# Identification of tissue types through urban morphometrics

The previous chapter defined the building blocks for morphometric analysis and proposed a conceptualisation needed for a comprehensive description of urban form, namely via relational framework of urban form. That can be now implemented in the next step towards a taxonomy of urban form - the algorithmic identification of urban tissue types as operational taxonomic units.

This chapter aims to provide theoretical and practical grounds to the method of automatic detection of distinct types of urban tissues. While similar research has been done before [@araldi2019; @berghauserpont2019; @bobkova2019; @schirmer2015], it was never linked to the coherent theory of morphometrics and numerical taxonomy, nor it was both inclusive in terms of a spectrum of characters used within a model and the spatial extent (see Chapter 2). Following pages present a method which aims to be inclusive in these terms and at the same time automatised and efficient to allow for examination of large datasets spanning across metropolitan regions.

Structurally, this chapter is divided into two main sections. The first one is focusing on the methodological propositions and briefly discuss theoretical grounds of the whole approach based on numerical taxonomy (section ), following with specification of two types of morphometric characters used within this research - primary (section ) and contextual (section ). The second part of methods (section ) outlines the cluster analysis and introduces a Gaussian Mixture Model clustering together with the issues of the selection of the optimal number of clusters and scalability of the method.

The second major section () applies the methods on the case study of Prague, Czechia and presents the results of all steps, eventually presenting the urban tissue types detected using cluster analysis based on morphometric assessment.

## Methodological proposition

The automatic detection of urban tissue types consists of multiple procedural steps detailed in the following section. It first requires specification of the principle of the recognition itself, followed by the design of actual methodological steps, starting from identification of morphometric characters for individual elements, finishing with the clustering algorithm detecting clusters. The structure of the method is reflected in the structure of the following sections.

### Tissue type as a homogenous cluster

Building on the propositions outlined in chapters 5 and 6, it is possible to define the hypothesis of urban tissue type recognition based on relational framework presented in the previous chapter. The working hypothesis hence stands as follows:

**Urban tissue types can be recognised by empirical measuring of the physical structure of urban fabric represented by *the relational analytical framework of urban form* in the form of homogenous clusters.**

The concept of urban tissue discussed in the previous chapter is fundamental for the understanding of the structure of cities we live in but at the same time a bit elusive in what *distinct* in the definition means. How much distinct two parts of the urban fabric needs to be to become different tissues? Who makes the decision, and based on what ground? While some have partial answers to these questions [@kropf1996], one remains unanswered. How to consistently identify urban tissues across metropolitan areas in an automatised, algorithmic way?

This research attempts to answer that question using the recognition of clusters based on the principles of numerical taxonomy (see chapter 3). Whilst previous chapters identified urban tissue as an OTU of urban form; we first have to identify those. For the cluster analysis recognising tissue types, we focus on a dual feature building-tessellation cell as the smallest entity of urban form. The whole cluster recognition is then based on the assumption that features recognised as a part of the same cluster (*species*) are, in fact, elements of the single urban tissue (where continuous) or of multiple individuals of the same kind of urban tissue (where discontinuous).

The urban form is full of exceptions from the pattern. Individual plots follow the different development process and are, in some cases, amalgamated or split [@conzen1960]. That does not happen to the rest of the same tissue at the same time (while it might or might not later), causing the constant emergence of exceptions from the pattern. The proposed method is working with two kinds of characters - primary and contextual to overcome the issue of exceptions.

The primary characters are those focusing on the individual elements and their relationships as identified in the relational framework (Chapter 6). These are mostly following what a character would be in biology. A typical example could be building height or area. Both are specific to each building. In the context of plots with subsequent internal development, buildings in the head and the tail of the plot might have significantly different values, i.e. exceptions from a continuous pattern.

As primary characters, by definition, do not describe the pattern but rather its elements, they are not optimal input for pattern detection algorithms. The second type of characters, contextual, has been designed specifically to turn values captured by primary characters into values describing the characters’ tendency in the area - to describe the pattern. As such, values are spatially lagged and can be used as an input for the cluster analysis. In the end, the data captured by contextual characters are used to cluster individual building-tessellation cell entities to statistically homogenous clusters, each capturing distinct kind of urban tissue.

The following section will detail the use of primary characters, contextual characters and the clustering method itself.

### Morphometric characters

The main scope of this research is not to develop new morphometric characters (even though there are some to fill the gaps), but to use existing knowledge in urban morphometrics and combine it in a systematic framework providing a comprehensive description of urban form. Chapter 4 mapped in detail the existing characters used across the field and the resulting database (Appendix A4) and classification is the basis for selection and definition of primary characters and to some extent, even contextual characters.

#### Primary characters

Primary characters describe different elements and their relationships as are identified within the relational framework of urban form. Building on the definition of the term *primary* from Oxford English Dictionary [@oxfordenglishdictionarya], we can define primary characters within the context of this cluster analysis as *characters occurring first in a sequence of methodological steps capturing individual features of urban form elements and their fundamental relations*. The link to the relational framework is crucial here as it defines which relations are meant and later reflected in the whole recognition model.

The choice of characters affects the result of cluster analysis; therefore, a potential embedded bias can cause a distortion of clustering results towards aspects occurring multiple times or other adverse effects. For that reason, specific principles of characters selection were defined.

