Delineating urban areas using building density

Marie-Pierre de Bellefon[*] [†]
INSEE, *Paris School of Economics*

Pierre-Philippe Combes[*][§]
University of Lyon and Sciences Po

Gilles Duranton[*][‡]
University of Pennsylvania

Laurent Gobillon[*][¶]
Paris School of Economics-CNRS

Clément Gorin[¶]
University of Lyon

13 November 2019

ABSTRACT: We develop a new dartboard methodology to delineate urban areas using detailed information about building location, which we implement using a map of all buildings in France. For each pixel, our approach compares actual building density after smoothing to counterfactual smoothed building density computed after randomly redistributing buildings. We define as urban any area with statistically significant excess building density. Within urban areas, extensions to our approach allow us to distinguish 'core' urban pixels and detect centres and subcentres. Finally, we develop novel one- and two-sided tests that provide a statistical basis to compare maps with different delineations, which we use to assess the robustness of our approach and to document large differences between our preferred delineation and the corresponding official one.

Key words: urban area definition, dartboard approach, Jaccard indices

JEL classification: C14, R12, R14

[*] This work is supported by the Zell Lurie Center for Real Estate at the Wharton School. We appreciate the comments from two anonymous referees, the editor Stuart Rosenthal, Miquel-Angel Garcia-Lopez, Tom Holmes, Tomoya Mori, Esteban Rossi-Hansberg, Elisabet Viladecans, and participants at conferences and seminars. We are also grateful to Juan Vargas for helping with satellite data. The views expressed here are those of the authors and not of any institution they may be associated with.

[†] INSEE and Paris School of Economics, 48 Boulevard Jourdan, 75014 Paris, France (email: mariepierre.debellefon@gmail.com).

[§] University of Lyon, CNRS, GATE-LSE UMR 5824, 93 Chemin des Mouilles, 69131 Ecully, France and Sciences Po, Economics Department, 28, Rue des Saints-Pères, 75007 Paris, France (e-mail: ppcombes@gmail.com; website: https://www.gate.cnrs.fr/ppcombes/). Also affiliated with the Centre for Economic Policy Research.

[‡] Wharton School, University of Pennsylvania, 3620 Locust Walk, Philadelphia, PA 19104, USA (email: duranton@wharton.upenn.edu); website: http://real-faculty.wharton.upenn.edu/duranton/. Also affiliated with the National Bureau of Economic Research and the Center for Economic Policy Research.

[¶] PSE-CNRS, 48 Boulevard Jourdan, 75014 Paris, France (e-mail: laurent.gobillon@psemail.eu; website: http://laurent.gobillon.free.fr/). Also affiliated with the Centre for Economic Policy Research and the Institute for the Study of Labor (IZA).

[¶] University of Lyon, CNRS, GATE-LSE UMR 5824, 93 Chemin des Mouilles, 69131 Ecully, France (e-mail: gorin@gate.cnrs.fr; website: https://www.gate.cnrs.fr/spip.php?article818).

1. Introduction

We develop a new dartboard methodology to delineate urban areas using detailed information about building location, which we implement using a map of all buildings in France. For each pixel, our approach compares actual building density after smoothing to counterfactual smoothed building density computed after randomly redistributing buildings. We define as urban any area with statistically significant excess building density. Within urban areas, extensions to our approach allow us to distinguish 'core' urban pixels and detect the centres and subcentres of urban areas. Finally, we develop novel one- and two-sided tests that provide a statistical basis to compare maps with different delineations, which we use to assess the robustness of our approach and to document large differences between our preferred delineation and the corresponding official one.

Delineating urban areas is important for at least two reasons. First, urban research obviously needs to define its object. Extant administrative units such as municipalities do not generally constitute self-contained, functionally autonomous units.[1] Second, inappropriately defined units may lead to a variety of biases. Urban areas that are defined too narrowly or too broadly may fall foul of the modifiable areal unit problem (MAUP) by, for instance, misstating the extent of urban sprawl or by missing important positive or negative spatial spillovers of urban policy interventions.[2]

To delineate urban areas, the first key choice regards what to consider to define functionally integrated units: flows of commuters (or perhaps other flows) or some form of proximity between people or between buildings. Flows of commuters are meant to capture integrated labour markets while morphological approaches that rely on physical proximity or contiguity arguably reflect a broad set of interactions.[3] Although we believe both types of definitions are legitimate, our approach falls into this second category. Some of the innovations we make with our statistical approach are fairly straightforward to implement with a morphological

[1] In many countries, as cities grew, they would directly annex surrounding municipalities. This process of amalgamation has stopped for a variety of reasons, from mayors willing to keep their job to richer municipalities resisting fiscal integration with their poorer neighbours.

[2] Inappropriate definitions may also affect perceptions and consequently policies more broadly. For instance, Latin American countries appear unusually highly urbanised for their level of GDP per capita when using national definitions of what is urban. In turn, this apparent over-urbanisation of Latin America was accepted as fact and fed a long-standing skepticism towards urbanisation and cities on the continent. More systematic and comparable definitions using satellite data show that the 'over-urbanisation of Latin America' is, to a large extent, an artefact of lax definitions that categorise even small villages as urban (Roberts, Blankespoor, Deuskar, and Stewart, 2017).

[3] For definitions that rely on commuting flows, the term 'metropolitan area' may be more appropriate.

approach categorising pixels on a map but would be more difficult to adapt to flows of commuters.

To develop and implement our approach, we face four main challenges. The first is to avoid arbitrary thresholds. Official definitions typically aggregate arbitrarily-defined administrative units using a set of ad hoc rules mandating, among others, pre-defined urban cores, minimum population thresholds, minimum distances between constructions, or minimum shares or numbers of commuters, etc. While the use of thresholds is unavoidable for any approach that seeks to discretise a continuous territory into urban and rural areas, the main decisions that underlie our delineation are grounded either in maximisation criteria or in standard statistical thresholds associated with our dartboard methodology. This is our first innovation.

Our second challenge is to gather appropriate data to implement our approach. We use high-resolution data describing the built environment of an entire country. Data about individual buildings are preferable to population data since (residential) population data primarily describe where people sleep at night. As a result, an approach using population data may fail to classify as urban large central business districts devoid of residents. To avoid counterfactual distributions with buildings in the middle of bodies of water or on the peak of the highest mountains, we also need detailed data to describe the natural environment. Finally, the computation of population for the urban areas that we delineate also requires high-resolution data for population. We gathered these data for France. Although such detailed and comprehensive data are unfortunately not yet available for many countries, in robustness checks we implement our approach using less than ideal but more readily available data, including the builtup footprint of pixels or their residential population.

Any delineation of urban areas should also be consistent with a description of their internal geography. This is our third challenge. To go beyond an urban vs. rural classification of pixels, we propose extensions of our approach that allow us to distinguish highly urbanised, or ‘core’, urban pixels from regular urban pixels. Within urban areas, these extensions to our approach also allow us to isolate centres, subcentres, and their surrounding areas.

Our final challenge is to provide a statistically-grounded approach to compare different delineations, such as delineations generated by different variants of our approach or our preferred delineation and official ones. Whether two settlements form a single unified urban area or two separate ones may depend on a few ‘joining’ pixels which may be close to the threshold of being urban. This type of situation may sometimes lead to the delineation of a single urban area and sometimes to the delineation of two. It is desirable to assess how much of the differences between two delineations is due to sampling instead of true methodological differences. Being able to assess the importance of sampling is our last main innovation.

Our work contributes to the literature that seeks to define urban areas, and more generally any form of spatial units. A long-standing concern in the literature has been to provide a rigourous definition of urban or metropolitan areas, first relying on a notion of central places (Berry, 1960, Fox and Kumar, 1965), then integrated local labour markets (Berry, Lobley, Goheen, and Goldstein, 1969, Kanemoto and Kurima, 2005, Duranton, 2015), contiguous development (Rozenfeld, Rybski, Gabaix, and Makse, 2011), or various forms of spatial interactions measured, in particular, with land prices (Bode, 2008, Corvers, Hensen, and Bongaerts, 2009). A second concern in the literature has been to develop robust approaches with minimal data requirements so that urban areas can be delineated in a comparable manner over several countries (Hall and Hay, 1980, Cheshire and Hay, 1989). Within spatial units, there is also a recurring interest in the formal definition of urban centres and counts of subcentres (see McDonald, 1987, Giuliano and Small, 1991, McMillen, 2001, and related litterature).

There has been a renewed interest in delineating urban areas in the recent past. Concerns about urbanisation and cities in policy and development circles (e.g., Asian Development Bank, 2019, CAF Development Bank of Latin America, 2017, Ferreyra and Roberts, 2018, for the World Bank) have led to a number of attempts to delineate urban areas for comparative purpose using night-time lights from satellite data (Ch, Martin, and Vargas, 2018, Davis, Dingel, and Miscio, 2020), a combination of night- and day-time lights (Baragwanath-Vogel, Goldblatt, Hanson, and Khandelwal, 2020) or gridded population data (Dijkstra, Florczyk, Freire, Kemper, and Pesaresi, 2019, Henderson, Kriticos, and Nigmatulina, 2020, Moreno-Monroy, Schiavina, and Veneri, 2019). Comparability across countries imposes some limitations to the methodology being adopted and the data being used. We can label these approaches as 'wide but shallow'.

New sources of data, sometimes unique to particular countries, have given instead some impetus for 'deep but narrow' approaches. Like us, Arribas-Bel, Garcia-Lopez, and Viladecans-Marsal (2019) exploit a detailed map of all buildings for Spain. They develop a clustering algorithm to classify cities. A different approach is taken by Galdo, Li, and Rama (2019) who use a variety of data sources combined with human judgement for a small subsample of locations in India. Human judgement is then mechanically replicated for the whole of India. Bosker, Roberts, and Park (2018) propose another type of 'deep but narrow' approach. They use commuting data together with many other sources to look at differences in the delineation of urban areas for Indonesia across a broad variety of approaches. In the spirit of Briant, Combes, and Lafourcade (2010), they also explore the implications of different delineations for the estimation of a number of urban relationships. While our main approach belongs to this second group, we also propose a simplified version of what we do that can be implemented using widely available gridded data.

Our work is also related to a large literature in spatial statistics that relies on dartboard counterfactuals. Much of that work is concerned with detecting spatial concentration from the distribution of distances between its objects of interests, such as establishments within the same industry (Duranton and Overman, 2005). Unfortunately, we cannot adapt this type of approach to buildings since it would only tell us whether buildings are spatially concentrated (and at which spatial scale) but not whether a specific group of buildings forms an urban area. There is a literature that attempts to detect clusters of particular sectors of economic activity in adjacent areas. See Mori, Nishikimi, and Smith (2014) for a recent development. A key endeavour in this literature is to isolate a single or multiple clusters by grouping contiguous discrete regions. Our approach uses instead 'arbitrarily small' spatial units and relies on detecting excess smoothed density. While more demanding in terms of data, this allows us to treat geographic space as a quasi-continuum and bypass the difficult computations associated with finding the best cluster of regions or the best set of clusters. Billings and Johnson (2012) propose an approach closer in spirit to ours but they use it to assess industrial specialisation instead of clusters.

2. Data

Our main source of data is the 2014 BD TOPO from the French Geographical Institute (IGN). This is a three-dimensional vectorial representation of the French territory with a one-metre precision containing information on all buildings, including their footprint, height, and use. This dataset is a key component of the large-scale geographical reference for the country. It integrates a variety of pre-existing sources from IGN, satellite images, and the French cadastral information.

Table 1 reports some descriptive statistics for the 33,400,921 buildings in 'mainland' France, which includes a number of small nearby islands but not Corsica nor overseas territories. Unsurprisingly, there is much variation around the mean footprint of 153 m^2 and the mean

Table 1: Descriptive statistics on buildings

	Min.	25th pctl	Med.	Mean	75th pctl	95th pctl	99th pctl	Max.	St. dev.
Surface (m^2)	0.2	44	93	153	152	421	1,224	579,352	509
Volume (m^3)	0.4	186	466	1,054	848	3,044	10,870	14,483,811	7,336

Notes: Authors' calculations for 33,400,921 buildings from BD TOPO. We eliminated 5 buildings with zero footprint in the data.

Table 2: Descriptive statistics for pixel building density (volume and footprint)

Built area	Min.	25th pctl	Med.	Mean	75th pctl	95th pctl	99th pctl	Max.	St. dev.
Raw									
Surface (m^2/pixel)	0	0	0	412	29	2,380	7,211	582,501	1,551
Share Built (%)	0	0	0	1.03	0.07	5.95	18.03	1,456.25	3.88
Volume (m^3/pixel)	0	0	0	2,839	99	13,580	51,723	14,511,414	16,877
Smoothed									
Surface (m^2/pixel)	0	95	192	397	363	1,405	4,211	22,571	806
Share Built (%)	0	0.24	0.38	0.99	0.91	3.51	10.53	56.43	2.02
Volume (m^3/pixel)	0	544	1,106	2,725	2,159	9,225	33,073	443,669	8,061

Notes: Authors' calculations from BD TOPO using 12,403,734 pixels. To keep buildings lumpy, we allocate each building to the pixel that includes the largest share of its area. In extremely rare cases, this leads to a builtup footprint density that exceeds one for a pixel.

volume of 1,054 m^3 per building. The largest building in France is the Peugeot car assembly line near Sochaux, which is several kilometre long, has footprint of nearly 0.6 km^2, and an average height of 25 metres. Overall, the footprint of all buildings in France represents 0.93% of the area of mainland France with an average height of 6.90 metres, which corresponds to about two stories.

Our approach requires the rasterisation of the information about actual buildings to work with pixels. To keep the implementation computationally manageable, we divide the French territory into pixels of 200 metres by 200 metres, which we designed to match those used by the French national statistical institute (INSEE).[4] We then compute the 'building density' of each pixel. For our baseline approach, we use the volume of builtup space in each pixel to measure building density. In a variant, we also use the footprint of all buildings in each pixel.

Table 2 reports descriptive statistics about building density. We note that 74% of pixels are unbuilt. Even at the 95th percentile of the distribution of pixel building footprints, only 5.95%

[4] Pixel sizes are approximate because of tiny variations arising from the curvature of the earth. We also note that INSEE only considers pixels with positive population whereas our grid is complete.

