more than 50,000 jobs. In fact, only the biggest urban areas have more than one employment center with more than 10,000 jobs. This is the case, for example, of the cities of Barcelona (nine employment centers), Valencia (five), Madrid (four) and Málaga (three). This evidence is in line with the polycentric structure of these urban areas as reported and analyzed by (Garcia-López, 2010, 2012). To illustrate how the algorithm works to delineate employment centers at a smaller scale, Figure 7 shows the nine biggest delineated centers in the urban area of Barcelona.

Table 5: Delineated employment centers in the urban areas

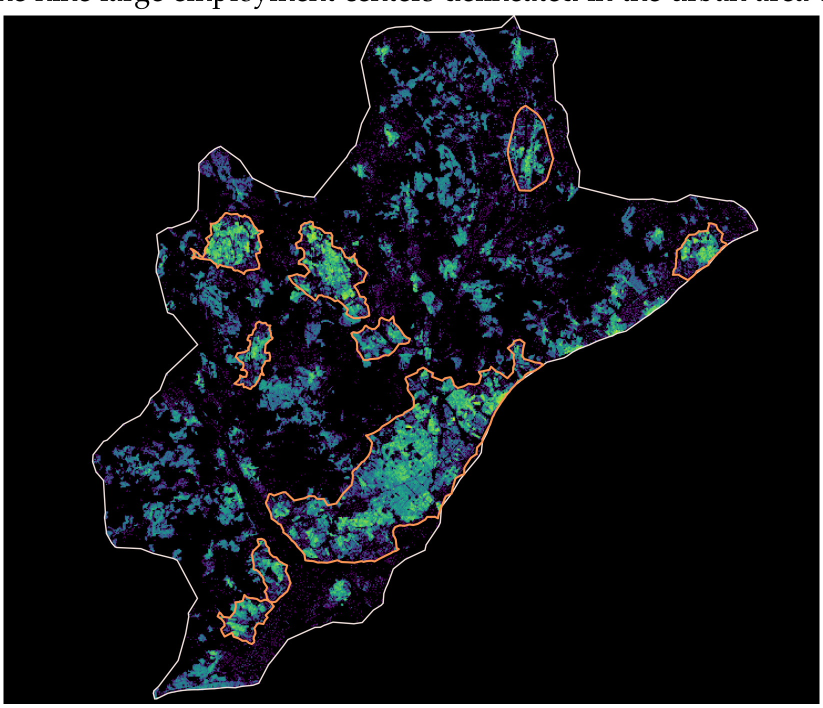
Panel A: Counts and areas for all delineated employment centers

	Counts		Areas
Buildings	3,632,133	Horizontal area = $\sum$(Building footprint)	873 km^2
Percentage of buildings in urban areas	63.5 %	Percentage of horizontal area in urban areas	54.7 %
Units within buildings	21,042,024	Vertical area = $\sum$(Footprint $\times$ floors)	3,064 km^2
Percentage of units in urban areas	79.1 %	Percentage of vertical area in urban areas	62.9 %
Average number of floors	3.5	Delineated surface	3,020 km^2
		Percentage of total surface of urban areas	13.4 %

Panel B: Number of employment centers by size

	Employment centers		Urban areas
All	2,056	in	717
Employment $\leq$ 2,500 jobs	1,420	in	503
2,500 $<$ Employment $\leq$ 5,000	347	in	297
5,000 $<$ Employment $\leq$ 10,000	193	in	169
10,000 $<$ Employment $\leq$ 25,000	70	in	63
25,000 $<$ Employment $\leq$ 50,000	19	in	17
Employment $>$ 50,000	7	in	7

Figure 7: The nine large employment centers delineated in the urban area of Barcelona



5. Comparing our delineated urban areas

Comparing our urban area delineation results, obtained with a machine learning method based on the geolocation of the country's buildings, with previous delineations performed for Spain is far from straightforward. The main reason for this is that all previous methodologies have taken the municipality (i.e. the political administrative local entity) as their starting unit of analysis. In such studies, urban areas were built by aggregating surrounding municipalities to a central one. Spain has 8,131 municipalities, most of them quite small (in fact, 90% have fewer than 5,000 inhabitants) and with considerable variation in terms of their surface area. This suggests that these methodologies are likely to be much less precise and that their results cannot be treated at the same scale as ours. Despite this, it is nevertheless interesting to compare our results with those obtained using these different methodologies.

In recent years, there have been a few attempts to aggregate municipalities into urban areas. The Statistical Atlas of Urban Areas (Atlas Estadístico de las Áreas Urbanas), published at fairly regular intervals, by the Ministry of Public Works defines 91 urban areas. The main limitation of the methodology employed, besides its use of the administrative borders of the municipalities as its starting unit, is that it only considers an urban area if the central city has more than 50,000 inhabitants. After identifying these big central municipalities, neighboring municipalities with sufficient economic links and sufficient numbers of commuters between the two units are added. Employing a very similar methodology, the AUDES research Project (AUDES Areas Urbanas de España[6]), conducted in 2010, represented another attempt to delineate urban areas. The aim of this project was to delineate all the urban areas in Spain (not just the biggest ones) using commuting and urban contiguity patterns. As a result, 261 urban areas were defined. The restriction imposed by the AUDES project was that an urban area had to have a minimum population of 20,000 inhabitants.

Addressing a different objective, and drawing solely on commuting data from the 2011 Census, Feria and Martínez-Bernabéu (2016) defined Local Labor Markets for Spain by adapting the procedure employed by the Office of Management and Budget for the US Census[7]. The main limitation of this particular exercise was that the initial population threshold imposed by the authors was 100,000 inhabitants and for this reason they identified just 41 local labor markets.

As discussed, comparing our urban area delineations with existing ones is conceptually challenging but empirically possible. By way of an initial approach, Panel B in Table 6 ranks the 10 largest urban areas delineated by the AUDES methodology that we consider to have the fewest limitations, as well as being the one for which we can access most data. We include our results in Panel A and, in Panel C, the outcomes corresponding to the municipalities' administrative borders, which we consider informative. For each urban area and method, we show the population (in thousands of inhabitants), the surface (in km^2) and the vertical area (in km^2).

[6] AUDES project. Documentation and open data available at http://alarcos.esi.uclm.es/per/fruiz/audes/ (accessed March 2018).

[7] For a detailed description of the methodology, see Feria et al. (2015).

Table 6: Main urban areas identified in Spain using different delineation methods

	Panel A: A-DBSCAN				Panel B: AUDES				Panel C: Municipalities			
		Population ('000 inh)	Surface (km^2)	Vertical (km^2)		Population ('000 inh)	Surface (km^2)	Vertical (km^2)		Population ('000 inh)	Surface (km^2)	Vertical (km^2)
1	Madrid	4,516	691	412	Madrid	6,143	4,124	689	Madrid	3,314	605	298
2	Barcelona	4,376	1,191	512	Barcelona	4,348	1,506	567	Barcelona	1,758	99	232
3	Valencia	1,655	628	218	Valencia	1,641	1,135	227	Valencia	967	137	83
4	Sevilla	1,214	635	149	Sevilla	1,227	1,467	163	Sevilla	715	141	77
5	Málaga	840	309	100	Bilbao	948	860	n.a.	Zaragoza	695	974	88
6	Zaragoza	647	88	60	Zaragoza	747	2,290	116	Málaga	573	395	65
7	Murcia	620	427	99	Málaga	647	588	79	Murcia	462	886	70
8	Sta Cruz	585	356	79	Murcia	573	1,155	98	Palma	421	209	58
9	L. Palmas	579	356	65	Palma	544	998	89	Hospitalet	412	14	18
10	Granada	514	342	75	Sta Cruz	512	608	68	L. Palmas	406	102	40
		14,967	4,667	1,769		17,332	14,731	2,096		9,723	3,562	1,029
UAs	717				296				8,131			

On average, a comparison of our delineations with those performed by AUDES shows that our method delineates cities that contain less population (up to 15% less). Likewise, the pattern that emerges when considering area as opposed to population is similar, and if anything slightly more restricted boundaries than those identified by AUDES (our urban areas have a surface area that is one third less). Indeed, proportionally, these differences are larger than when considering population. This outcome is probably the consequence of our working with buildings instead of basing the aggregation on the municipal units. However, it is interesting to see how this approach affects some of the biggest urban areas. For example, while the population assigned to Madrid's urban area by AUDES is 40% greater than that delineated by our algorithm (6.1 vs 4.5M, respectively), the area within the city's boundary is almost seven times larger for AUDES than for our delineation (4,124 vs 691 km^2, respectively). In the case of the urban area of Barcelona, the outcome is quite distinct. Both methods delineate an area of similar population (around 4.3M inhabitants) but the surface assigned by AUDES is 26% greater than that assigned when using our method (1,506 vs 1,191 km^2). Interestingly, the surface of the administrative area of Barcelona is just 99 km^2. In contrast, the urban areas of Zaragoza and Murcia are especially extensive according to the administrative delineation of their borders (974 and 886 km^2, respectively), but their surface sizes are much smaller according to our delineation method (88 and 427 km^2, respectively). As for the vertical land of the urban areas, even this is greater, on average, according to the AUDES delineation, although the difference between the two methods is smaller than that for their respective surface areas.

Because our method takes buildings as its basic unit of analysis and does not impose any ancillary geography to calculate densities, the boundaries it generates do not include any low density spaces, which abound in the peripheries of Spain's municipalities. However, it should be borne in mind that these conclusions are based on a small subset of cities. To provide confirmation, we would need to expand the analysis to the entire set of delineations. In the sections that follow, we report different exercises aimed at comparing more accurately the outcomes of the different delineation methods.

5.1 Rand index and overlapping

In this section, we employ the adjusted Rand index (see Section 3) to compare our delineated urban areas with those of the AUDES project and the Spanish municipalities. As discussed, the index developed by Hubert and Arabie (1985) measures the ratio of agreements (buildings assigned to the same urban area in two delineations) over the total number of buildings, and its values range from 0, dissimilarity or no overlap, to 1, maximum similarity or complete overlap. The ratio between our algorithm and AUDES is 0.350, while that with the municipalities is 0.003, indicating that the former provides the closest definition to our own delineations, while the municipalities is least similar[8].

Another way to compare the three delineations is simply by analyzing the extent to which, and just where exactly, the alternative delineations (AUDES and municipalities) are included within the boundaries defined by our algorithm. Figure 8 presents two histograms showing the number of AUDES urban areas and Spanish municipalities included within our A-DBSCAN boundaries (Figures 8a and 8c, respectively) and two maps showing their location (Figures 8b and 8d, respectively). In line with the evidence presented in the paragraph above, A-DBSCAN delineations coincide much more closely with the AUDES boundaries than they do with the municipalities. This is to be expected, given that AUDES are groups of municipalities linked by their common geography and commuting flows. However, the figure provides evidence that our method is able to approximate these same boundaries using a quite distinct approach.