##### Principles of character selection and definition

The idea of morphometric recognition of tissue types is based on mixture problem, and the selection of morphometric characters then build on the principles used within the selection of taxonomic characters in biology, as defined by @sneath1973. Building on the biological experience brings methodological grounds to the selection, and it is expected that a final set of characters selected according to these rules will describe an urban form suitable for recognition of clusters. However, the validity of the set is still only hypothetical, unlike the validity of individual characters which is tested throughout the selection process.

Selection strategy is tied to the classification of morphometric characters into categories as defined in chapter 4 and, more importantly, to the relational framework of urban form. There are three top-level aims for the selection of primary characters. The set should:

1. **Capture structural complexity of urban form by covering all categories of morphometric characters:**
   * dimension
   * shape
   * spatial distribution
   * intensity
   * connectivity
   * diversity

Each category captures different aspects of urban form, which all should be incorporated to derive a complex description of urban patterns. However, as different categories tend to focus on different scales and elements (see chapter 3), not all are likely to be equally represented. That is not an issue, rather a consequence of the nature of characters and the aim of the recognition model.

1. **Capture all fundamental elements of urban form**

In this case, in the context of the relational framework, these are:

* building
* street network
* morphological cell

Urban form is composed of multiple elements. Hence all fundamental ones should be captured. Here the attempt is to use as little of input data as possible, to extend the applicability of the whole model. Other elements (e.g., plot, open space, greenery) could be included, and the resulting model would likely be more precise, but the availability of such data is limited. This research uses only the three elements of urban form defined in the relational framework (coming from two data sources as MT is generated); hence this aim is focused on these only.

1. **‌Capture cross-scale complexity of urban form by covering all meaningful topological scales**

The relational framework defines three topological scales:

* single/small
* medium
* large

For tissue type recognition, not all scales are equally meaningful, as the spatial extent of tissues is usually restricted. However, S, M and L are all relevant for the scale of urban tissue and should all be represented. The city and its urban form are composed of complexities occurring on different scales. Capturing them all together within the single model allows the description of cross-scale complexity needed for systematic morphometric characteristics of built-up patterns.

To fulfil the aims, the relational framework comes to help with defined subsets as a combination of elements and scales, combining second and third aim into a single solution. Each of the subsets represents specific relations between specific elements, hence covering all subsets will help the pursuit of complex description. Then, having subsets, meaningful characters for each subset should be identified. The following procedure directly builds on the @sneath1973 to determine a methodical approach to the selection of the final set of morphometric characters. Steps of selection and elimination should follow this sequence:

1. **Extract all characters used in relevant literature**

The starting point should be a wide range of characters used within relevant literature, as such characters are already tested, and it is expected that they bear significant meaning in the description of urban form. This extraction has already been done in Chapter 4, so the resulting database of morphometric characters can be directly used. This database works as the principal source of characters. Due to its extent, it is expected that the majority of possible characters is included (keeping in mind identified gaps).

1. **Select characters using data intended to be used within each subset**

Not all characters are based on the same data sources used within this research and relational framework (see Chapter 6 for details). Some can be adapted (e.g., a morphological cell can be, in some cases, used as a spatial unit where plot is normally used), but some are based on the different sources of data. Characters which could not be used within subsets of the relational framework are then excluded from the initial selection.

1. **Adapt characters to fit the framework**

Those characters which are applicable, but are not readily available to be used within the relational framework should be adapted to fit the framework. It comprises mostly translation of plot-based characters to cell-based and metric-based characters into topology-based. Adaptation should be made with a sense of the meaning of each character which should not be significantly changed. Otherwise, its foundation in literature would be questionable, and a character should be seen as a newly developed one.

1. **Eliminate logical correlations**

Logically correlated characters should be omitted. Otherwise, the feature which is causing the correlation could distort the results of the clustering. Fully correlated characters caused by the causality (because A equals 1, B will be 1) have to be excluded, and only one should be kept. Partial logical correlation depends on the nature of other factors that are affecting character. If they reflect variation, we can include them. Also, “*characters that are tautological - those that are true by definition as well as those that are based on properties known to be obligatory - should not be included.*” [@sneath1973, p.104]

1. **Eliminate ineffective characters**

Due to the nature of the analysis, working with large-scale data, the process of measuring has to be computationally efficient. Some of the characters are not easily measurable, and it has to be evaluated whether the value of the characters would balance the difficulty of implementation and/or computational demand. Examples of such characters could be those based on expensive generative elements like axial maps [@hillier1996] or topological skeleton [@schirmer2015], or characters which implementation details are improperly documented in the literature.

1. **Add characters to minimise gaps in subsets**

The database of characters showed an imbalance of different categories and pointed out gaps, especially in the measuring of diversity. Moreover, the relational framework brings some subsets, which are often overlooked in the existing literature. While the overall balance between subsets and categories is only theoretical and would not reflect different nature and importance of different subsets, each of them should be sufficiently covered to capture the complex phenomenon of urban form structure. That may involve the development of new characters.

1. **Exclude invariant characters**

Some characters might be invariant over the entire sample. Those should not be included as they are not bearing any morphometric value. However, this exclusion is an ongoing process because it depends on actual measured values. Moreover, the invariance in one sample of data does not mean that the character overall is not valuable.

1. **Limit empirical correlation**

When we have the evidence that more than one factor affects two correlated characters within a study, regardless of whether this evidence comes from within a study or from outside, we would include both characters; otherwise, we would employ only one. We assume that at least some independent sources of the variation in any empirical correlation unless we have reason to believe otherwise.