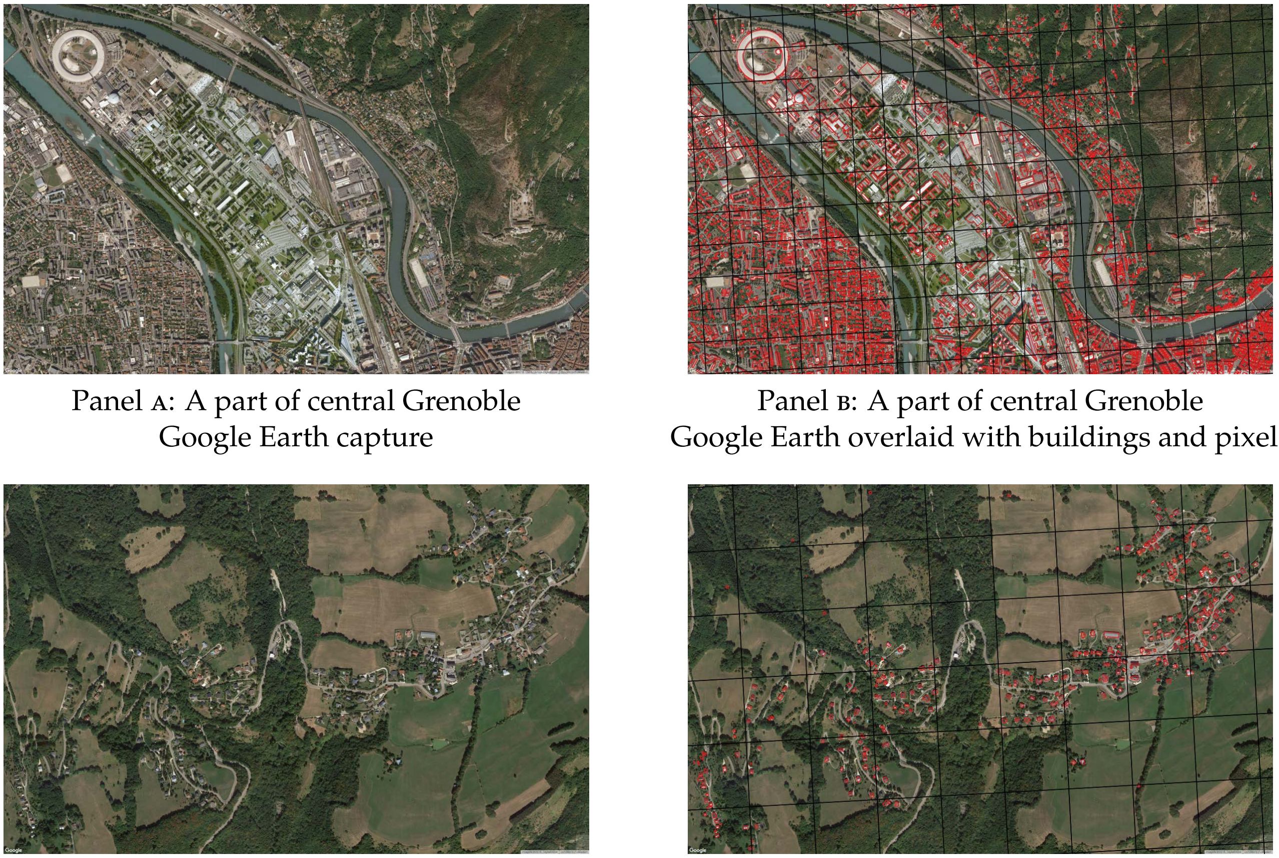
of a pixel is built up. It is only at the far-right tail of the distribution that we observe intensely built pixels. At the 99^{th} percentile, a pixel is 18.0% built up. More generally, the distribution of buildings across pixels is highly skewed with a Gini coefficient of 0.927 for builtup volume and 0.909 for builtup area.

We illustrate our data work with the city of Grenoble for reasons that will become clear below. Panel A of figure 1 shows a Google Earth capture of a section of central Grenoble, which centres on its 'scientific polygon' located at the confluence of the Isère and Drac rivers. The large round building at the northwestern end is the European Synchrotron Radiation Facility, a particle accelerator. In panel B, we overlay the picture of panel A with the BD TOPO data for buildings and pixel boundaries. We note from this panel that the overlap of buildings between BD TOPO and Google Earth is near perfect.[5] Panels C and D of the same figure repeat the same exercise for a rural area on the outskirts of Grenoble. Again the building overlap is extremely good. The exceptions are some isolated buildings in the BD TOPO which do not appear in the Google Earth capture. As it turns out, these buildings exist but are hidden under the canopy.

Our approach involves randomly redistributing buildings across pixels. Some pixels are difficult or impossible to build upon because they are covered by a body of water, have an extremely steep slope, or have a high elevation. Information about bodies of water can be retrieved from the BD CARTHAGE. Data about elevation is obtained from BD ALTI. From this last source, we can also compute a measure of mean slope for each pixel. See Appendix A for further details.

Among pixels which contain at least one building, we determine the 99^{th} percentile for the share of the pixel covered by water (42.4%), the elevation (1,213 metres), and the average slope (21.0%). We then consider all pixels with either a proportion of water or an elevation or a slope above the 99^{th} percentile to be non-buildable.[6] Figure 2 represents non-buildable

Figure 1: BD TOPO: Illustrations for Grenoble



Panel A: A part of central Grenoble
Google Earth capture

Panel B: A part of central Grenoble
Google Earth overlaid with buildings and pixels

Panel C: Rural area near Grenoble
Google Earth capture

Panel D: Rural area near Grenoble
Google Earth overlaid with buildings and pixels

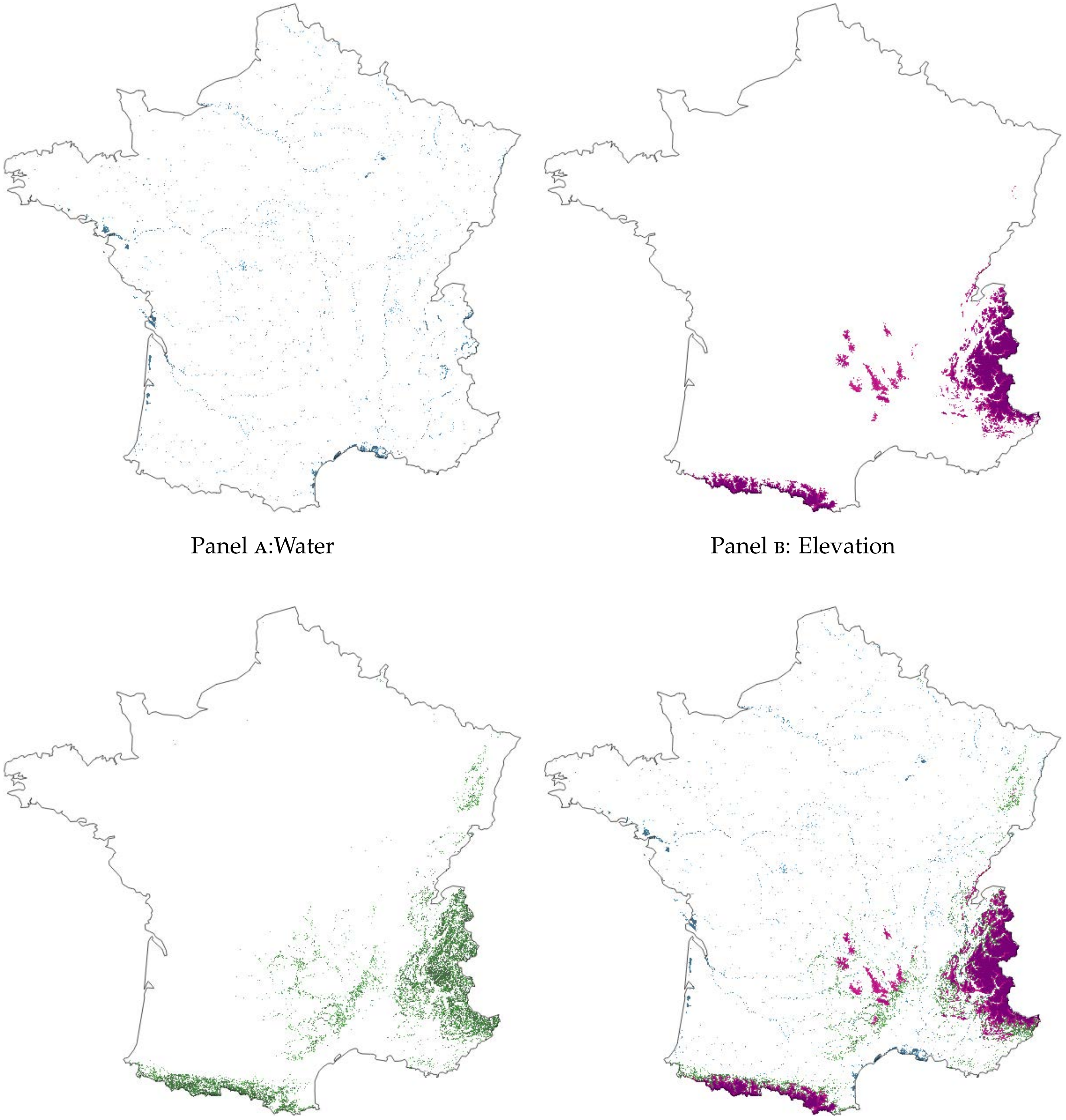
pixels according to these criteria taken separately or together.

To illustrate our treatment of the data, we return to the Grenoble region in figure 3. Panel A overlays the data about buildings from BD TOPO for Greater Grenoble on top of a Google Earth capture. Panel B represents our final data. The map shows both individual buildings and building density of buildable pixels. It also shows empty buildable pixels and non-buildable pixels covered by water, with steep slopes, or with high elevation. We chose Grenoble for our illustration because it is the only city in France with population above

[5] Most grey areas not highlighted in red in the peninsula are parking lots. A careful inspection reveals one building close to the rail tracks that does not appear in the BD TOPO data. This building was torn down. A Google Earth update posterior to the production of figure 1 shows this area as a construction site while Google Streetviews, updated even more recently, shows some new constructions (as of June 2018).

[6] We do not consider maxima as they impose little to no restrictions due to a small number of exceptional cases such as high altitude observatories or tiny islands that were built up to host a jail or defense facilities. Overall, this led us to discard 8.2% of all pixels and we end up with 12,403,734 buildable pixels. While high-elevation and steep-sloped pixels are unsurprisingly concentrated around the Alps and the Pyrenees, pixels covered with water are more evenly spread out but nonetheless follow expected patterns and underscore major rivers and lakes.

Figure 2: Non-buildable areas



Panel A:Water

Panel B: Elevation

Panel C: Slope

Panel D: Water, elevation and slope

Notes: Maps produced using BD CARTHAGE (water) and BD ALTI (elevation and slope).

half a million surrounded by mountains and thus all three types of non-buildable pixels are well-represented.

Finally, we use geolocalised population data from INSEE originally collected for fiscal purposes. These data are readily available for the pixels we use.

3. Delineating urban areas: methodology

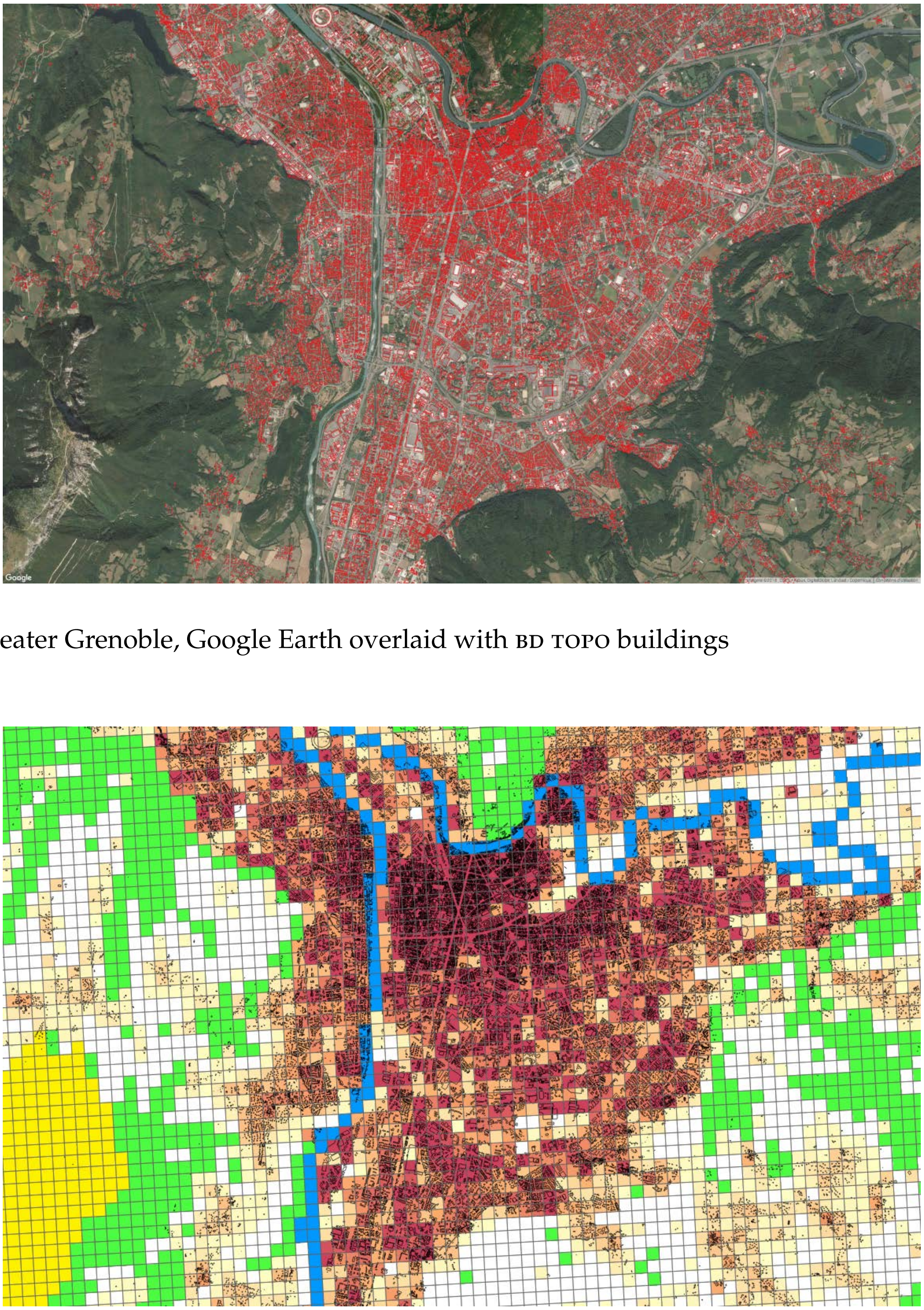
Our analysis is conceptually simple. We first compute building density for each pixel as described above from the volume of all buildings attached to the pixel. The second step is to smooth building densities across pixels using a kernel. In the third step, we then generate counterfactual building densities by randomly redistributing buildings across buildable pixels and smooth these counterfactual building densities just like we smooth the actual density for a given delineation. In the fourth step, we consider that a pixel is urban if its actual smoothed density is above the 95^{th} percentile of the distribution of counterfactual smoothed densities computed for that pixel. Urban areas are finally defined as sets of contiguous urban pixels. We name these urban areas after the most populated municipality they overlap with. We return to this last issue in further details below.

As it turns out, this process delineates several thousand urban areas. To avoid having to limit this large set using an arbitrary population threshold, we also define highly urbanised areas that we refer to urban cores. To do this, we repeat our analysis and randomly redistribute all urban buildings *across all urban pixels*. We then consider that a pixel is part of an urban core if its smoothed density is above the 95^{th} percentile of the distribution of smoothed densities computed from this second set of counterfactuals. This procedure allows us to distinguish between urban areas that contain one or more cores from those that do not have one.

Finally, we detect the centre and subcentres of urban areas. To achieve this, we repeat again our analysis and randomly redistribute each and every urban buildings *within their own urban area*. We then select all pixels with a smoothed density above the 95^{th} percentile of the distribution of counterfactual smoothed densities in their urban areas. In each urban area, the largest set (in terms of builtup volume) of contiguous selected pixels defines the centre (or central area) while other sets of contiguous pixels define subcentres.

We now describe some of these steps in greater details.

