Delineating urban areas using building density

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

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

Gilles Duranton[*‡]
University of Pennsylvania

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

27 October 2018

Abstract: We develop a new dartboard methodology to delineate urban areas using detailed information about building location, which we implement using a map of 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. We also define the urban cores of these urban areas in a similar manner. Finally, we develop novel one- and two-sided tests to provide a statistical basis to compare maps with different delineations, which we use to document the robustness of our approach and 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 Miquel-Angel Garcia-Lopez, Tomoya Mori, Esteban Rossi-Hansberg, Elisabet Viladecans, and participants at conferences and seminars. 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).

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 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. We also define the urban cores of these urban areas in a similar manner. Finally, we develop novel one- and two-sided tests to provide a statistical basis to compare maps with different delineations, which we use to document the robustness of our approach and 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. We believe both types of definitions are legitimate. 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. Our approach falls into this second category.[3] We rely on data about buildings rather than population since (residential) population data may fail to capture where people are during the day.

[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 on the continent. More systematic and comparable definitions using satellite data show that the 'over-urbanisation of Latin America' is an artefact of lax definitions that categorise even small villages as urban (Roberts, Blankespoor, Deuskar, and Stewart, 2017).

[3] For functional definitions, the term metropolitan area may be more appropriate.

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 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 some settlements form a single unified urban area or two separate ones may depend on a few joining locations which may be close to the threshold of being urban. Statistically, a perfectly reasonable approach may sometimes delineate a single urban area in a region while, some other times, it may delineate two. Hence, it is desirable to assess how much of the differences between two delineations is due to sampling. Alternatively, the difference between one integrated urban area and two separate nearby urban areas may reflects true methodological differences. Being able to assess the importance of sampling is our second main innovation.

Our third challenge is more mundane. We need to retain computational feasibility. We smooth a large number of buildings over a large number of pixels in a three-dimensional space and repeat the exercise many times for counterfactual distributions. Doing this for both our baseline approach and for a number of variants can be computationally overwhelming. Comparing maps using the approach we develop below can be equally daunting from a computational point of view. While demanding, our computations only require weeks, not years, of computer time. Our methodology is thus able to satisfy this implementation challenge.

Finally, we need appropriate high-resolution data describing the built environment of an entire country. We also need detailed data to describe the natural environment to avoid buildings being distributed in the middle of bodies of water or on the peak of the highest mountains in our counterfactual distributions. 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.

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).

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., 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, 2018), a combination of night- and day-time lights (Baragwanath-Vogel, Goldblatt, Hanson, and Khandelwal, 2018) or gridded population data (Dijkstra, Florczyk, Freire, Kemper, and Pesaresi, 2018, Henderson, Kriticos, and Nigmatulina, 2018, Veneri, Boulant, Moreno-Monroy, and Royuela, 2018). 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 Arribas-Bel, Garcia-Lopez, and Viladecans-Marsal (2018), our work belongs to this second group. A different approach is taken by Galdo, Li, and Rama (2018) 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 both difference in 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.

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 about whether a statistically significant concentration of buildings is observed (and at which spatial scale) but not whether a specific group of buildings in an area is spatially concentrated. 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 difficulty 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 to ours but they use it to assess industrial specialisation instead of clusters.

In the remainder of this paper, section 2 presents our data and our preparatory data work. Section 3 describes our methodology to delineate urban areas. Section 4 provides descriptive evidence about our baseline delineation. Section 5 introduces our methodology to assess the similarity between maps in the context of a comparison between our approach and the delineation proposed by the French statistical institute. Section 6 makes a number of further comparisons between maps arising from variants of our approach. Finally, section 7 proposes some concluding thoughts.

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. This dataset is a key component of the large-scale geographical reference for the country and it integrates a variety of pre-existing sources from IGN, satellite images, and the French cadastral information. It contains information on all buildings, including their footprint, height, and use.

Table 1 reports some descriptive statistics for the 33,960,665 buildings in 'mainland' France (which includes a number of small nearby islands but not Corsica nor overseas territoires). Unsurprisingly, there is much variation around the mean footprint of 153 m^2 and the mean volume of 1,057 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

Table 1: Descriptive statistics on buildings

	Min.	25^{th} pctl	Med.	Mean	75^{th} pctl	95^{th} pctl	99^{th} pctl	Max.	St. dev.
Surface (m^2)	0.2	44	93	153	151	421	1,224	579,352	508
Volume (m^3)	0.6	186	465	1,057	849	3,054	10,920	14,483,810	7,325

Notes: Authors' calculations for 33,960,665 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.	25^{th} pctl	Med.	75^{th} pctl	95^{th} pctl	99^{th} pctl	Max.	St. dev.
Raw (m^2)	0	0	0	0	2,155	6,929	579,352	1,484
— (%)	0	0	0	0	5.39	17.3	1,448	3.71
— (m^3)	0	0	0	0	12,417	49,149	14,483,810	27,982
Smoothed (m^2)	0	69	171	343	1,377	4,192	23,104	808
— (%)	0	0.17	0.43	0.86	3.44	10.5	57.8	2.02
— (m^3)	0	394	984	2032	9,011	32,801	456,697	8,056

Notes: Authors' calculations from BD TOPO using 13,628,277 pixels. To keep buildings lumpy, we attribute each building to the pixel that includes the largest share of its area. In extremely rare cases, this leads to a builtup density that exceeds one.

average height of 25 metres. Overall, the footprint of all buildings in France represents 0.94% 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 76% of pixels are unbuilt. Even at the 95^{th} percentile of the distribution of pixel building footprints, only 5.4% 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 17.3% built up. More generally, the distribution of buildings across pixels is highly skewed with a Gini coefficient of 0.933 for builtup volume

[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.

and 0.918 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 by 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 are located high in altitude. 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] Overall, this led us to discard 8.2% of all pixels and we end up with 12,506,581 buildable pixels. Figure 2 represents non-buildable pixels according to these criteria taken separately or together. While high-elevation and steep-sloped pixels are unsurprisingly concentrated around the Alps and the Pyrenees, pixel covered with water are more evenly spread out but nonetheless follow expected patterns and highlight major rivers and lakes.

[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 which 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.

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

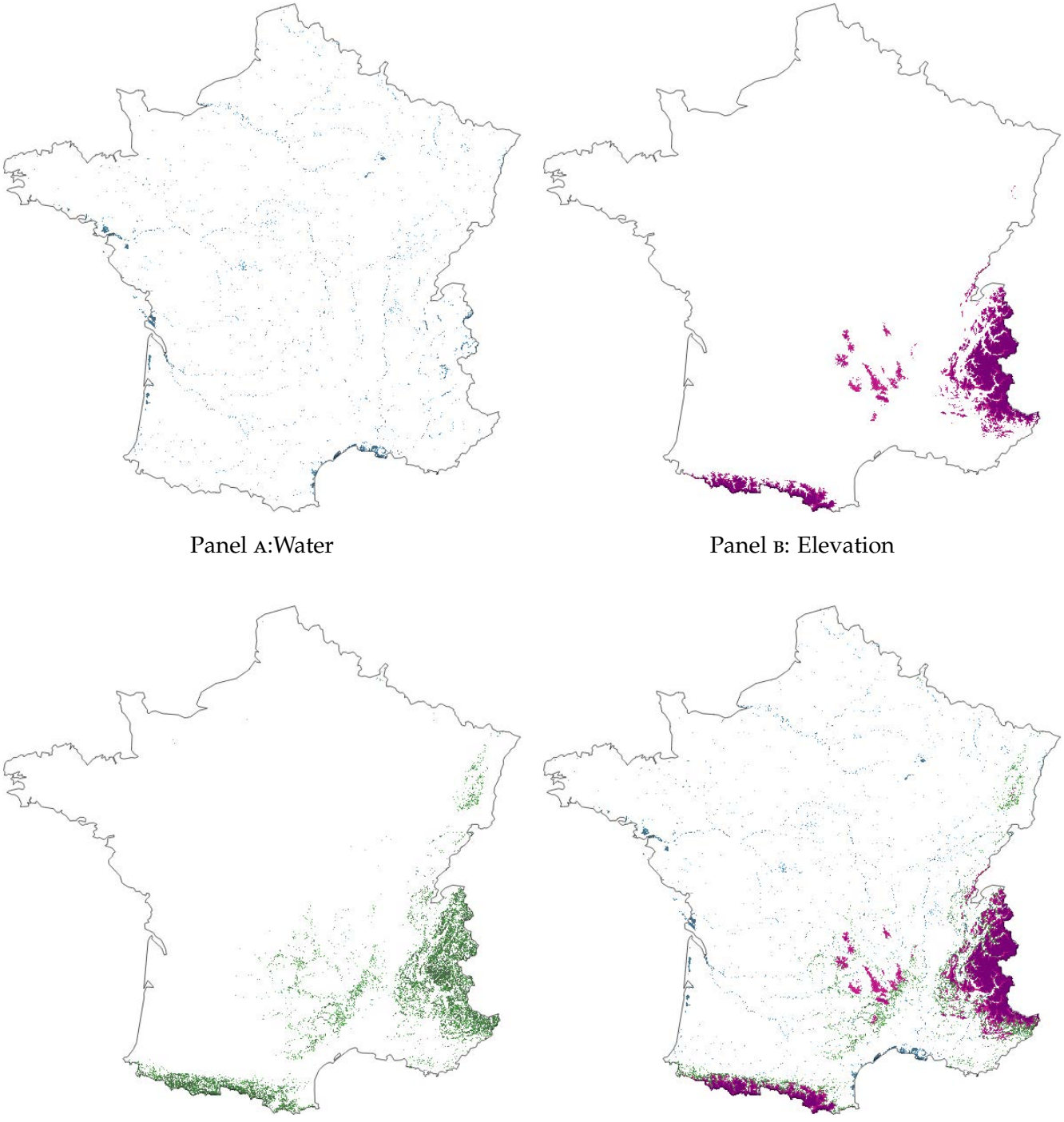
Panel C: Rural area near Grenoble
Google Earth capture

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

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 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.