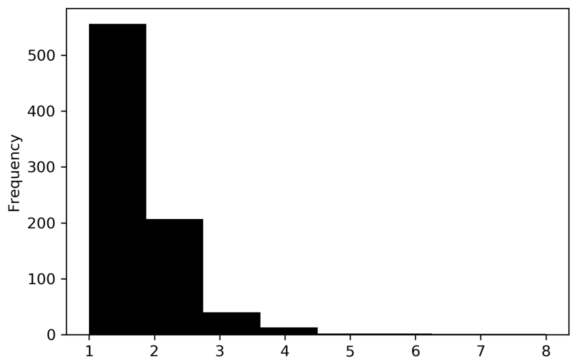
Geographically, our maps also reveal a number of clear patterns. Our delineated urban areas containing parts of more than two AUDES are disproportionately located in the Mediterranean coast. We interpret this in terms of the type of urban development present in this region compared to that in the rest of Spain. The Mediterranean region is much more developed than the rest of the country. The density of urban development is also much higher, as can be discerned from Figure 2. This pattern results in A-DBSCAN identifying larger contiguous areas in which the building density is above the threshold required for an area to be considered urban. In turn, this makes our definitions of Barcelona, Valencia, or Alicante, among others, larger than their AUDES counterparts. In contrast, in the center of the country, a much sparser region with well delimited towns and cities, most urban areas delineated by A-DBSCAN contain only one AUDES.

[8] The index between AUDES and municipalities is 0.007 and, as a result, the degree of similarity between both delineations is also quite low.

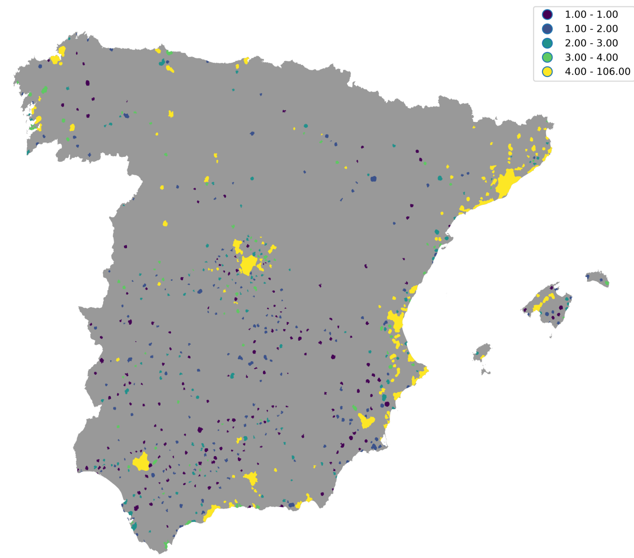
Figure 8: Overlap between different urban area delineations

Panel A: AUDES urban areas included in A-DBSCAN delineations.

(a)

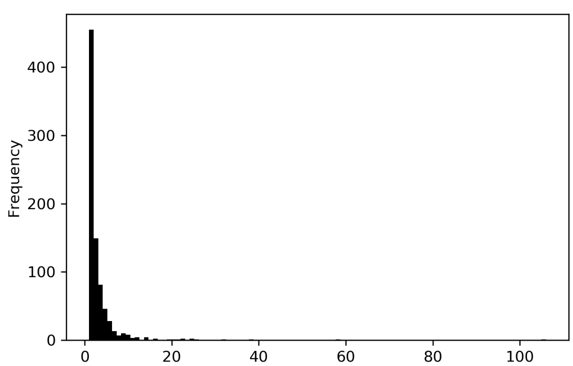


(b)

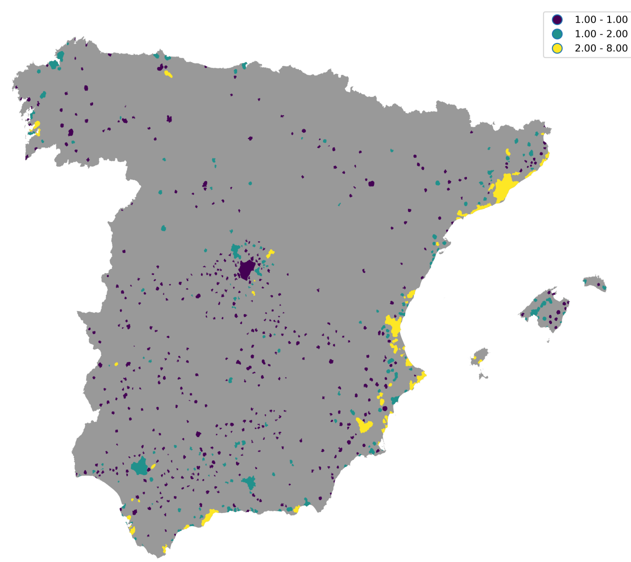


Panel B: Municipalities included in A-DBSCAN delineations.

(c)



(d)



5.2 *City size distribution*

Zipf's law suggests that a country's city size distribution can be approximated with a Pareto distribution with shape parameter equal to one. The higher (lower) the Pareto exponent, the more equally (unequally) distributed is the city system. The power law implies that, in a system of cities, the largest city is roughly twice the size of the second largest city, about three times the size of the third largest city, and so on. Indeed, since the seminal work of Gabaix (1999) and Eeckhout (2004), an enormous amount of city size distribution literature has been published (see Nitsch, 2005 and Arshad et al., 2018 for a review of this literature).

The evidence reported by this literature is mixed. In some cases the law holds precisely but, in others, the outcome lies some distance from the unit parameter. The variety of results seems to be attributable to the city definition employed and, as a consequence, to the heterogeneity in the city samples used to perform the tests. Given this situation, it is interesting to compare the city size distribution by simulating an exercise performed by Rozenfeld et al. (2011) in which different definitions of city within the same country are taken into consideration. Thus, in the following paragraphs, we seek to determine whether the city size distribution in Spain depends on the definition of the units of analysis.

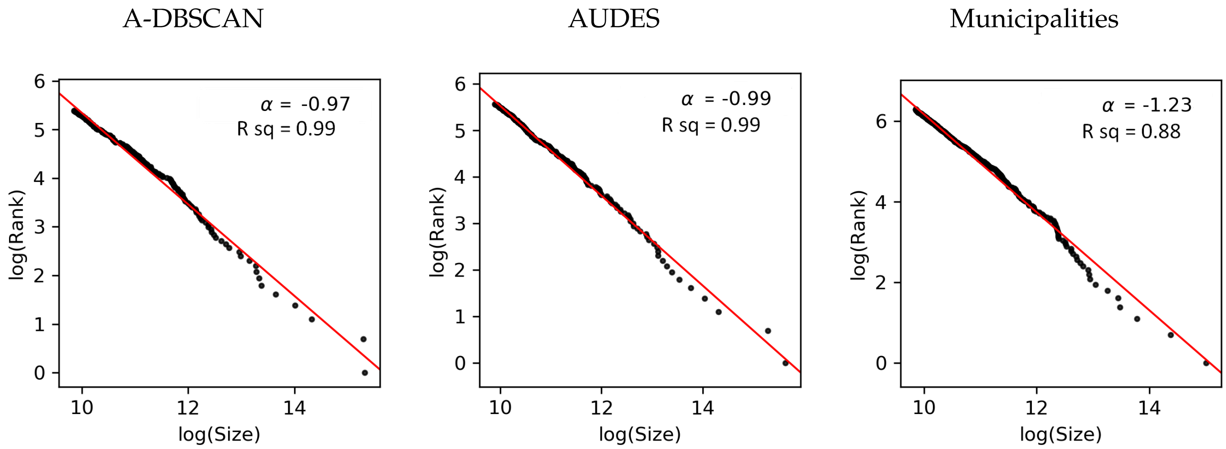
Figure 9 plots the log-ranks against the log-sizes for the urban area boundaries created by: 1) our algorithm A-DBSCAN, 2) AUDES commuting-based patterns, and 3) the administrative municipalities. Panel A shows the plots for the three delineations using population as the measure of each urban area's size[9]. The estimated Pareto exponent is negative and very close to 1 for both our delineated urban areas and those from the AUDES project (-0.97 and -0.99, respectively). As discussed above, if the estimated value of the Pareto exponent is equal to one, then Zipf's law is confirmed as holding exactly for these two delineations. In contrast, the estimated parameter for the administrative municipalities is -1.23, indicating that Zipf's law does not fit in this case and also that the distribution of the population across these units is more unequal. However, the relationship does fit the log-linear specification quite well in all three cases (with an R^2 of 0.99, 0.99 and 0.88, respectively). Interestingly, in all three delineations the slope fits very well in the upper tail of the distribution.

Most of the evidence tests Zipf's law by comparing city sizes and their populations (with the notable exception of Rozenfeld et al., 2011). However, our data allow us to replicate the analysis using the surface (Panel B) and the vertical land (Panel C) of the urban areas. In the case of the former, our results indicate that Zipf's law holds only for the urban areas produced by our delineation (with an estimated Pareto exponent of -1.09 and an R^2 of 0.98). Interestingly, for the bigger urban areas (the upper tail of the distribution) the linear fit is not perfect. For the AUDES urban areas and the municipalities the estimated values for the coefficients lie far from the unit (-0.78 and -0.61, respectively) and the R^2 values are smaller (0.79 and 0.61, respectively). In both cases, the slopes show a log-quadratic, as opposed to a log-linear, relationship between surface-rank size. When the size of the urban areas is measured in terms of the vertical land (Panel C), the estimated coefficients for both our urban delineated areas and the AUDES areas are very close to -1 (-1.04 and -1.09, respectively). In the case of the municipalities, whose distribution presents a clear concave shape, the coefficient is -0.67 clearly indicating that Zipf's law does not hold for this delineation. A detailed inspection of the shape of the distribution of the AUDES urban areas shows that the biggest city in terms of vertical land is a clear outlier. This is not the case for our delineated urban areas, which present a more continuous pattern.

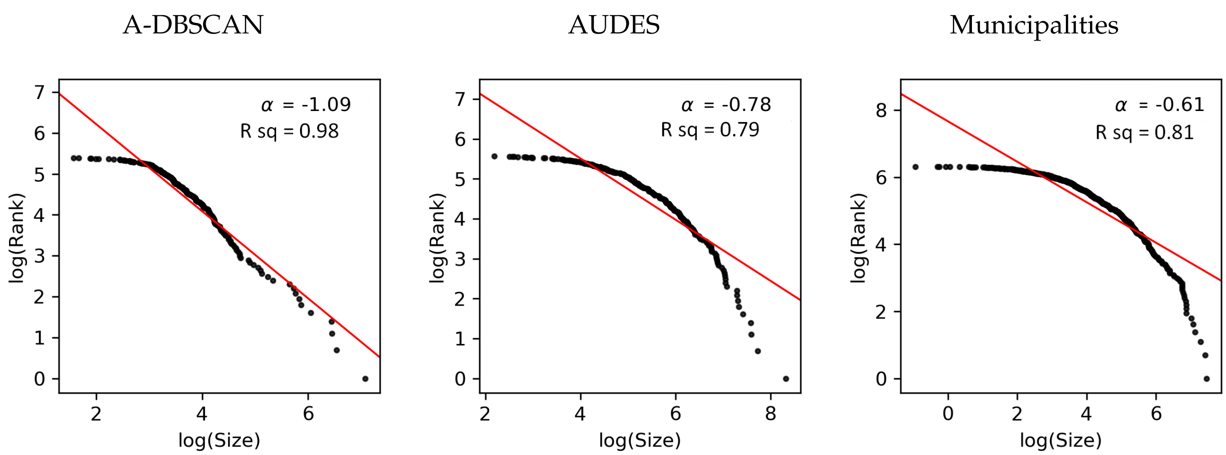
[9]The AUDES delineation only reports urban areas with more than 20,000 inhabitants. For the comparisons to be meaningful, we do the same for our delineated urban areas and the administrative municipalities.