Figure 3: Non-buildable areas and data treatment: Illustrations for Grenoble



Panel A: Greater Grenoble, Google Earth overlaid with BD TOPO buildings

Panel B: Greater Grenoble, buildings (in black), builtup densities (shades of orange), buildable but unbuilt areas (in white) and non-buildable areas (rivers in blue, steep slopes in green, and high elevations in yellow).

Smoothing building density

After computing building density for each pixel directly from the data as described in section 2, we smooth this density across pixels. Smoothed building density for pixel j with coordinates (x_j,y_j) is given by:

$$\widehat{z}_j = \frac{1}{\sum_i K_h(d_{ij})} \sum_i K_h(d_{ij}) z_i \, , \tag{1}$$

where z_i is the building density for pixel i, $d_{ij} = \sqrt{\left(x_j - x_i\right)^2 + \left(y_j - y_i\right)^2}$ is the distance between pixels i and j, and K_h is a kernel with bandwidth h.

We need to choose a type of kernel and a bandwidth h. Our choice of kernel is dictated by computational reasons. To avoid using too many (all) pixels for smoothing, we choose a bisquare kernel verifying:

$$K_h\left(d_{ij}\right) = \left[1 - \left(\frac{d_{ij}}{h}\right)^2\right]^2 1\{d_{ij} < h\} \, ,$$

such that weights are zero after a given distance h. This choice of kernel allows us to drastically reduce the size of our computations since the building density of each pixel is smoothed over hundreds of other pixels instead of millions.

For the choice of bandwidth, we face the following tradeoff. Taking a large bandwidth will lead to over-smoothed data, which in turn will make it difficult to identify differences between more or less intensely builtup areas. Taking a small bandwidth will instead lead to under-smoothed data, which will in turn make it difficult to define homogeneous areas.

To decide on a bandwidth, we use the following generalised cross-validation criterion. We first compute the smoothed building density of each pixel *net of its own contribution*. That is, we amend the computation described by equation (1) to exclude the pixel at hand from the summation. For each bandwidth h, we measure the fit between actual building density and smoothed building density (excluding own pixel contribution) using a pseudo-R^2. Because of our focus on urban areas, each pixel receives a weight proportional to its builtup volume. We then determine the optimal bandwidth so that it maximises the fit between actual and smoothed density using the cross-validation technique of Bowman and Azzalini (1997) implemented with the `sm` R package.

We note the following issue about the choice of bandwidth. In areas with very few buildings, it is best to fully smooth them out to predict a landscape mostly devoid of buildings on most pixels. In these areas, which represent a large part of the country, the optimal bandwidth can be large and unsuitable for more densely builtup areas. To avoid taking an overly large bandwidth that optimally fits mostly rural areas, one can instead divide the country into smaller tiles of, for instance, 100 kilometres by 100 kilometres and compute the optimal bandwidth for each. Then, we can take the median optimal bandwidth across all tiles as our preferred value.

As the literature acknowledges (e.g., Bowman and Azzalini, 1997), the optimal bandwidth can vary a lot across tiles of identical size. For 100-kilometre tiles, the optimal bandwidth varies between 0.36 and 12.6 kilometres. Consistent with our conjecture above, larger optimal bandwidths are found for tiles in less intensely built-up areas of the country. For the same reason, there is even more variation in optimal bandwidth for smaller tiles. This said, there is less variation across tile sizes for the median bandwidth. It is for instance close to 0.9 kilometre for 50-kilometre tiles, 1.7 kilometres for 100-kilometre tiles, and 3.3 kilometres for 200-kilometre tiles. We use a bandwidth of 2 kilometres as our default but report below robustness checks using 1 and 3 kilometres, respectively. As we show below, the choice of bandwidth for the range we consider mostly affects the smaller urban areas.

Smoothing obviously reduces the skew of the distribution of building densities. As reported in table 2, after smoothing only 0.1% of buildable pixels are 'unbuilt' instead of 74% for raw density. Conversely, the pixel at the 99^{th} percentile of the smoothed density is 10.5% builtup instead of 18.0% in raw data. As a result of smoothing, the Gini coefficient is 0.679 for smoothed builtup density in volume instead of 0.927 for raw builtup density.

Counterfactual building densities

Next, we generate counterfactual building densities for the entire country. To do this, we randomly redistribute all existing buildings across all buildable pixels with equal probability. This redistribution of all buildings without replacement is equivalent to a full 'reshuffling of the deck'. For each counterfactual distribution of buildings and for each pixel, we compute its counterfactual building density just like we computed its actual building density above. We then smooth each counterfactual building density across pixels like we smoothed actual building density. For each pixel and each counterfactual distribution, we thus end up with a smoothed counterfactual building density.

For our baseline delineation, we perform this procedure 10,000 times. As we show below, sampling errors barely affect our results. For variants of our baseline and robustness checks, we only perform 100 counterfactuals to limit computations.

Table 3: Descriptive statistics for the 95th percentile of smoothed building density

Q95 distribution	Min.	25th pctl	Med.	Mean	75th pctl	95th pctl	99th pctl	Max.	St. dev.
Surface (m^2/pixel)	8	491	508	503	527	558	586	768	47
Share Built (%)	0.02	1.23	1.27	1.26	1.32	1.39	1.46	1.92	0.12
Volume (m^3/pixel)	63	3,652	3,853	3,854	4,084	4,508	4,907	8,454	442

Notes: Authors' calculations from BD TOPO from 13,628,277 pixels. This table reports various moments of the distribution of pixels for the 95th percentile of counterfactual building densities.

Detecting excess building density

For each pixel, we can now measure its actual smoothed building density relative to its distribution of counterfactual smoothed building densities. We call 'urban', a pixel for which the actual smoothed building density is above the 95th percentile of *its* distribution of counterfactual smoothed building densities.[7] We refer to the other pixels as 'rural'. Finally, we define an 'urban area' as a set of contiguous urban pixels.[8]

We note that we compute a different distribution of counterfactual building densities for each pixel. If all pixels were buildable and in absence of geographic irregularities, we would be able to compute the distribution of counterfactual building densities for a single representative pixel and use this distribution for all pixels. With buildings of the same size, we could even use a normal approximation and apply a simple formula to compute the density threshold for a pixel to be defined as urban.[9] However, the presence of non-buildable pixels, the irregular geography of the country, and heterogeneity in building sizes make such shortcuts problematic. Table 3 documents how the 95th percentile of the counterfactual distribution of building densities varies. These are 'percentiles of the 95th percentile', equal for instance to 3,652 m^3 of builtup volume for the 25th percentile and 4,084 m^3 for the 75th percentile.

[7] We use the 95th percentile for our baseline results but we also consider alternative thresholds at the 75th and 99th percentiles in supplementary results.

[8] We ensure that no urban area is divided because of a river.

[9] We can think of building density as the outcome of a binomial distribution which we can approximate by a normal distribution given the large number of draws. However, such normal approximation is unlikely to work well in any case given the skew in the distribution of building sizes and the fact that the probability of receiving any given building is equal to the inverse number of pixels and is thus close to zero. For the smoothed distribution of buildings, this formula will be fairly involved since it needs to account for smoothing across pixels.

Although modest, this variation should not be ignored.[10] Thus, we work with pixel-specific percentiles computed from a full set of counterfactual distributions of buildings. Below, we assess how using a common threshold for a 'representative' pixel affects our results.

Finally, we want to measure how much sampling matters in our approach and provide a statistical basis to the comparison of our baseline delineation with some variants or with alternative delineations, such as the official delineation from INSEE. To achieve this, we can use the 10,000 counterfactual distributions drawn for our baseline delineation as 100 sets of 100 counterfactual distributions. We can then generate a different delineation from each set of 100 replications and measure how these delineations vary relative to a median delineation.[11]

Urban cores

Before turning to our results, we note the following. The approach described so far will lead to the delineation of a large number of urban areas, ranging from major metropolitan areas hosting a million or more buildings to villages with no more than a few hundred buildings. This result simply reflects the fact that most buildings are much closer to each other than a random assignment would predict. Recall that for the 'median' pixel in table 3 it only takes a smoothed building density of 9.6%, corresponding to a builtup footprint of 508 m^2, or a building volume of 3,853 m^3 for a pixel to qualify as urban. The latter threshold corresponds to slightly less than 0.1 m^3 of builtup volume for a squared-meter of land.

While we think it is useful to delineate all statistically significant peaks of building density and be able to study them, for many applications ranging from the analysis of the scarcity of land for housing to the agglomeration of production establishments in the same location(s), we would like to focus on larger urban areas and more densely builtup pixels.

To do this, a first possibility would be to raise the significance level to define what is urban above 95%. Unfortunately, this approach still generates many small urban areas. Another possibility would be to add a minimum size threshold for the number of buildings, the

[10] In part, this variation in the thresholds across pixels arises from our sampling of a finite number of counterfactuals. However, mapping this variation shows that it mainly reflects the uneven geography of buildable pixels in France. The 95^{th} percentile of the distribution of counterfactual building densities is lower for pixels that are surrounded by non-buildable pixels from which they receive nothing from the smoothing of counterfactual distributions. This 95^{th} percentile is even equal to zero for non-buildable pixels for which the distance to the nearest buildable pixels is more than the smoothing bandwidth and thus can never receive a strictly positive building density.

[11] For this 'median' delineation, a pixel is classified as urban when it is urban in more than 50% of the 100 delineations.

builtup volume, or the population to qualify as urban area. This would be arbitrary. Instead, we propose the following approach. After applying the methodology described above and classifying all pixels in the country into urban and rural, we discard rural pixels and their buildings. We then repeat the same dartboard approach as previously and generate one hundred counterfactual redistributions *of all buildings located in urban pixels across all urban pixels*. After smoothing as previously, we say that a given pixel is part of an 'urban core' if its observed density is above the 95^{th} percentile of the distribution of counterfactual densities computed for that pixel. We can then focus on urban areas that contain at least an urban core. As we argue below, urban cores are also of independent interest. Note that we can repeat the same procedure again to isolate urban cores of higher order.

Centres and subcentres

An important advantage of our methodology is that it can readily be extended to analyse the internal geography of urban areas. We propose the following approach. In each urban area (as previously delineated), we generate one hundred counterfactual redistributions of the buildings of this urban area across all its pixels. After smoothing as previously, we select the pixels whose observed builtup density is above the 95^{th} percentile of the distribution of counterfactual builtup densities computed for that pixel. The largest contiguous set of selected pixels in each urban area defines its central area or centre. The other sets of contiguous selected pixels form subcentres.

The key difference between cores and centres or subcentres is that cores are defined relatively to all urban pixels in the country whereas centres and subcentres areas are defined relatively to their own urban areas. As it turns out, cores are overwhelmingly found in larger urban areas whereas small urban areas often have a statistically significant centre.

4. An anatomy of French urban areas

General results

We now describe more fully the output of our baseline delineation approach, which defines 5,771 urban areas representing 12.9% of all pixels (or 14.1% of all buildable pixels) in mainland France. Total urban area population is 50,417,382 or 80.5% of the population of mainland

Table 4: Descriptive statistics for urban areas, baseline delineation

Pixel distribution	Min.	25th pctl	Med.	Mean	75th pctl	95th pctl	99th pctl
Panel A: All urban areas (5,771)							
Population	0	437	1,018	8,736	2,446	14,319	11,325,097
Area	0.04	2	4	12	8	32	3,618
Population density	0	196	305	358	453	821	3,431
Panel B: Urban areas with at least an urban core (812)							
Population	0	4,578	8,815	54,023	22,092	158,511	11,325,097
Area	0.12	13	21	58	41	180	3,618
Population density	0	307	434	486	601	1,018	3,130
Panel C: Urban areas without no urban core (4,959)							
Population	0	380	840	1,321	1,658	4,193	24,630
Area	0.04	1	3	4	6	12	98
Population density	0	184	287	337	423	787	3,431
Panel D: INSEE urban units (2,216)							
Population	622	3,240	4,784	21,743	8,897	55,227	10,613,004
Area	2	20	34	56	59	156	2,864
Population density	7	108	171	233	277	630	3,706

Notes: Population is from 2015; area in km^2; population density is the number of inhabitants per km^2.

France. Urban pixels host 69.8% of all buildings. Tables 6 and 7 in Appendix B provide a summaries of our main results.

Descriptive statistics about the size of the urban areas obtained in the baseline delineation are reported in panel A of table 4. Because our approach defines a large number of urban areas, they tend to be small in their large majority. Population is 437 at the first quartile, while area is 2 km^2. At the 95th percentile of the distribution of urban areas, population is still only 10,319.[12] While our approach classifies only a small minority of pixels as urban, it also appears that it only takes a modest concentration of buildings to generate statistically significant excess builtup density.

Panels B and C of table 4 report descriptive statistics similar to panel A but distinguish between urban areas with a core (or several) and those without. The contrast between the two groups of urban areas is striking. There are 812 urban areas with at least an urban core vs. 4,959 urban areas without. Urban areas with a core have a much higher population. They

[12] Our approach delineates a small number of urban areas without residents. These are mainly isolated airports and nuclear power plants. Although devoid of residents who call these 'urban areas' home during the night, some of these buildings or groups of buildings host a large population during the day.

host on average 54,023 inhabitants instead of 1,321 for urban areas without a core. Overall, urban areas with a core host 70.1% of the French population and occupy 8.8% of the French territory, while urban areas without a core account for 4.0% of the French territory and 10.4% of the French population. Our distinction between urban areas with and without a core does a particularly good job to distinguish the upper tail of the distribution of French urban areas from the rest.

