

A PARTITION-FREE SPATIAL CLUSTERING THAT PRESERVES TOPOLOGY: APPLICATION TO BUILT-UP DENSITY

Gaëtan Montero, Geoffrey Caruso,
Mohamed Hilal, Isabelle Thomas

REPRINT | 3210

CORE

Voie du Roman Pays 34, L1.03.01

B-1348 Louvain-la-Neuve

Tel (32 10) 47 43 04

Email: lidam-library@uclouvain.be

<https://uclouvain.be/en/research-institutes/lidam/core/core-reprints.html>

A partition-free spatial clustering that preserves topology: application to built-up density

Gaëtan Montero^a, Geoffrey Caruso^b, Mohamed Hilal^c, Isabelle Thomas^a

Accepted in Journal of Geographical System (september 2022)

^a Centre of Operations Research and Econometrics, Université catholique de Louvain, Belgium

^b Department of Geography and Spatial Planning, University of Luxembourg (UL) and Luxembourg Institute of Socio-Economic Research (LISER), Urban development and Mobility, Luxembourg

^c CESAER, INRAE, l'Institut Agro, Univ. Bourgogne-Franche-Comté, France

Abstract

Urban density is central to urban research and planning and can be defined in numerous ways. Most measures of urban density however are biased by arbitrary chosen spatial units at their denominator and ignore the relative location of elementary urban objects within those units. We solve these two problems by proposing a new graph-based density index which we apply to the case of buildings in Belgium. The method includes two main steps. First, a graph-based spatial descending hierarchical clustering (SDHC) delineates clusters of buildings with homogeneous inter-building distances. A Moran scatterplot and a maximum Cook's distance are used to prune the minimum spanning tree at each iteration of the SDHC. Second, within each cluster, the ratio of the number of buildings to the sum of inter-building distances is calculated. This density of buildings is thus defined independently of the definition of any basic spatial unit (BSU) and preserves the built-up topology, i.e. the relative position of buildings. The method is parsimonious in parameters and can easily be transferred to other punctual objects or extended to account for additional attributes.

Keywords: Density; Topology; Graph; Moran scatterplot; Buildings

1 Introduction

Density is the ratio of a mass - typically a number of individuals, jobs, or buildings - to a given two-dimensional reference, i.e. an area. In urban and planning practices, different density metrics are used (Churchman, 1999; Longley & Mesev, 2002; Angel et al., 2021) to represent the functioning of a city and, more particularly, to hint at the sustainability or liveability of urban forms (Pauleit & Duhme, 2000; Boyko & Cooper, 2011; Ewing et al., 2018; Rinkinen et al., 2021; Martino et al., 2021). Density metrics, however, are subject to two major issues: they use a reference area which is not necessarily related to research goals, and they ignore the relative locations of spatial elements within that reference area, thus aggregate without making sure of the internal homogeneity of the area. These problems were already stressed within the axiomatic approach to geographical space of Beguin & Thisse (1979) who showed that the elementary area considered to measure a density cannot be separated from the metric of the relative location of places, namely the topology.

We draw from this inherent property of geographic space and from more recent work by Caruso et al. (2017) to propose a method that computes a topology-based density index. The idea is to compute the index without using a reference surface but a graph that connects spatial objects (points) via edges. By weighting the edges of the graph with the distance between points, we can obtain a graph that preserves the relative position of the points with respect to each other. One can then cut the

graph in order to obtain groups of points whose relative distances are homogeneous, thus having a homogeneous base (topology) for computing an aggregate index such as density. We suggest a novel spatial descending hierarchical clustering (SDHC) method to cut the graph, where the most locally dissimilar edges are removed iteratively. Rather than an exogeneous cutoff, we use the Cooks statistics on Moran scatterplot’s regressions, thus directly using local topological information at each stage.

We focus here on the case where points are buildings. Buildings, together with plots and streets, are the elementary constituents of an urban space (Moudon, 1997). We illustrate our method by proposing a buildings’ density index, which is a key index in urban planning. Hence, we move the measurement of the buildings’ density from an area-based problem to a graph-based problem. Our density metric preserves the relative location, i.e. the topology of the buildings.

We can see this research as a contribution to a strand of the urban literature where density is complemented by morphological indices (e.g. Galster et al., 2001; Berghauser Pont & Haupt, 2005; Sémécurbe et al., 2019; Fleischmann, 2019, 2021). Contrary to existing work in this domain, we emancipate from the definition of a reference spatial unit. We can also see our work as extending another line of research where topological graphs are used to capture the local spatial organisation of buildings (e.g. Caruso et al., 2017; Wu et al., 2018, for recent examples). Compared to this second strand, we move a step forward by proposing to use Moran’s scatterplots not only for describing but also for cutting the graph, and by adding a density measure.

In the next section, we position our work within broader urban morphology research and with regards to recent methods applied to graphs of buildings. We then present data inputs and the different steps of our methodology in Section (3), then the results of an application to all buildings located in Belgium (Section 4). We discuss our findings in Section 5 and conclude in Section 6.

2 Background

2.1 Density metrics and the urban space

The interplay between forms, structures and processes is central to urban science and for understanding how cities are made up: *‘from form we infer processes that create the structures we see in cities, thus enabling us to build models of these processes, that in turn will simulate forms’* (Batty, 2013, p.79). The physical part of the city, essentially buildings, roads and plots, reveals the presence of inhabitants, activities and hence movements (flows), which themselves relate to planning and normative issues, such as what is the “good” city form (Lynch, 1984) according to its given environmental context and use? How to create a sustainable city? How to transform a city to a more sustainable one? For a while, the idea was to reduce distances and increase density to save energy and space (Williams et al., 2000). Today, many studies and authors (see for example Berghauser Pont et al., 2021, for a review of the impacts of densification) bring nuance to this statement by showing that city-compaction and densification is not always the way to more sustainability. However, metrics of “densities” remain central in debates on how to deliver ‘better’ cities at least in policy arena.

Two main approaches are used to study urban forms: the discrete object approach in discrete areas and the network approach (Berghauser Pont, 2021). In the first approach, researchers measure morphometric characters such as size, shape, intensity, etc. (see Fleischmann, Romice, & Porta, 2021, for a review) of discrete elements such as buildings, streets and plots in more or less complex discrete areas (see for example Berghauser Pont & Haupt, 2005; de Bellefon et al., 2019; Arribas-Bel et al., 2019; Godoy-Shimizu et al., 2021; Fleischmann, Feliciotti, et al., 2021). The morphometric characters are then often further associated with the evaluation of the urban form in terms, for example, of liveability (Martino et al., 2021), waste production, traffic volume, water and energy consumption (Pauleit & Duhme, 2000), heat island and the flow of air (Boyko & Cooper, 2011), the urban vitality (Bobkova et al., 2017), etc. It is not merely the density of these structures that matters, however, but also their geometry at specific scales (Schirmer & Axhausen, 2015). Batty (2013, p.180) suggests

that spatial interactions and the functioning of connections within cities “*need to be physically rooted in the detailed geometry of buildings*”. Similarly for Gehl (1987, p.83), building density “says nothing conclusive about whether human activities are adequately concentrated. The design of buildings in relation to relevant human dimensions is crucial”.

The network approach does not focus on discrete objects but on systems of objects. Researchers study particularly the street network which implies the study of network structure, connectivity, centrality, hierarchy, etc. (Marshall et al., 2018). The Space Syntax movement, initiated by Hillier’s seminal work (Hillier, 1996), is one of the precursors of this approach with attention given to the relative position of lines, while more recent publications rather focus on ubiquity across cities and the massive use of data (see particularly Boeing, 2017). The study of the spatial configuration of street elements allows the measure of the urban form and its impacts. For example, Berghauser Pont et al. (2019) study the centrality of the road network and its impact on pedestrian movement in three cities. The studies of street networks predominantly use methods based on graph (Marshall et al., 2018). In the studies of networks, the graph can be a primal graph (for example streets intersections are the nodes of the graph and streets are the edges) or, as it is the case for Space Syntax, it can be a dual graph (streets becomes nodes and streets intersections are edges) (Porta et al., 2006).

In the same line of research than Space Syntax, the present study work with dual connectivity graphs. As inspired by Caruso et al. (2017), Euclidean segments between buildings are computed as the edges of a primal graph. The idea is then to characterise those edges according to their connectivity. A dual connectivity graph is then computed with the Euclidean segments as nodes. The connectivity between the nodes, expressed by the edges, is function of the presence of nodes (buildings) in the primal graph.

Leaving aside the problem of flows (e.g. Andrienko et al., 2010; Hurvitz et al., 2014), switching from physical structures to functions often leads researchers and planners to switch to areal objects, and count population or activities over a given surface (buildings, parcels, grid cells, etc.). An a priori selected reference (basic) spatial unit (BSU) is often used to measure a density index or a more complex metric. A good example is the set of urban metrics proposed by Galster et al. (2001) to outcompete density while measuring sprawl. Each of their eight indices, not just the average density, requires a count of population or land use category over a set of exogenous grid cells, before being further aggregated over an urban region. While structures and arrangements of population and land uses are definitely picked up at the scale of an entire city or neighbourhood, there is still an aggregation process beforehand, and hence information loss, depending on the resolution and placement of the grid or depending on the original recording units (e.g. census tracts). These zonal and scale effects, known as *Modifiable Areal Unit Problem (MAUP)* (Openshaw, 1983), inevitably bias the measures of urban form such as density (Zhang & Kukadia, 2005). In order to avoid the biases due to the use of surfaces, we here computed metrics associated to networks, following earlier works of Flahaut et al. (2003) or Okabe et al. (2009).

Overall, the density of buildings is a major indicator for urban planning, but can only properly be used when two conditions are met: (i) density is complemented with indicators describing the relative spatial organisation of buildings and (ii) measurement biases due to the use of basic spatial units (BSU) are overcome.

2.2 Methodological advances

Regarding the first of the two conditions above, progress has been made since Galster et al. (2001) especially towards deriving composite multi-factorial indicators. The use of multiple attributes can depict how urban densities are experienced by inhabitants and users (Teller, 2021). Caruso et al. (2017) reviewed some of these multi-indicator contributions, mostly directed to capturing sprawl. They usually consider urban land pixels, not buildings. Teller (e.g. 2021); Godoy-Shimizu et al. (e.g.

2021); Berghauser Pont & Haupt (e.g. 2005); Araldi & Fusco (e.g. 2019) are examples of multi-factorial methods applied to building’ distribution in relation to density. However, these indicators are defined from discrete zone (BSU) and therefore do not meet the second condition.

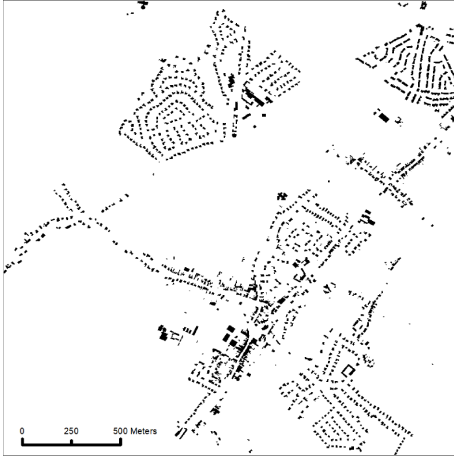
Regarding the second condition, i.e. avoiding the bias of a basic spatial unit (BSU), two strategies have been adopted so far: (i) spatial or statistical smoothing across the BSUs and (ii) clustering individual data with standard (k-means, etc.) or more complex classification methods (artificial neuronal networks methods, etc.). de Bellefon et al. (2019) is a recent example of smoothing. The authors use a kernel function applied on a very fine grid resolution to avoid losing information about the internal spatial organisation while aggregating. In essence however, kernels still depend on an *a priori* defined surface (bandwidth), even if it is often optimised for each case study (similarly to geographically weighted regression approaches with optimum bandwidth (Brunsdon et al., 1996)). Although exceptions exist such as Araldi & Fusco (2019) (among others) in many cases, smoothing methods are applied uniformly across space and may not fit the local composition everywhere. Let us take Figure 1a as an example of a spatial pattern of buildings. If we compute density using a smoothing grid (kernel) (figure 1b), we can see that the topology is lost and that the grid prevents from detecting groups of buildings with a similar topology. Referring to (ii), a recent example of density-based spatial clustering is the A-DBSCAN (*Approximate - Density-Based Spatial Clustering of Applications with Noise*) by Arribas-Bel et al. (2019), building on earlier work by Ester et al. (1996). Buildings are grouped according to a density criterion in order to draw city boundaries. Similar to other density-based methods, A-DBSCAN is not parameter free (minimal number of buildings and maximal distance between them in a cluster) and results can thus greatly vary from one user to another. Furthermore, these two criteria are applied uniformly over the study area, which therefore prevents the detection of locally specific patterns. If we now apply A-DBSCAN (figure 1c) to our example, we see that buildings with different visual topologies are clustered into one large group (blue). Hence, with this method we can see that the topology is partly lost.