Figure 9: Zipf plots for different measures of city size

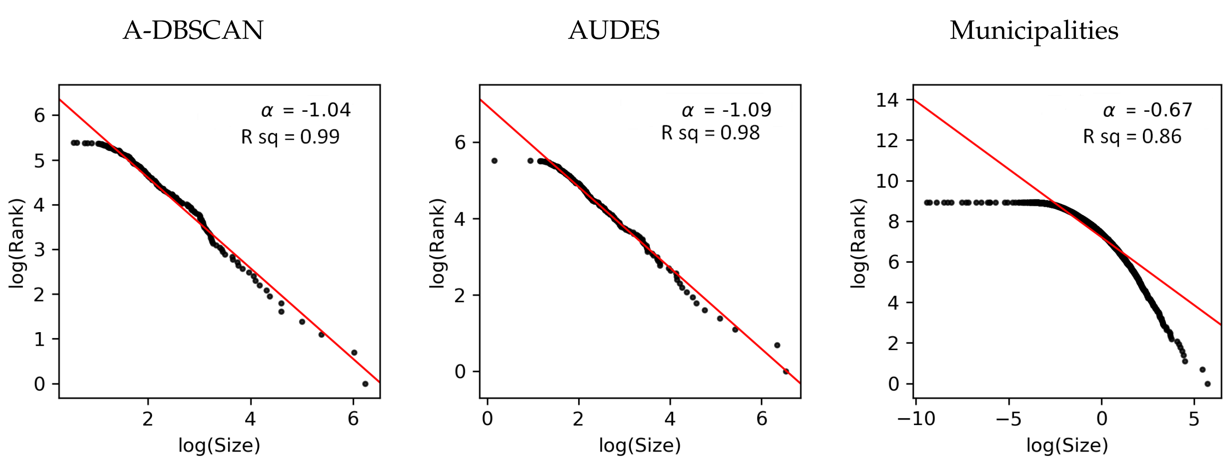
Panel A: Population



Panel B: Surface



Panel C: Vertical



Notes: Following Gabaix and Ibragimov (2011), we substract 0.5 from the rank to improve the estimation.

5.3 *City size and transportation*

As has been well documented, one of the main problems urban economists face is the so-called 'Modifiable Areal Unit Problem' (MAUP) (Openshaw, 1984). As Briant et al. (2010) have shown, empirical results can change when different spatial units are adopted. In this subsection, we compare our delineated urban areas with AUDES urban areas and Spanish municipalities by studying how city size relates to transportation[10].

Based on empirical studies analyzing the impact of transportation on city size measured in terms of population (Duranton and Turner, 2012, Garcia-López et al., 2015, Garcia-López et al., 2018, Baum-Snow et al., 2019) and land area (Brueckner and Fansler, 1983, Garcia-López, 2019), we can estimate the following equation:

$$\ln(\text{City size}) = \alpha_0 + \sum_i (\alpha_{1,i} \times \ln(\text{Transportation}_i)) + \sum_j (\alpha_{2,j} \times \text{Controls}_j) \quad (1)$$

We consider three dependent variables to measure city size. First, in line with tradition, we consider the size of the city in terms of population using the log of the 2010 population (inhabitants). Second, we take into account the physical size of the city's surface, that is, its size in terms of land area (horizontal dimension) with the log of city surface (km^2). Finally, in line with recent studies by Ahlfeldt and McMillen (2018), Brueckner et al. (2017) and Liu et al. (2018), we also study the vertical dimension of the city by using the log of vertical land area (km^2).

Our main explanatory variables are related to transportation. Specifically, we consider the 2010 log of the length of the highway network (km), and the 2010 log of the length of the railroad network (km). To compute these, we use GIS maps of the road system and the railroad network in Spain that form part of the Büro für Raumforschung, Raumplanung und Geoinformation (RRG) GIS Database. The related empirical literature (see the survey by Duranton and Puga, 2015) considers these variables as proxies for transportation costs (and road congestion).

Since the different versions of the monocentric model show that socioeconomic, geographical and historical characteristics also shape city size, we include controls related to these features. In Appendix C we discuss in detail the different control variables and we report summary statistics for all variables using cities from the three delineation methods (Table C.1).

Table 7 reports the results of estimating Equation (1) by Ordinary Least Squares (OLS). Results for population are in columns 1, 2 and 3; those for surface in columns 4, 5 and 6, and those for vertical land in columns 7, 8 and 9. As discussed, we always control for socioeconomic, geographical and historical characteristics. A qualifier is important here, however: Since the results are based on OLS estimates, they only show correlations and not causal effects between city size and transportation variables.

[10]Here, we do not consider urban areas and municipalities in the Balearic and Canary Islands, the Basque Country and Navarra, and Ceuta and Melilla because of a lack of data for most of our explanatory variables.

Table 7: City size and transportation

Dependent variable:	ln(Population)			ln(Surface)			ln(Vertical area)		
Delineation:	ADBSCAN [1]	AUDES [2]	Muni [3]	ADBSCAN [4]	AUDES [5]	Muni [6]	ADBSCAN [7]	AUDES [8]	Muni [9]
ln(Length of highways)	0.343[a]	0.186[a]	0.219[a]	0.223[a]	0.158[a]	0.120[a]	0.311[a]	0.211[a]	0.199[a]
	(0.031)	(0.038)	(0.014)	(0.022)	(0.045)	(0.010)	(0.011)	(0.027)	(0.035)
ln(Length of railroads)	0.232[a]	0.049	0.140[a]	0.157[a]	0.028	0.077[a]	0.184[a]	-0.003	0.110[a]
	(0.030)	(0.035)	(0.013)	(0.022)	(0.038)	(0.009)	(0.010)	(0.027)	(0.031)
Socioeconomy	✓	✓	✓	✓	✓	✓	✓	✓	✓
Geography	✓	✓	✓	✓	✓	✓	✓	✓	✓
History	✓	✓	✓	✓	✓	✓	✓	✓	✓
Region FE	✓	✓	✓	✓	✓	✓	✓	✓	✓
Adjusted R^2	0.80	0.77	0.76	0.70	0.81	0.60	0.79	0.77	0.71
Observations	674	219	7,450	674	219	7,450	674	219	7,450

Notes: Robust standard errors in parentheses. Socioeconomic variables are the log of the average income and the share of population with a college degree. Geographical variables are altitude, the terrain ruggedness index, land area overlying aquifers, the length of rivers, annual median and range of precipitation and median and range of temperatures for winter and summer months. Historical variables are dummy variables for cities that were Roman settlements and Medieval major towns. [a], [b] and [c] indicate significant at 1, 5, and 10 percent level, respectively.

The results for our delineated urban areas show positive and significant relationships between population and highways and railroads (column 1): The larger both transportation networks, the larger the size of the city in terms of population. Results for AUDES (column 2) and municipalities (column 3) are in the same direction but present some notable differences. First, the estimated coefficients are significantly smaller for these two alternative definitions. Second, the effect of railroads is not significant in the AUDES regression.

When we consider the size of the city in terms of land surface (columns 4, 5 and 6), we obtain similar results: Positive relationships showing that cities with larger highway and railroad systems tend to have larger land areas. However, in the case of our delineation (column 4), the estimated coefficient is significantly larger than those corresponding to AUDES (column 5) and municipalities (column 6). Furthermore, the effect of railroads is again not significant in the AUDES sample.

The results for the vertical size of the city (columns 6 to 9) are in line with the above outcomes. In general, they show that larger highway and railroad systems are positively related to larger vertical developments. The largest effects are found for our delineated urban areas. Once more the effect of railroads is not significant in the AUDES regression.

In short, these results show that larger transportation networks can be related to larger cities in terms of their population, surface and vertical land. Our preferred results are those related to our delineated urban areas: First, because they are in line with the theory on urban spatial structure (Alonso, 1964, Mills, 1967, Muth, 1969, Brueckner, 1987, Duranton and Puga, 2015) and, second, because the other two delineations, AUDES and municipalities, seem to underestimate the effects.

6. Conclusions

Empirical research in Urban Economics has to address the challenge of identifying the best geographical unit of analysis for measuring what constitutes a city. But all too often information is provided solely for a city's local administrative/political units, even though there might be a clear consensus that such an approach fails to capture its real scope. To solve this problem, in recent years, and drawing on increasingly more sophisticated data, various methodologies have been developed to delineate urban areas.

In the paper, we present a new method for delineating urban areas based on very precise geolocated data for all the buildings in Spain. Using machine learning tools, we design and calculate a distance-based clustering algorithm that defines 717 urban areas containing three quarters of the whole population in less than 5% of the territory. Detailed information about the buildings allows us to better characterize the structure of the city in terms of its verticality and the location of its residential and non-residential activities. The algorithm can also be used to delineate the employment centers within these urban areas. When comparing our delineated urban areas with other delineations, we find that our urban areas are better measured, being more similar in this regard to commuting-based delineations than to areas delimited by administrative boundaries.

Our delineations are superior to those obtained using these other two methodologies because we do not include large areas of undeveloped land, which serves only to reduce the city's overall density. Because our delineated urban areas are spatially continuous collections of buildings rather than exogenous aggregations, such as grid cells or administrative boundaries, we believe that they better reflect the idea of an urban agglomeration based on a high concentration of inhabitants and firms. Technically, it should be stressed that our algorithm is robust to marginal changes in the data and that our results are computationally scalable in large datasets with millions of observations. Thus, one of the main advantages of our method is that, with the appropriate information, it can be replicated for other countries.

The use of our delineated urban areas as a unit of analysis for urban research is feasible when statistical information is available in a geocoded format. But, today, some information continues to be provided at the administrative level. Moreover, in some research fields, including Political Economy, Public Economics and Public Policy Evaluation, it is important to maintain the administrative borders in the analysis given that they continue to delimit the political-decision unit. However, in both instances, and with some simple adjustments, our urban areas can also be used. It remains our contention that a better definition of just what constitutes a city would improve the results of empirical analyses in Urban Economics and serve to guide policy makers when taking decisions that need to take into account the precise scope of the urban area.

References

Ahlfeldt, G. M. and McMillen, D. (2018). Tall buildings and land values: Height and construction cost elasticities in chicago, 1870 – 2010. *The Review of Economics and Statistics*, 100(5):861–875.

Alonso, W. (1964). *Location and Land Use. Toward a General Theory of Land Rent*. Cambridge, MA: Harvard University Press.

Arshad, S., Hu, S., and Ashraf, B. N. (2018). Zipf's law and city size distribution: A survey of the literature and future research agenda. *Physica A: Statistical Mechanics and its Applications*, 492(C):75–92.

Baragwanath, K., Goldblatt, R., Hanson, G., and Khandelwal, A. K. (2019). Detecting urban markets with satellite imagery: An application to India. *Journal of Urban Economics*.

Baum-Snow, N., Henderson, J. V., Turner, M. A., Zhang, Q., and Brandt, L. (2019). Does investment in national highways help or hurt hinterland city growth? *Journal of Urban Economics*, Forthcoming.

BEIS (2017). Spatial clustering: identifying industrial clusters in the UK. Technical report. Department for Business, Energy & Industrial Strategy.