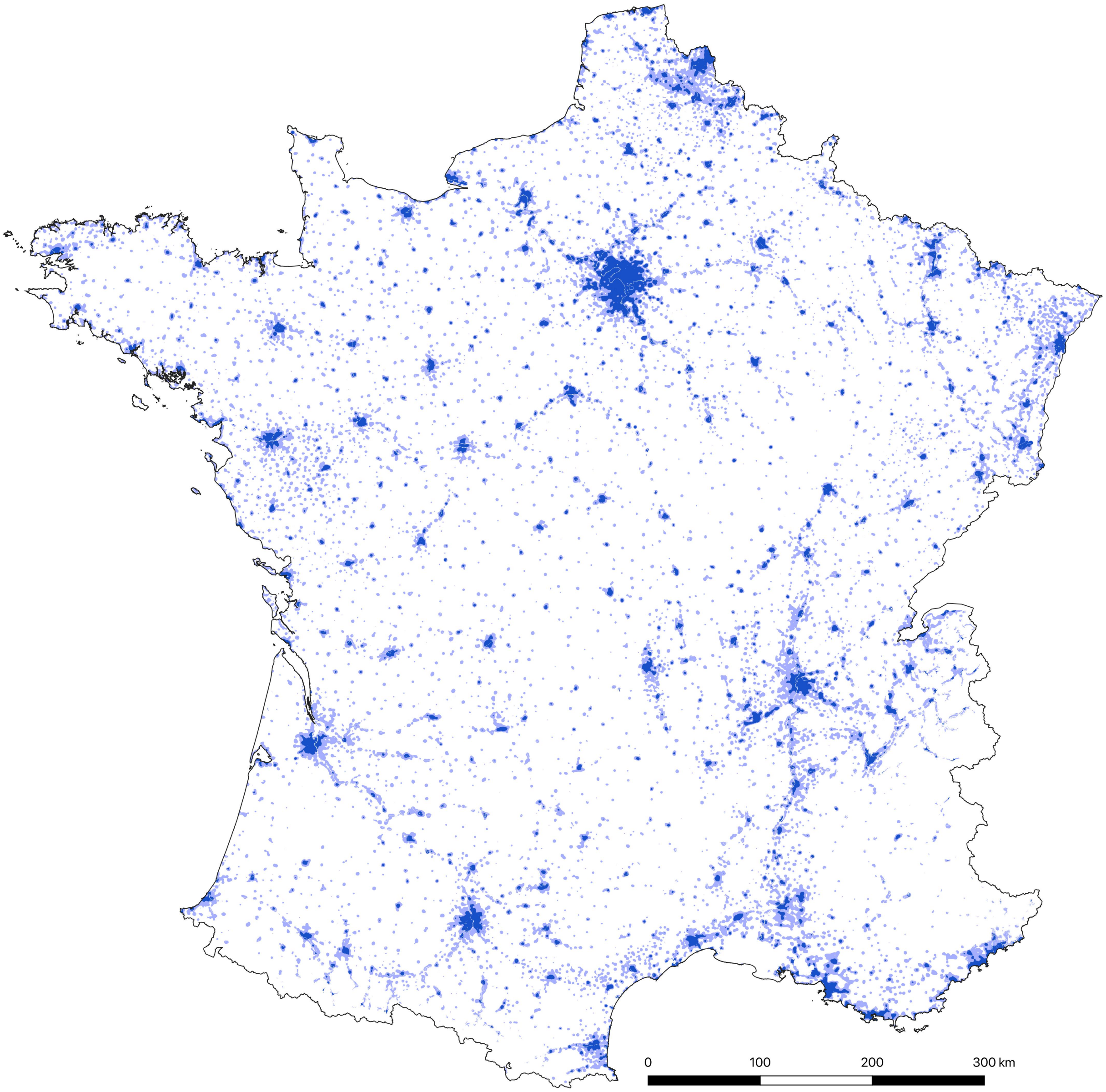
Urban cores represent 18.8% of all urban pixels or 2.4% of pixels in mainland France. This small fraction of pixels nonetheless hosts 51.6% of the population of mainland France. That more than half of the French population and 40.2% of all buildings are located on less than 3% of the French territory is revealing of a high level of spatial concentration. If we repeat our approach to detect second-order urban cores (statistically significant cores among cores), we isolate 0.5% of the area of the country but still 25% of its population. After another iteration, we find that 10% of the French population lives on the 0.1% of pixels that form 99 third-order urban cores. Eventually, after 8 iterations, we are left with only a part of central Paris as the most 'urban' part of the country.

Figure 4 is a map of the urban areas delineated by our approach. Unsurprisingly, the largest French cities are all clearly apparent. We also observe a high density of urban pixels along the coasts and major rivers. The third main feature of this map is that urban pixels are much less prevalent in the mountainous areas of the country.

A close-up on four urban areas

The four panels of figure 5 represent close-ups on the regions of Paris, Lille, Marseille, and Grenoble. Paris is the largest urban area in France. Lille and Marseille are also among the largest four. Grenoble is a smaller city, which we used above to illustrate our treatment of the data. These four cities also differ in interesting ways for our purpose. Starting with Paris in panel A, the urban area of Paris looks highly monocentric and centred on the municipality of Paris. The urban area branches out in four directions following the river Seine and its two main tributaries, Oise and Marne. There are also many small urban areas that surround the urban area of Paris. We finally note that the urban cores of Paris cover a large majority of the urban area. Not only is building density in the urban area of Paris significantly higher than that in the rest of the country but it is also significantly higher than that in the rest of urban

Figure 4: Urban areas in France

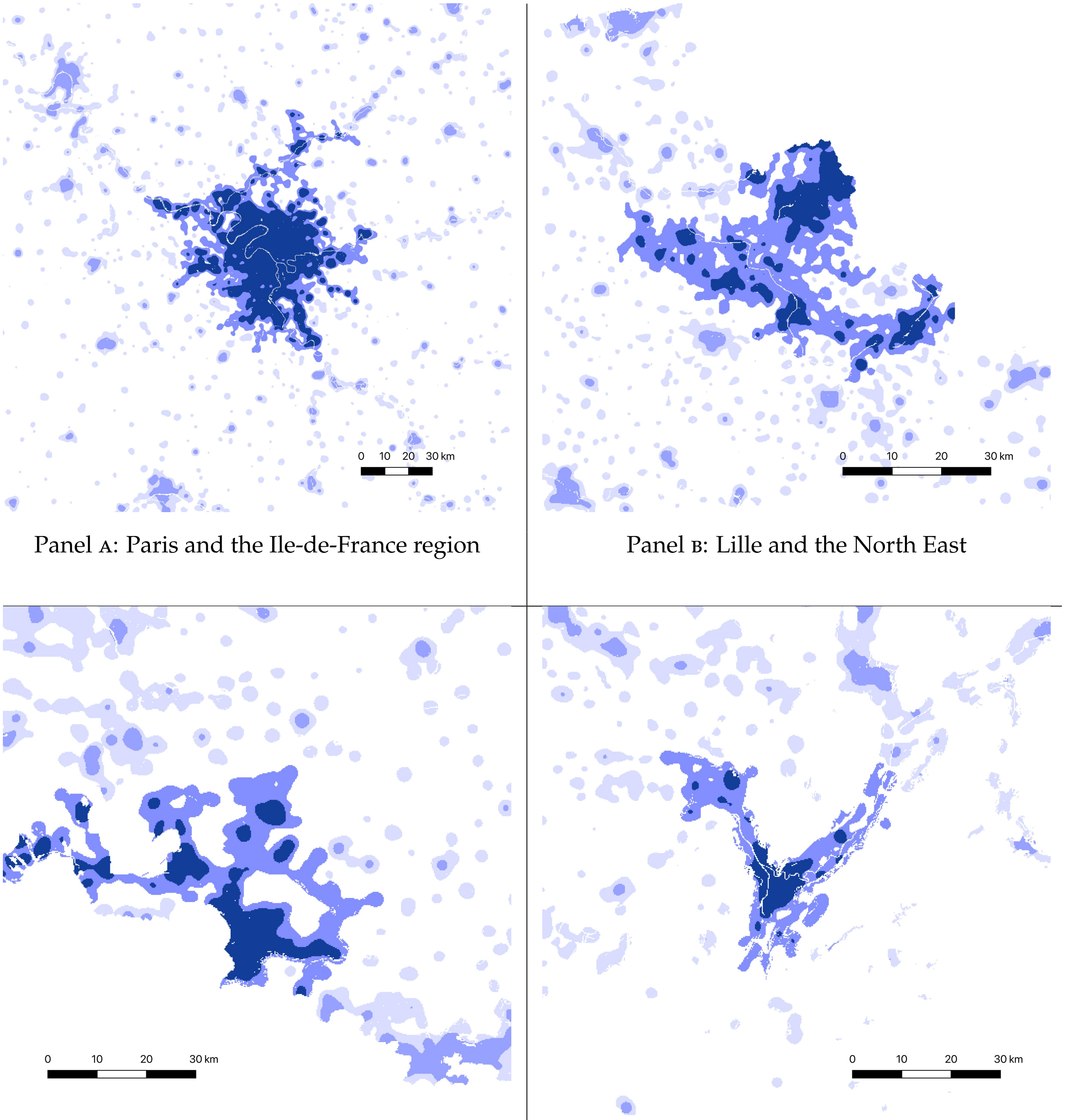


Notes: Urban areas in medium blue (medium grey). Urban cores in dark blue (dark grey).

France. This feature is also true for all large French urban areas.

The Lille urban area, represented in panel B of figure 5, is morphologically extremely different from Paris. It aggregates several large municipalities including Lille itself, Roubaix,

Figure 5: Urban areas in four regions



Panel A: Paris and the Ile-de-France region

Panel B: Lille and the North East

Panel C: Marseille and the South East

Panel D: Grenoble and the Alpine region

Notes: Paris, Lille, Marseille and Grenoble urban areas in medium blue (medium grey) and their urban cores in dark blue (dark grey). Other urban areas in very light blue (very light grey) and their urban cores in light blue (light grey).

Tourcoing, Douai, Lens, Valenciennes or Arras (not to mention a Belgian part for which we do not have data). These municipalities are tightly integrated and for many of them also belong to the same contiguous core.

The Marseille region in panel C is a good example of a difficult natural geography with the Mediterranean Sea to the South, a large lagoon to its west and a number of small mountains immediately west, north and east of the city. The core areas are centered around Marseille itself, Vitrolles next to the Berre lagoon, and Aix-en-Provence. Several relatively large distinct urban areas exist in the same region such as Avignon to the north or Toulon to the east. Finally, Grenoble in panel D is much more compact as it is surrounded by high mountains. The urban area of Grenoble is also Y-shaped by its two rivers, the Isère and the Drac. We conclude that despite extremely different geographies and underlying morphologies, our approach is able to robustly isolate large urban areas.

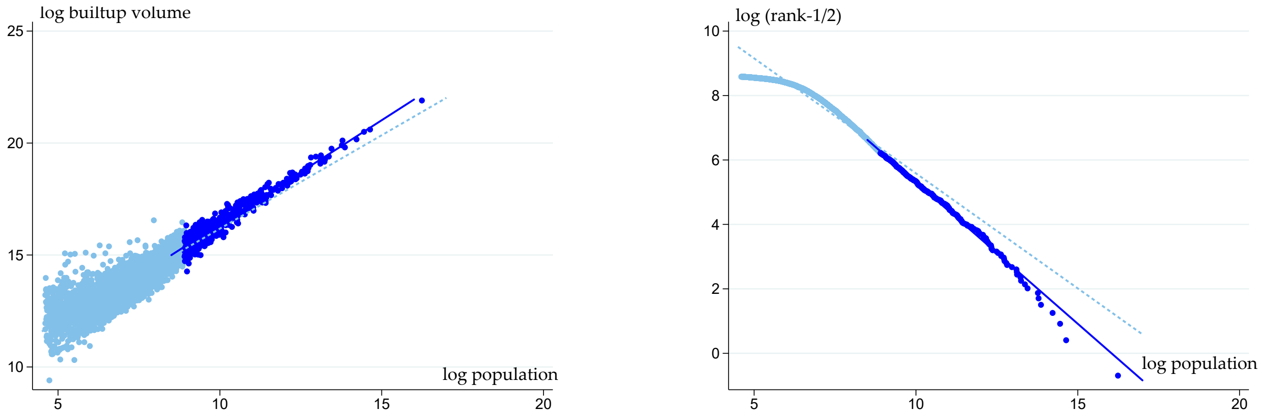
Population vs. buildings and the size distribution of urban areas

As made clear by table 4, there is considerable heterogeneity in buildings and population among the urban areas we delineate. To explore this heterogeneity further, we first focus on the relationship between the population of urban areas and their physical extent, measured by their area and their builtup volume. We then turn to Zipf's law to explore the distribution of urban area population. We note that these two explorations are also of independent interest since Zipf's law has attracted intense academic scrutiny (see for instance Duranton and Puga, 2014, for a critical review of the literature). The relationship between the size and density of urban areas and their population has also received much attention (see Ahlfeldt and Pietrostefani, 2019, for a full synthesis).

Panel A of figure 6 plots the population of the urban areas of our baseline delineation and their builtup volume. The relationship between these two variables is tight, especially for larger urban areas. For all urban areas with a population above 100, the elasticity of the builtup volume with respect to population is 0.84. We estimate a slightly higher elasticity of 0.93 when we focus on the 500 largest urban areas in population. For this latter group, log population explains 93% of the variance of log builtup volume. If instead of the builtup volume, we use the builtup area, we find elasticities with respect to population of 0.84 for all urban areas with population above 100 and 0.89 for the 500 largest urban areas.

Finding elasticities below one should be expected since a greater population makes land scarcer and leads to higher prices for land and builtup space (see Combes, Duranton, and Gobillon, 2019, for evidence from French cities). In turn, higher prices lead to a reduction in the consumption of land and builtup space per person. Finding a slightly higher population

Figure 6: Population and buildup volume and Zipf's Law for urban areas



Panel A: Builtup volume and population

Panel B: Zipf's Law

Notes: Dark blue (black) points represent the 500 most populated urban areas. Light blue (medium gray) points represent other urban areas with more than 100 inhabitants.

elasticity for builtup volume than for builtup area is also natural since we expect buildings to be taller where the population is larger and land is scarcer. The difference between the two elasticities for builtup volume and builtup area is nonetheless modest in France where building height is tightly regulated.

As far as we know, there is no prior estimate in the literature for the elasticity of the builtup volume with respect to population. For the elasticity of the builtup area with respect to population, our estimates are higher than the elasticities of around 0.6 reported by Ahlfeldt and Pietrostefani (2019) for a large sample of world cities.[13]

Turning to Zipf's law, we rank urban area population from the largest to the smallest and regress log rank on log city population. This is a non-standard regression. Gabaix and Ibragimov (2011) underscore the existence of a small sample bias in the estimation of the coefficient on log city population. To avoid this bias, they recommend using the log of the rank minus one half as the dependent variable:

$$\log(\text{Rank-}1/2) = \beta_0 - \xi \log \text{Population} + \epsilon \,. \tag{2}$$

The estimated coefficient ξ corresponds to the shape parameter of the Pareto distribution of urban area population. Zipf's law (after Zipf, 1949) corresponds to $\xi = 1$. This implies

[13] This difference is not due to specific features of our delineation. We find even slightly higher elasticities in France when using official urban areas. For the 500 largest official urban areas, the elasticities of the builtup volume and area with respect to population are 0.92 and 0.86, respectively.

that the expected size of the second largest city is half the size of that of the largest, that the expected size of the third largest city is a third of that of the largest, etc.

Panel B of figure 6 plots this relationship for all urban areas with population above 100. With this extremely low threshold, we estimate a coefficient of 0.71, arguably a large deviation from Zipf's law. The main reason for this deviation is the existence of a thin lower tail of urban areas with a small population. For all urban areas with a core, we estimate a higher Zipf's coefficient of 0.81. This coefficient is still dragged down by a smaller number of small urban areas with a core. When we focus on the 500 largest urban areas, we estimate a coefficient of 0.88, much closer to Zipf's law.[14] If we only consider the 200 largest urban areas, the Zipf coefficient rises to 0.93 and to 1.00 if we focus only on the 100 largest urban areas.

Centres and subcentres

We now describe our results for the internal geography of French urban areas. We find that 4,481 among 5,771 urban areas, or 78% of them, have a centre as defined above. The largest number of centres and subcentres is 71 for Paris. We also count 32 centres (and subcentres) for Lyon and 16 for Marseille. Although high, these numbers are not out of line relative to previous literature. Using coarser data, McMillen and Smith (2003) identify 47 centres and subcentres for Los Angeles and 39 for New York using a procedure that detects significant spikes of employment density and requires them to have a significant effect on nearby employment density to qualify as a centre or subcentre.

Overall central and subcentral areas in French urban areas represent 26.9% of urban pixels and host 70.1% of the population of French urban areas. Altogether, the (unique) centres of all urban areas represent 19.6% of urban pixels in France and 28.1% of the urban population.

The first interesting feature we note is the tendency for larger and more populated urban area to have more subcentres. We estimate an elasticity of the number of centres and subcentres with respect to urban area population of 0.17.[15] This elasticity is modestly increasing with city population. For the 500 largest urban areas, we estimate it to be 0.44 and the R^2 of the regression is 0.58. Although the results are hard to compare directly, McMillen and Smith

[14] If we use builtup volume instead of population to rank cities for the dependent variable and use log builtup volume as explanatory variable, we estimate a coefficient of 0.99.