Figure 2: Non-buildable areas



Panel A:Water

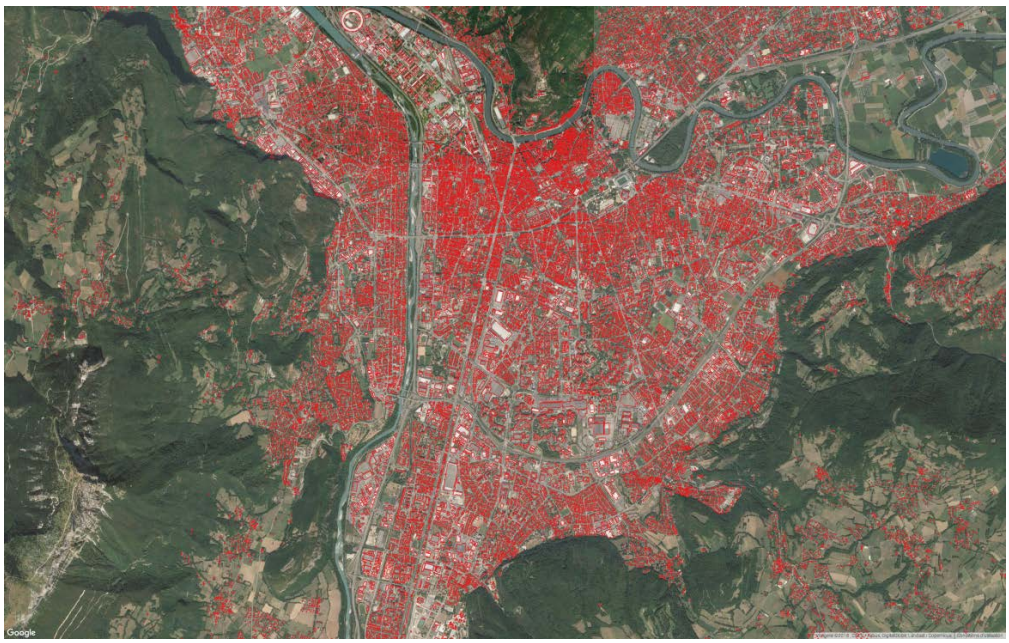
Panel B: Elevation

Panel C: Slope

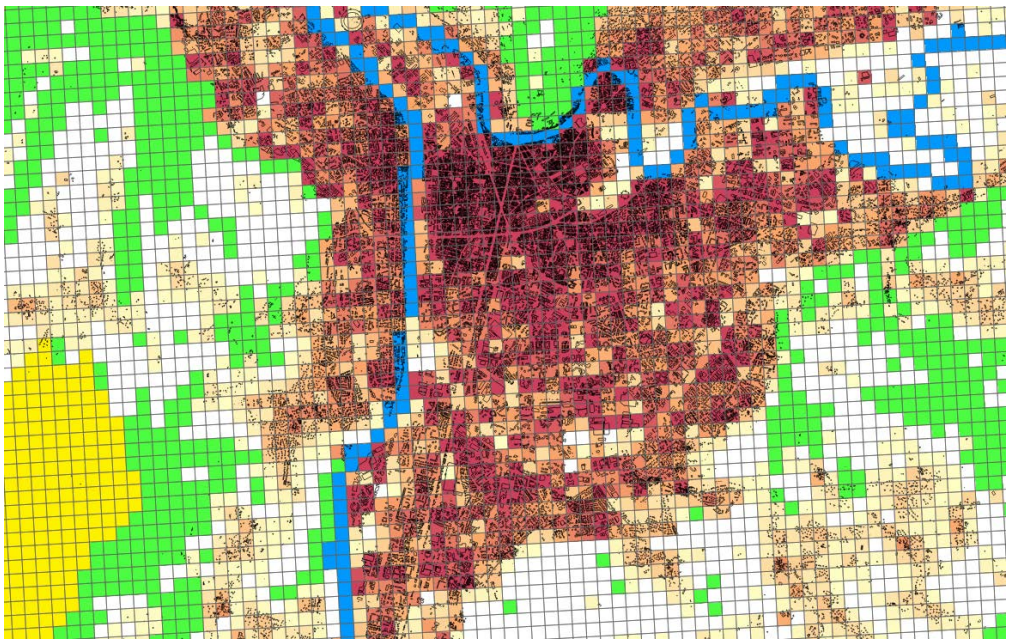
Panel D: Water, elevation and slope

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

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).

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. 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.

As it turns out, this process delineates several thousand urban areas. To avoid having to shorten this long list using an arbitrary population threshold, we also define urban cores. We do so by replicating our analysis a second time for urban pixels only. Our second set of counterfactuals randomly redistributes all buildings in urban pixels 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. Finally, this procedure allows us to distinguish between urban areas that contain one or more cores from those that do not have one.

We now describe some of these steps in greater details.

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:

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

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 split France into partially overlapping tiles of 100 kilometres by a 100 kilometres. In practice, our 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 in the case of a Gaussian kernel.[7] At the same time, we retain a sufficient overlap between our tiles to smooth consistently and avoid any loss of mass from smoothing across tiles.

For our choice of bandwidth, we note the following tradeoff. Taking a large bandwidth will lead to over-smoothed data and make it difficult to identify differences between more or less intensely builtup areas. Taking a small bandwidth will instead lead to under-smoothed data and make it hard to define homogeneous areas. To decide on a bandwidth, we use the following generalised cross-validation criterion. We first compute the 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: $\widehat{z}_j = \sum_{i \neq j} K_h(d_{ij}) z_i$. Next, for each bandwidth h and each tile, we measure the fit between actual building density and smoothed building density (excluding own pixel contribution) using a pseudo-R^2. Then, we determine the optimal bandwidth for each tile as the one that maximises the fit between actual and smoothed density. Finally, we take the median optimal bandwidth across all tiles as our preferred value.[8] Applying this procedure, we end up with a bandwidth of 1.97 kilometres.[9] Smoothing obviously reduces the skew of the distribution of building densities. After smoothing only 3.3% of pixels are 'unbuilt' instead of 76% for raw density. Conversely, the pixel at the 99^{th} percentile of the smoothed density is 10.5% builtup instead of 17.3% in raw data. As a result of smoothing, the Gini coefficient is 0.71 for smoothed builtup density in volume instead of 0.93 for raw 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

[7] In addition, each tile can be processed independently (and in parallel).

[8] We consider tiles separately and take the median and not the average because the optimal bandwidth can be fairly large in some (mountainous) tiles with low building density. 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.

[9] The optimal bandwidth when using building footprint is 1.80 kilometres.

of the deck'. We repeat this procedure to generate one hundred counterfactual building densities for the country.

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. We end up with one hundred smoothed 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 95^{th} percentile of *its* distribution of counterfactual smoothed building densities.[10] We refer to the other pixels as 'rural'. Finally, we define an 'urban area' as a set of contiguous urban pixels.[11]

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 discontinuities and obstacles, 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 formula to compute the density threshold for a pixel to be defined as urban.[12] However, the presence of non-buildable pixels and the irregular geography of the country make such shortcuts problematic. Table 3 documents how the 75^{th}, 95^{th}, and 99^{th} percentiles of the counterfactual distribution of building densities vary across pixels and reports various moments for these thresholds. For instance, the 95^{th} percentile of the counterfactual distribution of building densities to be designated as urban is 3,959m^3 for the median pixel, whereas the 25^{th} percentile is at 3,706 m^3 and the 75^{th} percentile is at 4,228 m^3. Put differently, these are percentiles of the distributions of the 95^{th} percentile of (the distributions of) counterfactual building densities, that is, 'percentiles of percentiles'. While modest at the 95^{th} percentile used to classify a

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

[11] We verify that no urban area is divided because of a river.

[12] 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.

pixel as urban, this variation should not be ignored. We show below that it makes a sizeable difference to the results in some cases.

In part, this variation in the thresholds across pixels arises from our sampling of a finite number of counterfactuals. However, figure 4 shows that this variation also, and perhaps 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.

Although relatively small, the differences across the columns of table 3 cannot be neglected. Thus, we must work with pixel-specific percentiles computed from a full set of counterfactual distributions of buildings instead of computing statistics for a 'representative' pixel or even using a normal approximation.