Bellefon, M.-P. d., Combes, P.-P., Duranton, G., Gobillon, L., and Gorin, C. (2019). Delineating urban areas using building density. Mimeo.

Birant, D. and Kut, A. (2007). St-dbscan: An algorithm for clustering spatial–temporal data. *Data & Knowledge Engineering*, 60(1):208–221.

Borah, B. and Bhattacharyya, D. (2004). An improved sampling-based dbscan for large spatial databases. In *Intelligent Sensing and Information Processing, 2004. Proceedings of International Conference on*, pages 92–96. IEEE.

Briant, A., Combes, P.-P., and Lafourcade, M. (2010). Dots to boxes: Do the size and shape of spatial units jeopardize economic geography estimations? *Journal of Urban Economics*, 67(3):287–302.

Brueckner, J. K. (1987). The structure of urban equilibria: A unified treatment of the Muth-Mills model. In Mills, E. S., editor, *Handbook of Regional and Urban Economics*, volume 2, chapter 20, pages 821–845. Elsevier, 1 edition.

Brueckner, J. K. and Fansler, D. A. (1983). The economics of urban sprawl: Theory and evidence on the spatial sizes of cities. *Review of Economics and Statistics*, 65(3):479–482.

Brueckner, J. K., Fu, S., Gu, Y., and Zhang, J. (2017). Measuring the stringency of land use regulation: The case of China's building height limits. *The Review of Economics and Statistics*, 99(4):663–677.

Burchfield, M., Overman, H. G., Puga, D., and Turner, M. A. (2006). Causes of sprawl: A portrait from space. *The Quarterly Journal of Economics*, 121(2):587–633.

Büchel, K. and von Ehrlich, M. (2019). Cities and the structure of social interactions: Evidence from mobile phone data. Mimeo.

Campello, R. J., Moulavi, D., and Sander, J. (2013). Density-based clustering based on hierarchical density estimates. In *Pacific-Asia conference on knowledge discovery and data mining*, pages 160–172. Springer.

Cascajo, R., Monzón, A., Romero, C., and Ruiz-de Galarreta, J. (2018). Informe 2016 OMM. Report, Observatorio de la Movilidad Metropolitana.

Chowdhury, P. K. R., Bhaduriy, B. L., and Mckee, J. J. (2018). Estimating urban areas: New insights from very high-resolution human settlement data. *Remote Sensing Applications Society and Environment,*, 10:93–103.

Dingel, J. I., Miscio, A., and Davis, D. R. (2019). Cities, lights, and skills in developing economies. *Journal of Urban Economics*.

Duranton, G. (2015). Delineating metropolitan areas: Measuring spatial labour market networks through commuting patterns. In Watanabe, T., Usegi, I., and Ono, A., editors, *The Economics of Interfirm Networks*, pages 107–133. Springer, Tokyo.

Duranton, G. and Puga, D. (2014). The growth of cities. In Aghion, P. and Durlauf, S., editors, *Handbook of Economic Growth*, volume 2 of *Handbook of Economic Growth*, pages 781–853. Elsevier, Amsterdam.

Duranton, G. and Puga, D. (2015). Urban Land Use. In Duranton, G., Henderson, J. V., and Strange, W. C., editors, *Handbook of Regional and Urban Economics*, volume 5 of *Handbook of Regional and Urban Economics*, pages 467–560. Elsevier.

Duranton, G. and Turner, M. A. (2012). Urban Growth and Transportation. *The Review of Economic Studies*, 79(4):1407–1440.

Edelsbrunner, H., Kirkpatrick, D. G., and Seidel, R. (1983). On the shape of a set of points in the plane. *IEEE Transactions of Information Theory*, 29(4):551–559.

Eeckhout, J. (2004). Gibrat's law for (all) cities. *American Economic Review*, 94(5):1429–1451.

Ester, M., Kriegel, H.-P., Sander, J., and Xu, X. (1996). Density-based spatial clustering of applications with noise. In *Int. Conf. Knowledge Discovery and Data Mining*, volume 240.

Fariñas, J. C. and Huergo, E. (2016). Demografía empresarial en españa: tendencias y regularidades. Working Paper Estudios sobre la Economía Española 2015/24, FEDEA.

Feria, J. M., Casado-Díaz, J. M., and Martínez Bernabéu, L. (2015). Inside the metropolis: the articulation of spanish metropolitan areas into local labor markets. *Urban Geography*, 36(7):1018–1041.

Feria, J. M. and Martínez-Bernabéu, L. (2016). La definición y delimitación del sistema metropolitano español. *Ciudad y territorio: Estudios territoriales*, 187:9–24.

Gabaix, X. (1999). Zipf's law for cities: An explanation. *The Quarterly Journal of Economics*, 114(3):739–767.

Gabaix, X. and Ibragimov, R. (2011). Rank- 1/2: a simple way to improve the ols estimation of tail exponents. *Journal of Business & Economic Statistics*, 29(1):24–39.

Garcia-López, M.-A. (2010). Population suburbanization in Barcelona, 1991-2005: Is its spatial structure changing? *Journal of Housing Economics*, 19(2):119–132.

Garcia-López, M.-A. (2012). Urban spatial structure, suburbanization and transportation in barcelona. *Journal of Urban Economics*, 72(2):176–190.

Garcia-López, M.-A. (2019). All roads lead to rome ... and to sprawl? evidence from European cities. *Regional Science and Urban Economics*, Forthcoming.

Garcia-López, M.-A., Hemet, C., and Viladecans-Marsal, E. (2017a). How does transportation shape intrametropolitan growth? An answer from the Regional Express Rail. *Journal of Regional Science*, 57(5):758–780.

Garcia-López, M.-A., Hemet, C., and Viladecans-Marsal, E. (2017b). Next train to the polycentric city: The effect of railroads on subcenter formation. *Regional Science and Urban Economics*, 67(C):50–63.

Garcia-López, M.-A., Holl, A., and Viladecans-Marsal, E. (2015). Suburbanization and highways in Spain when the Romans and the Bourbons still shape its cities. *Journal of Urban Economics*, 85(C):52–67.

Garcia-López, M.-A., Pasidis, I., and Viladecans-Marsal, E. (2018). Amphitheaters, cathedrals and operas: The role of historic amenities on suburbanization. Working Paper DP13129, CEPR.

Giuliano, G., Redfearn, C. L., Agarwal, A., Li, C., and Zhuang, D. (2007). Employment concentrations in Los Angeles, 1980-2000. *Environment and Planning A*, 39(12):2935–2957.

Giuliano, G. and Small, K. (1991). Subcenters in the los angeles region. *Regional Science and Urban Economics*, 21(2):163–182.

Henderson, J. V., Squires, T., Storeygard, A., and Weil, D. (2018). The global distribution of economic activity: Nature, history, and the role of trade. *The Quarterly Journal of Economics*, 133(1):357–406.

Hortas-Rico, M. and Onrubia, J. (2016). Renta personal de los municipios espanñoles y su distribución, annños 2004 a 2006 y actualización de 2007. Working Paper Estudios sobre la Economía Española 2016/11, FEDEA.

Hu, Y., Gao, S., Janowicz, K., Yu, B., Li, W., and Prasad, S. (2015). Extracting and understanding urban areas of interest using geotagged photos. *Computers, Environment and Urban Systems*, 54:240–254.

Hubert, L. and Arabie, P. (1985). Comparing partitions. *Journal of classification*, 2(1):193–218.

Instituto Nacional de Estadística (2018). Nota de prensa del 12 de abril de 2018 sobre la Encuesta Continua de Hogares año 2017. http://www.ine.es/prensa/ech_2017.pdf.

Karami, A. and Johansson, R. (2014). Choosing DBSCAN parameters automatically using differential evolution. *International Journal of Computer Applications*, 91(7).

Khan, K., Rehman, S. U., Aziz, K., Fong, S., and Sarasvady, S. (2014). Dbscan: Past, present and future. In *Applications of Digital Information and Web Technologies (ICADIWT), 2014 Fifth International Conference on the*, pages 232–238. IEEE.

Liu, C. H., Rosenthal, S. S., and Strange, W. C. (2018). The vertical city: Rent gradients, spatial structure, and agglomeration economies. *Journal of Urban Economics*, 106:101 – 122.

Louail, T., Lenormand, M., Cantu Ros, O. G., Picornell, M., Herranz, R., Frias-Martinez, E., Ramasco, J. J., and Barthelemy, M. (2014). From mobile phone data to the spatial structure of cities. *Scientific Reports*, 4:5276.

Lv, Y., Ma, T., Tang, M., Cao, J., Tian, Y., Al-Dhelaan, A., and Al-Rodhaan, M. (2016). An efficient and scalable density-based clustering algorithm for datasets with complex structures. *Neurocomputing*, 171:9–22.

McMillen, D. P. (2001). Nonparametric employment subcenter identification. *Journal of Urban Economics*, 50(3):448–473.

McMillen, D. P. and Smith, S. C. (2003). The number of subcenters in large urban areas. *Journal of Urban Economics*, 53(3):321–338.

Mills, E. S. (1967). An aggregative model of resource allocation in a metropolitan area. *American Economic Review*, 57:197–210.

Muñiz, I., Garcia-López, M.-A., and Galindo, A. (2008). The effect of employment sub-centres on population density in Barcelona. *Urban Studies*, 45(3):627–649.

Muth, R. F. (1969). *Cities and Housing: The Spatial Pattern of Urban Residential Land Use*. Chicago, IL: University of Chicago Press.

Nitsch, V. (2005). Zipf zipped. *Journal of Urban Economics*, 57(1):86–100.

Openshaw, S. (1984). The modifiable areal unit problem. *Concepts and techniques in modern geography*.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., and Duchesnay, E. (2011). Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830.

Redfearn, C. L. (2007). The topography of metropolitan employment: Identifying centers of employment in a polycentric urban area. *Journal of Urban Economics*, 61(3):519–541.

Riley, S. J., DeGloria, S. D., and Elliot, R. (1999). A terrain ruggedness index that quantifies topographic heterogeneity. *Intermountain Journal of Sciences*, 5(1–4):23–27.

Rozenfeld, H. D., Rybski, D., Gabaix, X., and Makse, H. A. (2011). The area and population of cities: New insights from a different perspective on cities. *American Economic Review*, 101(5):2205–25.

Wang, Q., Phillips, N. E., Small, M. L., and Sampson, R. J. (2018). Urban mobility and neighborhood isolation in america's 50 largest cities. *Proceedings of the National Academy of Sciences*.

Appendix A. The A-DBSCAN algorithm: Technical details

This section describes in detail A-DBSCAN, the novel methodology developed to delineate urban areas. The algorithm is based on an extension of the popular machine learning algorithm DBSCAN (Density-Based Spatial Clustering of Applications with Noise), originally developed by Ester et al. (1996). The intuition behind the initial proposal is rather straightfoward. DBSCAN requires two parameters to be set, the minimum number of points (`min_pts`) and the distance threshold (ϵ), and labels every observation in a dataset as either core, border or noise. 'Core' points contain at least a minimum number `min_pts` of observations within a maximum distance threshold ϵ; 'border' observations fall within the distance threshold of a core point, but they do not contain themeselves the minimum number of points within that distance of themselves; finally, observations are labeled as 'noise' if they are neither core, nor border. Since the original proposal, the algorithm has grown significantly in popularity, and several extensions, focusing on different contexts (e.g. Borah and Bhattacharyya, 2004, Birant and Kut, 2007), limitations (e.g. Campello et al., 2013, Karami and Johansson, 2014), and challenges (e.g. Lv et al., 2016), as well as a wide range of applications have been proposed. A comprehensive review is beyond the scope of this section (see Khan et al., 2014, for a recent account), but it is important to note that, although most of the contributions have stemmed from the machine learning and data mining literature, more recently, social scientists and policy makers have shown growing interest as well (e.g. Hu et al., 2015, BEIS, 2017, Wang et al., 2018).