[15] Because nearly a quarter of urban areas do not have a centre, we use the inverse hyperbolic sine ($arsinh(.)$) instead of a natural logarithm to measure the number of centres and subcentres. This function is defined in zero and closely approximates the natural logarithm of $2x$ for $x > 1$.

(2003) find a similar but weaker tendency for the number of subcentres to increase with city population.

The second notable feature regarding centres and subcentres is their highly uneven size distribution within urban areas. The population of the central area is usually much larger than that of the largest subcentral area. To document this feature, we estimate again equation (2) for the population of centres and subcentres within urban areas.[16] The Zipf's coefficient we estimate is 0.30. This is much more uneven than than the size distribution of cities for which is coefficient is close to one. Measuring the importance of centres and subcentres with their builtup volume or their builtup area yields comparable results.

The third key feature we evidence is that the central population is roughly proportional to urban area population. Regressing log population in the central area on log urban area population for the 4,245 urban areas with a centre and a positive population estimates a coefficient of 0.95. A similar coefficient is obtained when we measure urban areas and their centre using their builtup volume or their builtup area. When we restrict ourselves to the largest 500 urban areas, we estimate an elasticity of 1.16 for population. Thus, the population of an urban area's centre increases faster than its total population in the upper tail of the distribution. We estimate similar coefficients for the population of all centres and subcentres (in regressions containing an indicator variable for each rank).

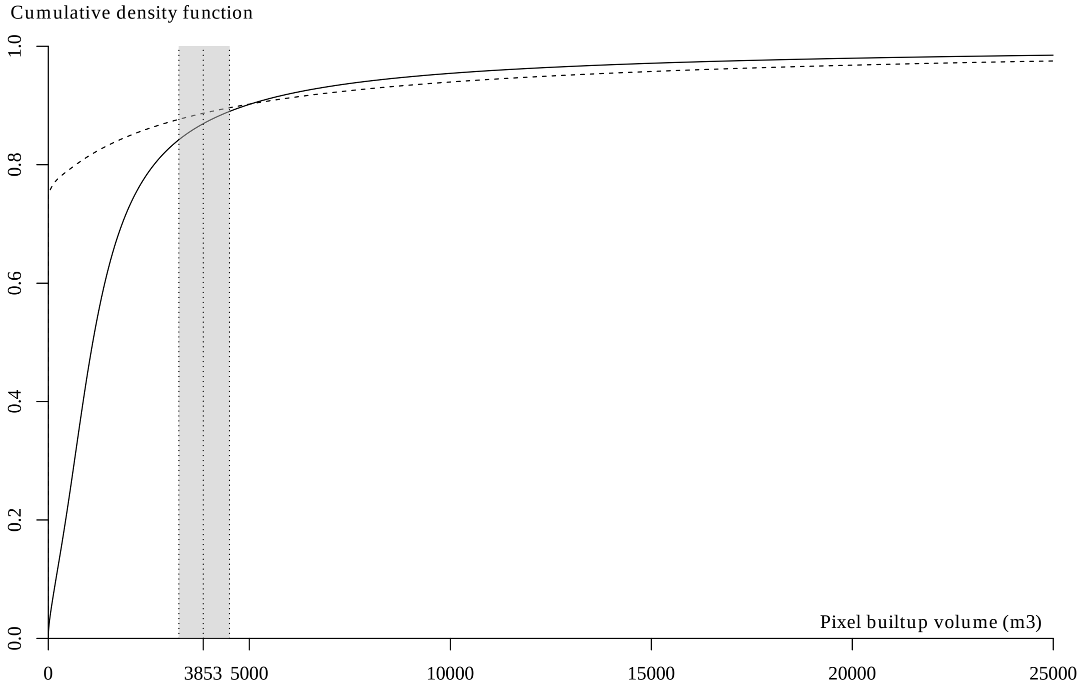
Obviously rural, obviously urban, and marginally urban pixels

To understand many of the comparisons discussed below, it is important to take another look at our baseline results in light of the large differences in pixel builtup density documented in table 2. Figure 7 plots the cumulative distribution functions of smoothed and unsmoothed pixel building densities (measured in m^3 per pixel). On the horizontal axis, we also plot the 95th percentile of the distribution of smoothed densities (for the median pixel as per table 3).

The cumulative distribution functions of figure 7 have three starkly different regions. For low building densities, the cumulative distribution function is extremely steep. This reflects the facts that a large majority pixels are empty of buildings (for unsmoothed building density) or close to empty (for smoothed building density). These pixels are 'obviously' rural. For high building densities, the cumulative distribution function is instead nearly horizontal

[16] We rank subcentral areas within each urban area (the central area is first by definition). We then pool the observations and estimate the Zipf regression with city fixed effects.

Figure 7: Distribution of building densities



Notes: Authors' calculations. Pixel building density (in m^3) on the horizontal axis and cumulative distribution function on the vertical axis. The dotted curve represents raw building density and the plain curve represents smoothed building density. On the horizontal axis, 3,853 m^3 corresponds to the median across pixels of the threshold for a pixel to be urban. This threshold is the 95th percentile of the counterfactual densities and varies across pixels (as per table 3). The grey band around 3,853 m^3 represents the 5th and 95th percentiles of the distribution of the threshold to be urban across pixels.

which reflects the extreme skew of the distribution of buildings across pixels in the upper tail. Put differently, in this flat part of the cumulative distribution, we find a small minority of 'obviously' urban pixels. They typically belong to urban cores. Given our purpose, the existence of these two extreme regions is good news since we expect reasonable delineations to agree on 'obviously' urban and 'obviously' rural pixels.

However, between these two main regions, there is an intermediate region where the cumulative distribution function is very concave as it transitions from near-vertical to near-horizontal. This region captures a middle class of 'marginally urban' pixels found in smaller settlements or at the periphery of larger settlements. We note that the 95th percentile of the distribution of smoothed densities for the median pixel falls into this region. Despite its concavity, the cumulative distribution function of pixel building density in figure 7 does not exhibit any obvious kink. In the absence of such kink, there is no natural threshold and any binary classification into urban and rural pixels will thus have to wrestle with how to treat this middle class. Put differently, small methodological differences will generate different delineations with disagreements at the periphery of urban areas and in smaller settlements.

Three further remarks are in order. First, the existence of a middle class of marginally urban pixels may be read as a call to define richer classifications with more categories. While such classifications may be needed for some purposes, this would not solve the issue at hand. Two kinks in the cumulative distribution function of pixel building density would be needed to cleanly define this middle class. We do not observe them. Second, while the middle class of marginally urban pixels is small relative to 'obviously rural' pixels, it seems large relative 'obviously urban' pixels. Figure 7 suggests that perhaps 80% of pixels are obviously rural and maybe 5 to 10% are obviously urban. This leaves 10 to 15% of marginally urban pixels. Third, while the set of marginally urban pixels may be geographically larger than the set of 'obviously urban' pixels, it may host only a small fraction of the population relative to 'obviously urban' pixels. Hence while delineations may differ a lot on the fraction of pixels they deem to be urban, the differences will be smaller for population.

5. Comparing our delineation with INSEE's urban units

In this section, we compare the outcome of our delineation approach to the official delineation performed by the French statistical institute (INSEE). We also use this comparison to introduce our formal metrics and tests to compare delineations.

INSEE's *urban units and a first informal comparison*

INSEE provides a delineation of urban areas which, like our approach, relies on a morphological zoning. INSEE 'urban units' (unités urbaines) are aggregates of French municipalities characterised by a total population of more than 2,000 inhabitants living in a continuously builtup area with no more than 200 metres between any two buildings.[17]

Panel D of table 4 above provides some descriptive statistics for INSEE urban units. We first note that there are 2,216 INSEE urban units while our approach delineates 5,771 urban areas. The much greater number of urban areas in our delineation occurs because even fairly small settlements can exhibit statistically significant building density, whereas INSEE imposes

[17] INSEE also defines functional metropolitan areas using commuting patterns, which it names urban areas (aires urbaines) instead of metropolitan areas. These areas are built around a core urban unit with at least 10,000 workers and iteratively aggregate other municipalities provided they send at least 40% of their workers to the core and other municipalities already aggregated to the core. References to alternative approaches are given in the introduction.

a lower bound of 2,000 inhabitants. Only 1,724 of our urban areas have more than 2,000 inhabitants.

Then, the comparison between panels A and D of table 4 further shows that INSEE urban units also have a greater physical extent. This difference in physical extent is even greater than the difference in population so that the population density of INSEE urban units is a third lower relative to our urban areas at the mean. Altogether and despite their much smaller number, the 2,216 INSEE urban units cover 22.9% of the French territory instead of 12.9% for the 5,771 urban areas in our baseline delineation.

When we focus the comparison on the 812 urban areas with an urban core, we find that, relative to INSEE urban units, our approach selects larger settlements in terms of population despite their generally smaller physical extent. To illustrate this last feature, figure 8 in Appendix C duplicates figure 5 above for the same four regions of Paris, Lille, Marseille, and Grenoble but also represents INSEE urban units. Despite the different morphologies of these urban areas, everywhere we can observe the greater physical extent of INSEE urban units.

Urban Jaccard indices

We now introduce indices to assess the extent to which spatial units on two given maps coincide. The indices essentially measure the intersection (or overlap) of urban pixels between two maps relative to their union. These indices are variants of Jaccard indices (Jaccard, 1902), which we also refer to as similarity indices.

We start with Jaccard indices that measure the extent to which urban pixels overlap on two different maps. We refer to these indices as *urban* Jaccard similarity. Denote the set of urban pixels on map $j \in \{1,2\}$ as U^j and its cardinal as $\left|U^j\right|$. The urban Jaccard similarity is computed as:

$$J_U^{12} \equiv \frac{|U^1 \cap U^2|}{|U^1 \cup U^2|}\,. \tag{3}$$

This index measures the proportion of pixels that are urban in the two maps among pixels that are urban on either of the two maps. It varies between zero, when there is no intersection among urban pixels on the two maps, and one, when all urban pixels on the two maps are confounded. Note that the calculation of urban Jaccard similarity excludes pixels that are rural in both maps.[18]

The index described in equation (3) is extremely flexible since it can be used to compare any two binary classifications. In particular, we can assess the similarity between the official INSEE map of urban units and urban pixels in our baseline delineation. We find $J_U = 0.345$. This is low but recall that INSEE urban units cover 22.9% of mainland France vs. 12.9% for the urban areas we delineate. Even if all the urban pixels of our delineation were also classified as urban by INSEE, the similarity would be only equal to $0.129/0.229 = 0.563$.

If we restrict the comparison to pixels that belong to the 812 urban areas with a core in our delineation (and also limit urban units to the 1,147 with a core as defined by INSEE), we compute a urban Jaccard similarity of $J_U = 0.363$. Although we delineate many more urban areas, the effect of this on the similarity between the two maps appears modest.

Sampling could potentially explain much of the difference between our map and INSEE's map of urban units and the low Jaccard indices we compute from the comparison between these two maps. To assess the importance of sampling variation, we construct confidence intervals for the index described by equation (3). We want to compare a map generated with the dartboard approach proposed here to the official map of urban units proposed by INSEE. Because there is no statistical variation for this second map (or if there is some, we do not know it), we take this second map as exogenously given and we perform a one-sided test based on the variation of the first map generated by our dartboard approach.

We divide our 10,000 counterfactual redistributions into 100 groups of 100 counterfactuals to generate 100 'smaller versions' of our baseline delineation that each classify every pixel as either rural or urban. For each of these 'small' delineations, we compute the similarity with the INSEE map using equation (3). We take as reference the median value of the urban Jaccard similarity and compute a confidence interval around this value. This procedure amounts to bootstrapping our index relying on the same dartboard approach.

We find extremely small standard errors around the urban Jaccard indices that we estimate. For our baseline delineation, we compute a standard error slightly above 0.0001 around our index $J_U = 0.345$. For a single small replication, the largest deviation relative to our reported

[18] After defining R^j the set of rural pixels on map j, we could instead measure the overlap as $(|U^1 \cap U^2| + |R^1 \cap R^2|)/N$ where N is the total number of pixels on each map. It is easy to see that when there is no overlap among urban pixels between the two maps, this index is equal to $(N - |U^1| - |U^2|)/N$. When $|U^1|$ and $|U^2|$ are both small, the index is close to one because of the strong overlap between rural pixels. We prefer to use a Jaccard index defined by equation (3) which is more easily interpretable for our purpose.

value for this index is only about 0.0002 and thus only affects its third decimal.[19]

To explain such small standard errors, a first possibility is that our classification of individual pixels is very stable and not subject to sampling variation. We can assess this explanation by using again 'smaller versions' of our baseline delineations and then compute the Jaccard similarity between pairs of these 'small' delineations. We find that the resulting urban Jaccard similarity is on average equal to 0.906. Such differences were to be expected given our choice of a 5% threshold. Alternatively, the average Jaccard index between smaller versions of our baseline and the baseline is equal to 0.935. Interestingly, using 500 sets of counterfactuals instead of only 100 leads to a Jaccard similarity with our baseline delineation of 0.963. This result shows that working with 100 counterfactuals is typically good enough to achieve a satisfactory level of precision in our context.

This said, deviations of about 5% around our baseline delineation do not fully explain the tiny standard errors that we obtain in the comparison with the urban units delineated by INSEE. The main explanation is that the sampling variation we have for our pixels mostly cancels in out with the law of large numbers when computing Jaccard indices.[20]

We can also compute Jaccard indices for population. By weighting each pixel in equation (3) by its population we can measure the overlap in urban population between two maps instead of the mere spatial overlap. For the comparison between the official INSEE map of urban units and urban pixels in our baseline delineation, we find a population-weighted Jaccard index of 0.868. This is obviously much higher than the value of 0.345 computed above without weighting for population. If we restrict the comparison to pixels that belong to the 812 urban areas with a core in our delineation (and also limit urban units to the 1,147 with a core as defined by INSEE), we compute a urban Jaccard similarity of 0.847 when weighting pixels by their population instead of 0.363.

These much higher values of Jaccard indices when we weight pixels by population reflect an important feature. While our baseline delineation differs quite a lot from INSEE's delineation regarding which pixels are urban, the two delineations nevertheless largely agree regarding more heavily populated pixels and capture essentially the same population. The

[19] For all urban pixels that belong to an urban area with a core, the standard error for our index $J_U = 0.363$ is again small, slightly below 0.0018.

[20] The following complication is worth keeping in mind. The Jaccard Index is a ratio. In particular, pixels which are at the margin of being urban may be classified by our approach as urban or rural depending on sampling, while INSEE's map mainly classifies these 'marginal' pixels as urban. Hence, at the numerator of urban Jaccard indices, the cardinal of the subset of INSEE's urban pixels that we also classifies as urban may not be sensitive to sampling even though the exact subset is. In this case, note that the denominator is also essentially constant. This may not be the case in general.

disagreements are mostly about marginally urban pixels at the periphery of larger urban areas or smaller settlements. As made clear by the results reported in tables 6 and 7 this result applies more generally to all our comparisons.