We can also draw an interesting conclusion from a comparison across rows in table 3. While the 75^{th}, 95^{th}, and 99^{th} percentiles of the counterfactual distribution of building densities obviously differ, the differences between them are also modest relative to the differences observed in the actual distribution. For the median pixel, the 75^{th} percentile of the counterfactual distribution of building densities is at 3,031 m^3 (or 1.09% of the pixel built when measuring density with building footprint), the 95^{th} percentile is at 3,959 m^3 (or 1.30% built), and the 99^{th} percentile is at 5,205 m^3 (or 1.54% built). Table 2 reports that for smoothed building density the bottom quartile is 394 m^3 (or 0.17% built) while building density for a pixel at the 95% percentile is 9,011 m^3 (or 3.44% built). That is, the former pixel will be classified as rural and the latter will be classified as urban, regardless of the threshold we use, 75^{th}, 95^{th}, or 99^{th} percentile. The extreme nature of the distribution of building densities documented in table 2 implies that our delineations of urban areas will be only moderately sensitive to the exact threshold we use.

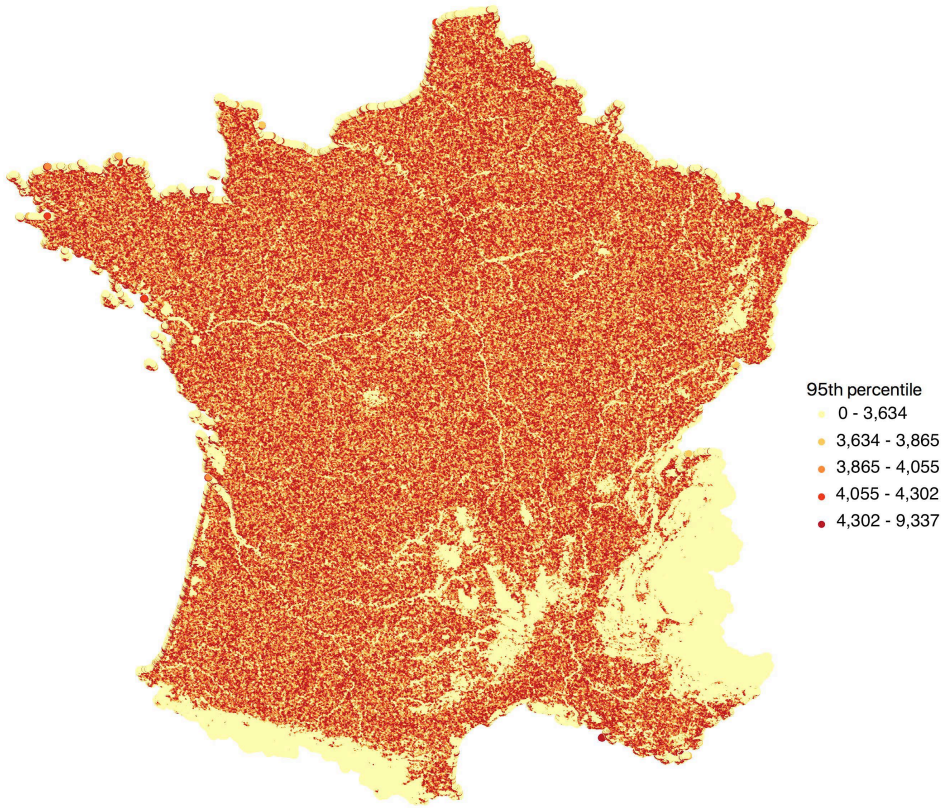
To gain further insight about this important point, figure 5 plots the cumulative distribution functions of smoothed and unsmoothed pixel building densities (measured in m^3 per pixel). Consistent with the very uneven distribution of buildings across pixels in France, these cumulative distribution functions have three starkly different regions. For low building densities, the cumulative distribution function is extremely steep. This reflects the facts that

Table 3: Descriptive statistics on smoothed building density thresholds

Pixel distribution	Min	25^{th} pctl	Med	75^{th} pctl	95^{th} pctl	99^{th} pctl	Max	Std. dev.
Density threshold to be defined as urban:								
Q75 (m^2)	0	427	437	445	456	465	613	71
Q95 (m^2)	0	499	521	542	580	615	949	87
Q99 (m^2)	0	568	615	676	818	1,002	6,510	140
Q75 (%)	0	1.07	1.09	1.11	1.14	1.41	1.53	0.18
Q95 (%)	0	1.25	1.30	1.36	1.45	1.54	2.37	0.21
Q99 (%)	0	1.42	1.54	1.69	2.05	2.51	16.2	0.35
Q75 (m^3)	0	2,932	3,031	3,109	3,222	3,314	4,363	701
Q95 (m^3)	0	3,706	3,959	4,228	4,722	5,190	9,337	976
Q99 (m^3)	0	4,556	5,205	6,098	8,337	11,409	152,606	2,003

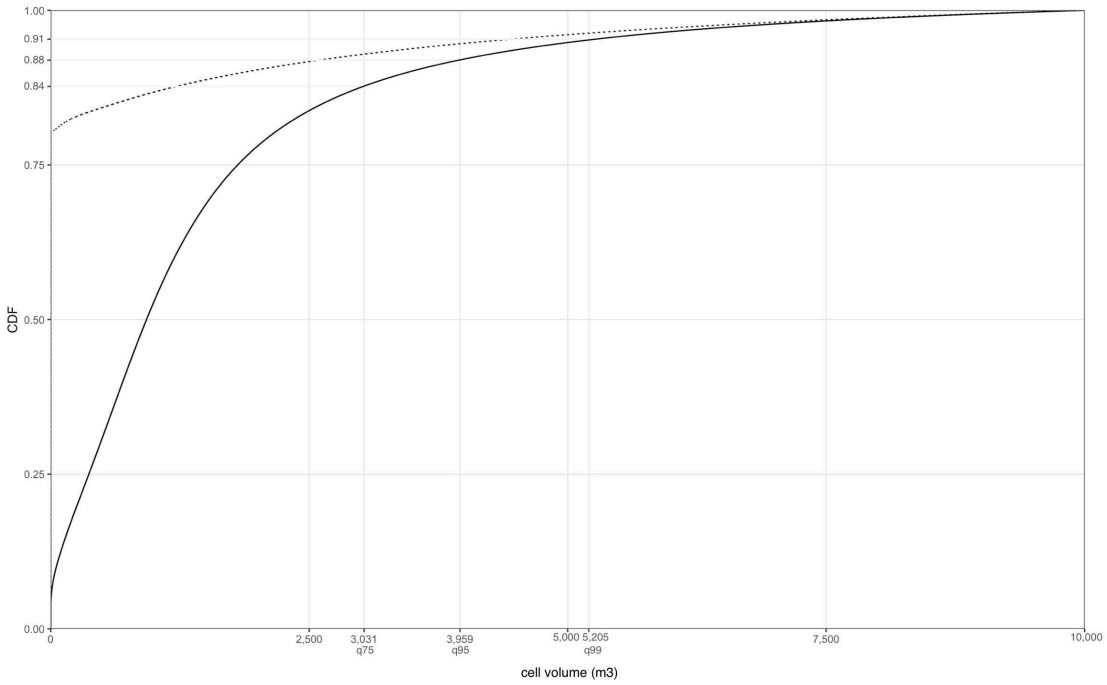
Notes: Authors' calculations from BD TOPO from 13,628,277 pixels. This table reports various moments of the distribution of pixels for three urban thresholds (75^{th}, 95^{th}, and 99^{th} percentiles). From row 8, we can read that a pixel at the bottom quartile in the distribution of all pixels has its 95^{th} percentile of counterfactual building densities at 3,706 m^3 (column 2). The corresponding 95^{th} percentile threshold for the top quartile is 4,228 m^3.

Figure 4: 95^{th} percentile of the distribution of counterfactual smoothed building densities in France



Notes: Authors' calculations. Building density is in m^3 per pixel.

Figure 5: 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.

a large majority pixels are empty of buildings (for unsmoothed building density) or close to empty (for smoothed building density). For high building densities, the cumulative distribution function is instead nearly horizontal which reflects the extreme skew of the distribution of buildings across pixels in the upper tail. In between these two main regions, there is a small intermediate region where the cumulative distribution function is very concave as it transitions from near-vertical to near-horizontal.

On the horizontal axis of figure 5, we also plot the 75^{th}, 95^{th}, and 99^{th} percentile of the distribution of smoothed densities (for the median pixel as per table 3). As could be expected, these thresholds are all in the intermediate region of the cumulative distribution function of pixel building densities. The 75^{th} percentile corresponds to 84% of all pixels being classified as rural while the corresponding figures are 88% and 91% for the 95^{th} and 99^{th} percentiles, respectively.[13] Although these figures obviously differ, their range is limited. Taking the 99^{th} percentile instead of the 95^{th} leads to only a three percentage point difference in the share of

[13]These figures are only rough approximations. As made clear above, each pixel requires to be compared to its own threshold at any level of statistical confidence.

urban pixels.

We draw some important lessons from figure 5. First, the steep part of the cumulative distribution function of pixel building densities indicates a large majority of 'obviously' rural pixels while the flatter part indicates a small minority of 'obviously' urban pixels. Given our purpose, this is good news since we expect all reasonable thresholds to fall into the intermediate region of the cumulative distribution function of pixel building density and agree on 'obviously' urban and 'obviously' rural pixels.