1. **Exclude characters which cannot capture patterns**.

Some characters may show random-like spatial distribution, meaning that the geographical location has no relationship to the actual value. These characters do not have the value within this framework, as they are not able to describe a spatial pattern. To test the capability of each character to capture such patterns is used spatial autocorrelation analysis based on global Moran’s I [@moran1950notes]. Those characters without a significant autocorrelation should be excluded as they do not bear any value in the process of identification of tissue types.

1. **Balance scales and uniqueness of values**.

The set of taxonomic characters has to be balanced regarding the scale as well as *uniqueness* of values. Some of the initially identified characters are possible to measure on different topological scales. Due to the logical correlations between them, only one has to be used. The selection is trying to use the most appropriate in terms of the meaning of the character (which might be more suitable to the street edge than a block, for example). It also aims to limit the characters with limited uniqueness of values. Because the values are always stored on the smallest scale, the values of characters measured on the block scale are shared among all elements in the block. The intention is to limit those characters to a minimum.

The data on selection itself, starting from the database retrieved from chapter 4 is available as Appendix 7.1. It illustrates the selection process determining which characters should be part of the final set. The following section describes the final set of 74 primary characters only.

##### Identified set of primary characters

Based on the principles described in the section above, the following morphometric characters compose the final set of primary characters.

1. **Area of a building** is denoted as

and defined as an area covered by a building footprint in m2 .

1. **Height of a building** is denoted as

and defined as building height in m measured optimally as weighted mean height (in case of buildings with multiple parts of different height). It is a required input value not measured within the morphometric assessment itself.

1. **Volume of a building** is denoted as

and defined as building footprint multiplied by its height in m3.

1. **Perimeter of a building** is denoted as

and defined as the sum of lengths of the building exterior walls in m.

1. **Courtyard area of a building** is denoted as

and defined as the sum of areas of interior holes in footprint polygons in m2.

1. **Form factor of a building** is denoted as
2. .

It captures three-dimensional unitless shape characteristic of a building envelope unbiased by the building size [@bourdic2012].

1. **Volume to façade ratio of a building** is denoted as
2. .

It captures the aspect of the three-dimensional shape of a building envelope able to distinguish building types, as shown by @schirmer2015. It can be seen as a proxy of volumetric compactness.

1. **Circular compactness of a building** is denoted as

where is an area of minimal enclosing circle. It captures the relation of building footprint shape to its minimal enclosing circle, illustrating the similarity of shape and circle [@dibble2017].

1. **Corners of a building** is denoted as

where is defined as a vertex of building exterior shape with an angle between adjacent line segments 170 degrees. It uses only external shape (shapely.geometry.exterior), courtyards are not included. Character is adapted from [@steiniger2008] to exclude non-corner-like vertices.

1. **Squareness of a building** is denoted as

where is the deviation of angle of corner from 90 degrees and is a number of corners.

1. **Equivalent rectangular index of a building** is denoted as
2. $ERI\_{blg} = \sqrt{{a\_{blg}} \over {a\_{blgB}}} \* {p\_{blgB} \over p\_{blg}}$

where is an area of a minimal rotated bounding rectangle of a building (MBR) footprint and its perimeter of MBR. It is a measure of shape complexity identified by @basaraner2017 as the shape characters with the best performance.

1. **Elongation of a building** is denoted as

where is length of MBR and is width of MBR. It captures the ratio of shorter to the longer dimension of MBR to indirectly capture the deviation of the shape from a square [@schirmer2015].

1. **Centroid - corner distance deviation of a building** is denoted as

where is a distance between centroid and corner and is mean of all distances. It captures a variety of shape. As a corner is considered vertex with angle < 170º to reflect potential circularity of object and topological imprecision of building polygon.

1. **Centroid - corner mean distance of a building** is denoted as

where is a distance between centroid and corner . It is a character measuring a dimension of the object dependent on its shape [@schirmer2015].

1. **Solar orientation of a building** is denoted as

where is an orientation of the longest axis of bounding rectangle in a range 0 - 45. It captures the deviation of orientation from cardinal directions. There are multiple ways of capturing orientation of a polygon. As reported by @yan2007, @duchene2003 assessed five different options (longest edge, weighted bisector, wall average, statistical weighting, bounding rectangle) and concluded a bounding rectangle as the most appropriate. Deviation from cardinal directions is used to avoid sudden changes between square-like objects.

1. **Street alignment of a building** is denoted as

where is a solar orientation of the building and is a solar orientation of the street edge. It reflects the relationship between the building and its street, whether it is facing the street directly or indirectly [@schirmer2015].

1. **Cell alignment of a building** is denoted as

where is a solar orientation of tessellation cell. It reflects the relationship between a building and its cell.

These seventeen characters are capturing aspects of individual building (topological context 0). Following are measuring aspects of tessellation cells on the same level.

1. **Longest axis length of a tessellation cell** is denoted as

where is a diameter of the minimal circumscribed circle around the tessellation cell polygon. The axis itself does not have to be fully within the polygon. It could be seen as a proxy of plot depth for tessellation-based analysis.

1. **Area of a tessellation cell** is denoted as

and defined as an area covered by a tessellation cell footprint in m2.

1. **Circular compactness of a tessellation cell** is denoted as

where is an area of minimal enclosing circle. It captures the relation of tessellation cell footprint shape to its minimal enclosing circle, illustrating the similarity of shape and circle.