To avoid the vanishing of topologies due to spatial aggregation, Zhang & Kukadia (2005) suggest creating BSUs that make sense with regards to the initial spatial organisation of the disaggregated data. This goal is pursued by density-based or graph-based clustering methods (Wu et al., 2018; Deng et al., 2011). This is done by Fleischmann et al. (2020) or by Schirmer & Axhausen (2015) who perform clustering, local spatial statistics, and spatial smoothing within the topological constraints of building-based tessellation adjacencies or street network topology. In both publications, BSUs are adapted in size and shape to the urban intensity. We pursue the same objective in this paper by presenting a method that is able to provide BSUs that make spatial sense in terms of distribution of buildings (i.e. topology) using a graph-based clustering method.

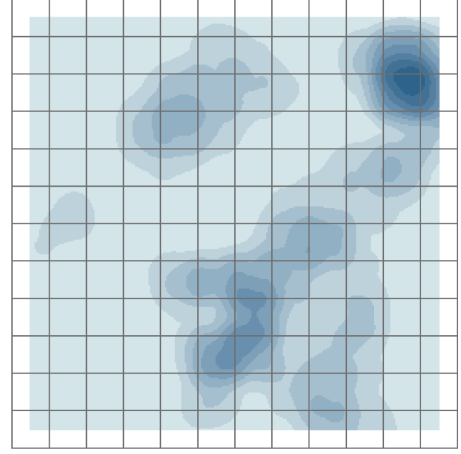
In graph-based methods, the nodes are typically the buildings (primal graph) and the edges the inter-building segments computed from their centroid or from their edge. The advantage is to conserve information about the absolute location of buildings as well as their relative location (Anders et al., 2001; Assunção et al., 2006; Caruso et al., 2017; Wu et al., 2018). Most graph-based methods start with a Minimum Spanning Tree (MST), which is easily partitioned into subgraphs, i.e. clusters. Each time one edge is removed, two distinctive subgraphs are created. If n edges are removed from the initial MST, $n + 1$ clusters of buildings appear. Different strategies are available to determine which edges should be removed to perform the spatial clustering. Caruso et al. (2017) remove all edges larger than an *a priori* threshold fixed at 200 metres. But this unique threshold cannot fit the local spatial organisation of buildings everywhere (figure 1d). Zahn (1971); Yu et al. (2014), based on Gestalt theory, remove the edges that are the most different in the set of their contiguous edges according to three parameters determined *a priori* (p_1 the number of contiguous neighbours; p_2 a ratio between the length of an edge and the average length of its neighbours; p_3 a ratio based on the difference between the length of an edge and the average length of its neighbours to their standard deviation). Results will depend upon these thresholds. While figure 1e shows interesting results overall, it can be seen that the pattern of buildings to the north-east of the area is very poorly captured by the method.

Assunção et al. (2006) use a *Spatial 'K'luster Analysis by Tree Edge Removal* (SKATER) where they iteratively suppress the edge which, once suppressed, minimises the sum of the intra-cluster variances. In this case, the number of final clusters has to be fixed *a priori*.

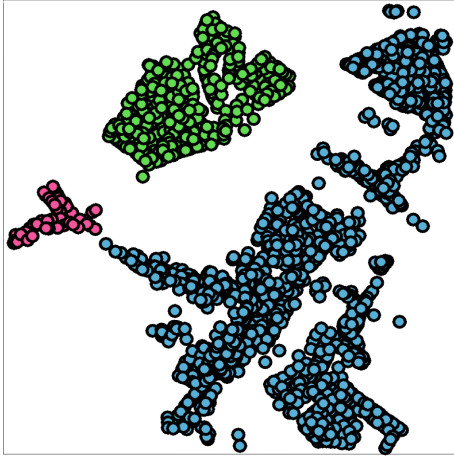
Following the axiomatic of Beguin & Thisse (1979), we know that the denominator of our density metric must preserve the relative positions of buildings everywhere in space. We thus follow graph-based approaches where inter-building distances are used to build and then to prune the MST. We propose a new segmenting (clustering) algorithm that does not require an exogeneous threshold to be applied uniformly across the area, nor a number of clusters to be defined beforehand. Our strategy, inspired by the LISA approach of Caruso et al. (2017), is to compute a Moran scatterplot of inter-building distances for each graph (subgraph) and remove the main outlier for segmenting, rather than applying a distance threshold. As a result, the segmentation can be different across the area and catch distinct local topologies. Density is then computed, based on these topologically homogeneous clusters. The reader can already see that other morphometric indices than density could be calculated on these clusters.



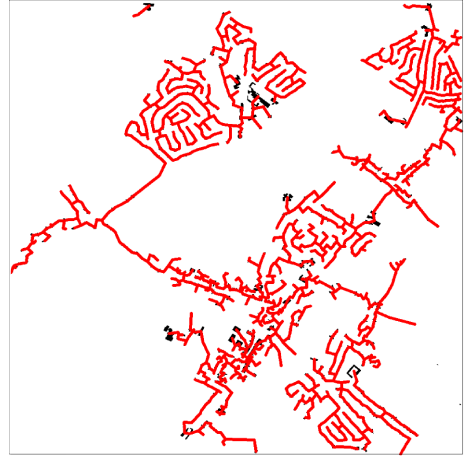
(a) Built-up footprints



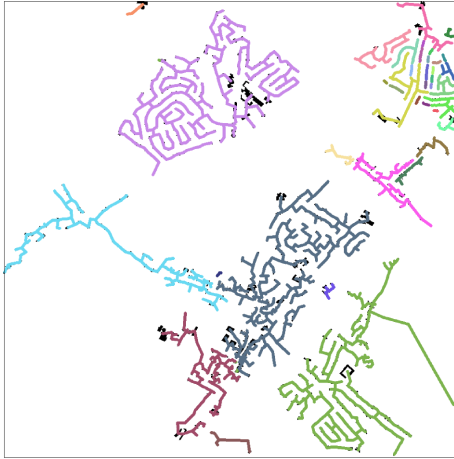
(b) Smoothing with a kernel function (bandwidth = 250 m.) based on a grid



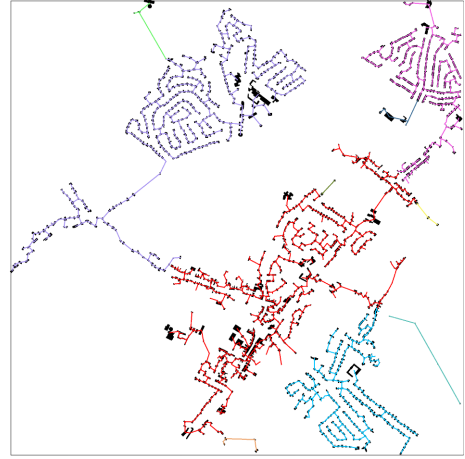
(c) Clustering : A-DBSCAN (distance = 200 m. ; min nodes = 30)



(d) Clustering : fixed distance threshold (distance $\leq$ 200 m.)



(e) Clustering : based on Gelstat theory ($p_1 = 2$, $p_2 = 2$, $p_3 = 3$)



(f) Clustering : SKATER (number of cluster = 10)

Figure 1: Different methods to calculate the density of buildings or to perform a spatial clustering of a built-up pattern

3 Materials and methods

3.1 Data input and study area

Our method is applied to all buildings located in Belgium. Data are provided by the Land Registry Administration of Belgium (© 2018 Administration Générale de la Documentation Patrimoniale). All buildings are used regardless of their function: each house (detached or semi-detached), office building, shop, garage, church or factory is kept in the database. In order to avoid the noise generated by very small buildings such as garden sheds, all built polygons smaller than $12m^2$ were removed from the database as was done by [Montero et al. \(2021\)](#). The database includes 5,726,804 buildings, which further leads us to chunk the data for computation (see Appendix A). Figure 2 shows the study area and a zoomed map of the buildings' footprint.

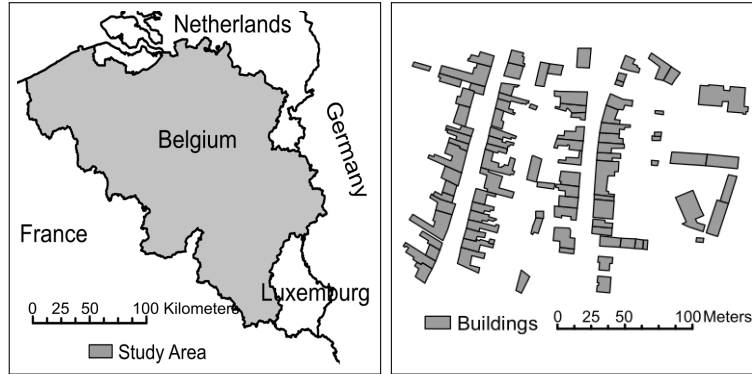


Figure 2: Study area (left) and an example of the footprint of buildings (right)

3.2 Methods

Our objective is to create a topology-based density index that preserves the local spatial organisation of the buildings. Hence, after pre-processing, our method comprises two main steps : a clustering method (step I) leading to topology consistent groups and a density computation (step II). The process and the outcome of each step are summarised in figure 3 for a toy example (figure 3a).

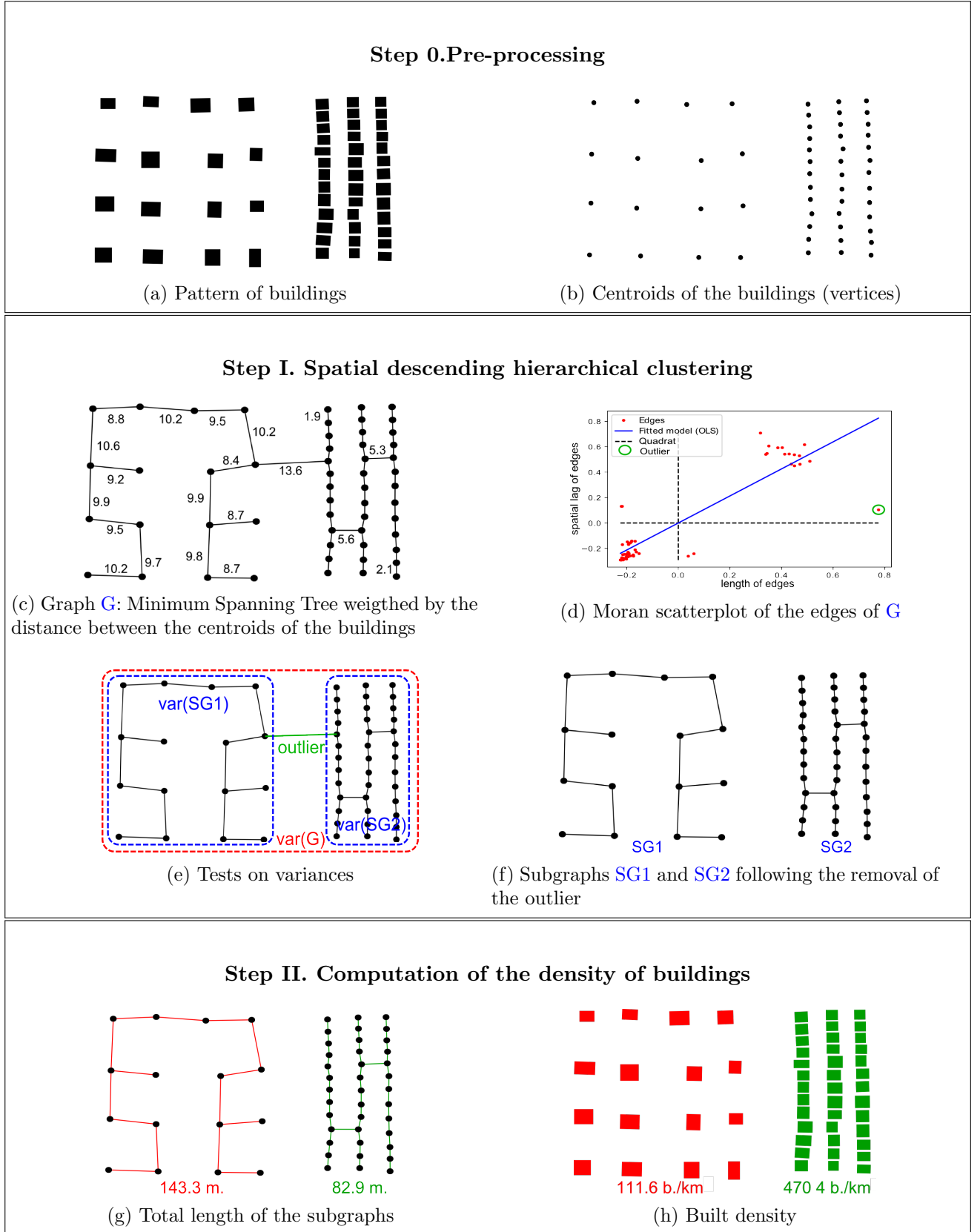


Figure 3: Application of the method on a simple spatial toy structure of buildings

3.2.1 Step 0 - Pre-processing.

The input data consists of polygons. Although the size and shape of the polygons can be heterogeneous across space and further impact the distribution of inter-building distances, we ignore those shapes by retrieving centroids (step 0, see figure 3b). The distance between buildings is here the distance between the centroids of the buildings. We could have followed Yu et al. (2014), who measure the actual distance between buildings but this would not be appropriate in our case. Indeed, we want to measure the density of buildings per km of graph. By using distance between centroids, the user of the metric can then say "When I walk 1 km along the graph, I encounter x buildings". With the actual distance between buildings, this interpretation of the metric would no longer hold. Rather, the user should imagine teleporting him/her self from one end of the buildings to the other as he/she travels along the graph. Moreover, the real distance between two buildings may be zero (e.g. terraced house) which leads to an indeterminacy (denominator is null).