However, there is a middle class of 'marginally urban' pixels in between 'obviously' rural and 'obviously urban' pixels. While, the cumulative distribution function of pixel building density in figure 5 is very concave, it does not exhibit any obvious kink. In the absence of such kink, any binary classification into urban and rural pixels will thus have to wrestle with how to treat this middle class as it needs to define a density threshold. Put differently, the existence of 'marginally urban' pixels that are intrinsically hard to classify implies that small methodological differences are expected to generate different delineations.

Pixels that belong to this middle class of marginally urban pixels are found in smaller settlements or at the periphery of larger settlement since we observe a slowly declining building density away from city centres (Combes, Duranton, and Gobillon, 2019). This is not specific to the geography of French cities. We thus expect this middle class of marginally urban pixels to be found everywhere. More concretely, this implies that disagreements between delineations are expected to concern the periphery of urban areas and smaller settlements.

Three further remarks are in order. First, this 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 are needed to cleanly define this middle class. Since there is no single kink in the cumulative distribution function of pixel building density that allows us to neatly separate urban from rural pixels, there will not be two. Second, this middle class of marginally urban pixels is small relative to 'obviously rural' pixels but seems large relative 'obviously urban' pixels. Figure 5 suggests that perhaps 80% of pixels are obviously rural and maybe 5 to 10% are obviously urban. This leaves 10 to 15% of pixels in the middle class 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 widely in terms of the fraction of pixels they deem to be

urban, the differences will be relatively smaller for population.

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 1.30%, corresponding to a builtup footprint of 521 m^2, or a building volume of 3,959 m^3 to qualify as urban.

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 study 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. To avoid arbitrary minimum size thresholds defined in terms of number of buildings or population, 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 an urban pixel across all buildable 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.

Note that, to define urban cores, we redistribute urban buildings across all urban areas rather than only within their own urban area. While counterfactuals generated from redistributions within urban areas are useful to define centres as significant peaks of building density, they are not useful for our purpose. Even tiny urban areas may have statistically significant centres despite low building density. On the other hand, a large homogeneously-built urban area may lack a statistically significant peak. Instead, we want to define statistically significant peaks of building densities that can be compared across all urban areas.

4. An anatomy of French urban areas

We now describe more fully the output of our baseline delineation approach. Our approach defines 7,223 urban areas that represents 11 % of all pixels (or 12 % of all buildable pixels)

Table 4: Descriptive statistics on pixel built area

Type of urban area	Min.	25^{th}	Med.	Mean	75^{th}	95^{th}	Max.
Panel A: All urban areas (7,223)							
Population	0	271	781	6,810	1,930	10,562	10,932,880
Area	0.04	0.92	2.7	8.9	6	23	3,616
Population density	0	208	331	382	489	892	3,825
Panel B: Urban areas with a core (695)							
Population	0	5,674	10,563	60,127	26,148	176,300	10,932,880
Area	0.04	14	21	60	42	197	3,616
Population density	0	368	504	553	687	1,098	3,023
Panel C: Urban areas without a core (6,528)							
Population	0	233	671	1,134	1,437	3,872	22,746
Area	0.04	0.76	2.3	3.4	4.7	10	61
Density	0	197	314	364	465	856	3,825
Panel D: INSEE urban units (2,231)							
Population	606	3,300	4,817	22,493	8,846	57,809	10,730,549
Area	1.5	20	33	57	140	346	2,854
Population density	7	116	183	250	286	679	3,760

Notes: Population is from the 2013 census; area in km^2; population density is the number of inhabitants per km^2.

in mainland France. Total urban area population is 49,188,740 or 75% of the population of mainland France. Descriptive statistics for all urban areas 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 at the 95^{th} percentile of the distribution of urban areas is still only 10,562.[14] While our approach classifies only a small minority of parcels as urban, it also appears that it only takes a modest concentration of buildings to generate statistically significant excess builtup density.[15]

Panels B and C of table 4 report similar descriptive statistics but distinguish between urban areas with a core and those without. The contrast between the two groups of urban areas is striking. There are only 695 urban areas with a core vs. 6,528 urban areas without a core. However, urban areas with a core have a much higher population. They host on average 60,127 inhabitants instead of 1,134 for urban areas without a core. Overall, urban areas with a

[14]Our approach delineates a small number of urban areas without residents. These are mainly isolated airports and nuclear power plants. Although devoid of residents that call these 'urban areas' home during the nights, some of these buildings or groups of buildings host lots of workers and passengers during the day.

[15]47.9% of 36,248 French municipalities in mainland France have at leat one urban pixel.

core host 64 % of the French population and occupy 7.7% of the French metropolitan territory, while urban areas without a core account for 4.1 % of the French metropolitan territory and 11 % of the French population. Our distinction between urban areas with and without a core does a particularly good job at distinguishing the upper tail of French urban areas from the others.

Figure 6 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, perhaps more surprisingly, along major rivers. The third main feature of this map is that urban pixels are much less prevalent in the mountainous areas of the country.

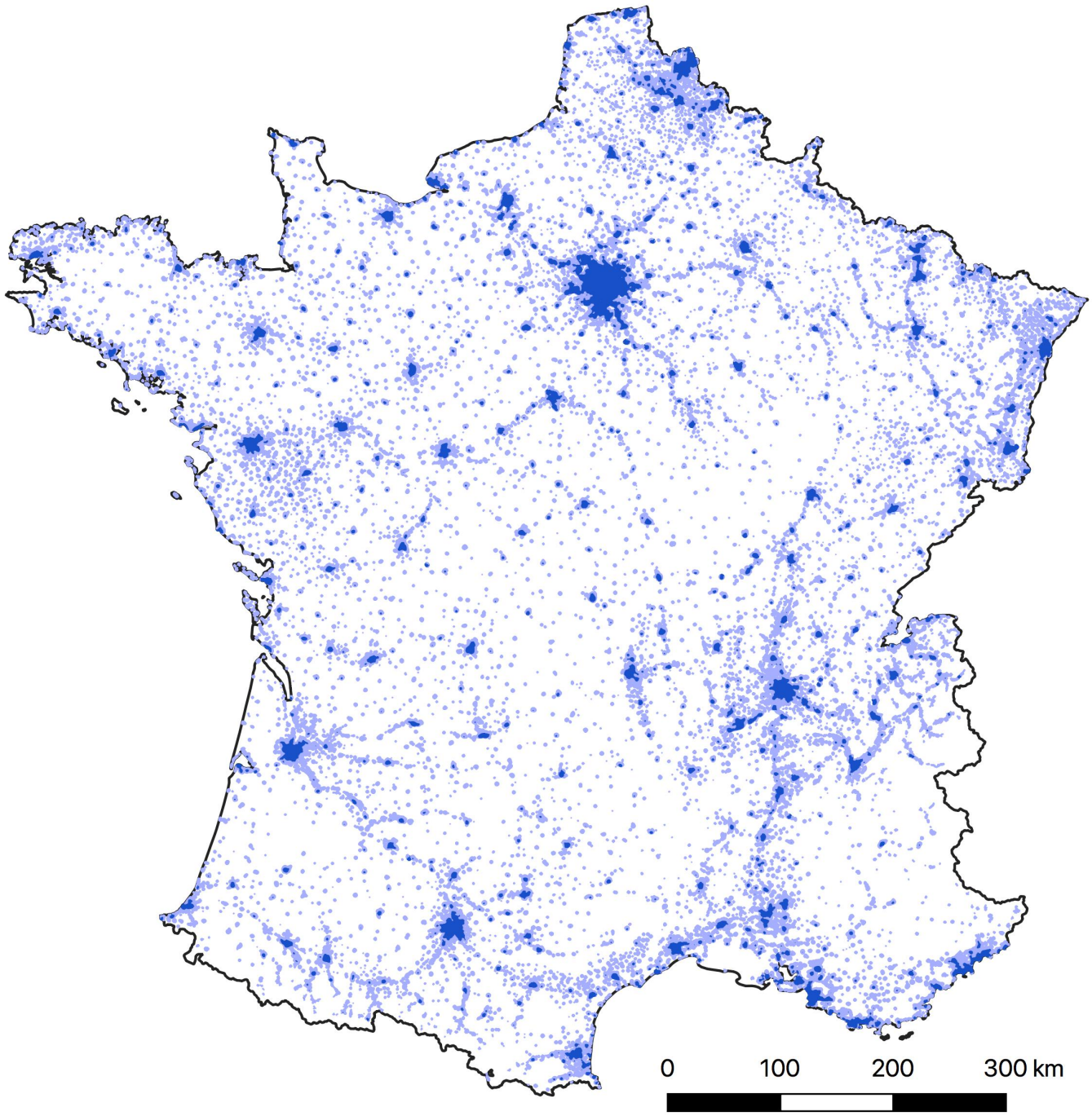
The four panels of figure 7 represent close-ups on the regions of Paris, Lille, Marseille, and Grenoble.[16] Starting with Paris in panel A, the urban area of Paris looks highly monocentric and centred on the municipality of Paris. The urban area branches out in four directions following the river Seine 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 'core' area of Paris covers a large majority of its 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 France. This feature is also true for all large French urban areas.