1. **Equivalent rectangular index of a tessellation cell** is denoted as
2. $ERI\_{cell} = \sqrt{{a\_{cell}} \over {a\_{cellB}}} \* {p\_{cellB} \over p\_{cell}}$

where is an area of the minimal rotated bounding rectangle of a tessellation cell (MBR) footprint and its perimeter of MBR. It is a measure of shape complexity identified by @basaraner2017 as a shape character of the best performance.

1. **Solar orientation of a tessellation cell** is denoted as

where is an orientation of the longest axis of bounding rectangle in a range 0 - 45. It captures the deviation of orientation from cardinal directions.

1. **Street alignment of a building** is denoted as

where is a solar orientation of tessellation cell and is a solar orientation of the street edge. It reflects the relationship between tessellation cell and its street, whether it is facing the street directly or indirectly.

1. **Coverage area ratio of a tessellation cell** is denoted as

where is an area of a building and is an area of related tessellation cell [@schirmer2015]. Coverage area ratio (CAR) is one of the commonly used characters capturing *intensity* of development. However, the definitions vary based on the spatial unit.

1. **Floor area ratio of a tessellation cell** is denoted as

where is a floor area of a building and is an area of related tessellation cell. Floor area could be computed based on the number of levels or using an approximation based on building height.

1. **Length of a street segment** is denoted as

and defined as a length of a LineString geometry in metres [@dibble2017; @gil2012].

1. **Width of a street profile** is denoted as

where is width of a street section i. The algorithm generates street sections every 3 meters alongside the street segment, and measures mean value. In the case of the open-ended street, 50 metres is used as a perception-based proximity limit [@araldi2019].

1. **Height of a street profile** is denoted as

where is mean height of a street section i. The algorithm generates street sections every 3 meters alongside the street segment, and measures mean value [@araldi2019].

1. **Height to width ratio of a street profile** is denoted as

where is mean height of a street section i and is the width of a street section i. The algorithm generates street sections every 3 meters alongside the street segment, and measures mean value [@araldi2019].

1. **Openness of a street profile** is denoted as

where is a sum of section lines (left and right sides separately) intersecting buildings and total number of street sections. The algorithm generates street sections every 3 meters alongside the street segment.

1. **Width deviation of a street profile** is denoted as

where is width of a street section i and is mean width. The algorithm generates street sections every 3 meters alongside the street segment.

1. **Height deviation of a street profile** is denoted as

where is height of a street section i and is mean height. The algorithm generates street sections every 3 meters alongside the street segment.

1. **Linearity of a street segment** is denoted as

where is Euclidean distance between endpoints of a street segment and is a street segment length. It captures the deviation of a segment shape from a straight line. It is adapted from @araldi2019.

1. **Area covered by a street segment** is denoted as

where is an area of tessellation cell belonging to the street segment. It captures the area which is likely served by each segment.

1. **Buildings per meter of a street segment** is denoted as

where is a number of buildings belonging to a street segment and is a length of a street segment. It reflects the granularity of development along each segment.

1. **Area covered by a street node** is denoted as

where is an area of tessellation cell belonging to the street node. It captures the area which is likely served by each node.

1. **Shared walls ratio of adjacent buildings** is denoted as

where is a length of a perimeter shared with adjacent buildings and is a perimeter of a building. It captures the amount of wall space facing the open space [@hamaina2012a].

1. **Alignment of neighbouring buildings** is denoted as

where is the solar orientation of a building and is the solar orientation of building on a neighbouring tessellation cell. It calculates the mean deviation of solar orientation of buildings on adjacent cells from a building. It is adapted from @hijazi2016.

1. **‌Mean distance to neighbouring buildings** is denoted as

where is a distance between building and building on a neighbouring tessellation cell. It is adapted from @hijazi2016. It captures the average proximity to other buildings.

1. **Weighted neighbours of a tessellation cell** is denoted as

where is a number of cell neighbours and is a perimeter of a cell. It reflects granularity of morphological tessellation.

1. **Area covered by neighbouring cells** is denoted as

where is area of tessellation cell within topological distance 1. It captures the scale of morphological tessellation.

1. **Reached cells by neighbouring segments** is denoted as

where is number of tessellation cells on segment within topological distance 1. It captures accessible granularity.

1. **Reached area by neighbouring segments** is denoted as

where is an area covered by a street segment within topological distance 1. It captures an accessible area.

1. **Degree of a street node** is denoted as

where is an edge of a street network between node and node . It reflects the basic degree centrality.

1. **Mean distance to neighbouring nodes from a street node** is denoted as

where is a distance between node and node within topological distance 1. It captures the average proximity to other nodes.

1. **Reached cells by neighbouring nodes** is denoted as

where is number of tessellation cells on node within topological distance 1. It captures accessible granularity.

1. **Reached area by neighbouring nodes** is denoted as

where is an area covered by a street node within topological distance 1. It captures an accessible area.

1. **Number of courtyards of adjacent buildings** is denoted as

where is a number of interior rings of a polygon composed of footprints of adjacent buildings [@schirmer2015].

1. **Perimeter wall length of adjacent buildings** is denoted as

where is a length of an exterior ring of a polygon composed of footprints of adjacent buildings.

1. **Mean inter-building distance between neighbouring buildings** is denoted as

where is a distance between building and building on a tessellation cell within topological distance 3. It is adapted from @caruso2017. It captures the average proximity between buildings.