City Jaccard indices

In table 5 we compare the population and ranking of the 20 largest urban areas with their corresponding INSEE urban units. While we postpone the discussion of the last column of this table, we can note that despite some differences in rankings, 17 of our top 20 urban areas have their corresponding urban unit in INSEE's top 20 ranking and the population counts are surprisingly close for a large majority of urban areas. The main exception is the urban area of Lille. With its population of 2.31 million, this is unambiguously the second largest urban area in France in our baseline delineation. According to INSEE, the urban unit of Lille ranks fourth with only slightly less than a million inhabitants. This gap is consistent with our discussion of panel B of figure 5 above. Our approach evidences a large continuous urban cover around Lille. Instead, INSEE delineates four separate urban units. We also observe some differences for Strasbourg, Metz, Perpignan, and a number of smaller urban areas for which we systematically obtain a greater population. This occurs because our approach often aggregates together areas that the INSEE delineation treats as separate urban units. Hence, even though our approach is more conservative at the extensive margin relative to INSEE's delineation, it also has a tendency to aggregate more into larger urban areas at the intensive margin.

More generally, while it is informative to measure to what extent the urban pixels on two different maps overlap, it is also important to measure to what extent the spatial units delineated on these two maps coincide. To understand the difference between these two notions, consider the example of the urban area of Lille. As represented in panel B of figure 5, our approach delineates a large integrated urban area. According to INSEE's delineation, the urban unit of Lille is much smaller and is only one urban unit in a group of several independent urban units located close to each other. Although many pixels are 'urban' according to both our delineation and INSEE's, they are partitioned differently. We want to be able to take this into account when making comparisons between maps.

To do this, we must take a stand regarding the 'identity' of the spatial units across maps. To return to the example of the region of Lille in panel B of figure 5, our approach delineates a

Table 5: Descriptive statistics on pixel built area

Rank	Urban area	Population	Density	INSEE urban unit population	INSEE rank	Jaccard by urban area
1	Paris	11,326,600	3,123	10,613,004	1	0.398
2	Lille	2,314,310	1,159	980,148	4	0.170
3	Lyon	1,894,178	1,117	1,574,158	2	0.329
4	Marseille	1,521,720	1,428	1,530,694	3	0.323
5	Nice	1,051,676	1,477	955,474	5	0.407
6	Bordeaux	1,020,664	769	857,528	7	0.284
7	Toulouse	974,258	1,015	879,968	6	0.387
8	Strasbourg	693,598	843	420,648	13	0.193
9	Nantes	641,775	1,131	599,328	8	0.362
10	Grenoble	580,128	873	486,534	10	0.319
11	Toulon	576,310	1,236	563,868	9	0.345
12	Metz	501,168	821	267,163	20	0.246
13	Rouen	497,320	1,052	436,680	12	0.334
14	Montpellier	491,727	1,218	386,374	14	0.333
15	Avignon	481,079	532	461,764	11	0.343
16	Saint-Etienne	423,426	941	357,690	15	0.329
17	Perpignan	367,972	585	188,027	30	0.227
18	Mulhouse	359,378	785	237,299	23	0.270
19	Rennes	340,389	1,179	286,712	18	0.335
20	Nancy	340,218	1,020	252,584	21	0.280

Notes: Population is from 2015; area in km^2; density is the number of inhabitants per km^2. Urban Jaccard similarity by urban area as per equation (6)

large urban area that we naturally, but perhaps loosely at this stage, call "Lille". In the same region, INSEE delineates several urban units, only one of which it calls "Lille". While we want to measure the overlap between our Lille and INSEE's Lille, how do we know these spatial units are appropriately named? With our delineation, the urban area of Lille also includes Valenciennes, the centre of a distinct INSEE urban unit with a population above 300,000. Had we named our urban area "Valenciennes" instead of Lille, we would want to compute the overlap of our large urban area (now called Valenciennes) with the "Valenciennes" urban unit delineated by INSEE. Put differently, we need to know when a spatial unit is the 'same' across two maps that delineate them differently.

To define the identity of the urban areas delineated by our approach, we proceed as follow. As already mentioned, we name each urban area after the municipality with the largest population it overlaps with.[21] Hence, we name the large urban area at the extreme northern end of the country "Lille" because Lille is, among all municipalities with which this urban area overlaps, the municipality with the largest population. This approach is consistent with the INSEE naming convention of urban units, which always uses the municipality with the greatest population either as the unique name or as the first name for its urban units.

More formally, for map $j \in \{1,2\}$, denote by U_k^j the subset of pixels in urban area $k \in \{1,...,K\}$, where K is the number of different spatial units on the two maps. This quantity K can be obtained by summing the number of spatial units on the first map and the number of spatial units on the second map that are not defined on the first map. If spatial unit k is absent from map j, obviously $U_k^j = \emptyset$. Then, it is also the case that $\left\{U_1^j,...,U_K^j\right\}$ is a partition of U^j. We can now define the *city* Jaccard (similarity) index:

$$J_C = \frac{\sum_{k \in K} \left|U_k^1 \cap U_k^2\right|}{|U^1 \cup U^2|}, \tag{4}$$

after dropping the superindex from indices that compare map 1 and map 2 to lighten the notations (by noting this index J_C instead of J_C^{12}).

The key difference between the urban Jaccard index defined by equation (3) and the city Jaccard index defined by equation (4) is the following. The urban Jaccard index 'counts' at the numerator all pixels that are urban in both maps while the city Jaccard similarity only counts them when they are part of the same urban area. Hence, $J_C \leq J_U$. More specifically, we can readily observe from equations (3) and (4) that:

$$J_C \equiv J_U \times P, \qquad \text{where } P \equiv \frac{\sum_{k \in K} \left|U_k^1 \cap U_k^2\right|}{|U^1 \cap U^2|}. \tag{5}$$

The city Jaccard index, which measures the overlap between urban pixels that belong to same urban area(s). It can be expressed as the product of the urban Jaccard index that measures the overlap between urban pixels and the ratio P of the sum of the overlap by spatial units to the overall overlap. This ratio can be interpreted as a measure of the propensity of the two maps to aggregate urban pixels into the same units. Put slightly differently, our narrow measure of overlap, the city Jaccard index J_C, is equal to the product of our broad measure of overlap, the urban Jaccard index J_U, and an overlap quality factor, P.

[21] In theory, we need a criterion for breaking ties in case two or more urban areas share the same largest municipality. In practice, this never happens. These largest municipalities of urban areas are either fully included into their urban area or only partially included with a rural remainder.

Note that we can also define a Jaccard index for each individual urban area k:

$$J_k \equiv \frac{|U_k^1 \cap U_k^2|}{|U_k^1 \cup U_k^2|}\,, \tag{6}$$

with $J_k = 0$ if urban area k is missing from either map. We can now show that the city Jaccard index in equation (4) can be decomposed into the weighted sum of individual Jaccard index J_k calculated for every spatial unit 1,...K. From equations (4) and (6), we can write:

$$J_C = \sum_{k \in K} s_k\, J_k\,, \qquad \text{where } s_k \equiv \frac{|U_k^1 \cup U_k^2|}{|U^1 \cup U^2|}\,. \tag{7}$$

Hence, the city Jaccard index is the weighted sum of the individual Jaccard index of every spatial unit where the weights are the share of pixels s_k that belong to this spatial unit in either map relative to the number of urban pixels across both maps. All these indices can be bootstrapped following the approach described above for urban Jaccard indices.

For the comparison between the official INSEE map of 2,216 urban units and our preferred delineation with all urban pixels of 5,771 urban areas, we obtain $J_C = 0.272$. For the comparison between the official INSEE map of urban units and our preferred delineation where we restrict ourselves to 1,147 INSEE urban units with a core and 812 urban areas with a core, we obtain $J_C = 0.289$.

The main point to note is that although these two indices are below their corresponding urban Jaccard indices obtained above that take the values of 0.345 (all urban pixels) and 0.363 (pixels that belong to urban areas with a core), the differences are not large. Using expression (5), this suggests a high propensity of the two maps to aggregate urban pixels into the same units close to 0.8.

Another way to understand those figures 0.272 and 0.289 for the city Jaccard similarities is to return to expression (7), which shows that the city Jaccard index can be decomposed into a sum of weighted individual Jaccard indices for urban areas. For the 20 largest urban areas, the last column of table 5 reports their city Jaccard index. The results confirm our visual impressions from earlier with a reasonably high Jaccard index for Paris of 0.398 and a much lower value for Lille of 0.170 while the indices for Marseille and Grenoble are in between these two cases at 0.323 and 0.319 respectively. For the largest 20 cities, the average (weighted) city Jaccard index is 0.341. This is slightly more than the overall value of either 0.272 for the overall city Jaccard index. This is because, at the lower end of the distribution, urban areas either overlap poorly or do not overlap at all, in which case the Jaccard similarity for them is zero.

When we compute standard errors for these city Jaccard indices using the same approach as described above, we obtain again values that are small, about 0.0007 for $J_C = 0.272$ (baseline delineation) and 0.0009 for $J_C = 0.289$ (all urban pixels that belong to an urban area with a core). We note nonetheless that these standard errors are more than seven times as large as those computed for urban Jaccard indices. As mentioned above, city Jaccard indices are expected to be more sensitive to sampling in the case of nearby groups of buildings which, depending on sampling, may or may not belong to the same urban area. At the same time, we note that this sensitivity remains minimal for any practical purpose in our case. Similarly, when we perform the two-sided test on 100 smaller versions of our baseline delineations (each using 100 counterfactuals), we obtain a city Jaccard $J_C = 0.903$, which is virtually equal to the urban Jaccard $J_U = 0.906$ with a slightly larger standard error of 0.004. This shows that the identity of our urban areas does not change across our delineations. They neither aggregate nor split.

In Appendix D, we propose alternative Jaccard indices that measure the tendency of pairs of pixels to belong to the same urban areas. Like with individual pixels, we can measure the tendency of pairs of pixels to be similarly classified as urban (a direct counterpart to J_U above) as well as the tendency of pairs of pixels to belong to a given urban area (a direct counterpart to J_C above). Because pairs of pixels can be measured as part of the same area over two different maps without having to define the identity of these areas, this also allows us to define a less conservative measure of similarity across cities. For instance, a pair of points where both points belong to 'Valenciennes' will be counted as part of the similarity between INSEE's delineation (which treats Valenciennes as separate from Lille) and our delineation (which integrates Valenciennes with Lille). Because these indices are more involved and their interpretation less straightforward, we only report and discuss them in Appendix D.

The role of base units

The imperfect overlap between our preferred delineation and INSEE's delineation of urban units may be caused, at least in part, by the fact that INSEE aggregates entire municipalities. Although French municipalities are 'small', they are still much larger than our base units, pixels of 200 metres by 200 metres. On average, a French municipality corresponds to nearly 400 pixels (or 16 square kilometers).

To assess the effect of the difference in the size of the underlying base units, we consider a variant of our delineation where we 'discretise' urban areas ex post. Starting from our baseline delineation, we classify an entire municipality as urban if 50% or more of its component pixels are urban. If not, we classify this municipality as rural. This is similar in spirit to the approach used by INSEE.[22]

When comparing our discretised delineation with INSEE's urban units, we obtain an urban Jaccard index of 0.291 for all urban areas and 0.223 for all urban areas with a core. These values for the urban Jaccard indices are actually lower than their respective values of 0.345 and 0.363 obtained above for the comparison with our baseline delineation. This worsening in the overlap between delineations arises because this discretisation is performed with a high threshold of 50%. This makes our delineation even more restrictive than previously as many municipalities are 'rounded down'. The effects of this rounding-down are dramatic. With this 50% threshold, we delineate only 945 urban areas instead of 5,771 in our baseline. Only 7.8% of pixels instead of 12.9% are now classified as urban.

It turns out that we can improve the overlap with INSEE's delineation with a lower discretisation threshold. If we classify as urban any municipality for which 25% or more of its pixels are urban, we end up with 17.2% of urban pixels and the urban Jaccard index for the comparison with INSEE's delineation rises to 0.465 when considering all urban pixels or to 0.561 when considering all pixels part of an urban area with a core. While the overlap between our delineation and INSEE's is still imperfect, these changes to the Jaccard indices caused by the discretisation of municipalities make it clear that the size of the underlying units to be aggregated plays an important role when delineating urban areas. This 25% threshold appears to maximise the similarity with INSEE's delineation.[23]

Another possible explanation for the limited overlap between our delineation and INSEE's delineation of urban units lies in our use of a 95% statistical threshold, which may lead to a stricter definition of what is urban with our approach relative to INSEE's. It is true that if we take an even more restrictive threshold of 99%, the Jaccard index between our delineation and INSEE's falls from 0.345 to 0.305. However, taking a much less conservative statistical threshold of 75% only increases the urban Jaccard index to 0.363.

Taking a less conservative threshold has three effects on our delineation. First, it leads to

[22] A municipality is classified by INSEE as part of an urban area if it is at least 50% urban (according to its own definition of urban of course). We nonetheless keep in mind that INSEE allows for distinct urban units to be adjacent whereas we always integrate adjacent urban municipalities into a single urban area.

[23] Discretising municipalities with a threshold of 25% also improves the city Jaccard indices, albeit less so. This is because 'rounding up' municipalities leads to further aggregation of urban areas. Our preferred delineation has already a tendency to aggregate urban areas more than INSEE's delineation, which gets magnified when using a low threshold for discretisation.

an expansion of the largest urban areas that are also delineated by INSEE. This contributes to improving the similarity between the two maps since, with a threshold of 95%, our urban areas are physically less extensive than those delineated by INSEE. However, a lower statistical threshold also leads to the expansion of urban areas that are not part of INSEE's delineation and to the delineation of new urban areas that are also absent from INSEE's delineation. These two effects lead to a worsening of the Jaccard index between the two maps. Overall, the first effect dominates, but only modestly. The lack of similarity between our delineation and INSEE's is thus not due to using a 95% threshold to define statistical significance.

6. Other comparisons

Urban areas defined with builtup volume vs. footprint

While our preferred approach to measure building density relies on cubic metres of building, an obvious alternative is to measure building density with squared metres of building footprint. Using builtup areas is perhaps closer to the definition used by INSEE to define urban units as well as other morphological definitions used elsewhere.[24]

Our most important result here is that measuring building density with builtup areas instead of builtup volumes leads to more urban areas that are physically larger. Overall, with density measured with buildup area, urban areas take 16.3 % of all pixels (instead of 12.9% with volumes) and host 85.0% of the French metropolitan population (instead of 80.5%). This greater physical extent of urban areas when using building footprints instead of building volumes is unsurprising since taller buildings tend to be located at the centre of urban areas. Peripheral areas with fewer and shorter buildings may thus still exhibit excess building density when measuring their footprint but not when using their volume. We thus end up with physically larger urban areas and more of them when using building footprint instead of building volume.

We can assess the difference between the volume- and footprint-based delineations more systematically using Jaccard indices as above. When we compare the two delineations using all urban pixels on both, we find $J_U = 0.754$. We note that this figure for the urban Jaccard index is close to the ratio of urban pixels across both delineations, which is equal to 0.791. This indicates that urban pixels when using a volume-based definition of building density are to

[24] We retain the same two-kilometre bandwidth as before. The impact of changing the bandwidth is explored below.

a large extent a subset of the urban pixels defined when using a footprint-based definition of building density. When we restrict ourselves to urban pixels that are part of urban areas with a core, this feature is even more striking since we find that $J_U = 0.693$ while the ratio of urban pixels is 0.715.

Turning to city Jaccard indices we compute $J_C = 0.262$ for all urban pixels and $J_C = 0.290$ for pixels part of an urban area with a core. The difference between the urban and the city Jaccard indices arises because the greater extent of urban pixels with a footprint-based definition also leads to more aggregated urban areas. This is most obvious in the case of Lille. When using building footprint to measure building density, the urban area of Lille aggregates more than 10 urban areas, including Arras with more than 100,000 inhabitants, that are delineated as independent urban areas using a volume-based definition of building density.