3.2.2 Step I - spatial descending hierarchical clustering (SDHC)

This step starts with a minimum spanning tree (MST) where the nodes are the centroids of the buildings, and the edges are the inter-building segments. Euclidean distances are used as weights (figure 3c) while computing the MST graph, which we denote by G . A descending hierarchical classification (SDHC) is then applied on G to iteratively define subgraphs SG by removing an edge out of each parent graph. The SDHC process is explained below and flow-charted on Figure 4. It is carried out on graphs of at least 30 vertices in order to avoid statistical problems due to small numbers.

A Moran scatterplot (Anselin, 1995) and a maximum Cook's distance are used to identify the edge that should be removed at each iteration of the SDHC. The Moran scatterplot (figure 3d) shows how the edges are spatially associated locally. One point represents one edge of the MST. On the x-axis is the length of an edge, i.e. the distance between two connected buildings, and on the y-axis, its spatial lag, i.e. the weighted average distance of its contiguous edges. We voluntarily restrict the computation of the spatial lag of an edge to its contiguous neighbours of order 1 because we want to detect break between direct neighbours. A topological depth greater than 1 would lead to a smoothing of the spatial lag of each edge. Details about the spatial lag computation can be found in Appendix B.

A linear model (OLS) is then estimated on the scatterplot, the slope of which indicates the global spatial autocorrelation level (Moran's I, see Anselin (1995)). In most cases, we expect a positive slope, meaning that a long (short) separation between two buildings is found in neighbourhoods where distances between buildings are long (short) on average, i.e. in topologically homogeneous cases. The slope will be insignificant in cases where the relative distance between buildings and thus the topology are more heterogeneous. As pointed out by Anselin (1996), *"points in the scatterplot that are extreme with respect to the central tendency reflected by the regression slope may be outliers in the sense that they do not follow the same process of spatial dependence as the bulk of the other observations."* We build on this property and use the maximum Cook's distance to identify the most extreme point (outlier) of each graph G , which is actually the edge to be removed to obtain two subgraphs ($SG1$ & $SG2$) (figure 3f) featuring two topologically distinct clusters of buildings.

In order to determine whether the removal of the outlier leads to the creation of more homogeneous subgraphs, tests of variance (Brown & Forsythe, 1974) between the parent graph (G) and each of the child subgraphs ($SG1$ & $SG2$) are performed (figure 3e). These tests measure whether the variance of the length of the edges in at least one of the two subgraphs is statistically different from that of the parent graph. If the null hypothesis (equality of variances) is rejected, the edge is removed, and the algorithm is re-run separately on each of the subgraphs. If the null hypothesis is not rejected, the edge is not removed, and the algorithm stops. Technically, in order to perform a relevant variance test, one needs first to make sure the parent graph is not very large (thus bearing a lot of heterogeneity) and

that each child graph is at least made of more than one point. Hence the following three conditions are used to determine whether the observed outlier is removed or not (figure 4):

- the total length of the MST is higher than 10,000 metres (see Appendix C for details) (1)
- one of the subgraphs is made of a single vertex (2)
- the variance of the length of the edges in at least one of the two subgraphs is statistically different from that of the graph before edge suppression (3)

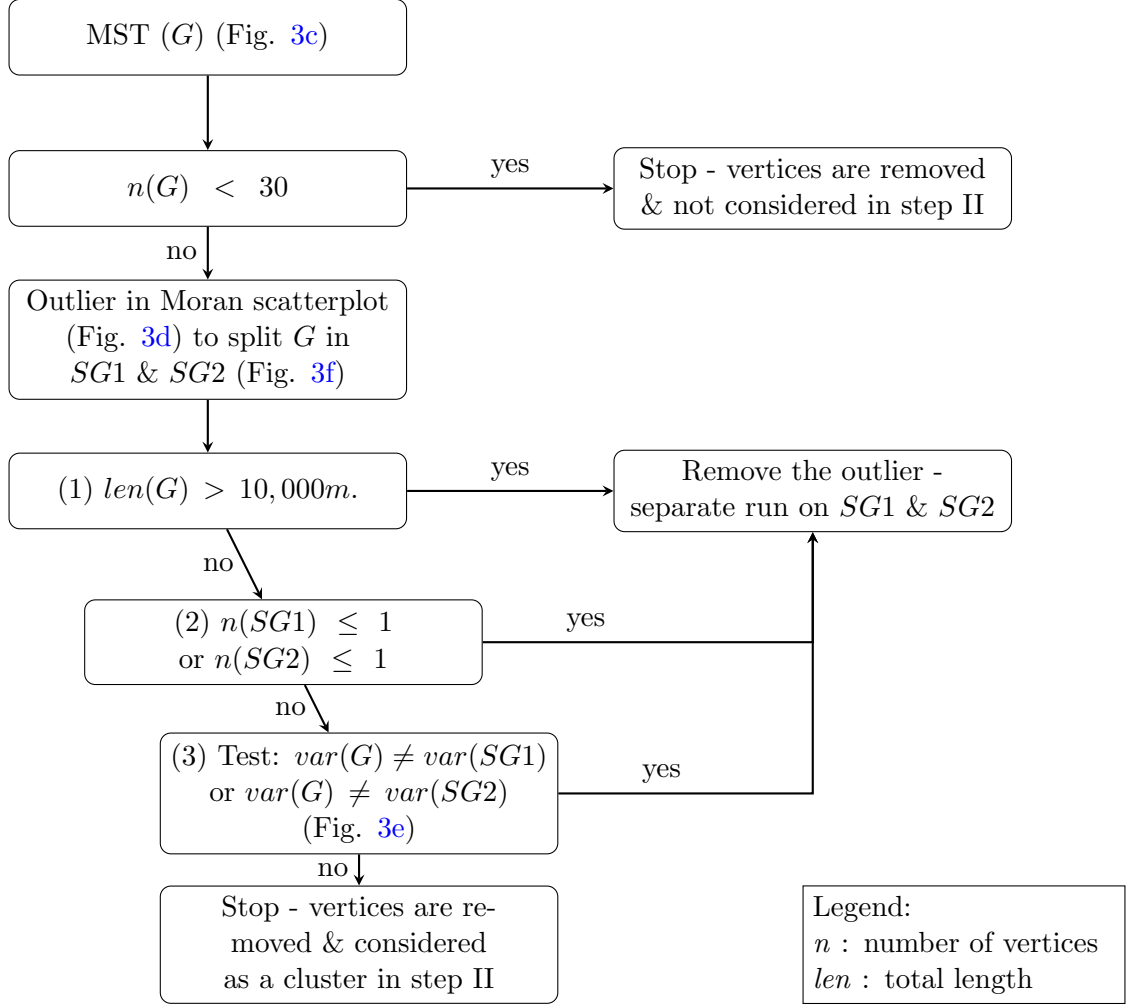


Figure 4: Step I: Flowchart

Step I is completed when no more edge can be removed (no significant outlier) from any of the subgraphs. Each subgraph is thus a topologically homogeneous cluster of buildings.

3.2.3 Step II - the topology-based density index

Let i be a cluster of buildings resulting from the SDHC above. D_i^* , the topology-based density of the cluster i , is then defined as the ratio of N_i , the number of buildings and L_i (figure 3g) the total length of the edges of the MST of i :

$$D_i^* = \frac{N_i}{L_i} \quad (1)$$

(figure 3g)

D_i^* is then a linear density (buildings per linear distance). It would be possible to use the square of L_i to obtain a surface version of the density. However, this transformation brings new biases (due

to the different lengths of L_i) without bringing any new information.

D_i^* can then be mapped onto each building of the corresponding graph (see figure 3h). Unlike common practice, the denominator is no longer an *a priori* chosen surface but the length of the shortest line connecting all buildings in a cluster. Contrary to grid based approaches, it is not constant over space in order to match the local spatial structure.

4 Results

4.1 The minimum spanning tree

The minimum spanning tree computed on all buildings located in Belgium describes the global topological structure of the built-up Belgian reality. As expected by the level of urbanisation of the country, the lengths between buildings in the MST are shorter than 50 m for a very large majority (95%) (see table 1). The histogram (figure 5) is right-skewed and has a bimodal distribution with a very strong peak around 6 metres, and a secondary peak around 20 metres. The left part of the histogram shows short edges which correspond to attached or very close buildings of small size. Note that the very small peak around 3 metres corresponds mostly to sets of contiguous building extensions. In the database, each extension is a small polygon (although larger than 12m²). A set of small extensions joined together then leads to the creation of small edges in the MST. The right part corresponds to edges between 15 and 50 metres characterizing more detached buildings. Edges longer than 50 metres (not illustrated in the histogram) are typical of more isolated buildings. The presence of a large peak at short distances combined with a second peak at medium distances and the absence of a peak at longer distances fully reflects Belgian urbanisation. This urbanisation is indeed characterised by numerous centres (towns, villages) connected with a strong level of suburbanisation and sprawl (Vanneste et al., 2008; Vandermotten et al., 2008; Vanderstraeten & Van Hecke, 2019). One would expect a much smaller second peak and a rural peak (long distance) for a less continuous urbanisation, as is the case for example in the Netherlands.

Mode	q0.25	Median	Mean	q0.75	q0.95	Max
6.00	8.20	15.24	20.85	23.54	49.29	4,378.99

Table 1: Descriptive statistics: length of the edges of the minimum spanning tree

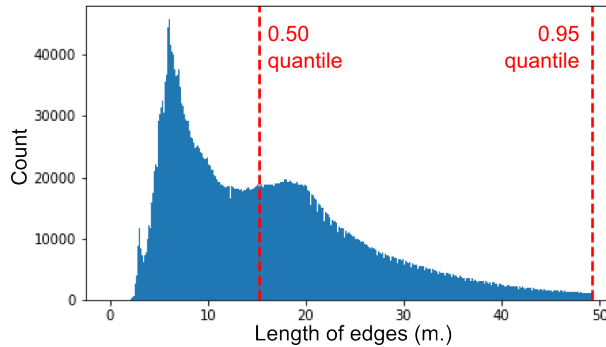


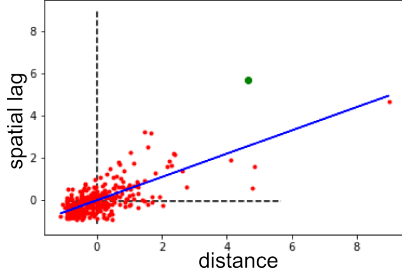
Figure 5: Distribution of the 95% smallest edges of the minimum spanning tree

4.2 Moran scatterplots and their outliers

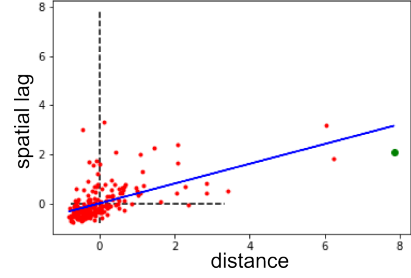
Out of the 260,359 Moran scatterplot regressions performed during the whole process, 93% show a positive and significant slope (Moran’s I). This demonstrates a high degree of homogeneity in the spatial distribution of buildings within the Belgian landscape. Indeed, at each iteration of the method and therefore at all scales, there would be very few abrupt discontinuities in the spatial distribution of

the buildings. This confirms the observations made earlier. Furthermore, the high level of significance of the OLS confirms the relevance of using Moran scatterplot to identify outliers.

We observe two types of outliers within the scatterplots: first, those corresponding to an edge surrounded by edges longer than expected by the global spatial autocorrelation (figure 6a) and second, those corresponding to an edge surrounded by edges much smaller than expected (figure 6b). An outlier separates two distinct topological forms (figure 7a) but in some cases it can simply isolate some remote buildings from the rest (figure 7b). An outlier therefore separates settlements, towns, city districts, villages, etc. or separates isolated and heterogeneous housing from a homogeneous structure of buildings (a farm on the outskirts of a village, a church in a city centre, etc.).

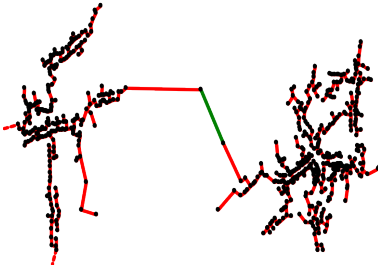


(a) Outlier is an edge surrounded by edges longer than expected by the OLS

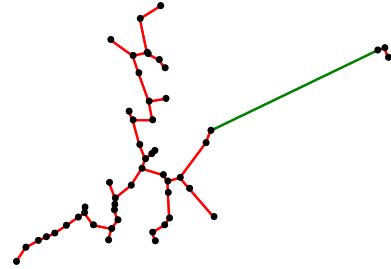


(b) Outlier is an edge surrounded by edges smaller than expected by the OLS

Figure 6: Two examples of typical Moran scatterplot (outlier in green)



(a) outlier separates two distinct morphological spaces



(b) outlier isolate some remote vertices from the rest

Figure 7: Two examples of graphs and related sub-graphs according to the outlier (in green)

75% of the removed edges have a length between 30 and 80 metres while the median length takes the value of 44 metres (table 2). The large observed range of removed edges shows that the iterative process implemented in the method allows the identification of clusters of different patterns at different scales and for different realities. Indeed, from the first removed edge up to the last one, the iterative process progressively splits the initial graph (all of Belgium) into a series of smaller graphs that outline nested clusters, each with its specific characteristics. The use of the Moran scatterplot allows the selection of the edge to be deleted taking into account these specificities. The method identifies the removed edge at each step. It is therefore possible to go back up the clustering tree to observe these different clusters at different scales. This would not have been possible when using a method based on an *a priori* defined threshold (Zahn, 1971; Yu et al., 2014; Caruso et al., 2017). Moreover, the median value (43 m.) shows that the discontinuity in buildings is in the majority of cases much lower than the one generally used by those studies (between 100 and 200 m.).