The Lille urban area, represented in panel B of figure 7, is morphologically extremely different from Paris. It aggregates several large municipalities including Lille itself, Roubaix, 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, with the exception of Arras and Valenciennes, they 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 mountain immediately north 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 and Salon-de-Provence to the north or Toulon to the east. Finally,

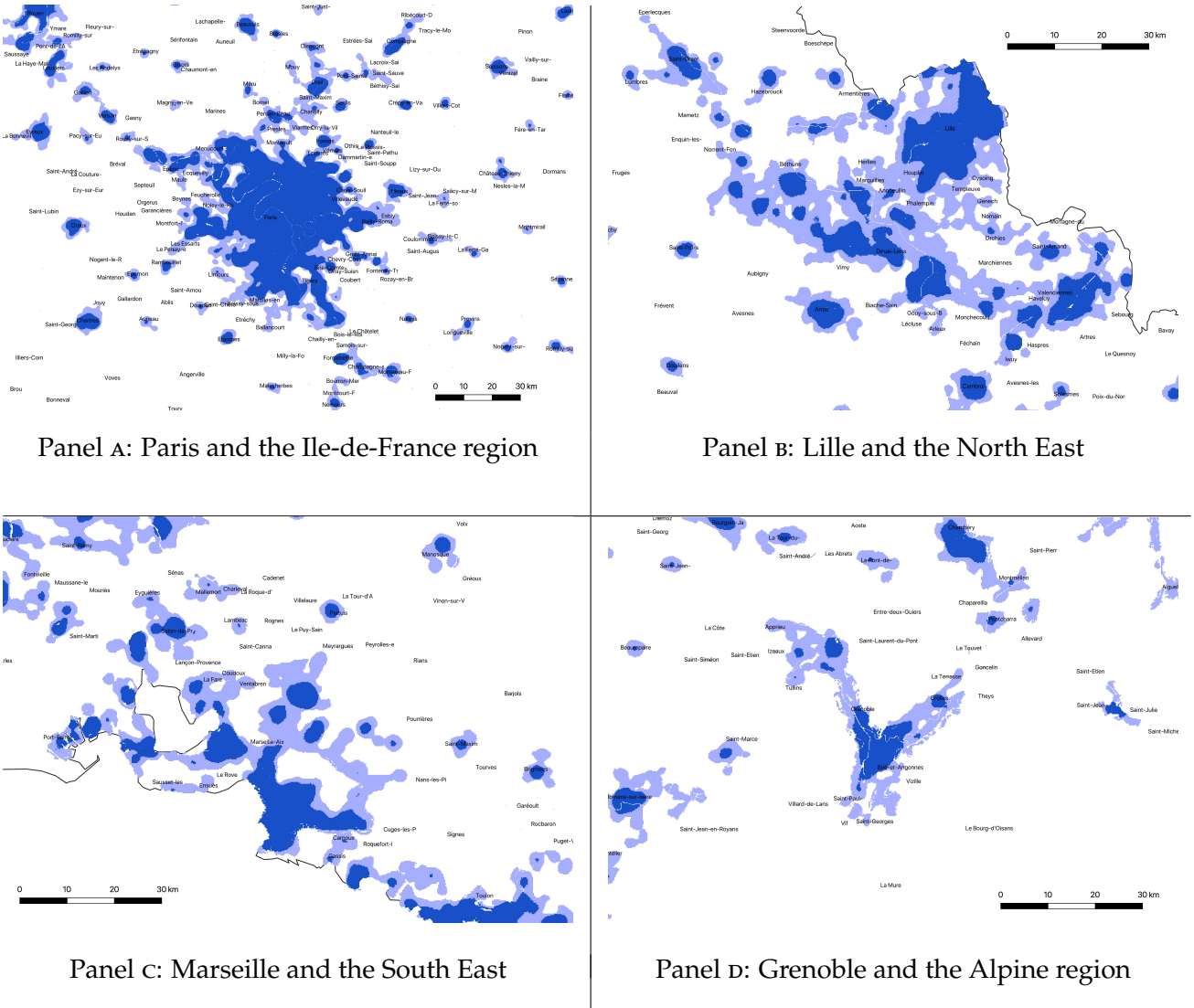
[16]Paris is the largest urban area in France. Lille and Marseille are also among the largest four. Grenoble is a smaller city which we used above to illustrate our treatment of the data. As made clear below, these four cities also differ in interesting ways for our purpose.

Figure 6: Urban areas in France



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

Figure 7: 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: Urban areas in yellow (light grey). Urban cores in red (dark grey).

Grenoble in panel D is much more compact as it is surrounded by 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.

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 contiguous aggregates of French municipalities characterised by an aggregate 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 2,231 INSEE urban units while our approach delineates 7,223 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 a lower bound of 2,000 inhabitants.

Then, the comparison between panels A and D of table 4 further shows that INSEE urban units have a much higher population. 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 2,231 INSEE urban units cover 22% of the French territory instead of 11% for urban areas in our baseline delineation.

When we focus the comparison on the 695 urban areas with a core, we find that, relative to INSEE urban units, our approach selects larger settlements in terms of population though their physical extent remains smaller for most quantiles. This is a general difference between INSEE urban units and our urban areas. Even for the same cities our delineation is less physically expansive. To illustrate this further, figure 8 in Appendix C duplicates figure 7 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.

[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 or to another municipality aggregated to the core. References to alternative approaches are given in the introduction.

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:

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

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 (2) 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 either all urban pixels or only to urban pixels that belong to an urban area with a core. In the first case with all urban pixels, we find $J_U = 0.319$. If we restrict the comparison of INSEE urban units to pixels that belong to the 695 urban areas with a core in our delineation, we compute a Jaccard similarity of $J_U = 0.298$.[19]

Two sources of discrepancy explain this imperfect overlap. First, recall that our approach delineates many more urban areas than there are INSEE urban units when we consider all urban areas and fewer when we consider only urban areas with a core. Overall, the value of the urban Jaccard similarity decreases marginally when we restrict ourselves to urban areas with a core. This restriction eliminates both many small areas that are urban with our delin-

[18] After defining R^j the set of rural pixels on map j, we could instead measure the overlap as $(\left|U^1 \cap U^2\right| + \left|R^1 \cap R^2\right|)/N$ 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 $(N - \left|U^1\right| - \left|U^2\right|)/N$. When $\left|U^1\right|$ and $\left|U^2\right|$ 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 (2) which is more easily interpretable for our purpose.

[19] When comparing with urban pixels that belong to an urban area with a core we use a similar restriction in the INSEE delineation and only consider 849 urban units with a core, as defined by INSEE.

eation but are rural according to INSEE as well as many smaller INSEE urban units.[20] Second, as already noted, INSEE urban units are physically much larger than the ones delineated with our approach.

As discussed above, an imperfect overlap between two maps leading to a Jaccard similarity below one may have two different causes. Methodological differences in the approach used to build these two maps provide a first obvious reason. As made clear above, our delineation approach differs from INSEE's. However, sampling could also explain differences between our map and INSEE's map of urban units. Two different sets of counterfactuals to generate our delineation will only lead to the same results asymptotically. If many pixels are at the margin of being urban, a Jaccard comparison using one map produced from one set of 100 counterfactuals for our delineation may be subject to sampling variation. To assess the importance of sampling variation, we propose to construct confidence intervals for the index proposed in equation (2).

We are in a situation where we wish to compare a map generated with the dartboard approach proposed here and 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.

This test consists in the following. We replicate the delineation approach described in section 3 100 times, which requires $100 \times 100 = 10{,}000$ counterfactuals. This generates 100 delineations that classify each pixel as either rural or urban. For each delineation, we can then compute the index described by equation (2). Finally, we take as reference the median value of the Jaccard similarity and compute a confidence interval around this value. In summary, this procedure amounts to bootstrapping our index relying on the same dartboard approach.

We find extremely small standard errors around the Jaccard indices that we estimate. For all urban pixels, we compute a standard error slightly below 0.0001 around our index $J_U = 0.319$. For a single replication, the largest deviation relative to our reported value for this index is only about 0.2% and thus only affects its third decimal. For all urban pixels that belong to an urban area with a core, the standard error for our index $J_U = 0.298$ is again small, slightly above 0.0001.

[20] When we consider only urban pixels that belong to an urban area with a core, we reduce the numerator of the Jaccard index in expression (2). We also reduce the denominator since we no longer consider some urban pixels that are classified as part of an urban unit by INSEE. This second effect slightly dominates the first so that the Jaccard similarity is marginally lower when we restrict ourselves to urban areas with a core.

To explain such small standard errors, a first possibility could be that our classification of individual pixels is very stable and not subject to sampling variation. While the random distribution of nearly 34 million buildings across nearly 14 million pixels leads to fairly large standard errors for the 95^{th} percentile of 100 draws for each individual pixel, most of that variation may wash out through smoothing across neighbouring pixels. Another possible explanation is that, despite smoothing, there may be a lot of variation at the level of individual pixels with our approach. However, this variation may essentially cancel out through the law of large numbers when computing the numerator of Jaccard indices.[21]

To distinguish between these two explanations, we assess the importance of sampling for our delineation by computing the Jaccard similarity between pairs of our baseline delineations, each obtained from 100 sets of counterfactual distributions. We find that the resulting Jaccard similarity is always above 0.996 which indicates that our classification of individual pixels is stable.