Because the way the minimum number of points and the maximum distance threshold criteria are combined, DBSCAN is able to use a density-based criteria to identify clusters of points (observations labeled as core or border), and this is performed without the need to rely on ancillary aggregations. However DBSCAN is "memory expensive" (Borah and Bhattacharyya, 2004), which means the algorithm becomes for all purposes computationally unfeasible when the size of the dataset grows to a certain point. Thus it is not scalable. Similarly, neither the original proposal nor, to the best of our knowledge, any of the extensions proposed to date, are robust in the sense described above. As an illustration with the initial proposal, it is entirely possible that two clusters are connected by a single "bridge" point that is close enough to border points of both clusters to merge them into one. In the context of our application, this would equate to two distinct urban areas merged into one due to a single building.

To overcome those two remaining requirements –scalability and inferential robustness– we develop an extension to the original DBSCAN algorithm, which we term 'Approximate-DBSCAN' (A-DBSCAN). The intuition behind our approach is to create several replications of approximate solutions, and keep only those labellings where a large majority of the replications agree on a given label, considering as noise cases where not enough replications agree on a specific label. Approximate solutions are obtained by calculating the exact DBSCAN on a small random subsample, and then assigning to the remainder points the label of their nearest observation in the considered subsample. A more detailed description of our algorithm can be expressed as follows (a formal derivation of the algorithm is included in Algorithm 1):

1. Split the dataset into two random subsets, one of them potentially smaller than the other

(e.g. 10%/90%).

2. Run the original DBSCAN algorithm on one of the subsamples to obtain a set of cluster labels for each point in that subset.
3. For each point in the second subset, assign the label of the nearest point in the first subset.
4. Save the entire set of labels as a single candidate solution.
5. Repeat steps 1-4 a reasonably large amount of times (e.g. 1,000), obtaining several candidate solutions.
6. Align labels across candidate solutions so a given label represents the same cluster in each of the replications. This can be done in the following way:
   (a) Set the solution with most clusters as reference.
   (b) For each cluster in the remaining solutions, find the nearest cluster in the reference and assign its label. In this context, nearest can be expressed as the shortest distance between the centroids of the two clusters.
7. For each point in the dataset, obtain the most common label and the proportion of times across candidate solutions where that label is assigned.
8. If the most common label appears a proportion of times that is smaller than a desired threshold (e.g. 90%), label the observation as noise; otherwise assign the most common label as the label for that point.

Since the core "clustering engine" of A-DBSCAN is DBSCAN, our extension carries with it all of the original benefits, including the independence of ancillary geographies and the density-based criterion. In addition, since A-DBSCAN only requires to run DBSCAN on a potentially small fraction of the data, the approach is much more scalable. For A-DBSCAN to work, the original DBSCAN algorithm needs to be applied to a subset that is large enough to capture the overall spatial structure of the point pattern represented in the initial dataset. Furthermore, our approach is also more robust to outliers and thus more likely to accurately capture the underlying clustering process. The final label a point receives is not the result of a single run, but includes information based on several independent runs. A-DBSCAN exploits the assumption that the initial dataset is large to its advantage, treating it as a "pool" of replications from the underlying data generating process (DGP), that can be used to construct not only one but several observed patterns, with less observations but with the same properties and spatial structure. By considering several random subsets that come from the same DGP, we obtain empirical distributions for each observation and each label. This approach allows us to evaluate the uncertainty behind each assignment and to only label as part of a cluster those cases where we have sufficient evidence. This is not possible in the traditional approach because only one assignment is carried out. The resulting label each point is assigned is thus robust to outliers, corner solutions, and other forms of noise.

Algorithm 1 A-DBSCAN Algorithm

procedure ADBSCAN(XY, ϵ, `min_pts`, pct_{thin}, thr)
  $D \leftarrow$ Dataset of point locations (XY coordinates)

  for r in R **do**:
    D_{thin}, $D_{extend} \leftarrow$ Split D into two random subsets of proportions pct_{thin} and $(1 - pct_{thin})$
    $L_{thin} \leftarrow$ DBSCAN(D_{thin}, ϵ, `min_pts` $\times$ pct_{thin})
    $KNN1_{thin} \leftarrow$ Fit the K-Nearest Neighbor regressor ($K = 1$): $L_{thin} = KNN1(D_{thin})$
    $L_{extend} \leftarrow$ Use $KNN1_{thin}$ on D_{extend}
    $L_r \leftarrow L_{extend} \cup L_{extend}$
  $L_{\bar{R}} \leftarrow$ ALIGN L_R
  for i in D **do**:
    Obtain the most common label cl_i and its frequency across R, pct_{cl}
    if $pct_{cl} < thr$, where thr is a set threshold **then** label cl as noise
  $L_{ADBSCAN} \leftarrow$ Final set of labels for each observation in D
procedure ALIGN(L_R)
  $L_{ref} \leftarrow$ Solution with most clusters idenfied
  for r in R; $r \notin L_{ref}$ **do**:
    for cluster cl_r in r **do**
      Assign to cl_r the label of the nearest cluster centroid in L_{ref}

Appendix B. The A-DBSCAN with different distance thresholds (ϵ)

Figure B.1: Delineated urban areas with different distance thresholds (ϵ)

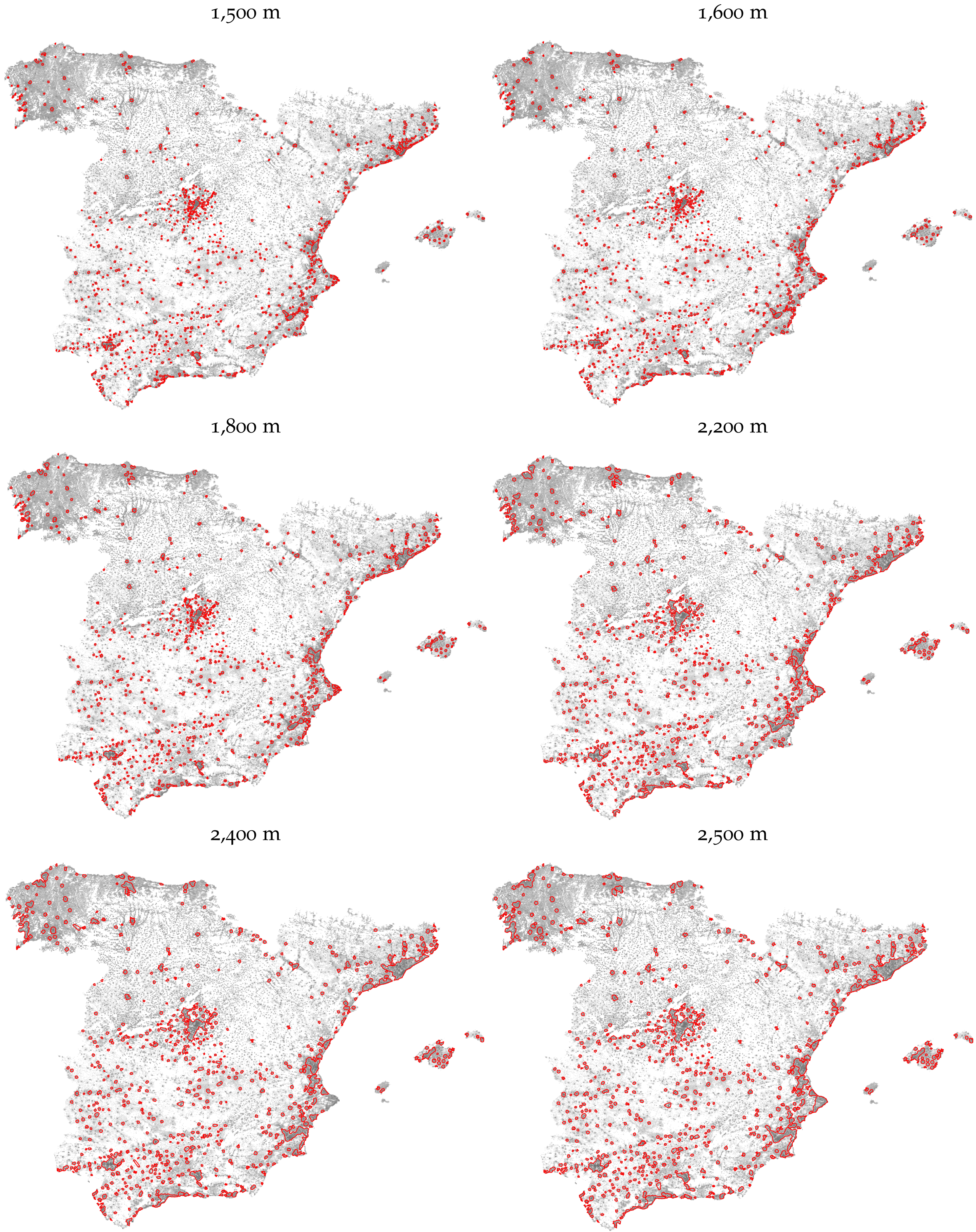


Table B.1: Delineated urban areas with different distance thresholds (ϵ)