Min.	q0.25	Median	Mean	q0.75	Max.
0.71	30.66	43.91	78.09	80.95	4,378.99

Table 2: Descriptive statistics: length (in meters) of the removed edges

4.3 Clusters of buildings

At the end of step I, the method discriminates 26,462 subgraphs (see figure D.6 in Appendix D). Over 95% of the subgraphs have a coefficient of variation of the length of the edges smaller than 1, which means that within a subgraph, the lengths of the edges are homogeneous (figure 8). Each subgraph can thus be considered as a topologically homogeneous cluster of buildings.

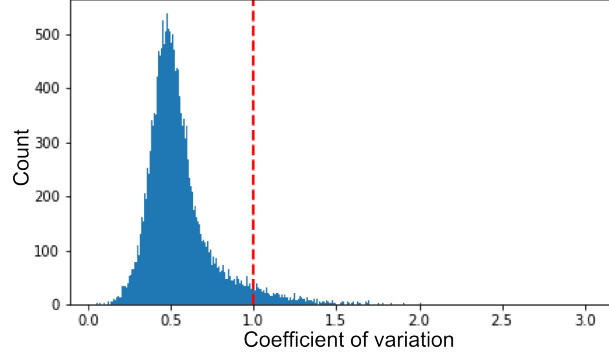


Figure 8: Histogram of the coefficients of variation of the edge lengths for all 26,462 subgraphs

Within each cluster, the variance of the inter-building distances is small, which results in the detection of built-up footprints characterised by a regular pattern of buildings (homogeneous topology). Let us now consider a first example illustrated in figure 9. It includes two regular neighbourhoods (A and B). A is a compact village with a radial morphological structure; B is made up of a regular alignment of buildings that forms a linear ribbon development. A second example is reported in figure 10 already used in section 2.2, south of Brussels, composed of eight homogeneous neighbourhoods (A:H). Each neighbourhood corresponds to a particular pattern of buildings, with a historical centre around the church (A), classical planned housing estates (B:F), and two more linear developments (G:H). Isolated buildings or heterogeneous groups of less than 30 buildings are left out (mainly isolated farms typical of the area).

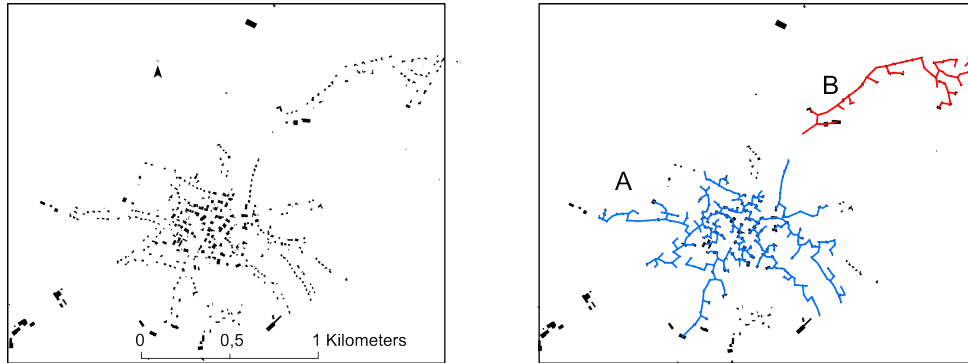


Figure 9: Example 1: Built-up footprints (left) and detected clusters (right) of a Belgian village (Ochamps)

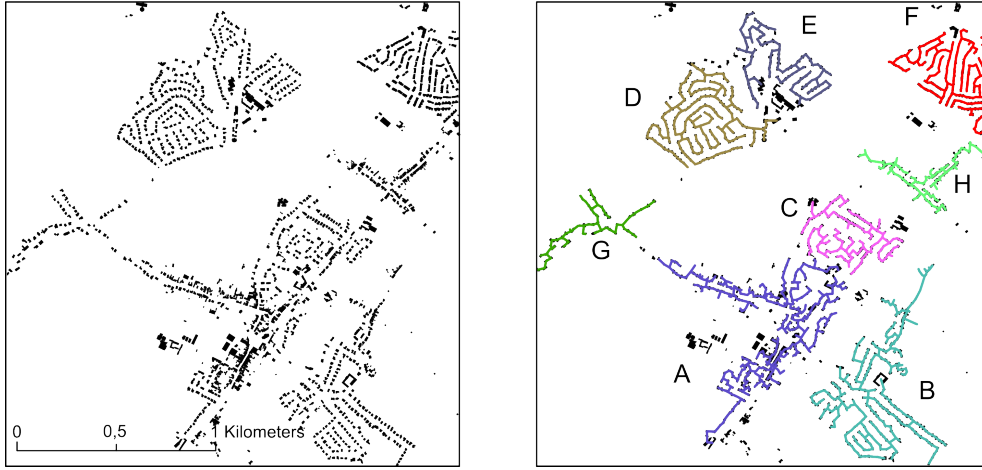


Figure 10: Example 2: Built-up footprints (left) and detected clusters (right) of a suburban settlement (Ophain-Bois-Seigneur-Isaac)

Our 26,462 clusters can now be considered to be topologically relevant Basic Spatial Units (BSU). Since each BSU are internally homogeneous in terms of distance between centroids, we can confidently use those units to compute index such as density, which characterise the spatial distribution of buildings centroids within each cluster.

4.4 Density of buildings

The topology-based density is now computed for each cluster. Each group of buildings with a homogeneous topology, has a specific density value. To explain this specificity of our method, we have compared our results with those obtained by a simple grid-based density smoothed by a kernel function with two examples (figures 11 and 12).

With our method, only one density value is computed by cluster when several values are needed with a grid. For example, in the case of Seneffe (figure 11), we identify seven clusters, with seven density values ranging from 19 to 86 buildings per km (figure 11b). The smoothed grid approach covers the area and compute density ranging from 0 to 31 buildings per hectare (figure 11c). While our method detects seven homogeneous buildings patterns with precise contours, the grid method suggests two or three main centres surrounded by less dense periphery located in the east and some shadows in the west. Similarly in the case of Genappe (figure 12), our method detects three distinct homogeneous buildings patterns (53, 58 and 100 b./km) while the grid method delivers different density values ranging from 0 to 57 b./ha) and showing a large centre in the west surrounded by a periphery that develops in a ribbon towards the east.

In each example, the different density values, associated with the different clusters, allow the identification of particular urban structures. In the case of Seneffe (figure 11b) the two clusters with a density of 73 b./km and 74 b./km include the buildings of the centre, consisting mainly of semi-detached buildings along a main axis from North to South. In the periphery of the centre, the cluster with a density of 60 b./km is formed of detached buildings while the cluster of 86 b./km includes much more attached buildings (public housing). The clusters around 30-50 b./km are associated with housing estates well separated from the centre with exclusively detached buildings. Large inter-building distances characterise the cluster with the lowest density (19 b./km) (industrial zone). In the case of Genappe (figure 12b), the cluster with a density of 100 b./km includes the buildings of the centre. In an extension of the centre, two clusters are identified with a density about 50-60 b./km consisting of well separated buildings. One of these clusters is a ribbons extension along a road from the centre (53 b./km) and the other is a more widely spread cluster assimilated to a district in the periphery of the centre (58 b./km).

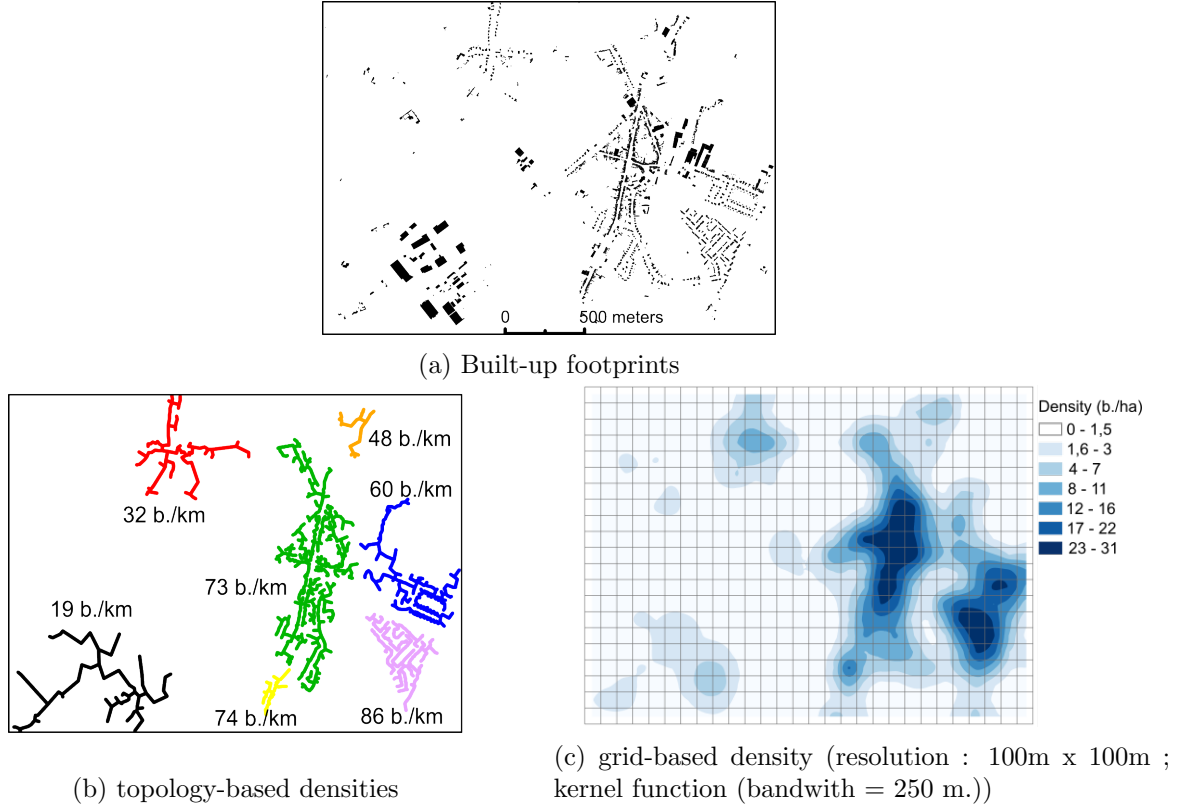


Figure 11: The case of Seneffe

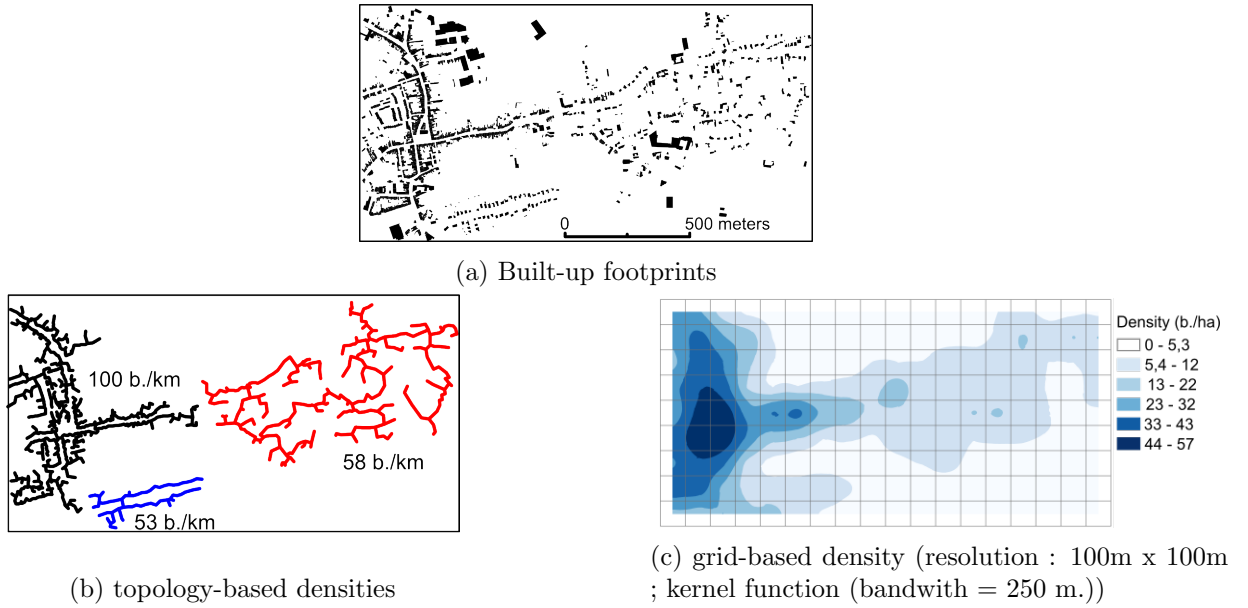


Figure 12: The case of Genappe

Given the definition of the topology-based density (the number of buildings divided by the length of the MST), there is a direct relation between the density value and the average length within a cluster ($\bar{l}_i$). In fact, inverting the equation 1,

$$D_i^* = \left(\frac{L_i}{N_i} \right)^{-1} \quad (2)$$

and because in a MST the number of edges is always equal to the number of points (N_i) minus 1,

$$\bar{l}_i = \frac{L_i}{N_i - 1} \approx \frac{L_i}{N_i} \quad (3)$$

Then D_i^* becomes:

$$D_i^* = \bar{l}_i^\beta \quad (4)$$

with $\beta \approx -1$ when N_i is large ($N_i \approx (N_i - 1)$). For all 26,462 clusters obtained in Belgium, the value of β can be estimated (OLS after logging both sides of equation 4). We obtain a value of β equal to -0.996 . This is indeed very close to 1 but shows a slight under(over) estimation for graphs of longer (smaller) average length. In the words of [Beguin & Thisse \(1979\)](#), we show that density (D_i^*) (of buildings in this case), cannot be separated from the metric of the relative location of places ($\bar{l}_i$) and that this relationship follows a simple power law.