These small standard errors are reassuring. Recall there is no obvious kink or discontinuity in the distribution of building density per pixel that would lead to a natural threshold to classify pixels as urban.[22] These is thus a sizeable proportion of 'marginally urban' pixels. However, the variation caused by sampling is much smaller. We can thus achieve precision in our delineations despite some ambiguity with respect to what may be classified as urban in the periphery of cities or with smaller settlements. This said, as we show below, this ambiguity caused by marginally urban pixels implies that delineations are sensitive to methodological choices and the exact criteria being used despite being robust to sampling.

In addition, these small standard errors also indicate that 100 sets of counterfactuals are generally enough to achieve a high level of precision in our computations since the gains in precision from using 100 times as many counterfactuals are minimal while the computational costs of doing so are high.

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	10,932,881	3,023	10,730,549	1	0.657
2	Lille	2,197,967	1,170	1,037,834	4	0.226
3	Lyon	1,777,944	1,115	1,627,937	2	0.485
4	Marseille	1,442,734	1,562	1,570,325	3	0.440
5	Nice	1,024,679	1,587	956,189	6	0.654
6	Toulouse	875,595	990	938,284	5	0.652
7	Bordeaux	831,453	1,004	893,384	7	0.466
8	Strasbourg	692,009	857	451,522	13	0.223
9	Nantes	587,495	1,136	628,718	8	0.534
10	Grenoble	520,445	1,234	518,495	10	0.483
11	Metz	498,052	873	292,007	22	0.319
12	Rouen	495,561	1,097	475,182	12	0.122
13	Toulon	486,616	1,524	570,591	9	0.354
14	Montpellier	460,796	1,216	421,031	15	0.493
15	Avignon	405,798	546	457,857	14	0.439
16	Mulhouse	344,118	882	252,001	26	0.405
17	Perpignan	342,646	656	201,282	33	0.325
18	Nancy	336,196	1,158	288,742	21	0.396
19	Saint-Etienne	329,258	1,107	379,791	16	0.334
20	Rennes	299,315	1,175	331,661	20	0.475

Notes: Population is from the 2013 census; area in km^2; density is the number of inhabitants per km^2. Jaccard similarity by urban area as per equation (5)

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, 16 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 2.2 million, this is unambiguously the second largest urban area

[21] 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.

[22] This is illustrated by figure 5 which shows a smooth cumulative distribution function for pixel building density. While the 95^{th} percentile of counterfactual building distribution for the median pixels reassuringly falls into its most concave region, there is no obvious kink that would provide a natural threshold to define which pixels are urban.

in France in our baseline delineation. According to INSEE, the urban unit of Lille ranks fourth with only about 1 million inhabitants. This gap is consistent with our discussion of panel B of figure 7 above. Our approach evidences a large continuous builtup area around Lille. Instead, INSEE delineates four separate urban units. We also observe some differences for Strasbourg, Metz 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 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 two different 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 7, 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 7, our approach delineates a large urban area that we naturally, but perhaps loosely at this stage, call "Lille". In the same region, INSEE delineates several urban units, 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? Our delineation of the urban area Lille also includes the city of Valenciennes. 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 (an urban unit distinct from Lille and much smaller than Lille in the INSEE delineation). Put differently, we need to know when a spatial unit is the 'same' across two maps which delineate them differently.

To define the identity of the urban areas delineated by our approach, we proceed as follow. We name each urban area after the municipality with the largest population it overlaps

with.[23] 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 (and it contains its centroid). This approach is consistent with the naming convention of urban units by INSEE, 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:

$$J_C = \frac{\sum_{k \in K} \left|U_k^1 \cap U_k^2\right|}{\left|U^1 \cup U^2\right|}, \tag{3}$$

after dropping the superindex from variables that compare map 1 and map 2 to lighten the notations, that is noting this index J_C instead of J_C^{12}.

The key difference between the urban Jaccard similarity defined by equation (2) and the city Jaccard similarity defined by equation (3) is the following. The urban Jaccard similarity '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 (2) and (3) that:

$$J_C \equiv J_U \times P, \qquad \text{where } P \equiv \frac{\sum_{k \in K} \left|U_k^1 \cap U_k^2\right|}{\left|U^1 \cap U^2\right|}. \tag{4}$$

The city Jaccard similarity, which measures the overlap between urban pixels that belong to same urban area(s), can thus 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 quality 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.

[23] 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:

$$J_k \equiv \frac{\left|U_k^1 \cap U_k^2\right|}{\left|U_k^1 \cup U_k^2\right|}, \tag{5}$$

with $J_k = 0$ if urban area k is missing from either map. We can now show that the city Jaccard similarity in equation (3) can be decomposed into the weighted sum of individual Jaccard similarity J_k calculated for every spatial unit 1,...K. From equations (3) and (5), we can write:

$$J_C = \sum_{k \in K} s_k J_k, \qquad \text{where } s_k \equiv \frac{\left|U_k^1 \cup U_k^2\right|}{\left|U^1 \cup U^2\right|}. \tag{6}$$

Hence, the city Jaccard similarity is the weighted sum of the individual Jaccard similarity 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 approached described above for the bootstrapping of urban Jaccard indices.

For the comparison between the official INSEE map of 2,231 urban units and our preferred delineation with all urban pixels of 7,223 urban areas, we obtain $J_C = 0.177$. For the comparison between the official INSEE map of urban units and our preferred delineation where we restrict ourselves to 849 INSEE urban units with a core and 695 urban areas with a core, we obtain $J_C = 0.182$.

The main point to note is that these two indices are well below the values of 0.319 (all urban pixels) and 0.298 (pixels that belong to urban areas with a core) for their corresponding urban Jaccard indices obtained above. Using expression (4), this suggests a fairly low quality for the propensity of the two maps to aggregate urban pixels into the same units, between 0.55 and 0.61.

Another way to understand those figures 0.177 and 0.182 for the city Jaccard similarities is to return to expression (6), 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 0.657 and a much lower value for Lille of 0.226 while the indices for Marseille and Grenoble are in between these two cases at 0.440 and 0.483 respectively. For the largest 20 cities, the average (weighted) city Jaccard similarity is 0.46. This is more than the overall value of either 0.177 or 0.182 for the overall city Jaccard similarity. In turn, this suggests that these low overall values

arise 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.[24]

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 0.0003 for both $J_C = 0.177$ (all urban pixels) and $J_C = 0.182$ (all urban pixels that belong to an urban area with a core). We note nonetheless that these standard errors are about three 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.

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

[24]Recall that we compare either 7,223 urban areas and 2,231 urban units or 695 urban areas and 849 urban units when considering only areas with a core.

as rural. This is similar in spirit to the approach used by INSEE.[25] When comparing our discretised delineation with INSEE urban units, we obtain an urban Jaccard index of 0.402 for all urban areas and 0.397 for all urban areas with a core. These values for the urban Jaccard indices are higher than their respective values of 0.319 and 0.298 obtained above for the comparison with our baseline delineation. This improvement in the overlap between delineations is consistent with the notion that part of the difference between our delineation and INSEE's delineation of urban units is due to their use of much larger discrete units.[26]

This said, Jaccard similarity indices of about 0.40 instead of about 0.30 are still indicative of large differences between the two delineations. Recall that our approach is generally more conservative than INSEE's and delineates physically less expansive urban areas. Our discretisation with a 50% threshold will improve the similarity with INSEE's delineation in some cases when a municipality is 'rounded up'. However, this discretisation will also worsen the similarity with INSEE's delineation when municipalities that INSEE classifies as urban but which, with our delineation, contain less than 50% of urban pixels and are thus 'rounded down'. 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 20% or more of its pixels are urban, the urban Jaccard index for the comparison with INSEE's delineation rises further to 0.562 when considering all urban pixels or to 0.553 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 20% threshold appears to maximise the similarity with INSEE's delineation. Alternative thresholds of 25, 15, and 10% yield marginally lower Jaccard indices relative to a 20% threshold.

Another possible explanation for the limited overlap between our delineation and INSEE's delineation of urban units may lie in our use of a 95% statistical threshold, which may lead to

[25] 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.

[26] We focus here on differences arising from the size of the underlying units but this is obviously not the only source of discrepancy between the two delineations. In particular, we measure density using builtup volumes, while INSEE relies on the distance between buildings, essentially a footprint criteria. When we implement our delineation using builtup footprint instead of volume to measure density, we compute a Jaccard similarity with INSEE's delineation of 0.35 for all urban pixels and 0.54 for pixels that belong to an urban area with a core. These higher values are due to the more expansive physical extent of our delineation using builtup footprints as shown below.

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 similarity between our delineation and INSEE's falls from 0.319 to 0.241. However, taking a much less conservative statistical threshold of 75% only increases the Jaccard similarity to 0.342.