1. **‌Building adjacency of neighbouring buildings** is denoted as

where is a number of joined built-up structures within topological distance three and is a number of buildings within topological distance 3. It is adapted from @vanderhaegen2017.

1. **Gross floor area ratio of neighbouring tessellation cells** is denoted as

where is a floor area ratio of tessellation cell and is an area of tessellation cell within topological distance 3. Based on @dibble2017.

1. **Weighted reached blocks of neighbouring tessellation cells** is denoted as

where is a number of blocks within topological distance three and is an area of tessellation cell within topological distance three.

1. **Area of a block** is denoted as

and defined as an area covered by a block footprint in m2.

1. **Perimeter of a block** is denoted as

and defined as lengths of the block polygon exterior in m.

1. **Circular compactness of a block** is denoted as

where is an area of minimal enclosing circle. It captures the relation of block footprint shape to its minimal enclosing circle, illustrating the similarity of shape and circle.

1. **Equivalent rectangular index of a block** is denoted as
2. $ERI\_{blk} = \sqrt{{a\_{blk}} \over {a\_{blkB}}} \* {p\_{blkB} \over p\_{blk}}$

where is an area of the minimal rotated bounding rectangle of a block (MBR) footprint and its perimeter of MBR.

1. **Compactness-weighted axis of a block** is denoted as

where is a diameter of the minimal circumscribed circle around the block polygon, is an area of a block and is a perimeter of a block. It is a proxy of permeability of an area.[@feliciotti2018]

1. **Solar orientation of a block** is denoted as

where is an orientation of the longest axis of bounding rectangle in a range 0 - 45. It captures the deviation of orientation from cardinal directions.

1. **Weighted neighbours of a block** is denoted as

where is a number of block neighbours and is a perimeter of a block. It reflects granularity of a mesh of blocks.

1. **‌Weighted cells of a block** is denoted as

where is a number of cells composing a block and is an area of a block. It captures the granularity of each block.

1. **Local meshedness of a street network** is denoted as

where is a number of edges in a subgraph, and is the number of nodes in a subgraph [@feliciotti2018]. A subgraph is defined as a network within topological distance five around a node.

1. **Mean segment length of a street network** is denoted as

where is a length of a street segment within a topological distance 3 around a segment.

1. **Cul-de-sac length of a street network** is denoted as

where is a length of a street segment within a topological distance 3 around a node.

1. **Reached cells by street network segments** is denoted as

where is number of tessellation cells on segment within topological distance 3. It captures accessible granularity.

1. **Node density of a street network** is denoted as

where is a number of nodes within a subgraph and is a length of a segment within a subgraph. A subgraph is defined as a network within topological distance five around a node.

1. **Reached cells by street network nodes** is denoted as

where is number of tessellation cells on node within topological distance 3. It captures accessible granularity.

1. **Reached area by street network nodes** is denoted as

where is an area covered by a street node within topological distance 3. It captures an accessible area.

1. **Proportion of cul-de-sacs within a street network** is denoted as

where is a node whiting topological distance five around a node. Adapted from [@boeing2017a].

1. **Proportion of 3-way intersections within a street network** is denoted as

where is a node whiting topological distance five around a node. Adapted from [@boeing2017a].

1. **Proportion of 4-way intersections within a street network** is denoted as

where is a node whiting topological distance five around a node. Adapted from [@boeing2017a].

1. **Weighted node density of a street network** is denoted as

where is a degree of a node within a subgraph and is a length of a segment within a subgraph. A subgraph is defined as a network within topological distance five around a node.

1. **Local closeness centrality of a street network** is denoted as

where is the shortest-path distance between and , and is the number of nodes within a subgraph. A subgraph is defined as a network within topological distance five around a node.

1. **Square clustering of a street network** is denoted as

where are the number of common neighbours of and other than (ie squares), and , where if and are connected and 0 otherwise [@PhysRevE.72.056127].

The final selection consists of 74 morphometric characters spanning across the subsets of the relational framework and covering all categories, even though not equally. The set is a result of the identification process proposed above. As such, it should provide an unbiased and non-skewed description of each of the elements.

#### Contextual characters

Looking at the primary characters and their spatial distribution, they could be abrupt and do not necessarily capture urban patterns as we would like them to (even though all capture some patterns as per spatial autocorrelation).

The characters defined above have to be expressed using their contextual, spatially lagged versions to become useful for pattern detection within the recognition model, which does not employ direct spatial constraints. *Context* here is defined as vicinity of each tessellation cell within three topological steps on morphological tessellation. That covers approximately 40 nearest neighbours (median 40, standard deviation ~13.4 based on Prague) providing a balance between the spatial extent large enough to capture a pattern and at the same time small enough not to over-smooth boundaries between different patterns (see Appendix 7.3 for sectional diagram analysis).

Within this method, four types of contextual characters are proposed. One is capturing a local central tendency and three capturing the properties of the distribution of values within the context. For each of the primary characters, each of the contextual is then calculated and then used within the clustering algorithm itself. The resulting set of used characters is then composed of 4 times 74 characters, giving 296 individual contextual characters.

##### Local central tendency

Statistics knows central tendency as a measure of a typical value for a probabilistic distribution [@weisberg1992central, p.2]. Based on a set of data of unknown distribution, central tendency aims to simplify the whole set into one representative number. In the case of morphometric characters, we can measure the central tendency of values of a single character across the whole case study, but that would not give us much information. As contextual characters are defined on three topological steps, it is proposed to measure *local central tendency*, thus a value unique for each building measured as a typical within its immediate context.