According to equation 4, the topology-based density only depends on the inter-building distances unlike a surface-based approach where the density can vary according to the area of the BSU without considering the distance between buildings. This might sound like a trivial result, but the use of surface-based densities cannot differentiate between two BSUs where the same number of buildings are located but where once is concentrated and once is dispersed. While others researchers would add additional metrics to capture this (e.g. [Galster et al., 2001](#); [Berghauser Pont & Haupt, 2005](#)), our density measure suffices.

Practically, if we know the value of the topology-based density, we can work out the relative spatial organisation of the buildings. Figure 13 illustrates four such cases. (a) A built density value higher than 100 b./km is computed on adjoining buildings or very close to each other (mean distances of less than 10 metres) as is the case in city centres (figure 13a). (b) A value between 50 b./km and 100 b./km is related to a topology of buildings relatively close to each other (mean distances between 10 m. and 20 m.) as may be the case, for example, in the periphery of cities or in smaller centres (figure 13b). (c) Relatively well-separated buildings such as in a peri-urban housing estate (mean distances between 20 m. and 50 m.) have densities in a range between 20 b./km and 50 b./km (figure 13c). Last but not least, (d) densities lower than 20 b./km reflect clusters of buildings with average distances greater than 50 m. (figure 13d).

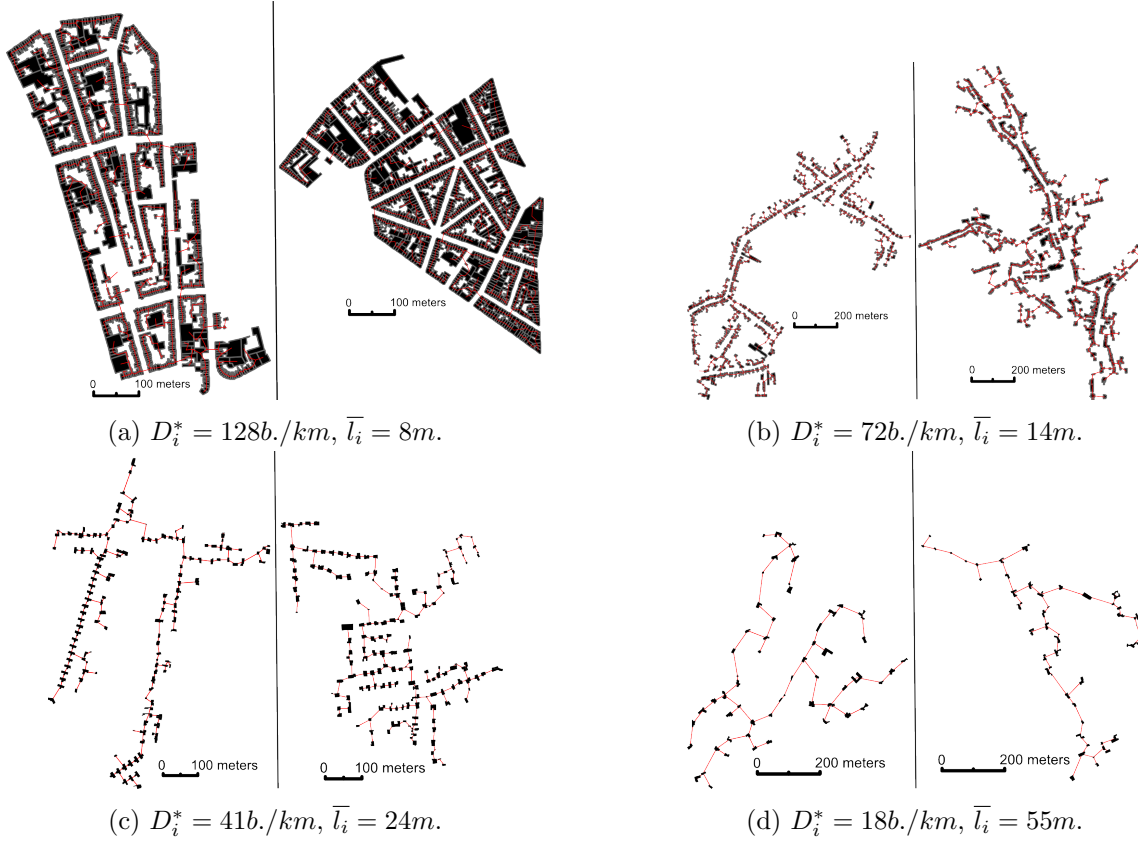


Figure 13: Typical cases of topology-based density according to inter-building distances

At the scale of the entire country, the newly computed topology-based density (see figure E.7 in Appendix E) shows a spatial structure that expresses urbanisation in Belgium (see section 4.1).

5 Discussion

5.1 Thematic contribution

We have proposed a spatial descending hierarchical clustering method that delineates clusters of buildings with homogeneous inter-building distances. Based on these clusters, we compute a topology-based density index where the denominator preserve the relative positions of buildings. We show (section 4.4) that the index eventually only depends on the average distance between buildings in each cluster. It is a strong advantage compared to standard surface-based densities where density depends on the delineation and definition of a reference area (BSU).

The numerator considered here is simply the number of buildings. Depending on the final objective of the measure, other numerators could equally be used such as the surface of buildings, their volume or height (see e.g. Yu et al., 2010). In our case, we could imagine using the surface area of the buildings in each cluster or the total surface area of their floors as the numerator. These indices would remain topology-based as the denominator does not change. The diversity of indices is therefore a function of the diversity of possible numerators. As shown by Wu et al. (2018), a large number of characteristics (distance, orientation, height, size, etc.) of buildings can easily be integrated into a graph. It is up to the planners to develop and use them according to their needs. In the same way, other parameters can be used as a basis for the SDHC. The association of a Cooks distance with a Moran Scatterplot can be used to distinguish the most different object in a cluster. Rather than having clusters with identical patterns in terms of distance between buildings, some could, for example, look at groups with similar building heights.

The only data input used here are the buildings. This can appear counterproductive since streets, squares, parks, and gardens are traditionally identified as important places in built-up realities (Gehl, 1987). However, as expressed in section 2, it is possible to link a large number of urban issues to the structure/proximity of buildings (energy consumption (Rinkinen et al., 2021), mental health (Sullivan & Chang, 2011), population estimates (Tomás et al., 2016), etc.). This is why we believe that focusing on the density of buildings can be relevant for urban space issues.

We are aware that the use of our graph-based index may appear difficult for urban policy makers as they are often used to work on an externally determined surface basis. However, we believe that it is sometimes necessary to change the approach because of biases and errors in measurement and interpretation induced by these surfaces. Moreover, measuring the density along a topological network (as proposed here) has also two practical advantages compared to a more classical surface approach. First, it is a more operational way to study the relationships/interactions between buildings and linear infrastructures. Linear infrastructures (electricity, gas, water) do not always follow roads and their planning could also benefit from our measure. Our method enables to determine which buildings are spatially connected and at which specific distance. Second, by using a distance-weighted graph, we find that our measure is more likely to lead to the study of the relationships and interactions that can exist between points (buildings). Indeed, the network approach allows us to distinguish points that are connected to each other while measuring the characteristics of each group. A parallel can be drawn with ecological research where graphs are used. For example, in the same way that Foltête (2017) delineates functional ecological zones with by means of landscape graphs, we should be able to better measure and therefore better understand the relationships between people living in different spaces of a city, or to understand how these different spaces are organized.

Our topology-based index is a contribution to increase the quality of the measurement and understanding of the morphology of the built space. We know that the topology of the buildings is only one aspect of the complexity of such space. Taking a multi-factorial approach, it would certainly be possible to develop and combine other indices with the topology-based density index as presented here. We already mentioned the addition of the third dimension (height of the buildings). But it is certainly also possible to adapt other indicators. The concentration index developed by Galster et al. (2001) could for example be adapted to measure whether or not, within a cluster, the buildings are rather aligned or form a block. Complementing the topology-based density with other indices could be the next step in this research.

5.2 Methodological limits

A first methodological limit is the use of the centroid of each building instead of the building footprint in the creation of the graph. We have seen that in our case, it is the distance between building centroids that must be considered in order to obtain a calculable and more easily interpretable measure. We note, however, that in some cases the interpretation of the measurement may lead to a poorer perception of the built environment. For example, a given spatial organization of centroids may reflect the location of small buildings that are far apart (i.e. isolated farms) or very large buildings that almost join (i.e. industrial zoning).

The use of a Cook's distance for identifying the outlier in the Moran scatterplot can also be discussed. There are many alternative graphical (scatterplot, boxplot, etc.) or analytical (standardised residuals, hat matrix, etc.) methods of detection of outliers in regression (Ampanthong & Suwattee, 2009). Analytical methods have the advantage that they do not require human visual interpretation. Cook's distance is pointed out by Ampanthong & Suwattee (2009) as one of the best indices for the detection of outliers in multiple regression. In our paper, we find it interesting because it combines both residual information (is a point far from the line?) and information on the influence of each point in the regression. It should be noted that the method identifies a single outlier (the observation for which the Cook's distance is maximum). If the outlier is clearly identifiable, the different indices

will converge. If several outliers are present and if they are not clearly identifiable, results might not converge. However, we did not encounter this figure in our empirical analyses, but are aware that it could happen. Another limitation that should be investigated in the future with the help of statisticians concerns the use of outliers in a regression whose slope is not significant. This happens very little in our case (5%) but could happen in a more important way if someone wants to work on a less homogeneous variable than the distance between centroid buildings.

Another methodological limit concerns the use of the 10,000 metres threshold as a constraint on the removal of an outlier. On the one hand, the objective of the method is not impacted by the value of the threshold. Whatever the threshold, the method always creates clusters whose topologies tend to be more and more homogeneous during the iterations. On the other hand, the threshold can modify the scale at which the method will stop. A high value will lead to the creation of large clusters (large length of graph) whereas a small value leads to the creation of very fine scale clusters. The threshold of 10,000 seems to be the most relevant for density measurements in Belgium but remains debatable.

Last but not least, the choice of variance test can also be a source of discussion. Indeed, it is important *a priori* to control the distribution of the populations tested when carrying out a test of variance (Box, 1953). We do not carry out this control systematically. However, we have noticed in a large majority of cases that the distribution of the length of the edges in the (sub)graphs was heavy-tailed (as shown in figure 5). Therefore, we opted for the Brown & Forsythe variance test. This test is the most appropriate for this type of distribution (it does not consider the most extreme 10% of the distribution) (Brown & Forsythe, 1974). In comparison with other tests, the Brown & Forsythe test gave the most visually appropriate results. One way to further improve the method may be to systematically assess the shape of the distributions to be tested and the application of the most appropriate test to each case.

6 Conclusion

Following Caruso et al. (2017) who sought to identify urban form patterns using methods based on graphs, and following Berghauser Pont & Haupt (e.g. 2005) who sought to measure buildings' density with distinct indices depending on spatial units, we develop here a method to obtain a built density index that preserves the topology of buildings. This means that we can now identify clusters of buildings with homogeneous inter-building distances, and we can further measure, for each cluster, the density of buildings while preserving information about their relative positions.

Our method works in two steps. After retrieving the centroids of the buildings, the first step in the method consists in a spatial descending hierarchical clustering (iterative approach). Based on a minimum spanning tree weighted by inter-building distances, a Moran scatterplot combined with a maximum Cook's distance are used to identify the edge of the MST that diverges most from its neighbours (outlier). This edge is removed if it meets several criteria; one of these is the inequality of the variances of the lengths of the edges with and without outlier. At the end of step 1, clusters of buildings with homogeneous inter-building distances are delineated. In the second step, the topology-based density is computed by dividing the number of buildings in a cluster by the total length of the MST connecting all buildings in that cluster.