Because a footprint-based definition of urban density leads to physically larger urban areas, we could expect it to lead to a map of urban areas that is closer to the official delineation of urban units by INSEE. The urban Jaccard index measuring the similarity between our footprint-based delineation and the official INSEE delineation is $J_U = 0.373$. This is only modestly larger than the urban Jaccard index of 0.345 measuring the similarity between our preferred delineation and the official delineation. For the city Jaccard index, we obtain $J_C = 0.262$ when using footprint instead of 0.272 when using builtup volumes. This worse similarity is due to the greater aggregation of urban areas when using footprint areas.

It is easy to see that the statistical approach to measure the significance of Jaccard indices described above readily generalises to comparisons between two maps produced by variants of our approach to perform a two-sided test. In this case, we can replicate each of the two maps using the bootstrap methodology just described. This generates 100 pairs of maps and we can compute a Jaccard index for each of these pairs and deduce again a confidence interval. Again, the standard errors computed from our simulations for this comparison are tiny, again of the order of 0.0002. As mentioned before, any two delineations we obtain from different sets of 100 counterfactual distributions of buildings overlap extremely closely when we use volume criteria to measure builtup density. The same result unsurprisingly holds for any two delineations that rely on the footprint of buildings. Hence, sampling typically affects at most the third digit of the indices reported above.

Urban areas defined using different statistical thresholds

The next comparison regards the effect of the statistical thresholds we use to define excess building density. Until now, we have focused on the standard 95% threshold for statistical significance.

To assess how this choice of threshold affects our delineation, we also replicate our approach with a 75% significance threshold and with a 99% significance threshold. When comparing our baseline delineation with a 95% threshold to the same delineation using a 75% threshold for statistical significance, we compute an urban Jaccard index of 0.756 for all urban pixels and 0.695 for pixels that belong to an urban area with a core. By construction, any pixel that is urban with a threshold of 95% for statistical significance is also urban with a threshold of 75%. In this special case, the urban Jaccard index thus represents the share of pixels that are classified as urban with a 95% threshold among those that are classified as urban with a 75% threshold (17% of all pixels). Turning to the comparison between delineations obtained with the 95 and 99% thresholds for statistical significance, we find an urban Jaccard index of 0.741 when considering all urban pixels and 0.699 when considering only urban pixels part of a urban area with a core.

For both the 75-95 and the 95-99 comparisons, the city Jaccard indices take lower values of 0.252 and 0.259, respectively. These lower values are unsurprising. Recall again that by equation (5) city Jaccard indices can be expressed as the product of urban Jaccard indices and an overlap quality factor. Considering a different threshold not only leads to a different proportion of urban pixels but it also leads to a different aggregation into urban areas, that is a lower overlap quality factor. For instance, while Marseille and Toulon are delineated as separate urban areas with a 95% threshold, they are part of the same integrated urban area with a 75% threshold. On the other hand, Lille and Valenciennes are part of the same urban area with a 95% threshold but get separately delineated with a 99% threshold.

Urban areas defined using different bandwidths

We now compare our baseline delineation to variants generated using alternative bandwidth of one and three kilometres instead of two. With a one-kilometre bandwidth, we delineate 21,020 urban areas representing 84.1% of the population. A shorter bandwidth implies that small clusters of buildings may no longer be smoothed out as previously and may appear now as statistically significant concentrations of buildings. Urban areas that are integrated with our preferred delineation may now be separated into several with a shorter bandwidth. For instance, Lille, Valenciennes, and Douai are all part of the same urban area with a two-kilometre bandwidth but form separate urban areas with a one-kilometre bandwidth. This shorter bandwidth also leads to the separation of Aix-en-Provence from Marseille.

This increase in the number of areas is also true for urban areas with a core. With a one-kilometre bandwidth, we delineate 2,221 of them instead of 812 with our preferred delineation. On the other hand, areas in close proximity to large clusters of buildings will receive less mass from smoothing and may no longer appear as significant builtup concentrations. A smaller bandwidth will thus lead to smaller urban areas. Overall, and despite delineating nearly four times as many urban areas, we end up with fewer urban pixels, 11.0% instead of 12.9%. Because the two effects we just highlighted are large, the urban Jaccard index between the two delineations is rather low at 0.597.

When we consider a longer bandwidth of three kilometres instead of two, we unsurprisingly observe the opposite. With a three-kilometre bandwidth, we delineate 2,891 urban areas, that is about half the 5,771 urban areas delineated with a two-kilometre bandwidth. Extent urban areas are also larger since 14.2% of pixels are classified as urban. The effect of increasing the bandwidth by one kilometre on the urban Jaccard index comparing with our baseline delineation, which is at 0.727, is nonetheless less dramatic than the effect of reducing our preferred bandwidth by one kilometre.

Urban areas defined with gridded builtup, population, or lights data

The data we use is a comprehensive map of all buildings. To our knowledge, this type of data exists for only one other country, Spain (Arribas-Bel *et al.*, 2019). While we expect this type of data to become more widespread, it will probably be many years before it is broadly and easily available for many countries. On the other hand, gridded population or land cover data is readily available for the entire world from a number of sources (see Henderson *et al.*, 2020, for a discussion). Here we assess how our approach fares when we adapt it to gridded data. Instead of redistributing buildings across pixels, we directly redistribute pixels and their builtup volume on the map. We then smooth cubic metres of building across pixels and proceed as previously.

When we redistribute the builtup volume of pixels directly, we delineate 5,294 urban areas, including 679 with a core, which occupy 10.2% of the territory and host 76.9% of the population. While this is slightly fewer urban areas, fewer urban pixels, and slightly less population than with our baseline population, the magnitudes are close. The urban Jaccard similarity index for this variant and our baseline delineation is 0.796. Because this variant generates 14.3% fewer urban pixel than our baseline, the highest value the urban Jaccard index could take is 0.857. This suggests that these two delineations are otherwise close.

To understand why redistributing pixels generates fewer urban pixels, consider the following heuristic reasoning. Pixels already aggregate buildings. The fact that some pixels contain many cubic meters of buildings, often from many buildings, is in most cases not due to chance. Our baseline delineation takes this variation into account through the redistribution of individual buildings. However, this variation is ignored when we redistribute pixels directly. To put it simply, redistributing pixels directly ignores a source of non-randomness and thus makes excess concentration of builtup volume harder to detect. The good news is that with a relatively fine grid of 200 metres by 200 metres, the bias is small.

Next, we can repeat the same exercise using population data on the same 200 by 200 metre grid. This delineates 4,147 urban areas with only 337 with a core. These urban areas occupy 7.9% of mainland France. Because a smaller fraction of France is urban relative to our baseline delineation, the urban Jaccard similarity index is much lower than previously at 0.560. The main reason for this lower level of similarity is the following. At the pixel level, there is much more variation in population than in builtup.[25] As result, it is harder to be above the 95^{th} percentile of the (smoothed) distribution of population than the 95^{th} percentile of builtup volume.

Finally, we replicate the same exercise once more but use information coming from satellite lights at night at the level of individual pixel. We use the same data as Ch *et al.* (2018) and recover a digital number that measures nightlight intensity for each pixel. We then proceed as previously but redistribute digital numbers across pixels instead of builtup volumes or populations. This procedure delineates 2,576 urban areas, including 664 with a core. These urban areas now occupy 16.8% of mainland France. This is more than our baseline delineation. The urban Jaccard index with our baseline is low at 0.402. Two effects are at play to explain this low Jaccard index. First, it is well-known that nightlights glow and this will mechanically lead to larger urban areas (see Baragwanath-Vogel *et al.*, 2020, for further evidence). Then it is also the case that the digital number that characterises pixel luminosity shows less variation

[25] For builtup volume, the standard deviation across pixels is slightly less than five times the mean whereas for population the standard deviation is more than seven times the mean. For nighlights below, the standard deviation is less three times the mean.

across pixels. As a result, the threshold statistical significance is also lower than for builtup volumes.

7. Conclusions

We propose a new approach to define urban areas. It relies on the most basic components of cities, individual buildings. Using a dartboard methodology, our approach naturally defines 'urban' as statistically significant excess building density. The main strength of our approach is to avoid (or at least minimise) the use of arbitrary criteria to define what is urban and what is rural. We rely instead on either optimality criteria or standard statistical thresholds. We also develop new formal tools to compare statistically different delineations on different maps.

While less than 1% of the French territory is covered by buildings, our preferred approach classifies about 12.9% of mainland France as urban and 80.5% of the French population is urbanised. Our approach delineates 5,771 urban areas, most of which are tiny. When we only consider urban areas with at least an urban core, that is urban areas with at least one pixel with excess building density relative to all urban pixels, the number of urban areas falls to 812. These urban areas cover less than 8.8% of the French territory but still host 70.1% of the population.

While some parts of the country such as the centres of large cities are obviously 'urban', others are clearly rural. However, building density in France (just like nearly everywhere) declines slowly as one moves away from the centre of cities or as one considers smaller settlements. Hence, there is no natural discontinuity in building density. At the same time, any attempt to partition the country into urban and rural needs to draw the line somewhere. Thus, minor methodological differences can lead to sizeable differences in the delineation of urban areas. For instance, defining building density with their footprint instead of their volume leads us to delineate more and physically larger urban areas that occupy 16.9% of the French territory instead of 12.9%. On the other hand, 'marginally urban areas' host only about 5% of the French population.

Our statistical tests allow us to make comparisons across maps. They indicate that the bounds around our preferred delineations are tight. While the choices made in the delineation approach matter, sampling issues do not when we make comparisons across approaches.

When we compare our preferred delineation with the official delineation of the French statistical institute (INSEE), we find that our approach tends to delineate either more urban areas (when we consider all of them) or fewer (when we restrict ourselves to urban cores) than the 2,231 urban units delineated by INSEE. We also find that our approach delineates physically smaller urban areas but, at the same time, has a stronger tendency to aggregate neighbouring urban centres. In part, INSEE's urban units are larger because they sum municipalities whereas our approach builds from tiny pixels.

We also extend our approach to document novel features of the internal geography of cities. However, many further applications are possible. Obviously, our approach could be used to delineate urban areas in other countries. We believe it could also be used for other classifications beyond urban-rural and at other spatial scales. For instance, it could be used to assess the clustering of retails stores in certain areas of a city by creating counterfactual distributions of existing stores instead of a counterfactual distributions of buildings. A variant of our approach which randomly redistributes characteristics such as race or income across households could be used to assess and describe social or racial segregation within cities.

References

Ahlfeldt, Gabriel M. and Elisabetta Pietrostefani. 2019. The economic effects of density: A synthesis. *Journal of Urban Economics* 111(0):93–107.

Arribas-Bel, Daniel, Miquel-Angel Garcia-Lopez, and Elisabet Viladecans-Marsal. 2019. Building(s and) cities: Delineating urban areas with a machine learning algorithm. Processed, University of Barcelona.

Asian Development Bank. 2019. Fostering growth and inclusion in asia's cities. Asian development outlook 2019 update.

Baragwanath-Vogel, Kathryn, Ran Goldblatt, Gordon Hanson, and Amit K. Khandelwal. 2020. Mixing satellite imagery to define urban markets: An application to India. *Journal of Urban Economics* forthcoming.

Berry, Brian, Joe Lobley, Peter G. Goheen, and Harold Goldstein. 1969. *Metropolitan Area definition: A re-evaluation of concept and statistical practice*. Washington, DC: US Bureau of the Census.

Berry, Brian J.L. 1960. The impact of expanding metropolitan communities upon the central place hierarchy. *Annals of the Association of American Geographers* 50(2):112–116.

Billings, Stephen B. and Erik B. Johnson. 2012. A non-parametric test for industrial specialization. *Journal of Urban Economics* 71(1):312–331.

Bode, Eckhart. 2008. Delineating metropolitan areas using land prices. *Journal of Regional Science* 48(1):131–163.

Bosker, Maarten, Mark Roberts, and Jane Park. 2018. Does definition matter? Metropolitan areas and agglomeration economies in a large developing country. Processed, World Bank.

Bowman, Adrian W. and Adelchi Azzalini. 1997. *Applied Smoothing Techniques for Data Analysis: The Kernel Approach with S-Plus Illustrations*. Oxford Statistical Science Series vol. 18. New York, NY: Oxford University Press.

Briant, Anthony, Pierre-Philippe Combes, and Miren Lafourcade. 2010. Does the size and shape of geographical units jeopardize economic geography estimations? *Journal of Urban Economics* 67(3):287–302.

CAF Development Bank of Latin America. 2017. Urban growth and access to opportunities: A challenge for Latin America. 2017 report on economic development.

Ch, Rafael, Diego Martin, and Juan F. Vargas. 2018. Measuring cities with night-time light data. Processed, CAF - Development Bank of Latin America.

Cheshire, Paul C. and Dennis Hay. 1989. *Urban Problems in Western Europe: An economic analysis*. London: Unwin Hyman.

Combes, Pierre-Philippe, Gilles Duranton, and Laurent Gobillon. 2019. The costs of agglomeration: House and land prices in French cities. *Review of Economic Studies* 86(4):1556–1589.

Corvers, Frank, Maud Hensen, and Dion Bongaerts. 2009. Delimitation and coherence of functional and administrative regions. *Regional Studies* 43(1):19–31.

Davis, Donald R., Jonathan I. Dingel, and Antonio Miscio. 2020. Cities, skills, and sectors in developing economies. *Journal of Urban Economics* forthcoming.

Dijkstra, Lewis, Aneta Florczyk, Sergio Freire, Thomas Kemper, and Martino Pesaresi. 2019. Applying the degree of urbanisation to the globe: A new harmonised definition reveals a different picture of global urbanisation. Processed, European Commission.

Duranton, Gilles. 2015. A proposal to delineate metropolitan areas in Colombia. *Economia & Desarrollo* 75(0):169–210.

Duranton, Gilles and Henry G. Overman. 2005. Testing for localization using micro-geographic data. *Review of Economic Studies* 72(4):1077–1106.

Duranton, Gilles and Diego Puga. 2014. The growth of cities. In Philippe Aghion and Steven Durlauf (eds.) *Handbook of Economic Growth*, volume 2. Amsterdam: North-Holland, 781–853.

Ferreyra, Maria Marta and Mark Roberts. 2018. *Raising the Bar for Productive Cities in Latin America and the Caribbean*. Washington, DC: World Bank.

Fox, Karl A. and T. Krishna Kumar. 1965. The functional economic area: Delineation and implications for economic analysis and policy. *Papers of the Regional Science Association* 15(1):57–85.

Gabaix, Xavier and Rustam Ibragimov. 2011. Rank-1/2: A simple way to improve the OLS estimation of tail exponents. *Journal of Business Economics and Statistics* 29(1):24–39.

Galdo, Virgilio, Yue Li, and Martin Rama. 2019. Identifying urban areas by combining human judgment and machine learning: An application to India. Processed, World Bank.

Giuliano, Genevieve and Kenneth A. Small. 1991. Subcenters in the Los Angeles region. *Regional Science and Urban Economics* 21(2):163–182.

Hall, Peter G. and Dennis Hay. 1980. *Growth centres in the European urban system*. London: Heinemann Educational Books.

Henderson, J. Vernon, Sebastian Kriticos, and Jamila Nigmatulina. 2020. Measuring urban economic density. *Journal of Urban Economics* forthcoming.