Commonly used measures of central tendency are mean, median or mode [@wilcox2003modern]. Each of them fits a different purpose. If one wants to use the arithmetic mean to determine central values, underlying distribution should not be skewed. Otherwise, outliers may significantly affect the resulting value. A mode is, by definition, not suitable for continuous variables like those obtained in primary characters. Median is the most robust of all, measuring the middle value. However, the robustness comes at a cost - the shape of a distribution is not reflected at all. Another option is to find a middle ground between easily distorted mean and robust median using truncated mean. Instead of computing arithmetic mean of the whole distribution, we can work with interquartile (smallest and largest 25% are omitted) or interdecile (smallest and largest 10% are omitted) range to minimise the outlier effect on the mean.

The distribution of values of individual characters vary and in some cases, tends to be skewed. As shown in Appendix 7.4 analysing the difference between mean, interdecile mean, interquartile mean and median (being equal to extremely truncated mean) on a selection of 8 characters, it is clear, that majority of data is somewhat asymmetric, causing volatility of mean, which should not be used in such cases. The question is then limited to the distinction between the median and truncated means (leaving aside midhinge and similar estimators). The data indicate that the difference between median and interquartile mean is minimal (but still present, e.g., in the case of *shared walls ratio*). As interquartile mean uses more information than the median, while being similarly robust to outliers, this research settles on implementation of the interquartile mean as a measure of local central tendency, denoted as

1. ,

where is selected primary character. Formula assumes sorted values.

##### Properties of a distribution

Apart from a local central tendency (in the geographical context sometimes present in literature also as spatial lag [@anselin2001spatial]), which aims to capture representative value, it is fundamental to understand how the actual distribution of values within the *context* looks like.

That could be approached in multiple ways. Three notable are 1) capturing the *diversity* of values within the local context, 2) measuring the *statistical dispersion* of values, and 3) measuring *similarity of a target and an actual distribution* of values, like in the case of inequality.

###### Diversity

While discussion on the importance of diversity has been central to urban discourse since the era of Jane Jacobs [@jacobs1961], there is not a very wide range of characters actually measuring diversity. The research focuses mostly on Simpson’s diversity index [@bobkova2017; @feliciotti2018], developed initially for categorical, not continuous variables and hence relying on pre-defined “bins” (classes of values). For example, @bobkova2017 use this index to measure the diversity of plot sizes, but their binning into intervals based on the actual case-specific values makes the comparability of outcomes limited: if we apply the same formula to another place, we will get different binning and different results. This appears to be a rather ubiquitous problem in applying Simpson’s diversity index, i.e., it is necessary to set a finite set of pre-established bins prior to undertaking the analysis. However, despite the need for urban morphometric analysis to produce comparable outcomes, it is challenging to ensure specific descriptiveness to “universal” pre-defined bins. The use of the Simpson’s diversity index in ecology is encouraged [@jost2006] because ecologists have a finite number of groups enabling them to pre-define all bins appropriately (moreover, bins are usually not defined on a continuous numerical scale). However, this is not often the case in urban morphology. Simpson’s diversity index and similar characters based on binning provide values specific to individual cases where binning is set and have to be interpreted as such.

Recent literature shows that there might be alternative ways to measure the diversity of morphological characters. @caruso2017 and @araldi2019 applied the Local Index of Spatial Autocorrelation (LISA) in the form of local Moran’s I, defined as “the weighted product of the difference to the mean of the value of a variable at a certain observation and the same difference for all other observations, with more weight given to the observations in close spatial proximity.” [@caruso2017, p.84] LISA aims to identify clusters of similar values in space, describing their similarity or dissimilarity, which could be seen as a proxy for diversity, but due to the limited number of significant categories (High-High, High-Low, Low-High, Low-Low), its application in this context is limited and somewhat reductionist.

###### Statistical dispersion

The second approach is to measure statistical dispersion, i.e., the ratio to which the distribution is stretched (wide distribution) or squeezed (narrow distribution). Together with the central tendency, dispersion is often used to describe the basic properties of distributions.

There are multiple ways of measuring dispersion. The most used are probably standard deviation, range or interquartile range as examples of *dimensional* (resulting value have the same units as the original character) measures. Dimensional measures of dispersion are the most common as they are generally easy to understand and interpret. Similarly to the measure of central tendency, all can be measured on the full range of values or a limited one, usually again as interquartile (IQ) or interdecile (ID) range. Dimensionless measures are not expressed in the same units as original characters, so while a dimensional measure of dispersion for building area will be in meters, dimensionless will have no units (the values are relative). Among dimensionless measures are the coefficient of variation (CoV) or quartile coefficient of dispersion (QCoD).

###### Distribution matching

The third approach focuses on the comparison of the actual distribution and the ideal distribution. One example is a test whether such a distribution follows the principle of the Power Law used by @salat2017. However, that is not a straightforward measurement, especially if the distribution is of a different shape; it is hard to quantify the relationship. Specific distribution is also embedded in the Gini index customarily used to measure inequality or indirectly in entropy-based indices like Theil index of inequality (a special case of the generalised entropy index) [@novotny2007measurement].