Taking a less conservative threshold has three effects on our delineation. First, it leads to an expansion of the largest urban areas that are also delineated by INSEE. This contributes to improving the similarity between the two maps since our urban areas are physically less extensive than those delineated by INSEE with a threshold of 95%. 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 similarity 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.

As discussed above, a second important difference between our delineation and INSEE's urban units is the greater propensity of our approach to aggregate builtup areas that are close to each other into a single urban area as, for instance, in the case of Lille. Discretising our delineation will in general have ambiguous effect on this difference. Using high thresholds to classify entire municipalities as urban can lead our approach to split hitherto integrated urban areas into separate urban areas. This occurs when two groups of urban pixels are joined by urban pixels that are part of a municipality that our discretisation classifies as rural. Using low thresholds may instead worsen this aggregation problem as municipalities now classified as urban may bridge between hitherto separate urban areas.

Recall that the city Jaccard index when comparing INSEE's delineation of urban units with our baseline delineation is $J_C = 0.177$ when considering all urban areas and $J_C = 0.182$ when considering only pixels that are part of an urban area with a core. When we discretise our delineation using a threshold of 50% of urban pixels for municipalities, the city Jaccard indices remain mostly unchanged at 0.171 and 0.173, respectively. When we use a lower threshold of 20%, the Jaccard indices are again barely affected at 0.182 and 0.186, respectively. Returning to equation (4), city Jaccard indices J_C can be decomposed into the product of their corresponding urban Jaccard indices J_U and an overlap quality factor. As discussed above, the municipal discretisation of our preferred delineation index leads to higher urban Jaccard indices. Since the city Jaccard indices are essentially the same, it must be that the discretisation of our preferred delineation worsens the overlap quality factor. In turn this oc-

curs because our discretisation tends, on average, to aggregate urban pixels into fewer urban areas. In 'dense' regions like around Lille or Marseille, the discretisation of municipalities magnifies the tendency of our approach to delineate large single urban areas while INSEE defines many adjacent urban units (as illustrated for instance by figure 8 in Appendix C).[27]

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.

Table 6 in Appendix B duplicates table 4 and report descriptive statistics for the urban areas delineated with this alternative definition. Figure 9 in Appendix C duplicates the four maps of figures 7 but also overlays these alternative urban areas over those defined using our preferred approach.

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 15 % of all pixels (instead of 11% with volumes) and host 80% of the French metropolitan population (instead of 75%). 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 $J_U = 0.79$. We note that this figure for the Jaccard similarity is also essentially equal to the ratio of urban pixels across both delineations. Consistent with the visual impression of the four illustrations of figure 9, this indicates that urban pixels when

[27]For instance, with a 20% threshold, the urban area of Marseille is now joined with Toulon to the east and Lille is joined with Saint Omer to the west.

using a volume-based definition of building density are essentially 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, we obtain roughly similar results with $J_U = 0.76$, which is consistent with the interpretation just given.

Turning to city Jaccard indices we compute $J_C = 0.49$ for all urban pixels and $J_C = 0.58$ 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 illustrated in panel B of figure 9. 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. As a result, the city Jaccard J_{Lille} is only 0.58. While this value is lower than the Jaccard index of 0.84 we find for Paris, Lille is not a pathological case. For instance, we find values of 0.61 for Nantes or Avignon and even 0.51 for Rennes.

Because a footprint-based definition of urban density leads to physically larger urban areas, we expect it to lead to a map of urban areas that is closer to the official delineation of urban units by INSEE. We can verify that this is the case since the urban Jaccard index for these two delineations is $J_U = 0.350$ for all urban pixels and the city Jaccard index is $J_C = 0.196$. Recall that when we compare our baseline delineation that defines building density with builtup volumes, we find lower indices of 0.319 and 0.177 respectively.[28]

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 0.0001. As mentioned before, any two delineations we obtain from different sets of 100 counterfactual distribution 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.

[28] When, in our delineation, we only consider urban pixels that belong to an urban area with a core we obtain $J_U = 0.540$ and $J_C = 0.199$ instead of 0.298 and 0.182 for our volume-based definition.

Urban areas defined using different thresholds

Our final comparison regards the effect of the statistical thresholds we use to define excess building density. Until now, we have focused on a 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 0.719 for all urban pixels and 0.673 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.

Our value of 0.719 for the urban Jaccard index corresponds to 15.5% of urban pixels with the 75% threshold and 11% of urban pixels with the 95% threshold. This figure of 0.719 is also consistent with a simple heuristic derived from figure 5. This figure indicates that the 75^{th} percentile of the counterfactual distribution of building density for the median pixel corresponds also to the 84^{th} percentile of the actual distribution of smoothed building density, while the 95^{th} percentile of the counterfactuals for the median pixel is around the 88^{th} percentile of the actual distribution.[29] As a first approximation, figure 5 thus implies that with a 75% threshold, 16% of pixels are urban (instead of 15.5% with our exact delineation), while with a 95% threshold about 12% of pixels are classified as urban (instead of 11% with our exact delineation) leading to an approximate Jaccard index of about 0.75, close to the exact figure of 0.719.

Turning to the comparison between delineations obtained with the 95 and 99% threshold for statistical significance, we find an urban Jaccard index of 0.594 when considering all urban pixels and 0.527 when considering only urban pixels part of a urban area with a core. These values are less consistent than previously with the figures implied by figure 5. In this figure, the 99^{th} percentile of the counterfactual distribution of building density for the median pixel corresponds to about the 91^{st} percentile of the distribution of actual buildup density. The

[29]This is only an approximation since, as described in table 3, each pixel faces a different value for any given threshold of statistical significance depending on its location, particularly its location relative to non-buildable pixels and coasts and borders.

approximation based on figure 5 thus suggests that 9% of pixels are urban with the 99^{th} percentile instead of 12% with the 95^{th} percentile. This thus implies an approximate similarity of about 0.75 instead of the value of 0.594 that we compute for the urban Jaccard index.

This discrepancy mainly occurs because our exact delineation classifies only 6.6% of pixels as urban when using a 99% threshold for each pixel instead of the 9% urban pixels implied by the use for all pixels of the 99^{th} percentile for the median pixel. In turn, this difference arises because the approximation based on the percentiles for the median pixels does not work very well for pixels that are close to the sea and rivers and thus have a lower significance threshold relative to the median pixel due to their closeness to non-buildable pixels. As shown by figure 6, many of these pixels are classified as urban with a 95% threshold. These pixels are then often rural with a 99% threshold. This difference illustrates the need for pixel-specific counterfactual distributions and thresholds instead of using a representative pixel.

For both the 75-95 and the 95-99 comparisons, the city Jaccard indices take lower values of 0.426 and 0.325, respectively. These lower values are unsurprising. Recall again that by equation (4) 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.

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 11% of mainland France as urban and 75% of the French population is

urbanised. Our approach delineates 7,223 urban areas, most of which are tiny. When we only consider urban areas with a 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 695. These urban areas cover less than 8% of the French territory but still host 64% 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 differences in the delineation approach will lead to 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 15% of the French territory instead of 11%. On the other hand, these 'ambiguous 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 extremely tight. While the choices made in the delineation approach matter, sampling issues do not.

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. This granularity allows us to detect patterns that were previously harder to detect. For instance, we find that excess building density tends to follow major rivers and not only coastal areas (Rappaport and Sachs, 2003). Consistent with this finding, we also find that small urban areas locate close to rivers and large urban areas expand along their main rivers.

Obviously, the sort of approach we develop here could be used to delineate urban areas in other countries. Our approach could also be extended to detect centres and subcentres within cities by considering random redistributions of buildings within urban areas instead of across them. We believe our approach 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

Arribas-Bel, Daniel, Miquel-Angel Garcia-Lopez, and Elisabet Viladecans-Marsal. 2018. Building(s and) cities: The role of transportation and geography. Processed, University of Barcelona.

Baragwanath-Vogel, Kathryn, Ran Goldblatt, Gordon Hanson, and Amit K. Khandelwal. 2018. Mixing satellite imagery to define urban markets: An application to India. Processed, University of California San Diego.

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.

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* forthcoming.

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. 2018. Cities, skills, and sectors in developing economies. Processed, Columbia University.

Dijkstra, Lewis, Aneta Florczyk, Sergio Freire, Thomas Kemper, and Martino Pesaresi. 2018. 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.

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.

Galdo, Virgilio, Yue Li, and Martin Rama. 2018. Identifying urban areas combing data from the ground and outer space: An application to India. Processed, World Bank.

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. 2018. Measuring urban economic density. Processed, London School of Economics.

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.

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.

Rappaport, Jordan and Jeffrey D. Sachs. 2003. The United States as a coastal nation. *Journal of Economic Growth* 8(1):5–46.

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.

Veneri, Paolo, Justine Boulant, Ana I. Moreno-Monroy, and Vicente Royuela. 2018. Urban agglomerations in the world. Testing a global identification. Processed, OECD.

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. The fraction of the pixel that is covered by water can be larger than one in extremely rare cases because of the overlap between river buffers and another body of water. We cap this fraction to one.

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

Localised Tax Revenues (INSEE, 2010). 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. Supplementary table

Table 6 duplicates table 4 when density is measured with footprint area instead of builtup volumes.