The method is applied to all buildings located in Belgium. Clusters with homogeneous inter-building distances are clearly identified. For example, some clusters refer to the organization in a compact village, others to a linear development along a road, or to a housing estate organization, etc. For each cluster, the value of the newly developed density index reflects the topology i.e. the relative position of buildings. A high (low) density will be measured when the distance between buildings is small (large). The topology-based density index is then only influenced by the relative position of the buildings (average inter-building distance). This is not the case for standard density measures, using

an *a priori* fixed surface. Topology-based density is therefore a quite useful index for measuring and understanding built-up patterns.

Acknowledgement

Geoffrey Caruso acknowledges support from the Luxembourg National Research Fund via the URBANFORMS project (INTER/MOBILITY/mobility/2020/14519030)

References

- Ampanthong, P., & Suwattee, P. (2009). A Comparative Study of Outlier Detection Procedures in Multiple Linear Regression. *Proceedings of the International MultiConference of Engineers and Computer Scientists*, 7.
- Anders, K.-H., Sester, M., & Fritsch, D. (2001). Analysis of settlement structures by graph-based clustering. , 10.
- Andrienko, G., Andrienko, N., Demsar, U., Dransch, D., Dykes, J., Fabrikant, S. I., ... Tominski, C. (2010, October). Space, time and visual analytics. *International Journal of Geographical Information Science*, 24(10), 1577–1600. Retrieved 2021-07-09, from <http://www.tandfonline.com/doi/abs/10.1080/13658816.2010.508043> (Number: 10) doi: 10.1080/13658816.2010.508043
- Angel, S., Lamson-Hall, P., & Blanco, Z. G. (2021, March). Anatomy of density: measurable factors that constitute urban density. *Buildings and Cities*, 2(1), 264–282. Retrieved 2021-05-19, from <http://journal-buildingscities.org/articles/10.5334/bc.91/> (Number: 1) doi: 10.5334/bc.91
- Anselin, L. (1995). Local indicators of spatial association—LISA. *Geographical analysis*, 27(2), 93–115. Retrieved 2016-05-26, from <http://onlinelibrary.wiley.com/doi/10.1111/j.1538-4632.1995.tb00338.x/full> (Number: 2)
- Anselin, L. (1996). The Moran scatterplot as an ESDA tool to assess local instability in spatial association. In *Spatial Analytical Perspectives on GIS* (Vol. 4).
- Araldi, A., & Fusco, G. (2019). From the street to the metropolitan region: Pedestrian perspective in urban fabric analysis. *Environment and Planning B: Urban Analytics and City Science*, 46(7), 1243–1263.
- Arribas-Bel, D., Garcia-López, M.-A., & Viladecans-Marsal, E. (2019, November). Building(s and) cities: Delineating urban areas with a machine learning algorithm. *Journal of Urban Economics*, 103217. Retrieved 2021-07-08, from <https://linkinghub.elsevier.com/retrieve/pii/S0094119019300944> doi: 10.1016/j.jue.2019.103217
- Assunção, R. M., Neves, M. C., Câmara, G., & Da Costa Freitas, C. (2006, August). Efficient regionalization techniques for socio-economic geographical units using minimum spanning trees. *International Journal of Geographical Information Science*, 20(7), 797–811. Retrieved 2019-06-24, from <http://www.tandfonline.com/doi/abs/10.1080/13658810600665111> (Number: 7) doi: 10.1080/13658810600665111
- Batty, M. (2013). *The New Science of Cities*. MIT Press.
- Beguín, H., & Thisse, J.-F. (1979). An Axiomatic Approach to Geographical Space. *Geographical Analysis*, 11(4), 325–341. Retrieved 2021-05-31, from <http://doi.wiley.com/10.1111/j.1538-4632.1979.tb00700.x> (Number: 4) doi: 10.1111/j.1538-4632.1979.tb00700.x
- Berghauser Pont, M. (2021). Measuring Urban Form. *28th International Seminar on Urban Form - ISUF 2021 : Urban Form For Sustainable and Prosperous Cities, Glasgow*.

- Berghauser Pont, M., & Haupt, P. (2005). The spacemate: Density and the Typomorphology of the Urban Fabric. *Nordisk Arkitekturforskning*, 18(4), 55–68.
- Berghauser Pont, M., Haupt, P., Berg, P., Alst  de, V., & Heyman, A. (2021). Systematic review and comparison of densification effects and planning motivations. *Buildings and Cities*, 2(1), 378–401. doi: <http://doi.org/10.5334/bc.125>
- Berghauser Pont, M., Stavroulaki, G., & Marcus, L. (2019). Development of urban types based on network centrality, built density and their impact on pedestrian movement. *Environment and Planning B: Urban Analytics and City Science*, 46(8), 1549–1564. Retrieved 2022-04-26, from <http://journals.sagepub.com/doi/10.1177/2399808319852632> doi: 10.1177/2399808319852632
- Bobkova, E., Marcus, L., & Pont, B. (2017). Multivariable measures of plot systems: describing the potential link between urban diversity and spatial form based on the spatial capacity concept. *Proceedings of the 11th Space Syntax Symposium, Lisbon, Portugal*, 16.
- Boeing, G. (2017, September). OSMnx: New methods for acquiring, constructing, analyzing, and visualizing complex street networks. *Computers, Environment and Urban Systems*, 65, 126–139. Retrieved 2021-07-08, from <https://linkinghub.elsevier.com/retrieve/pii/S0198971516303970> doi: 10.1016/j.compenvurbsys.2017.05.004
- Box, G. (1953). Non-normality and tests on variances. *Biometrika*, 40(3-4), 318–335. (Number: 3-4) doi: <https://doi.org/10.1093/biomet/40.3-4.318>
- Boyko, C. T., & Cooper, R. (2011, July). Clarifying and re-conceptualising density. *Progress in Planning*, 76(1), 1–61. Retrieved 2020-10-16, from <https://linkinghub.elsevier.com/retrieve/pii/S0305900611000274> (Number: 1) doi: 10.1016/j.progress.2011.07.001
- Brown, M. B., & Forsythe, A. B. (1974). Robust Tests for the Equality of Variances. *Journal of the American Statistical Association*, 69:346, 364–367. doi: <http://dx.doi.org/10.1080/01621459.1974.10482955>
- Brunsdon, C., Fotheringham, S., & Charlton, M. E. (1996). Geographically Weighted Regression: A Method for Exploring Spatial Nonstationarity. *Geographical Analysis*, 281–298. doi: <https://doi.org/10.1111/j.1538-4632.1996.tb00936.x>
- Caruso, G., Hilal, M., & Thomas, I. (2017, July). Measuring urban forms from inter-building distances: Combining MST graphs with a Local Index of Spatial Association. *Landscape and Urban Planning*, 163, 80–89. Retrieved 2017-08-09, from <http://linkinghub.elsevier.com/retrieve/pii/S0169204617300518> doi: 10.1016/j.landurbplan.2017.03.003
- Churchman, A. (1999). Disentangling the Concept of Density. *Journal of Planning Literature*, 13(4), 389–411. (Number: 4)
- de Bellefon, M.-P., Combes, P.-P., Duranton, G., Gobillon, L., & Gorin, C. (2019, December). Delineating urban areas using building density. *Journal of Urban Economics*, 103226. Retrieved 2020-06-02, from <https://linkinghub.elsevier.com/retrieve/pii/S0094119019301032> doi: 10.1016/j.jue.2019.103226
- Deng, M., Liu, Q., Cheng, T., & Shi, Y. (2011, July). An adaptive spatial clustering algorithm based on delaunay triangulation. *Computers, Environment and Urban Systems*, 35(4), 320–332. Retrieved 2019-06-24, from <https://linkinghub.elsevier.com/retrieve/pii/S019897151100024X> (Number: 4) doi: 10.1016/j.compenvurbsys.2011.02.003
- Ester, M., Kriegel, H.-P., & Xu, X. (1996). A Density-Based Algorithm for Discovering Clusters in Large Spatial Databases with Noise. In (Vol. 240, p. 6).

- Ewing, R., Hamidi, S., Tian, G., Proffitt, D., Tonin, S., & Fregolent, L. (2018, June). Testing Newman and Kenworthy's Theory of Density and Automobile Dependence. *Journal of Planning Education and Research*, 38(2), 167–182. Retrieved 2021-06-03, from <http://journals.sagepub.com/doi/10.1177/0739456X16688767> (Number: 2) doi: 10.1177/0739456X16688767
- Flahaut, B., Mouchart, M., Martin, E. S., & Thomas, I. (2003). The local spatial autocorrelation and the kernel method for identifying black zones A comparative approach. *Accident Analysis and Prevention*, 35, 991–1004.
- Fleischmann, M. (2019). momepy: Urban Morphology Measuring Toolkit. *Journal of Open Source Software*, 4(43), 1807. Retrieved 2022-04-26, from <https://joss.theoj.org/papers/10.21105/joss.01807> doi: 10.21105/joss.01807
- Fleischmann, M. (2021). The Urban Atlas: Methodological Foundation of a Morphometric Taxonomy of Urban Form. *Urban Design Studies Unit Department of Architecture University of Strathclyde, UK*, 511.
- Fleischmann, M., Feliciotti, A., Romice, O., & Porta, S. (2020). Morphological tessellation as a way of partitioning space: Improving consistency in urban morphology at the plot scale. *Computers, Environment and Urban Systems*, 80, 101441.
- Fleischmann, M., Feliciotti, A., Romice, O., & Porta, S. (2021). Methodological foundation of a numerical taxonomy of urban form. *Environment and Planning B: Urban Analytics and City Science*, 0(0), 1–17. Retrieved 2022-04-13, from <http://journals.sagepub.com/doi/10.1177/23998083211059835> doi: 10.1177/23998083211059835
- Fleischmann, M., Romice, O., & Porta, S. (2021). Measuring urban form: Overcoming terminological inconsistencies for a quantitative and comprehensive morphologic analysis of cities. *Environment and Planning B: Urban Analytics and City Science*, 48(8), 2133–2150. Retrieved 2022-04-13, from <http://journals.sagepub.com/doi/10.1177/2399808320910444> doi: 10.1177/2399808320910444
- Galster, G., Hanson, R., Ratcliffe, M. R., Wolman, H., Coleman, S., & Freihage, J. (2001, January). Wrestling Sprawl to the Ground: Defining and measuring an elusive concept. *Housing Policy Debate*, 12(4), 681–717. Retrieved 2019-03-06, from <http://www.tandfonline.com/doi/abs/10.1080/10511482.2001.9521426> (Number: 4) doi: 10.1080/10511482.2001.9521426
- Gehl, J. (1987). *Life between buildings: using public space* (Island press ed.).
- Godoy-Shimizu, D., Steadman, P., & Evans, S. (2021, February). Density and morphology: from the building scale to the city scale. *Buildings and Cities*, 2(1), 92–113. Retrieved 2021-04-29, from <http://journal-buildingscities.org/articles/10.5334/bc.83/> (Number: 1) doi: 10.5334/bc.83
- Hillier, B. (1996). *Space is the machine*.
- Hurvitz, P. M., Moudon, A. V., Kang, B., Saelens, B. E., & Duncan, G. E. (2014). Emerging Technologies for Assessing Physical Activity Behaviors in Space and Time. *Frontiers in Public Health*, 2. Retrieved 2021-07-09, from <http://journal.frontiersin.org/article/10.3389/fpubh.2014.00002/abstract> doi: 10.3389/fpubh.2014.00002
- Longley, P. A., & Mesev, V. (2002). Measurement of density gradients and space-filling in urban systems*. *Papers in Regional Science*, 81(1), 1–28. Retrieved 2020-06-02, from <http://doi.wiley.com/10.1111/j.1435-5597.2002.tb01219.x> (Number: 1) doi: 10.1111/j.1435-5597.2002.tb01219.x
- Lynch, K. (1984). *Good city form* (MIT press ed.).