###### Comparison of potential characters

To understand the properties and behaviour of potential characters capturing properties of distributions on the real morphometric data, wide selection of the most relevant characters from each group is analysed as a way of selecting the most appropriate ones to be used as contextual characters.

In terms of diversity measures, the key question is not which one should be used, either Simpson’s diversity index as in @bobkova2017 or Gini-Simpson diversity index as in @feliciotti2018, but how to define binning as that can significantly affect the resulting diversity values. For that reason, Simpson’s diversity is tested using *natural breaks* [@jenks1967data] (number of classes is based on the Goodness of Absolute Deviation Fit (GADF) [@rey2007]), *Head Tail breaks* [@jiang2013] and *quantiles* (5 and 10 bins). The reason for the inclusion of Simpson’s diversity index, even though it may not be fully comparable across cases is the fact that the recognition of tissue types is always local, always case-specific.

Dimensional characters capturing dispersion included in comparison are *standard deviation (SD)*, *range*, and *absolute deviations (median - MAD, average - AAD)*. Both standard deviation and range are measured for IQ, ID and unrestricted range of values. Included dimensionless characters are *coefficient of variation (CoV)*, *quartile coefficient of dispersion (QCoD)*.

The last group is represented by both *Gini index*, and *Theil index*; both measured for IQ, ID and unrestricted range of values.

Using four morphometric characters as test data - *area of a building*, *height of a building*, *coverage area ratio of tessellation cell* and *floor area ratio of tessellation cell*, all potential contextual characters listed above are measured on three topological steps around each building. Resulting spatial distribution is visually assessed to eliminate those unfit for pattern recognition, either for relative randomness of result or significant outlier effect (typically present in measures based on unrestricted range of values). Finally, a correlation matrix is used to identify potential overlaps and uniqueness of values leading to the selection of optimal contextual characters.

The results section of this assessment is available in Appendix 7.5.

###### Resulting selection of contextual characters

While the complete results of the analysis are available as Appendix 7.5, the main conclusions are as follows.

Because some of the values follow exponential (power-law or similar) distribution within the whole dataset, the binning method for Simpson’s diversity index has to acknowledge that. For that reason, HeadTail Breaks are the ideal method as it is specifically tailored to exponential distributions [@jiang2013]. Those characters which do not resemble exponential distribution should use natural breaks or similar classification method sensitive to the actual distribution, rather than quantiles, which may cause significant disruptions and very similar values may fall into multiple bins causing high diversity values in place where is not.

Within measures of statistical dispersion, IQ range and IQ standard deviation are better in capturing boundaries between types of development and are robust to outliers. Interquartile range was used by @dibble2017 and is easier to interpret. Due to its definition, CoV tends to infinity when the mean value tends to zero, being very sensitive to changes of the mean.

Theil index and Gini index are both used to assess inequality, but Theil index, unlike Gini, is decomposable to within-group inequality and between-group differences [@novotny2007measurement], making it more suitable for spatial analysis than Gini index would be. ID values used within the Theil index are better than other ranges as the resulting analysis is more sensitive, while outlier effect is still minimal. ID captures, for example, inner structures of blocks better than IQ, where such structures might be filtered out. It may help to distinguish between blocks with and without internal buildings.

The final selection of contextual characters is then composed of four distinct uncorrelated characters. Local central tendency is captured by *interquartile mean (IQM)* and describe the most representative value. The local measure of statistical dispersion is represented by *Interquartile range (IQR)* as dimensional character which expresses the range of values around IQM, indicating where the values mostly lie. IQR is denoted as

1. ,

where is third quartile of selected primary character and first quartile. Formula assumes sorted values. *Interdecile Theil index (IDT)* is denoted as

1. ,

where is selected primary character, and describes the (in)equality of distribution of values. Finally, *Simpson’s diversity index (SDI)* is denoted as

1. ,

where is richness expressed as number of bins, is the number of features the th type and is the total number of features. It captures the presence of various classes of values. Together, these four characters have a potential to describe spatial distribution of morphometric values within a set context.

For the clarity in terms of classification of contextual characters, IQM inherits the category from the primary parental character, while IQR, IDT and SDI all fall into *diversity* category.

After linking together primary and contextual characters, each of the primary 74 characters is represented by all four contextual, based on the values measured within three topological steps on morphological tessellation around each building. That gives 296 contextual characters in total, the set which is spatially autocorrelated by definition and hence can be used within the clustering method to identify distinct homogenous clusters representing tissue types. The fact that all input data for clustering are measured using this spatially lagged method ensures that spatial clusters should be geographically coherent and mostly continuous. The nature of data allows the use of spatially unconstrained clustering methods.

Importance of the proper selection of morphometric characters and the effect it may have on the overall results in not debatable. A robust method described above is employed starting from the selection of primary morphometric characters from literature and their adaptation to fit *relational framework of the urban form* and to minimise the potential error in selection. The resulting set of 74 characters is established to cover a wide range of descriptive features capturing urban form configuration from dimensions of individual elements, through spatial distribution to diversity. Four contextual characters are introduced to describe a local central tendency and variation in the area capturing morphological patterns, rather than a description of individual elements. These, combined, have the potential to capture the nature of each of the primary characters and their behaviour in the immediate spatial context.