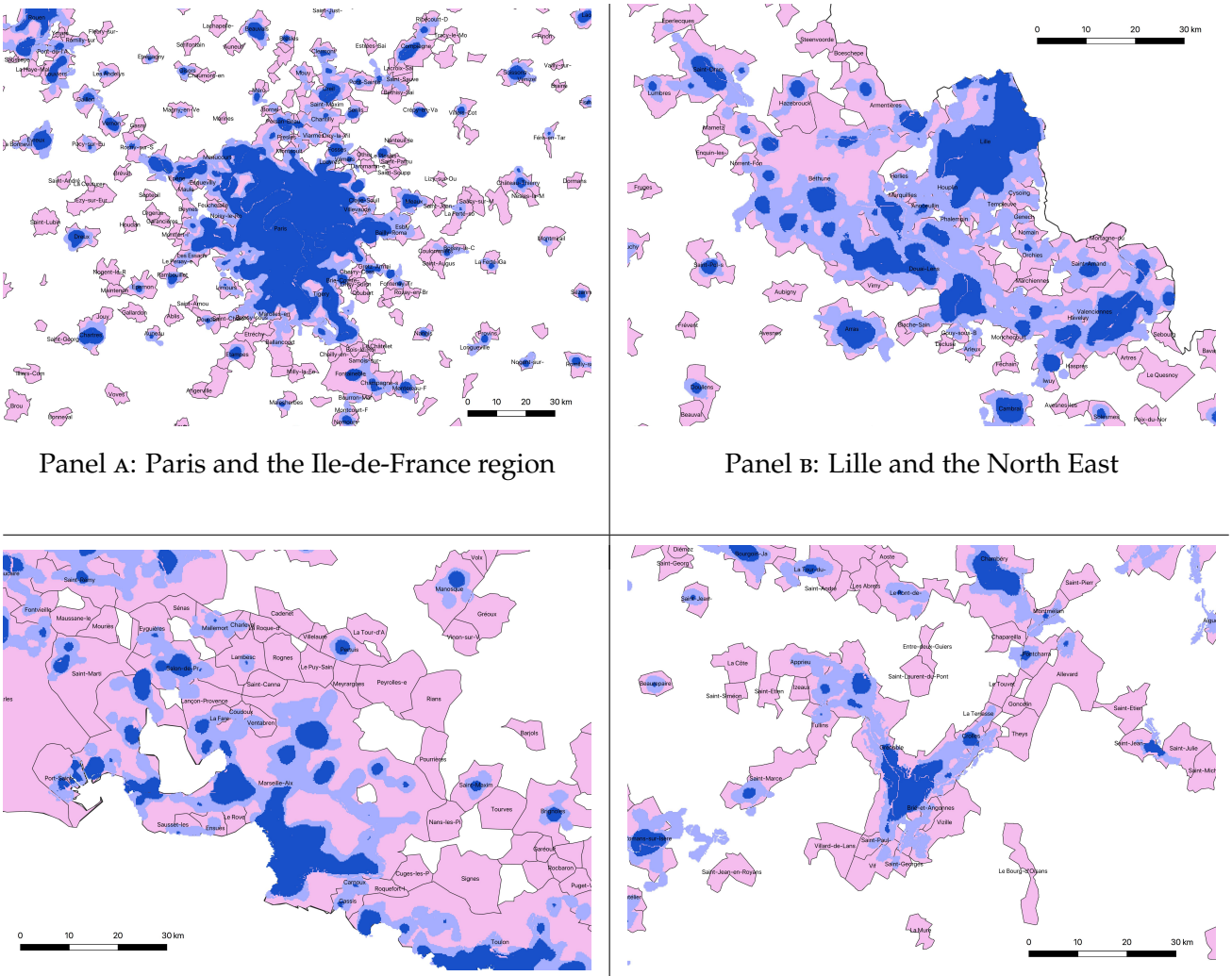
Table 6: Descriptive statistics on pixel built area when measuring building density with footprint area

Type of urban area	Min.	25^{th}	Med.	Mean	75^{th}	95^{th}	Max.
Panel A: All urban areas (8,482)							
Population	0	323	688	6,186	1,534	8,071	11,250,140
Area	0.04	1.04	2.68	9.87	5.96	23.4	4,229
Population density	0	187	257	298	352	642	3,700
Panel B: Urban areas with a core (1,025)							
Population	15	3,591	5,895	42,270	13,528	99,884	11,250,140
Area	0.04	11.6	18.4	58.9	37.3	203	4,229
Population density	1.82	272	347	388	458	730	2,616
Panel C: Urban areas without a core (7,457)							
Population	0	284	576	829	1,065	2,407	13,776
Area	0.04	0.88	2.24	3.14	4.32	8.84	69.7
Population density	0	179	243	285	334	622	3,700

Notes: Population is from the 2013 census; area in km^2; population density is the number of inhabitants per km^2.

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

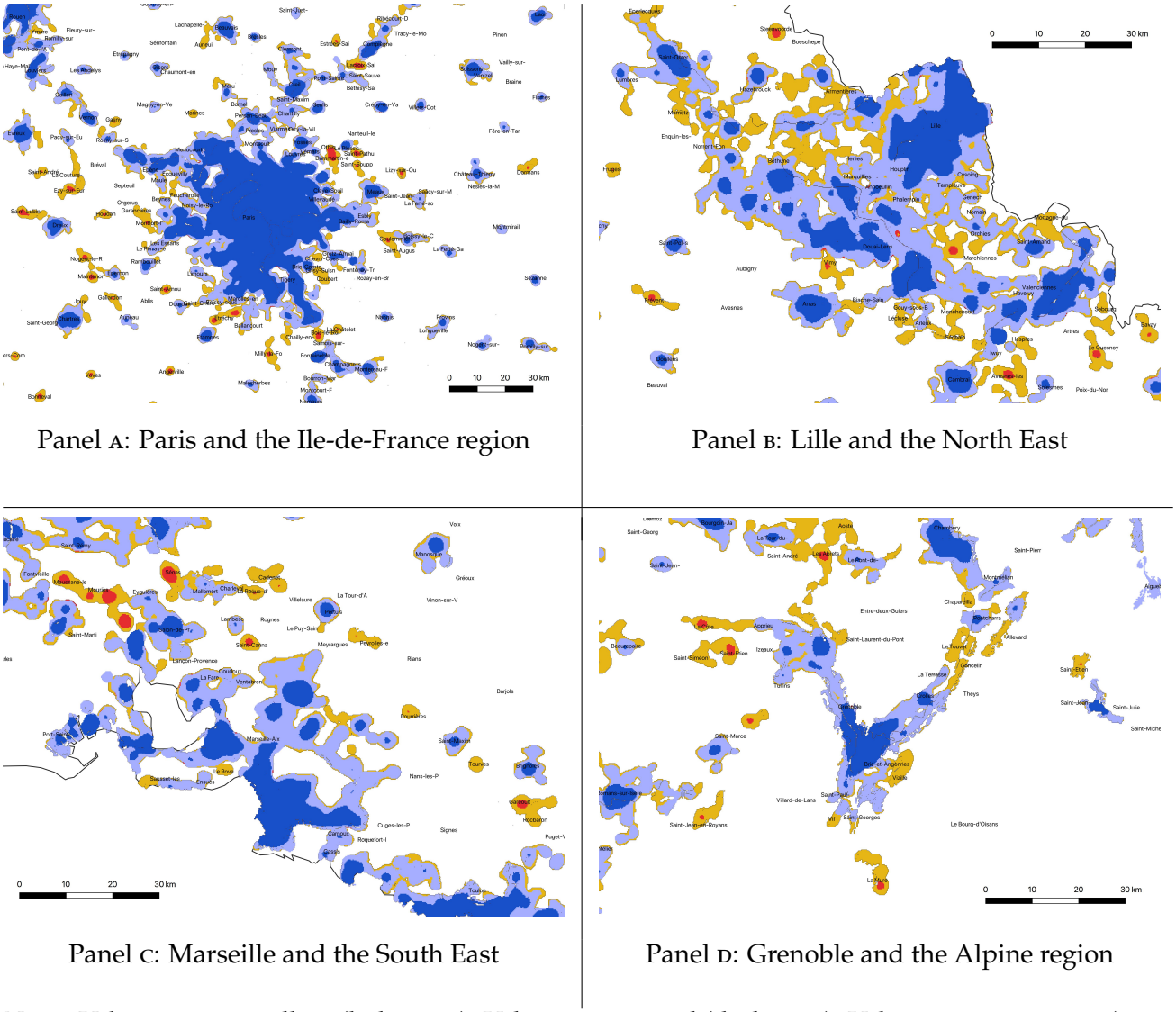
Panel B: Lille and the North East

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).

Figure 9: Comparing urban areas delineated with building density using volume vs. footprint area 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 yellow (light grey). Urban cores in red (dark grey). Urban units in mauve (very light grey). Urban areas defined using building footprint in yellow and their core in red.

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 (2). 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 (3) 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. Conceptually, the paired city Jaccard index also captures something different from the simple city Jaccard index. 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 understand 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 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 (3) 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 $J_{UP} = 0.240$, $J_{CP} = 0.545$, and $J_{CP2} = 0.510$.