Panel A: 1,500 m	Number of cities	Population	Percentage of Spain's pop
All	773	32,874,204	70.2 %
Population $\leq$ 5,000	98	359,085	0.8 %
5,000 $<$ Population $\leq$ 10,000	228	1,644,465	3.5 %
10,000 $<$ Population $\leq$ 25,000	240	3,813,255	8.1 %
25,000 $<$ Population $\leq$ 100,000	153	7,589,260	16.2 %
100,000 $<$ Population $\leq$ 500,000	48	9,255,530	19.8 %
Population $>$ 500,000	6	10,212,610	21.8 %
Panel B: 1,600 m	Number of cities	Population	Percentage of Spain's pop
All	759	32,716,224	69.9 %
Population $\leq$ 5,000	106	392,360	0.8 %
5,000 $<$ Population $\leq$ 10,000	225	1,620,100	3.5 %
10,000 $<$ Population $\leq$ 25,000	228	3,585,960	7.7 %
25,000 $<$ Population $\leq$ 100,000	145	7,015,350	15.0 %
100,000 $<$ Population $\leq$ 500,000	49	9,538,350	20.4 %
Population $>$ 500,000	6	10,564,105	22.6 %
Panel C: 1,800 m	Number of cities	Population	Percentage of Spain's pop
All	743	34,389,528	73.5 %
Population $\leq$ 5,000	116	428,190	0.9 %
5,000 $<$ Population $\leq$ 10,000	218	1,552,525	3.3 %
10,000 $<$ Population $\leq$ 25,000	212	3,286,870	7.0 %
25,000 $<$ Population $\leq$ 100,000	140	6,544,190	14.0 %
100,000 $<$ Population $\leq$ 500,000	49	9,159,175	20.0 %
Population $>$ 500,000	8	13,418,580	28.7 %
Panel D: 2,200 m	Number of cities	Population	Percentage of Spain's pop
All	698	35,666,120	76.2 %
Population $\leq$ 5,000	138	491,130	1.1 %
5,000 $<$ Population $\leq$ 10,000	212	1,502,455	3.2 %
10,000 $<$ Population $\leq$ 25,000	173	2,713,950	5.8 %
25,000 $<$ Population $\leq$ 100,000	121	5,740,090	12.3 %
100,000 $<$ Population $\leq$ 500,000	44	8,796,375	18.8 %
Population $>$ 500,000	10	16,422,120	35.1 %
Panel E: 2,400 m	Number of cities	Population	Percentage of Spain's pop
All	669	36,268,040	77.5 %
Population $\leq$ 5,000	139	489,910	1.1 %
5,000 $<$ Population $\leq$ 10,000	200	1,425,270	3.0 %
10,000 $<$ Population $\leq$ 25,000	164	2,634,635	5.6 %
25,000 $<$ Population $\leq$ 100,000	115	5,477,790	11.7 %
100,000 $<$ Population $\leq$ 500,000	41	8,849,540	18.9 %
Population $>$ 500,000	10	17,390,896	37.1 %
Panel F: 2,500 m	Number of cities	Population	Percentage of Spain's pop
All	669	36,886,244	78.8 %
Population $\leq$ 5,000	143	502,950	1.1 %
5,000 $<$ Population $\leq$ 10,000	198	1,400,340	3.0 %
10,000 $<$ Population $\leq$ 25,000	163	2,599,550	5.6 %
25,000 $<$ Population $\leq$ 100,000	115	5,583,065	11.9 %
100,000 $<$ Population $\leq$ 500,000	38	7,893,675	16.9 %
Population $>$ 500,000	12	18,906,664	40.4 %

Notes: In 2011, 46,815,916 inhabitants lived in Spain. Population is computed using population grid data (1$\times$1 km cells within the boundaries of our cities) from the 2011 Population Census.

Appendix C. Control variables and summary statistics

Table C.1: Summary statistics for the three delineation methods

	A-DBSCAN		AUDES		Municipalities	
	Mean	St. Dev.	Mean	St. Dev.	Mean	St. Dev.
Population (inhabitants)	47,838	261,519	143,494	528,340	6,432	50,802
Surface (km^2)	30.43	69.97	352.8	464.9	63.15	94.57
Vertical area (km^2)	6.66	28.32	20.73	63.23	1.05	5.30
Floors	2.80	0.38	2.95	0.40	2.57	0.43
Length of highway network (km)	5.25	23.87	31.64	72.34	2.08	6.70
Length of railroad network (km)	5.02	17.11	30.41	50.94	2.50	7.27
Median income (€)	10,823	5,415	13,523	2,228	1,162	3,128
Share of college degree	0.10	0.05	0.13	0.05	0.07	0.05
Altitude (km)	0.41	0.28	0.34	0.30	0.70	0.38
Terrain ruggedness index (km)	0.04	0.02	0.05	0.04	0.06	0.05
Land area overlying aquifers	19.91	56.23	212.9	319.1	32.47	62.18
Length of rivers (km)	6.40	15.11	53.73	71.59	10.03	17.18
Annual median precipitation (mm)	612.5	285.6	633.5	327.5	633.5	270.6
Annual range precipitation (mm)	177.9	227.7	372.8	412.4	351.3	322.4
December median temperature (°C)	8.53	2.34	9.26	2.56	6.07	2.53
January median temperature (°C)	7.72	2.31	8.44	2.49	5.29	2.50
February median temperature (°C)	8.94	2.16	9.51	2.33	6.51	2.47
March median temperature (°C)	11.28	1.98	11.61	2.16	8.83	2.45
June median temperature (°C)	20.87	1.89	20.61	2.10	18.56	2.74
July median temperature (°C)	24.41	2.18	23.91	2.31	22.14	2.96
August median temperature (°C)	24.29	2.19	23.95	2.40	21.90	3.00
September median temperature (°C)	21.22	2.06	21.17	2.29	18.70	2.84
December range temperature (°C)	0.85	0.80	1.89	2.23	1.64	1.67
January range temperature (°C)	0.85	0.82	1.76	2.14	1.55	1.62
February range temperature (°C)	0.80	0.77	1.86	2.21	1.61	1.68
March range temperature (°C)	0.79	0.76	2.04	2.38	1.82	1.87
June range temperature (°C)	0.74	0.68	1.96	2.32	1.73	1.79
July range temperature (°C)	0.67	0.61	1.76	2.01	1.49	1.50
August range temperature (°C)	0.69	0.63	1.82	2.10	1.58	1.61
September range temperature (°C)	0.74	0.69	1.98	2.34	1.78	1.86
Dummy for Roman settlements	0.24	0.43	0.53	0.50	0.06	0.24
Dummy for Medieval major towns	0.06	0.24	0.18	0.383	0.01	0.09
Observations	674	674	219	219	7,450	7,450

We control for socioeconomic characteristics by including the 2007 log of the median income (€) (based on the municipal estimates by Hortas-Rico and Onrubia (2016)). Using the 2011 Population Census grid data (1×1 km), we compute the share of population with a college degree.

Following Burchfield et al. (2006), we control for geography. First, using data from Spain's Digital Elevation Model we compute the altitude (km) and the terrain ruggedness index developed by Riley et al. (1999). Second, we compute the land area overlying aquifers (km^2) and the log of the length of the rivers (km) crossing each city. Finally, using the Átlas Climático Digital de la Península Ibérica, we compute the log of the annual median precipitation (mm) and the log of the annual range precipitation (mm). Similarly, we include the log of the median temperature (°C) and the log of the range temperature (°C) for December, January, February and March (winter), and for June, July, August and September (summer).

We follow Garcia-López et al. (2018) and add controls for history. We use information about Roman settlements and Medieval major towns from the Digital Atlas of Roman and Medieval Civilizations (DARMC) to classify the cities (and create dummies) according to their importance and origins (Roman settlement, Medieval city).

Finally, to consider the different dynamics and characteristics within Spain's geography, we add dummies for the regions where our urban areas are located.

2013

2013/1, Sánchez-Vidal, M.; González-Val, R.; Viladecans-Marsal, E.: "Sequential city growth in the US: does age matter?"
2013/2, Hortas Rico, M.: "Sprawl, blight and the role of urban containment policies. Evidence from US cities"
2013/3, Lampón, J.F.; Cabanelas-Lorenzo, P-; Lago-Peñas, S.: "Why firms relocate their production overseas? The answer lies inside: corporate, logistic and technological determinants"
2013/4, Montolio, D.; Planells, S.: "Does tourism boost criminal activity? Evidence from a top touristic country"
2013/5, Garcia-López, M.A.; Holl, A.; Viladecans-Marsal, E.: "Suburbanization and highways: when the Romans, the Bourbons and the first cars still shape Spanish cities"
2013/6, Bosch, N.; Espasa, M.; Montolio, D.: "Should large Spanish municipalities be financially compensated? Costs and benefits of being a capital/central municipality"
2013/7, Escardíbul, J.O.; Mora, T.: "Teacher gender and student performance in mathematics. Evidence from Catalonia"
2013/8, Arqué-Castells, P.; Viladecans-Marsal, E.: "Banking towards development: evidence from the Spanish banking expansion plan"
2013/9, Asensio, J.; Gómez-Lobo, A.; Matas, A.: "How effective are policies to reduce gasoline consumption? Evaluating a quasi-natural experiment in Spain"
2013/10, Jofre-Monseny, J.: "The effects of unemployment benefits on migration in lagging regions"
2013/11, Segarra, A.; García-Quevedo, J.; Teruel, M.: "Financial constraints and the failure of innovation projects"
2013/12, Jerrim, J.; Choi, A.: "The mathematics skills of school children: How does England compare to the high performing East Asian jurisdictions?"
2013/13, González-Val, R.; Tirado-Fabregat, D.A.; Viladecans-Marsal, E.: "Market potential and city growth: Spain 1860-1960"
2013/14, Lundqvist, H.: "Is it worth it? On the returns to holding political office"
2013/15, Ahlfeldt, G.M.; Maennig, W.: "Homevoters vs. leasevoters: a spatial analysis of airport effects"
2013/16, Lampón, J.F.; Lago-Peñas, S.: "Factors behind international relocation and changes in production geography in the European automobile components industry"
2013/17, Guío, J.M.; Choi, A.: "Evolution of the school failure risk during the 2000 decade in Spain: analysis of Pisa results with a two-level logistic mode"
2013/18, Dahlby, B.; Rodden, J.: "A political economy model of the vertical fiscal gap and vertical fiscal imbalances in a federation"
2013/19, Acacia, F.; Cubel, M.: "Strategic voting and happiness"
2013/20, Hellerstein, J.K.; Kutzbach, M.J.; Neumark, D.: "Do labor market networks have an important spatial dimension?"
2013/21, Pellegrino, G.; Savona, M.: "Is money all? Financing versus knowledge and demand constraints to innovation"
2013/22, Lin, J.: "Regional resilience"
2013/23, Costa-Campi, M.T.; Duch-Brown, N.; García-Quevedo, J.: "R&D drivers and obstacles to innovation in the energy industry"
2013/24, Huisman, R.; Stradnic, V.; Westgaard, S.: "Renewable energy and electricity prices: indirect empirical evidence from hydro power"
2013/25, Dargaud, E.; Mantovani, A.; Reggiani, C.: "The fight against cartels: a transatlantic perspective"
2013/26, Lambertini, L.; Mantovani, A.: "Feedback equilibria in a dynamic renewable resource oligopoly: pre-emption, voracity and exhaustion"
2013/27, Feld, L.P.; Kalb, A.; Moessinger, M.D.; Osterloh, S.: "Sovereign bond market reactions to fiscal rules and no-bailout clauses – the Swiss experience"
2013/28, Hilber, C.A.L.; Vermeulen, W.: "The impact of supply constraints on house prices in England"
2013/29, Revelli, F.: "Tax limits and local democracy"
2013/30, Wang, R.; Wang, W.: "Dress-up contest: a dark side of fiscal decentralization"
2013/31, Dargaud, E.; Mantovani, A.; Reggiani, C.: "The fight against cartels: a transatlantic perspective"
2013/32, Saarimaa, T.; Tukiainen, J.: "Local representation and strategic voting: evidence from electoral boundary reforms"
2013/33, Agasisti, T.; Murtinu, S.: "Are we wasting public money? No! The effects of grants on Italian university students' performances"
2013/34, Flacher, D.; Harari-Kermadec, H.; Moulin, L.: "Financing higher education: a contributory scheme"
2013/35, Carozzi, F.; Repetto, L.: "Sending the pork home: birth town bias in transfers to Italian municipalities"
2013/36, Coad, A.; Frankish, J.S.; Roberts, R.G.; Storey, D.J.: "New venture survival and growth: Does the fog lift?"