- Marshall, S., Gil, J., Kropf, K., Tomko, M., & Figueiredo, L. (2018). Street Network Studies: from Networks to Models and their Representations. *Networks and Spatial Economics*, 18, 735–749.
- Martino, N., Girling, C., & Lu, Y. (2021, March). Urban form and livability: socioeconomic and built environment indicators. *Buildings and Cities*, 2(1), 220–243. Retrieved 2021-04-28, from <http://journal-buildingscities.org/articles/10.5334/bc.82/> (Number: 1) doi: 10.5334/bc.82
- Montero, G., Tannier, C., & Thomas, I. (2021). Delineation of cities based on scaling properties of urban patterns: a comparison of three methods. *International Journal of Geographical Information Science*, 35(5), 919–947. Retrieved 2021-05-25, from <https://www.tandfonline.com/doi/full/10.1080/13658816.2020.1817462> (Number: 5) doi: 10.1080/13658816.2020.1817462
- Moudon, A. V. (1997). Urban morphology as an emerging interdisciplinary field. *Urban Morphology*, 1(1), 3–10.
- Okabe, A., Satoh, T., & Sugihara, K. (2009). A Kernel Density Estimation Method for Networks, Its Computational Method and a GIS-Based Tool. *International Journal of Geographical Information Science*, 23, 7–32.
- Openshaw, S. (1983). The Modifiable Areal Unit Problem. *Concepts and Techniques in Modern Geography*, 38(38). (Number: 38)
- Pauleit, S., & Duhme, F. (2000, November). Assessing the environmental performance of land cover types for urban planning. *Landscape and Urban Planning*, 52(1), 1–20. Retrieved 2021-03-08, from <https://linkinghub.elsevier.com/retrieve/pii/S0169204600001092> (Number: 1) doi: 10.1016/S0169-2046(00)00109-2
- Porta, S., Crucitti, P., & Latora, V. (2006). The network analysis of urban streets: A dual approach. *Physica A: Statistical Mechanics and its Applications*, 369(2), 853 – 866.
- Rinkinen, J., Shove, E., & Smits, M. (2021, February). Conceptualising urban density, energy demand and social practice. *Buildings and Cities*, 2(1), 79–91. Retrieved 2021-04-28, from <http://journal-buildingscities.org/articles/10.5334/bc.72/> (Number: 1) doi: 10.5334/bc.72
- Schirmer, P. M., & Axhausen, K. W. (2015). A multiscale classification of urban morphology. *Journal of Transport and Land Use*, 9(1), 101–130. Retrieved 2022-04-26, from <https://www.jtlu.org/index.php/jtlu/article/view/667> doi: 10.5198/jtlu.2015.667
- Sullivan, W. C., & Chang, C.-Y. (2011). Mental Health and the Built Environment. In A. L. Dannenberg, H. Frumkin, & R. J. Jackson (Eds.), *Making Healthy Places* (pp. 106–116). Washington, DC: Island Press/Center for Resource Economics. Retrieved 2021-08-04, from <http://link.springer.com/10.5822/978-1-61091-036-1.7> doi: 10.5822/978-1-61091-036-1.7
- Sémécurbe, F., Tannier, C., & Roux, S. G. (2019, June). Applying two fractal methods to characterise the local and global deviations from scale invariance of built patterns throughout mainland France. *Journal of Geographical Systems*, 21(2), 271–293. Retrieved 2020-06-02, from <http://link.springer.com/10.1007/s10109-018-0286-1> (Number: 2) doi: 10.1007/s10109-018-0286-1
- Teller, J. (2021). Regulating urban densification: what factors should be used? *Buildings and Cities*, 2(1), 302–317. doi: <http://doi.org/10.5334/bc.123>
- Tomás, L., Fonseca, L., Almeida, C., Leonardi, F., & Pereira, M. (2016, December). Urban population estimation based on residential buildings volume using IKONOS-2 images and lidar data. *International Journal of Remote Sensing*, 37(sup1), 1–28. Retrieved 2021-08-04, from <https://www.tandfonline.com/doi/full/10.1080/01431161.2015.1121301> (Number: sup1) doi: 10.1080/01431161.2015.1121301

- Vandermotten, C., Halbert, L., Roelandts, M., & Cornut, P. (2008, October). European Planning and the Polycentric Consensus: Wishful Thinking? *Regional Studies*, 42(8), 1205–1217. Retrieved 2018-07-10, from <http://www.tandfonline.com/doi/abs/10.1080/00343400701874206> (Number: 8) doi: 10.1080/00343400701874206
- Vanderstraeten, L., & Van Hecke, E. (2019). Les régions urbaines en Belgique. *Belgeo*(1). Retrieved 2021-05-31, from <http://journals.openedition.org/belgeo/32246> (Number: 1) doi: 10.4000/belgeo.32246
- Vanneste, D., Thomas, I., & Vanderstraeten, L. (2008, September). The spatial structure(s) of the Belgian housing stock. *Journal of Housing and the Built Environment*, 23(3), 173–198. Retrieved 2021-07-23, from <http://link.springer.com/10.1007/s10901-008-9111-3> (Number: 3) doi: 10.1007/s10901-008-9111-3
- Williams, K., Burton, E., & Jenks, M. (2000). Achieving sustainable urban form: an introduction. *Achieving sustainable urban form, 2000*, 1–5.
- Wu, B., Yu, B., Wu, Q., Chen, Z., Yao, S., Huang, Y., & Wu, J. (2018, March). An Extended Minimum Spanning Tree method for characterizing local urban patterns. *International Journal of Geographical Information Science*, 32(3), 450–475. Retrieved 2019-03-06, from <https://www.tandfonline.com/doi/full/10.1080/13658816.2017.1384830> (Number: 3) doi: 10.1080/13658816.2017.1384830
- Yu, B., Liu, H., Wu, J., Hu, Y., & Zhang, L. (2010, December). Automated derivation of urban building density information using airborne LiDAR data and object-based method. *Landscape and Urban Planning*, 98(3-4), 210–219. Retrieved 2021-08-04, from <https://linkinghub.elsevier.com/retrieve/pii/S016920461000191X> (Number: 3-4) doi: 10.1016/j.landurbplan.2010.08.004
- Yu, B., Shu, S., Liu, H., Song, W., Wu, J., Wang, L., & Chen, Z. (2014, November). Object-based spatial cluster analysis of urban landscape pattern using nighttime light satellite images: a case study of China. *International Journal of Geographical Information Science*, 28(11), 2328–2355. Retrieved 2019-03-06, from <http://www.tandfonline.com/doi/full/10.1080/13658816.2014.922186> (Number: 11) doi: 10.1080/13658816.2014.922186
- Zahn, C. (1971, January). Graph-Theoretical Methods for Detecting and Describing Gestalt Clusters. *IEEE Transactions on Computers*, C-20(1), 68–86. Retrieved 2019-03-06, from <http://ieeexplore.ieee.org/document/1671676/> (Number: 1) doi: 10.1109/T-C.1971.223083
- Zhang, M., & Kukadia, N. (2005). Metrics of Urban Form and the Modifiable Areal Unit Problem. *Transportation Research Record: Journal of the Transportation Research Board*(1902), 71–79. (Number: 1902)

A Chunking of the database

To limit computation time, the database is divided into different chunks spatially adjacent to each other. However, this chunking produces border effects that might bias the results. Indeed, two buildings belonging potentially to the same morphological structure can be separated if they belong to two different chunks. To avoid these border effects, a second chunking process is applied to the same dataset (figure A.1). The method runs for each chunk separately. The results of the chunks from the chunking 1 are then combined but the subgraphs close (1,000 metres) to the chunks border are replaced by subgraphs from chunking 2 to give the buildings’ density of Belgium (section 4).

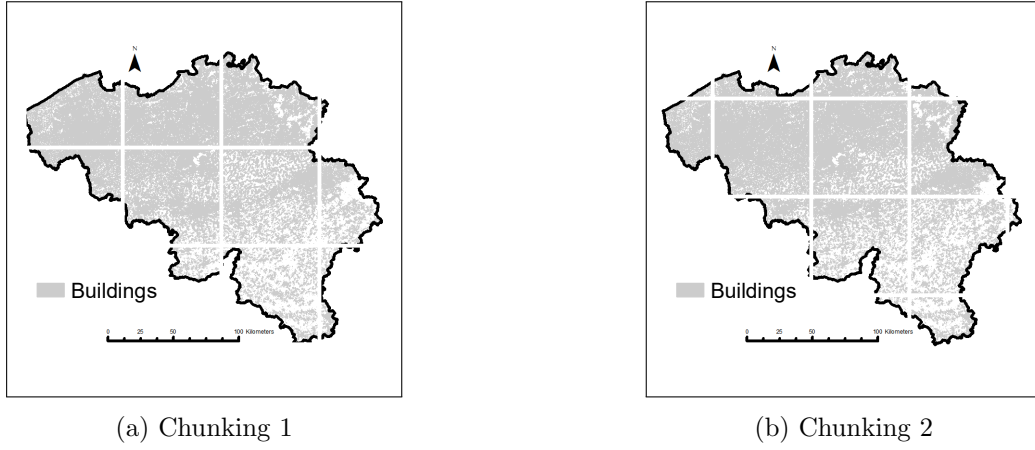


Figure A.1: The two chunkings

The final results of the method are not dependent on the initial chunking. Indeed, if each building is assigned the density value of the subgraph to which it belongs, then more than 92% of the buildings have exactly the same density with chunking 1 and chunking 2. If buildings within 1,000 metres of the chunking boundaries are not taken into account (boundary effect), the percentage of identical density between the two chunkings increases even more to 97%.

B Calculation of the Spatial Lag

The spatial lag is the product of the row standardised matrix of adjacency with the standardised vector of distance between the centroids of the buildings. The adjacency matrix (diagonal equal to zero) is obtained by the product of the incidence matrix with its transpose.

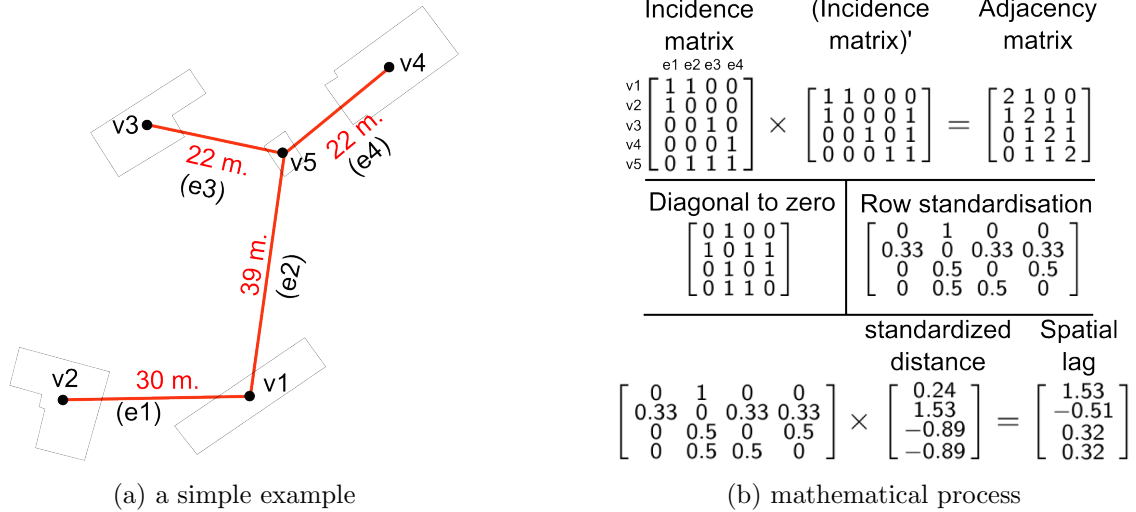


Figure B.2: Calculation of the spatial lag on a simple spatial toy structure of buildings

C Sensitivity to condition (1)

Condition (1) expresses the fact that if the total length of the graph is higher than 10,000 metres, the outlier is removed. With this condition, we ensure that the method detects fine-scale morphological features. We assume that a graph longer than 10,000 m is too heterogeneous with regard to its inter-building distances. It is therefore necessary to remove the outlier detected in the Moran scatterplot to allow the method to search for more homogeneous topological clusters as it does in condition (3). Note that the 10,000 m threshold was not selected at random: several values were tested

(see examples in Figures C.3, C.4 and C.5). With values larger than 10,000, the algorithm cannot distinguish different topological patterns, especially in the countryside. Conversely, with values lower than 10,000 the algorithm detects small local topological differences that are not relevant in terms of density measure.

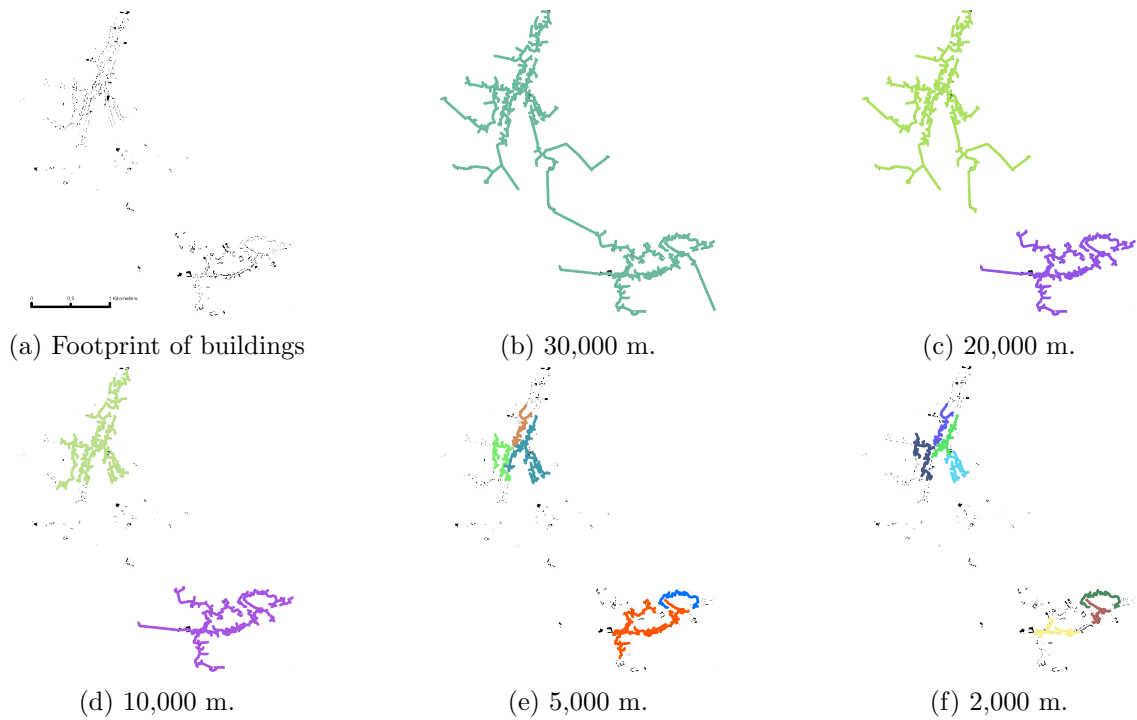


Figure C.3: Comparing different thresholds in condition (1) on a given pattern of buildings (a)