Jaccard, Paul. 1902. Lois de distribution florale dans la zone alpine. *Bulletin de la Société Vaudoise des Sciences Naturelles* 38(144):480–494.

Kanemoto, Yoshitsugu and Reiji Kurima. 2005. Urban employment areas: Defining Japanese metropolitan areas and constructing the statistical database for them. In Atsuyuki Okabe (ed.) *GIS-Based Studies in the Humanities and Social Sciences*. Boca Raton: Taylor & Francis, 85–97.

McDonald, John F. 1987. The identification of urban employment subcenters. *Journal of Urban Economics* 21(2):242–258.

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

McMillen, Daniel P. and Stefani C. Smith. 2003. The number of subcenters in large urban areas. *Journal of Urban Economics* 53(3):332–342.

Moreno-Monroy, Ana I., Marcello Schiavina, and Paolo Veneri. 2019. Urban economic agglomerations and their suburbanization patterns. A global analysis. Processed, OECD.

Mori, Tomoya, Koji Nishikimi, and Tony E. Smith. 2014. A probabilistic modeling approach to the detection of industrial agglomerations. *Journal of Economic Geography* 14(3):547–588.

Roberts, Mark, Brian Blankespoor, Chandan Deuskar, and Benjamin Stewart. 2017. Urbanization and development: Is Latin America and the Caribbean different from the rest of the world? World Bank Policy Research Working Paper 8019.

Rozenfeld, Hernán D., Diego Rybski, Xavier Gabaix, and Hernán A. Makse. 2011. The area and population of cities: New insights from a different perspective on cities. *American Economic Review* 101(5):2205–2225.

Zipf, George K. 1949. *Human behavior and the principle of least effort: an introduction to human ecology*. Cambridge (MA): Addison Wesley.

Appendix A. Further information about the data

BD CARTHAGE (IGN, 2006). The data describe all bodies of water in France. The river network is represented with lines and width categories (1: 0-15 metres, 2: 15-50 metres, 3: more than 50 metres). We reconstructed rivers by adding buffers around their lines (1: 10 metres, 2: 30 metres, 3: 50 metres). We then computed the water area for each pixel by summing sea, lakes, and river areas.

BD ALTI (IGN, 2015). The data report elevation continuously for the entire French territory from measurements made at least every 75 metres. For each pixel, we compute average elevation. We also construct the slope at each location and compute the mean slope for each pixel.

Localised Tax Revenues (INSEE, 2015). The French fiscal administration keeps a ledger of all households and their address to administrate income and residential taxes. The addresses of households are geolocalised by INSEE to assign households to pixels. We designed our pixels to match these INSEE pixels. A minor limitation of localised population data is that people living in retirement homes may have a fiscal address that differs from their actual residence. The same problem occurs with students. Homeless people will be missing altogether.

Appendix B. Summary table

Table 6: Descriptive statistics for our baseline delineation and its main variants

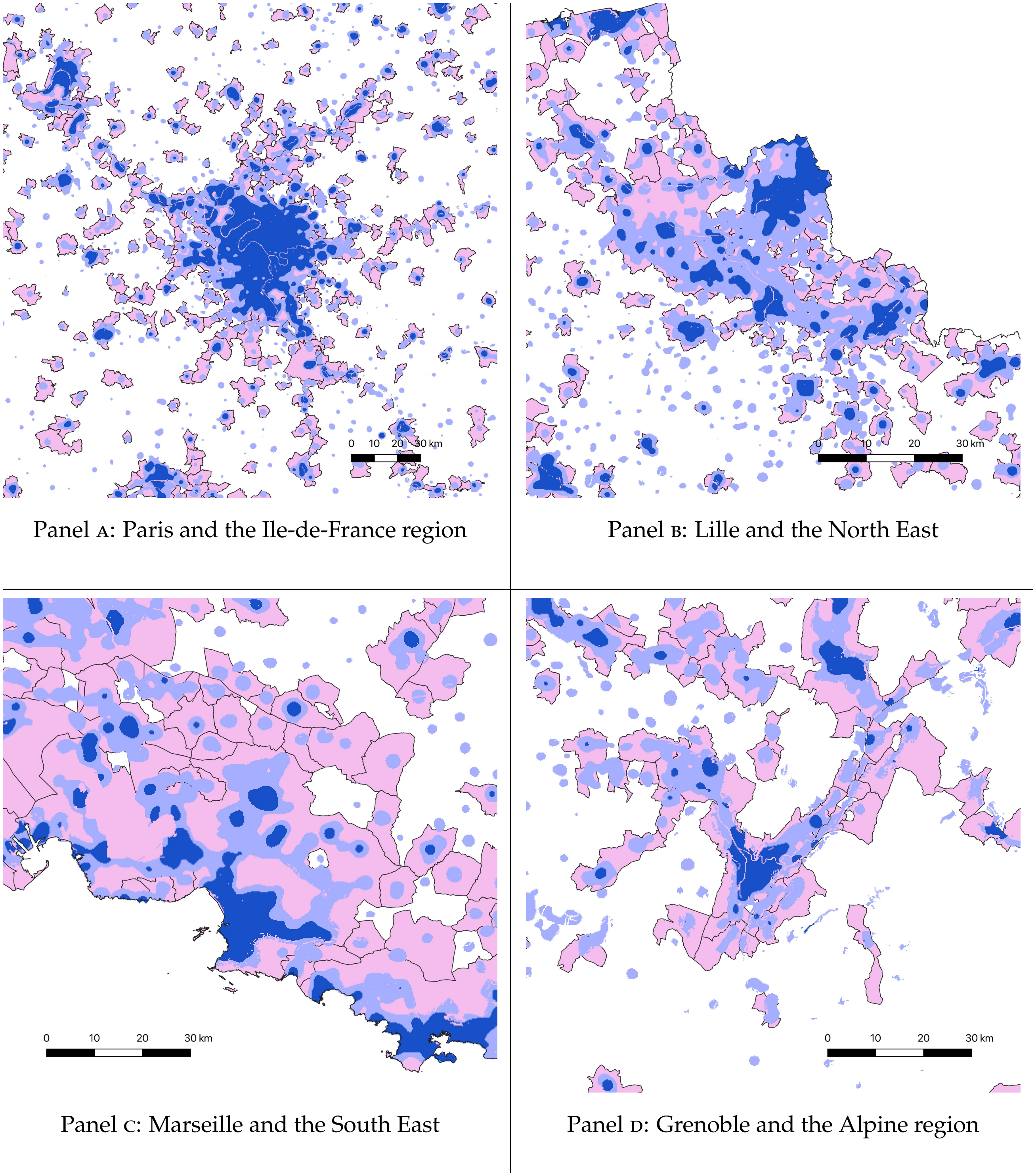
Type delineation	Number of UAs	J_U	J_U pop. weighted	% pixels	% build.	%pop
Panel A : Baseline delineation						
All UAs	5,771	–	–	12.9	69.8	80.5
UAs with core	812	–	–	8.8	57.1	70.1
UAs with core of degree 2	234	–	–	6.4	46.5	60.9
Core areas	1,228	–	–	2.4	40.2	51.6
Core areas of degree 2	353	–	–	0.5	17.1	24.6
Panel B : Other delineations						
INSEE Urban units	2,216	0.345	0.868	22.9	63.9	76.9
INSEE Urban units with core	1,147	0.363	0.847	15.7	54.8	68.7
Gridded volume UAs	5,294	0.796	0.955	10.2	65.8	76.9
Gridded volume UAs with core	679	0.773	0.938	6.8	52.5	65.7
Surface UAs	7,011	0.754	0.944	16.3	69.3	85.0
Surface UAs with core	1,552	0.693	0.908	12.3	58.5	76.8
Population UAs	4,147	0.560	0.916	7.9	56.7	75.4
Population UAs with core	337	0.518	0.859	4.9	42.9	61.0
Nightlights UAs	2,576	0.402	0.820	16.8	59.4	72.6
Nightlights UAs with core	664	0.425	0.849	15.1	57.7	71.5
Panel C : Other parameters						
Bandwidth 1km UAs	21,020	0.597	0.916	11.0	75.4	84.1
Bandwidth 1km UAs with core	2,221	0.591	0.903	6.8	59.0	70.7
Bandwidth 3km UAs	2,891	0.727	0.942	14.2	66.9	78.1
Bandwidth 3km UAs with core	561	0.706	0.935	10.4	56.6	69.9
Quantile 75 UAs	7,141	0.756	0.951	17.0	74.7	84.6
Quantile 75 UAs with core	1,792	0.695	0.916	12.7	64.7	76.5
Quantile 99 UAs	5,441	0.741	0.937	9.6	64.2	75.5
Quantile 99 UAs with core	580	0.699	0.911	6.2	50.5	63.8
Bootstrap 500 UAs	5,837	0.964	0.993	12.6	69.4	80.2
Bootstrap 500 UAs with core	930	0.962	0.990	8.6	56.8	69.7
Discretised 10% UAs	1,495	0.380	0.918	30.4	73.8	83.2
Discretised 10% UAs with core	635	0.503	0.956	16.7	59.8	72.2
Discretised 25% UAs	1,293	0.464	0.872	17.2	62.5	74.3
Discretised 25% UAs with core	622	0.560	0.939	12.7	55.8	68.8
Discretised 50% UAs	945	0.408	0.743	7.8	47.6	60.8
Discretised 50% UAs with core	574	0.532	0.833	7.1	45.8	59.3

Table 7: Descriptive statistics for comparisons between INSEE urban units and variants of our baseline delineation

Type delineation	Number of UAs	J_U	J_C	J_U pop. weighted	J_C pop. weighted
Reference					
INSEE Urban units	2,216	–	–	–	–
INSEE Urban units with core	1,147	–	–	–	–
Other delineations					
All UAs	5,771	0.345	0.272	0.868	0.759
UAs with core	812	0.363	0.289	0.847	0.754
Gridded volume UAs	5,294	0.317	0.267	0.874	0.792
Gridded volume UAs with core	679	0.321	0.273	0.838	0.771
Surface UAs	7,011	0.373	0.262	0.854	0.705
Surface UAs with core	1,552	0.403	0.290	0.839	0.711
Population UAs	4,147	0.290	0.249	0.884	0.813
Population UAs with core	337	0.265	0.227	0.815	0.757
Nightlights UAs	2,576	0.384	0.189	0.828	0.561
Nightlights UAs with core	664	0.388	0.201	0.809	0.566
Bandwidth 1km UAs	21,020	0.255	0.230	0.820	0.771
Bandwidth 1km UAs with core	2,221	0.292	0.265	0.834	0.789
Bandwidth 3km UAs	2,891	0.408	0.286	0.896	0.730
Bandwidth 3km UAs with core	561	0.394	0.275	0.843	0.702
Quantile 75 UAs	7,141	0.363	0.252	0.852	0.692
Quantile 75 UAs with core	1,792	0.395	0.278	0.837	0.695
Quantile 99 UAs	5,441	0.305	0.259	0.871	0.793
Quantile 99 UAs with core	580	0.301	0.258	0.826	0.762
Bootstrap 500 UAs	5,837	0.343	0.271	0.869	0.761
Bootstrap 500 UAs with core	930	0.360	0.287	0.846	0.754
Discretised 10% UAs	1,495	0.549	0.197	0.887	0.498
Discretised 10% UAs with core	635	0.505	0.288	0.851	0.628
Discretised 15% UAs	1,438	0.550	0.226	0.889	0.534
Discretised 15% UAs with core	632	0.498	0.295	0.851	0.637
Discretised 20% UAs	1,388	0.533	0.261	0.885	0.597
Discretised 20% UAs with core	621	0.491	0.303	0.851	0.664
Discretised 25% UAs	1,293	0.491	0.273	0.871	0.647
Discretised 25% UAs with core	622	0.474	0.303	0.848	0.675
Discretised 50% UAs	945	0.291	0.223	0.762	0.664
Discretised 50% UAs with core	574	0.353	0.281	0.796	0.705

Appendix C. Supplementary maps

Figure 8: Comparing urban areas with INSEE urban units in four regions



Panel A: Paris and the Ile-de-France region

Panel B: Lille and the North East

Panel C: Marseille and the South East

Panel D: Grenoble and the Alpine region

Notes: Urban areas in light blue (light grey). Urban cores in dark blue (dark grey). Urban units in mauve (very light grey).

Appendix D. Paired Jaccard indices

We can propose another approach to the computation of Jaccard indices. This approach relies on dealing with pairs of pixels instead of single pixels. We will refer extensively to the number of pairs of pixels that can be formed from a given set of pixels. To be clear, if a set has N pixels, the number of pairs is $N\,(N-1)/2$. For a given map $j \in \{1{,}2\}$, we introduce the set W^j that includes all pairs of pixels such that the two pixels constituting each pair are urban. Using counts of pairs of urban pixels, we can readily define the counterpart of the urban Jaccard index proposed in equation (3). The paired urban Jaccard index is given by:

$$J_{UP} \equiv \frac{\left|W^1 \cap W^2\right|}{\left|W^1 \cup W^2\right|}\,. \tag{A1}$$

For a given map $j \in \{1{,}2\}$, denote by W_k^j the subset of pairs within urban area $k \in \{1,...,K\}$. After denoting K^i the number of urban areas in map i, we can also write the paired city Jaccard index as:

$$J_{CP} \equiv \frac{\sum_{k\in K^1}\sum_{k'\in K^2}\left|W_k^1 \cap W_{k'}^2\right|}{\sum_{k\in K^1}\left|W_k^1\right| + \sum_{k'\in K^2}\left|W_{k'}^2\right| - \sum_{k\in K^1}\sum_{k'\in K^2}\left|W_k^1 \cap W_{k'}^2\right|}\,. \tag{A2}$$

Note that the paired city Jaccard is not the exact counterpart of the city Jaccard index defined in equation (4) since pairs of pixels that belong to the same spatial unit on both maps are counted regardless of the identity of this spatial unit. This is an important advantage because it allows us to bypass this issue of the identity of spatial units completely. In a sense, the paired city Jaccard index is less conservative since it counts pairs that belong to the same spatial units but these spatial units can be different across maps. To illustrate this point, we return to the example of Lille which we delineate as one large urban area whereas INSEE delineates several urban units in the same region. When we compute a simple (city) Jaccard index for this urban area, we only count the overlap between 'our' Lille and INSEE's Lille. With a paired Jaccard index, pairs where both pixels belong to Lille on our map and to, say, Valenciennes on INSEE's map will still be counted.

A more conservative possibility is to restrict the paired city Jaccard index to consider only pairs that belong to the same unit:

$$J_{CP2} \equiv \frac{\sum_{k\in K}\left|W_k^1 \cap W_k^2\right|}{\sum_{k\in K^1}\left|W_k^1\right| + \sum_{k'\in K^2}\left|W_{k'}^2\right| - \sum_{k\in K^1}\sum_{k'\in K^2}\left|W_k^1 \cap W_{k'}^2\right|}\,. \tag{A3}$$

This index is closer in spirit to the city Jaccard index described by expression (4) since it only sums across pairs that belong to the same city. It suffers nonetheless from the same drawback as the city Jaccard index in that it requires us to define again the identity of the units.

When assessing the similarity between our baseline delineation and INSEE's delineation of urban units through paired Jaccard indices, we find $J_{UP} = 0.293$, $J_{CP} = 0.631$, and $J_{CP2} = 0.610$.