2013/37, Giulietti, M.; Grossi, L.; Waterson, M.: "Revenues from storage in a competitive electricity market: Empirical evidence from Great Britain"

2014

2014/1, Montolio, D.; Planells-Struse, S.: "When police patrols matter. The effect of police proximity on citizens' crime risk perception"
2014/2, Garcia-López, M.A.; Solé-Ollé, A.; Viladecans-Marsal, E.: "Do land use policies follow road construction?"
2014/3, Piolatto, A.; Rablen, M.D.: "Prospect theory and tax evasion: a reconsideration of the Yitzhaki puzzle"
2014/4, Cuberes, D.; González-Val, R.: "The effect of the Spanish Reconquest on Iberian Cities"
2014/5, Durán-Cabré, J.M.; Esteller-Moré, E.: "Tax professionals' view of the Spanish tax system: efficiency, equity and tax planning"
2014/6, Cubel, M.; Sanchez-Pages, S.: "Difference-form group contests"
2014/7, Del Rey, E.; Racionero, M.: "Choosing the type of income-contingent loan: risk-sharing versus risk-pooling"
2014/8, Torregrosa Hetland, S.: "A fiscal revolution? Progressivity in the Spanish tax system, 1960-1990"
2014/9, Piolatto, A.: "Itemised deductions: a device to reduce tax evasion"
2014/10, Costa, M.T.; García-Quevedo, J.; Segarra, A.: "Energy efficiency determinants: an empirical analysis of Spanish innovative firms"
2014/11, García-Quevedo, J.; Pellegrino, G.; Savona, M.: "Reviving demand-pull perspectives: the effect of demand uncertainty and stagnancy on R&D strategy"
2014/12, Calero, J.; Escardíbul, J.O.: "Barriers to non-formal professional training in Spain in periods of economic growth and crisis. An analysis with special attention to the effect of the previous human capital of workers"
2014/13, Cubel, M.; Sanchez-Pages, S.: "Gender differences and stereotypes in the beauty"
2014/14, Piolatto, A.; Schuett, F.: "Media competition and electoral politics"
2014/15, Montolio, D.; Trillas, F.; Trujillo-Baute, E.: "Regulatory environment and firm performance in EU telecommunications services"
2014/16, Lopez-Rodriguez, J.; Martinez, D.: "Beyond the R&D effects on innovation: the contribution of non-R&D activities to TFP growth in the EU"
2014/17, González-Val, R.: "Cross-sectional growth in US cities from 1990 to 2000"
2014/18, Vona, F.; Nicolli, F.: "Energy market liberalization and renewable energy policies in OECD countries"
2014/19, Curto-Grau, M.: "Voters' responsiveness to public employment policies"
2014/20, Duro, J.A.; Teixidó-Figueras, J.; Padilla, E.: "The causal factors of international inequality in co2 emissions per capita: a regression-based inequality decomposition analysis"
2014/21, Fleten, S.E.; Huisman, R.; Kilic, M.; Pennings, E.; Westgaard, S.: "Electricity futures prices: time varying sensitivity to fundamentals"
2014/22, Afcha, S.; García-Quevedo, J,: "The impact of R&D subsidies on R&D employment composition"
2014/23, Mir-Artigues, P.; del Río, P.: "Combining tariffs, investment subsidies and soft loans in a renewable electricity deployment policy"
2014/24, Romero-Jordán, D.; del Río, P.; Peñasco, C.: "Household electricity demand in Spanish regions. Public policy implications"
2014/25, Salinas, P.: "The effect of decentralization on educational outcomes: real autonomy matters!"
2014/26, Solé-Ollé, A.; Sorribas-Navarro, P.: "Does corruption erode trust in government? Evidence from a recent surge of local scandals in Spain"
2014/27, Costas-Pérez, E.: "Political corruption and voter turnout: mobilization or disaffection?"
2014/28, Cubel, M.; Nuevo-Chiquero, A.; Sanchez-Pages, S.; Vidal-Fernandez, M.: "Do personality traits affect productivity? Evidence from the LAB"
2014/29, Teresa Costa, M.T.; Trujillo-Baute, E.: "Retail price effects of feed-in tariff regulation"
2014/30, Kilic, M.; Trujillo-Baute, E.: "The stabilizing effect of hydro reservoir levels on intraday power prices under wind forecast errors"
2014/31, Costa-Campi, M.T.; Duch-Brown, N.: "The diffusion of patented oil and gas technology with environmental uses: a forward patent citation analysis"
2014/32, Ramos, R.; Sanromá, E.; Simón, H.: "Public-private sector wage differentials by type of contract: evidence from Spain"
2014/33, Backus, P.; Esteller-Moré, A.: "Is income redistribution a form of insurance, a public good or both?"
2014/34, Huisman, R.; Trujillo-Baute, E.: "Costs of power supply flexibility: the indirect impact of a Spanish policy change"

2014/35, Jerrim, J.; Choi, A.; Simancas Rodríguez, R.: "Two-sample two-stage least squares (TSTSLS) estimates of earnings mobility: how consistent are they?"
2014/36, Mantovani, A.; Tarola, O.; Vergari, C.: "Hedonic quality, social norms, and environmental campaigns"
2014/37, Ferraresi, M.; Galmarini, U.; Rizzo, L.: "Local infrastructures and externalities: Does the size matter?"
2014/38, Ferraresi, M.; Rizzo, L.; Zanardi, A.: "Policy outcomes of single and double-ballot elections"

2015

2015/1, Foremny, D.; Freier, R.; Moessinger, M-D.; Yeter, M.: "Overlapping political budget cycles in the legislative and the executive"
2015/2, Colombo, L.; Galmarini, U.: "Optimality and distortionary lobbying: regulating tobacco consumption"
2015/3, Pellegrino, G.: "Barriers to innovation: Can firm age help lower them?"
2015/4, Hémet, C.: "Diversity and employment prospects: neighbors matter!"
2015/5, Cubel, M.; Sanchez-Pages, S.: "An axiomatization of difference-form contest success functions"
2015/6, Choi, A.; Jerrim, J.: "The use (and misuse) of Pisa in guiding policy reform: the case of Spain"
2015/7, Durán-Cabré, J.M.; Esteller-Moré, A.; Salvadori, L.: "Empirical evidence on tax cooperation between sub-central administrations"
2015/8, Batalla-Bejerano, J.; Trujillo-Baute, E.: "Analysing the sensitivity of electricity system operational costs to deviations in supply and demand"
2015/9, Salvadori, L.: "Does tax enforcement counteract the negative effects of terrorism? A case study of the Basque Country"
2015/10, Montolio, D.; Planells-Struse, S.: "How time shapes crime: the temporal impacts of football matches on crime"
2015/11, Piolatto, A.: "Online booking and information: competition and welfare consequences of review aggregators"
2015/12, Boffa, F.; Pingali, V.; Sala, F.: "Strategic investment in merchant transmission: the impact of capacity utilization rules"
2015/13, Slemrod, J.: "Tax administration and tax systems"
2015/14, Arqué-Castells, P.; Cartaxo, R.M.; García-Quevedo, J.; Mira Godinho, M.: "How inventor royalty shares affect patenting and income in Portugal and Spain"
2015/15, Montolio, D.; Planells-Struse, S.: "Measuring the negative externalities of a private leisure activity: hooligans and pickpockets around the stadium"
2015/16, Batalla-Bejerano, J.; Costa-Campi, M.T.; Trujillo-Baute, E.: "Unexpected consequences of liberalisation: metering, losses, load profiles and cost settlement in Spain's electricity system"
2015/17, Batalla-Bejerano, J.; Trujillo-Baute, E.: "Impacts of intermittent renewable generation on electricity system costs"
2015/18, Costa-Campi, M.T.; Paniagua, J.; Trujillo-Baute, E.: "Are energy market integrations a green light for FDI?"
2015/19, Jofre-Monseny, J.; Sánchez-Vidal, M.; Viladecans-Marsal, E.: "Big plant closures and agglomeration economies"
2015/20, Garcia-López, M.A.; Hémet, C.; Viladecans-Marsal, E.: "How does transportation shape intrametropolitan growth? An answer from the regional express rail"
2015/21, Esteller-Moré, A.; Galmarini, U.; Rizzo, L.: "Fiscal equalization under political pressures"
2015/22, Escardíbul, J.O.; Afcha, S.: "Determinants of doctorate holders' job satisfaction. An analysis by employment sector and type of satisfaction in Spain"
2015/23, Aidt, T.; Asatryan, Z.; Badalyan, L.; Heinemann, F.: "Vote buying or (political) business (cycles) as usual?"
2015/24, Albæk, K.: "A test of the 'lose it or use it' hypothesis in labour markets around the world"
2015/25, Angelucci, C.; Russo, A.: "Petty corruption and citizen feedback"
2015/26, Moriconi, S.; Picard, P.M.; Zanaj, S.: "Commodity taxation and regulatory competition"
2015/27, Brekke, K.R.; Garcia Pires, A.J.; Schindler, D.; Schjelderup, G.: "Capital taxation and imperfect competition: ACE vs. CBIT"
2015/28, Redonda, A.: "Market structure, the functional form of demand and the sensitivity of the vertical reaction function"
2015/29, Ramos, R.; Sanromá, E.; Simón, H.: "An analysis of wage differentials between full-and part-time workers in Spain"
2015/30, Garcia-López, M.A.; Pasidis, I.; Viladecans-Marsal, E.: "Express delivery to the suburbs the effects of transportation in Europe's heterogeneous cities"
2015/31, Torregrosa, S.: "Bypassing progressive taxation: fraud and base erosion in the Spanish income tax (1970-2001)"

2015/32, Choi, H.; Choi, A.: "When one door closes: the impact of the hagwon curfew on the consumption of private tutoring in the republic of Korea"
2015/33, Escardíbul, J.O.; Helmy, N.: "Decentralisation and school autonomy impact on the quality of education: the case of two MENA countries"
2015/34, González-Val, R.; Marcén, M.: "Divorce and the business cycle: a cross-country analysis"
2015/35, Calero, J.; Choi, A.: "The distribution of skills among the European adult population and unemployment: a comparative approach"
2015/36, Mediavilla, M.; Zancajo, A.: "Is there real freedom of school choice? An analysis from Chile"
2015/37, Daniele, G.: "Strike one to educate one hundred: organized crime, political selection and politicians' ability"
2015/38, González-Val, R.; Marcén, M.: "Regional unemployment, marriage, and divorce"
2015/39, Foremny, D.; Jofre-Monseny, J.; Solé-Ollé, A.: "'Hold that ghost': using notches to identify manipulation of population-based grants"
2015/40, Mancebón, M.J.; Ximénez-de-Embún, D.P.; Mediavilla, M.; Gómez-Sancho, J.M.: "Does educational management model matter? New evidence for Spain by a quasiexperimental approach"
2015/41, Daniele, G.; Geys, B.: "Exposing politicians' ties to criminal organizations: the effects of local government dissolutions on electoral outcomes in Southern Italian municipalities"
2015/42, Ooghe, E.: "Wage policies, employment, and redistributive efficiency"