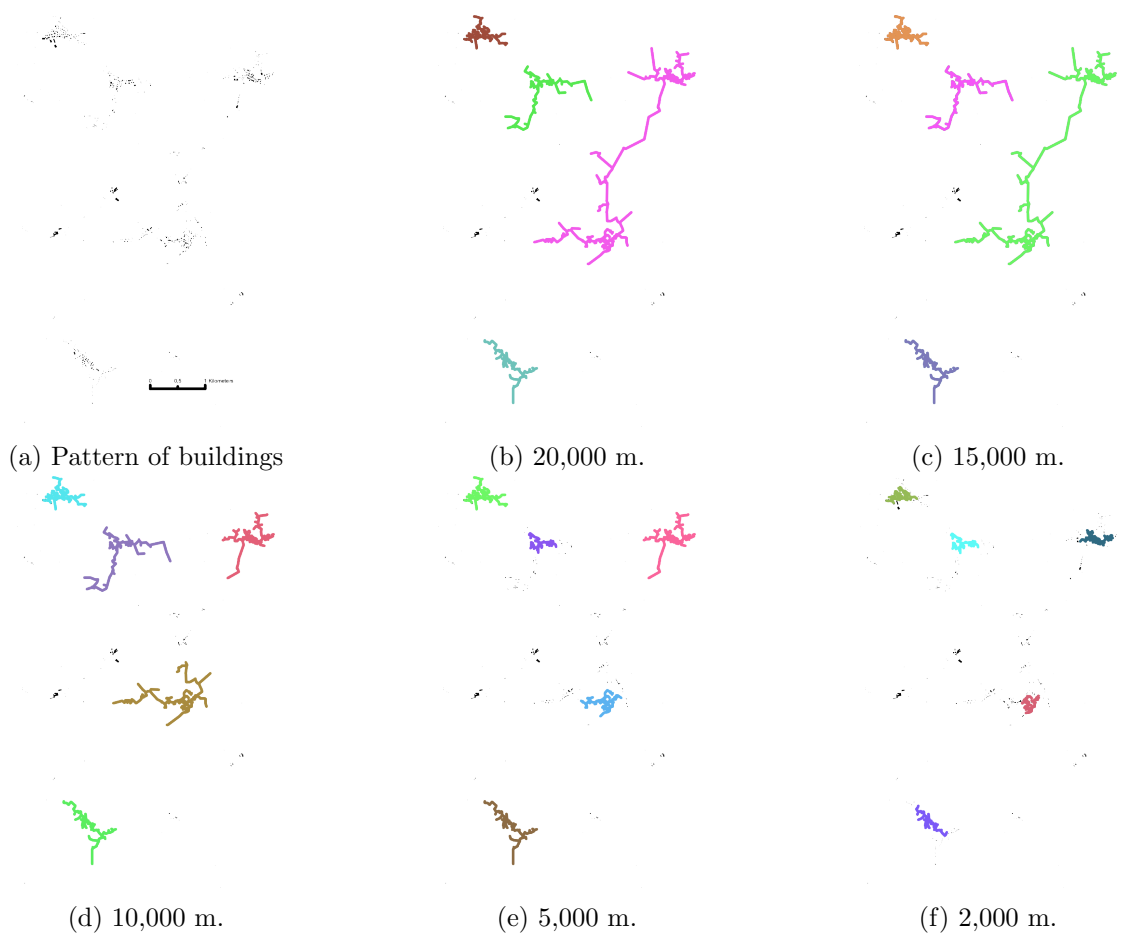


Figure C.4: Test of different value to condition (1) on a pattern of buildings

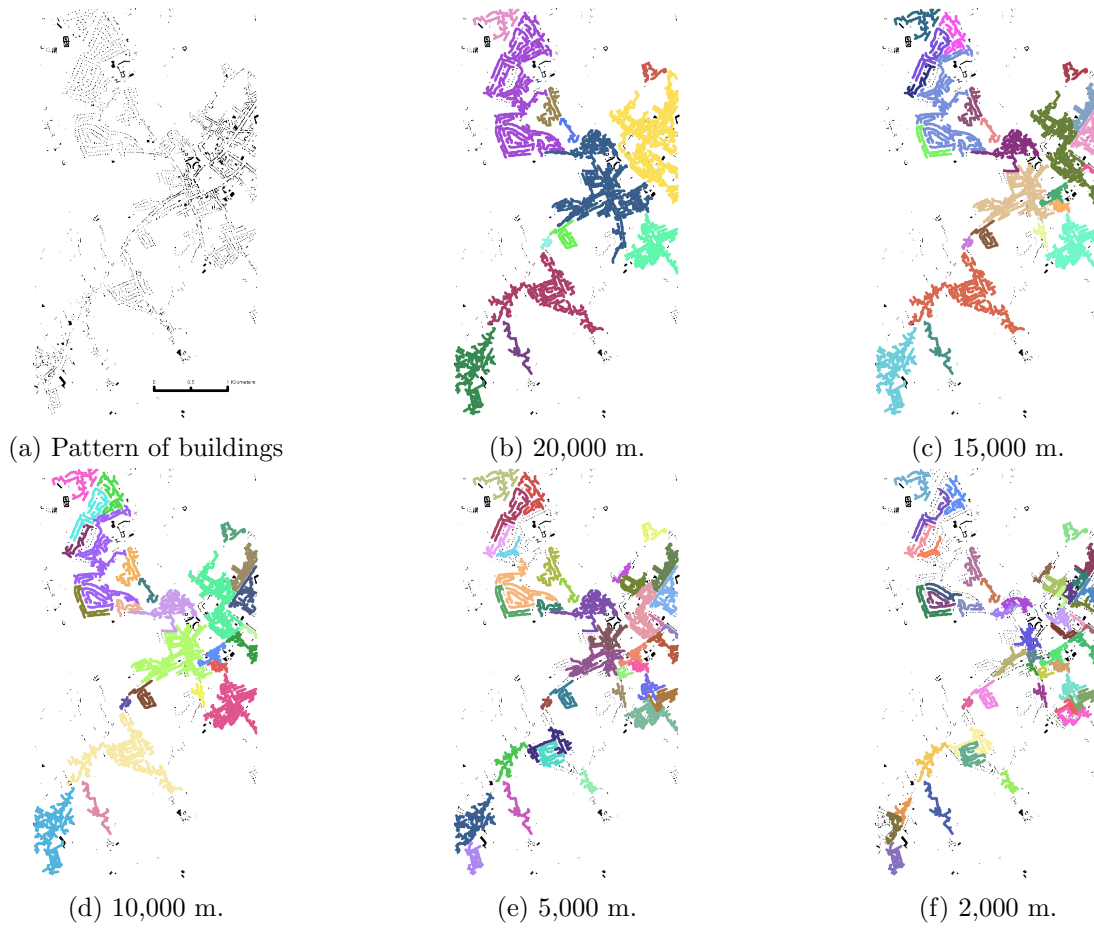


Figure C.5: Test of different value to condition (1) on a pattern of buildings

D The 26,462 clusters for Belgium according to the SDHC

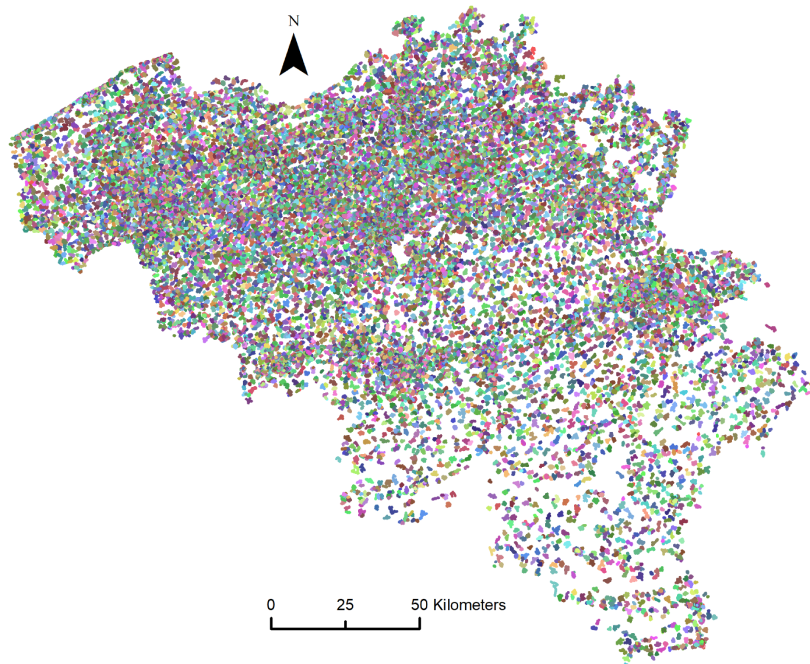


Figure D.6: The 26,462 clusters following spatial clustering

E The topology-based built density index computed for Belgium

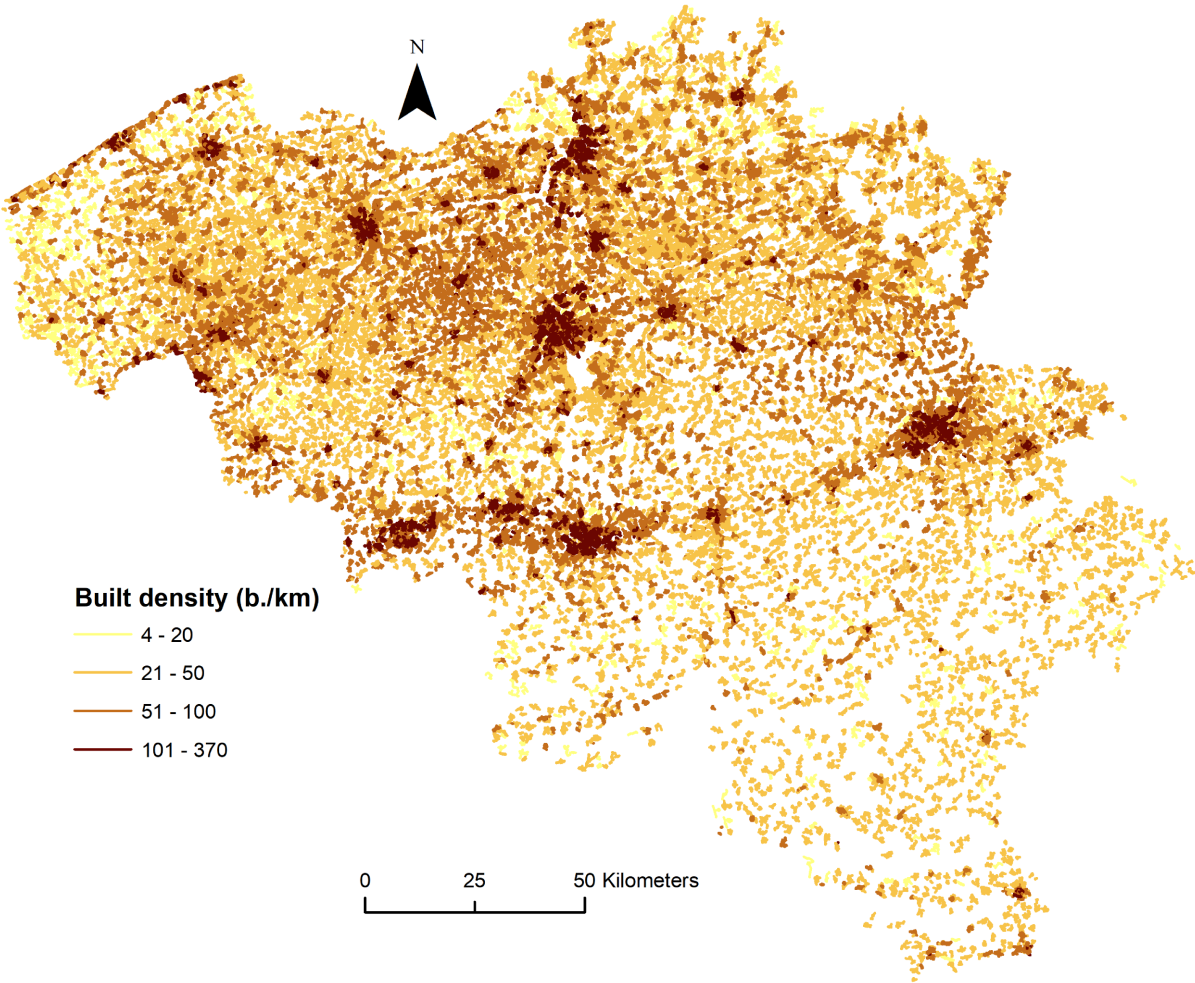


Figure E.7: The topology-based built density index computed for Belgium