2016

2016/1, Galletta, S.: "Law enforcement, municipal budgets and spillover effects: evidence from a quasi-experiment in Italy"
2016/2, Flatley, L.; Giulietti, M.; Grossi, L.; Trujillo-Baute, E.; Waterson, M.: "Analysing the potential economic value of energy storage"
2016/3, Calero, J.; Murillo Huertas, I.P.; Raymond Bara, J.L.: "Education, age and skills: an analysis using the PIAAC survey"
2016/4, Costa-Campi, M.T.; Daví-Arderius, D.; Trujillo-Baute, E.: "The economic impact of electricity losses"
2016/5, Falck, O.; Heimisch, A.; Wiederhold, S.: "Returns to ICT skills"
2016/6, Halmenschlager, C.; Mantovani, A.: "On the private and social desirability of mixed bundling in complementary markets with cost savings"
2016/7, Choi, A.; Gil, M.; Mediavilla, M.; Valbuena, J.: "Double toil and trouble: grade retention and academic performance"
2016/8, González-Val, R.: "Historical urban growth in Europe (1300–1800)"
2016/9, Guio, J.; Choi, A.; Escardíbul, J.O.: "Labor markets, academic performance and the risk of school dropout: evidence for Spain"
2016/10, Bianchini, S.; Pellegrino, G.; Tamagni, F.: "Innovation strategies and firm growth"
2016/11, Jofre-Monseny, J.; Silva, J.I.; Vázquez-Grenno, J.: "Local labor market effects of public employment"
2016/12, Sanchez-Vidal, M.: "Small shops for sale! The effects of big-box openings on grocery stores"
2016/13, Costa-Campi, M.T.; García-Quevedo, J.; Martínez-Ros, E.: "What are the determinants of investment in environmental R&D?"
2016/14, **García-López, M.A; Hémet, C.; Viladecans-Marsal, E.:** "Next train to the polycentric city: The effect of railroads on subcenter formation"
2016/15, Matas, A.; Raymond, J.L.; Dominguez, A.: "Changes in fuel economy: An analysis of the Spanish car market"
2016/16, Leme, A.; Escardíbul, J.O.: "The effect of a specialized versus a general upper secondary school curriculum on students' performance and inequality. A difference-in-differences cross country comparison"
2016/17, **Scandurra, R.I.; Calero, J.**: "Modelling adult skills in OECD countries"
2016/18, **Fernández-Gutiérrez, M.; Calero, J.**: "Leisure and education: insights from a time-use analysis"
2016/19, Del Rio, P.; Mir-Artigues, P.; Trujillo-Baute, E.: "Analysing the impact of renewable energy regulation on retail electricity prices"
2016/20, Taltavull de la Paz, P.; Juárez, F.; Monllor, P.: "Fuel Poverty: Evidence from housing perspective"
2016/21, Ferraresi, M.; Galmarini, U.; Rizzo, L.; Zanardi, A.: "Switch towards tax centralization in Italy: A wake up for the local political budget cycle"
2016/22, Ferraresi, M.; Migali, G.; Nordi, F.; Rizzo, L.: "Spatial interaction in local expenditures among Italian municipalities: evidence from Italy 2001-2011"
2016/23, Daví-Arderius, D.; Sanin, M.E.; Trujillo-Baute, E.: "CO2 content of electricity losses"

2016/24, Arqué-Castells, P.; Viladecans-Marsal, E.: "Banking the unbanked: Evidence from the Spanish banking expansion plan"
2016/25 Choi, Á.; Gil, M.; Mediavilla, M.; Valbuena, J.: "The evolution of educational inequalities in Spain: Dynamic evidence from repeated cross-sections"
2016/26, Brutti, Z.: "Cities drifting apart: Heterogeneous outcomes of decentralizing public education"
2016/27, Backus, P.; Cubel, M.; Guid, M.; Sánchez-Pages, S.; Lopez Manas, E.: "Gender, competition and performance: evidence from real tournaments"
2016/28, Costa-Campi, M.T.; Duch-Brown, N.; García-Quevedo, J.: "Innovation strategies of energy firms"
2016/29, Daniele, G.; Dipoppa, G.: "Mafia, elections and violence against politicians"
2016/30, Di Cosmo, V.; Malaguzzi Valeri, L.: "Wind, storage, interconnection and the cost of electricity"

2017

2017/1, González Pampillón, N.; Jofre-Monseny, J.; Viladecans-Marsal, E.: "Can urban renewal policies reverse neighborhood ethnic dynamics?"
2017/2, Gómez San Román, T.: "Integration of DERs on power systems: challenges and opportunities"
2017/3, Bianchini, S.; Pellegrino, G.: "Innovation persistence and employment dynamics"
2017/4, Curto-Grau, M.; Solé-Ollé, A.; Sorribas-Navarro, P.: "Does electoral competition curb party favoritism?"
2017/5, Solé-Ollé, A.; Viladecans-Marsal, E.: "Housing booms and busts and local fiscal policy"
2017/6, Esteller, A.; Piolatto, A.; Rablen, M.D.: "Taxing high-income earners: Tax avoidance and mobility"
2017/7, Combes, P.P.; Duranton, G.; Gobillon, L.: "The production function for housing: Evidence from France"
2017/8, Nepal, R.; Cram, L.; Jamasb, T.; Sen, A.: "Small systems, big targets: power sector reforms and renewable energy development in small electricity systems"
2017/9, Carozzi, F.; Repetto, L.: "Distributive politics inside the city? The political economy of Spain's plan E"
2017/10, Neisser, C.: "The elasticity of taxable income: A meta-regression analysis"
2017/11, Baker, E.; Bosetti, V.; Salo, A.: "Finding common ground when experts disagree: robust portfolio decision analysis"
2017/12, Murillo, I.P; Raymond, J.L; Calero, J.: "Efficiency in the transformation of schooling into competences: A cross-country analysis using PIAAC data"
2017/13, **Ferrer-Esteban, G.; Mediavilla, M.:** "The more educated, the more engaged? An analysis of social capital and education"
2017/14, Sanchis-Guarner, R.: "Decomposing the impact of immigration on house prices"
2017/15, Schwab, T.; Todtenhaupt, M.: "Spillover from the haven: Cross-border externalities of patent box regimes within multinational firms"
2017/16, Chacón, M.; Jensen, J.: "The institutional determinants of Southern secession"
2017/17, Gancia, G.; Ponzetto, G.A.M.; Ventura, J.: "Globalization and political structure"
2017/18, González-Val, R.: "City size distribution and space"
2017/19, García-Quevedo, J.; Mas-Verdú, F.; Pellegrino, G.: "What firms don't know can hurt them: Overcoming a lack of information on technology"
2017/20, Costa-Campi, M.T.; García-Quevedo, J.: "Why do manufacturing industries invest in energy R&D?"
2017/21, Costa-Campi, M.T.; García-Quevedo, J.; Trujillo-Baute, E.: "Electricity regulation and economic growth"

2018

2018/1, Boadway, R.; Pestieau, P.: "The tenuous case for an annual wealth tax"
2018/2, Garcia-López, M.À.: "All roads lead to Rome ... and to sprawl? Evidence from European cities"
2018/3, Daniele, G.; Galletta, S.; Geys, B.: "Abandon ship? Party brands and politicians' responses to a political scandal"
2018/4, Cavalcanti, F.; Daniele, G.; Galletta, S.: "Popularity shocks and political selection"
2018/5, Naval, J.; Silva, J. I.; Vázquez-Grenno, J.: "Employment effects of on-the-job human capital acquisition"
2018/6, Agrawal, D. R.; Foremny, D.: "Relocation of the rich: migration in response to top tax rate changes from spanish reforms"
2018/7, García-Quevedo, J.; Kesidou, E.; Martínez-Ros, E.: "Inter-industry differences in organisational eco-innovation: a panel data study"
2018/8, Aastveit, K. A.; Anundsen, A. K.: "Asymmetric effects of monetary policy in regional housing markets"

2018/9, Curci, F.; Masera, F.: "Flight from urban blight: lead poisoning, crime and suburbanization"
2018/10, Grossi, L.; Nan, F.: "The influence of renewables on electricity price forecasting: a robust approach"
2018/11, Fleckinger, P.; Glachant, M.; Tamokoué Kamga, P.-H.: "Energy performance certificates and investments in building energy efficiency: a theoretical analysis"
2018/12, van den Bergh, J. C.J.M.; Angelsen, A.; Baranzini, A.; Botzen, W.J. W.; Carattini, S.; Drews, S.; Dunlop, T.; Galbraith, E.; Gsottbauer, E.; Howarth, R. B.; Padilla, E.; Roca, J.; Schmidt, R.: "Parallel tracks towards a global treaty on carbon pricing"
2018/13, Ayllón, S.; Nollenberger, N.: "The unequal opportunity for skills acquisition during the Great Recession in Europe"
2018/14, Firmino, J.: "Class composition effects and school welfare: evidence from Portugal using panel data"
2018/15, Durán-Cabré, J. M.; Esteller-Moré, A.; Mas-Montserrat, M.; Salvadori, L.: "La brecha fiscal: estudio y aplicación a los impuestos sobre la riqueza"
2018/16, Montolio, D.; Tur-Prats, A.: "Long-lasting social capital and its impact on economic development: the legacy of the commons"
2018/17, Garcia-López, M. À.; Moreno-Monroy, A. I.: "Income segregation in monocentric and polycentric cities: does urban form really matter?"
2018/18, Di Cosmo, V.; Trujillo-Baute, E.: "From forward to spot prices: producers, retailers and loss averse consumers in electricity markets"
2018/19, Brachowicz Quintanilla, N.; Vall Castelló, J.: "Is changing the minimum legal drinking age an effective policy tool?"
2018/20, Nerea Gómez-Fernández, Mauro Mediavilla: "Do information and communication technologies (ICT) improve educational outcomes? Evidence for Spain in PISA 2015"
2018/21, Montolio, D.; Taberner, P. A.: "Gender differences under test pressure and their impact on academic performance: a quasi-experimental design"
2018/22, Rice, C.; Vall Castelló, J.: "Hit where it hurts – healthcare access and intimate partner violence"
2018/23, Ramos, R.; Sanromá, E.; Simón, H.: "Wage differentials by bargaining regime in Spain (2002-2014). An analysis using matched employer-employee data"

2019

2019/1, Mediavilla, M.; Mancebón, M. J.; Gómez-Sancho, J. M.; Pires Jiménez, L.: "Bilingual education and school choice: a case study of public secondary schools in the Spanish region of Madrid"
2019/2, Brutti, Z.; Montolio, D.: "Preventing criminal minds: early education access and adult offending behavior"
2019/3, Montalvo, J. G.; Piolatto, A.; Raya, J.: "Transaction-tax evasion in the housing market"
2019/4, Durán-Cabré, J.M.; Esteller-Moré, A.; Mas-Montserrat, M.: "Behavioural responses to the re)introduction of wealth taxes. Evidence from Spain"
2019/5, Garcia-López, M.A.; Jofre-Monseny, J.; Martínez Mazza, R.; Segú, M.: "Do short-term rental platforms affect housing markets? Evidence from Airbnb in Barcelona"
2019/6, Domínguez, M.; Montolio, D.: "Bolstering community ties as a means of reducing crime"
2019/7, García-Quevedo, J.; Massa-Camps, X.: "Why firms invest (or not) in energy efficiency? A review of the econometric evidence"
2019/8, Gómez-Fernández, N.; Mediavilla, M.: "What are the factors that influence the use of ICT in the classroom by teachers? Evidence from a census survey in Madrid"
2019/9, Arribas-Bel, D.; Garcia-López, M.A.; Viladecans-Marsal, E.: "The long-run redistributive power of the net wealth tax"