### Identification of clusters

The actual identification of urban tissue types is in principle statistical clustering of building/tessellation cell features with similar information about itself and its context. Moreover, clusters, in this case, need to be contiguous. As mentioned above, the solution of the contiguity issue is built in the design of contextual characters. All characters are spatially autocorrelated by design. There is a significant overlap between areas used for computation of contextual characters of two neighbouring cells that indirectly supports contiguity of clustering. However, this solution may result in less defined boundaries between two clusters, and every edge of the cluster needs to be interpreted as fuzzy rather than defined. Specific mitigation of over-smoothing of boundaries is embedded in the design of contextual characters as they are mostly based on truncated values, which not only eliminate outlier effect but also result in more defined boundaries.

The general principle of clustering is using the data to iteratively determine the optimal division of observed data into homogenous clusters. In case of probabilistic methods, this prediction can have associated probability that the chosen cluster is the correct one and have the probability of belonging to every other cluster.

Current progress in machine learning brings various methods to choose from. Every clustering method follows different principles and is able to identify different kinds of clusters. The most common is *k*-means clustering [@macqueen1967some] and its derivatives (*k*-medoid [@park2009simple], *k*-median [@jain1988algorithms] or Gaussian mixture models [@reynolds2009gaussian]). The algorithm divides observations into predefined *k* clusters based on the nearest mean value to minimise within-cluster variance based on squared Euclidean distances between observations [@reynolds2009gaussian]. As a result, clusters tend to be of a similar size. In the case of urban form, it is unlikely that each urban typology is equally present, rendering the use of *k*-means as less fit for the purpose. It is expected that cluster will be on unequal size and also of unequal density - clusters capturing rigid patterns will be more densely packed than those capturing more diverse areas. The clustering algorithm needs to take into account all these requirements stemming from the specificity of urban morphometric data. Moreover, every building is by definition part of some urban tissue, which could be very heterogeneous, meaning that algorithms expecting and identifying noise (in this case buildings which do not belong to any cluster) in the data like DBSCAN [@ester1996density], HDBSCAN [@mcinnes2017hdbscan] or OPTICS [@ankerst1999optics] are not ideal either.

#### Gaussian Mixture Model clustering

Clustering method which does reflect the nature of the problem is the Gaussian Mixture Model (GMM), a probabilistic derivative of *k*-means [@reynolds2009gaussian]. Unlike the *k*-means itself, it does not rely on squared Euclidean distances only but is based on the assumption that a Gaussian distribution represents each dimension of each cluster. Hence the cluster itself is defined by a mixture of Gaussians. Where k-means looks for clusters of similar extent, GMMs embedded expectation-maximization (EM) algorithm, which allows identification of different shapes. EM is an iterative method which starts from random points (like k-means) but can find the maximum likelihood of parameters of expected underlying Gaussians.

GMM is probabilistic clustering, which means that it defines *n* components (equal to k in k-means) and their expected underlying Gaussian distributions and then predicts the probability that each observation belongs to each cluster. The exemplar observation *A* can then belong to cluster *1* with the probability *0.6*, to cluster *2* with the probability *0.35* and to clusters *3 - 9* with probability *<0.01*, considering 9-component-GMM.

The result of GMM applied to the illustrative artificial dataset, as shown on figure , illustrates both resulting labelling, which correctly identifies known clusters, and underlying Gaussian distributions shown as ellipses, where the shade reflects the probability that the points in hyperspace belong to the selected cluster.

GMM clustering (4 components) of the illustrative artificial two dimensional (x, y) dataset containing four known clusters. All clusters are fairly successfully distinguished. The figure also shows underlying Gaussian distributions as ellipses reflecting the probability by the change of the shade.

Because in the first step of GMM, the seed points are placed randomly, this placement might affect the resulting model. This specificity makes GMM non-deterministic clustering, which means that each run will likely result in (slightly) different clusters. GMM has to be done repeatedly in several initializations, of which the best should be used to ensure the stability of the clustering.

Within the context of urban morphology, the method has been applied within a similar classification task by @jochem2020. Within this research, a sklearn.mixture.GaussianMixture() implementation of GMM within open-source python package scikit-learn v.0.22 [@pedregosa2011scikit] is used. Further details on the exact algorithm are available in scikit-learn documentation and code.

#### Levels of clustering resolution and its scalability

The ideal outcome of the tissue type recognition is each cluster as a distinct urban tissue type. However, the definition of urban tissue does not specify the threshold when two similar parts of the city are still the same tissue type and when they become a different one. This issue is mirrored in the clustering method. The ideal outcome of clustering is the optimal number of clusters based on the actual structure of the observed data. That might not be straightforward to determine as better-looking clustering (from the statistical, not visual perspective) might be just overfitted. Moreover, the relation between resulting clusters and urban tissues is always questionable as there is no ground truth for either of them. Detecting 5 large cluster in the whole Prague would likely be based on under-fitted model and cluster would not represent urban tissues in the traditional sense, but their aggregations. On the other hand, detecting 100 would likely represent the over-fitted model, and each cluster would be only a part of a tissue. It is expected that the statistically optimal number of clusters should be close to what we would typically call urban tissue. However, this link requires further interpretative work, which should happen based on the taxonomy of clusters to allow scale-dependent flexibility.

##### Number of components

Gaussian Mixture Model clustering requires, similarly to k-means, specification of a number of components of the model (i.e., clusters) before clustering. However, that number is usually not known, especially in the case of urban form. Assumptions can be made based on the expert knowledge, but that would limit the application and unsupervised nature of the whole process and go against the prepositions set in chapters 1 and 5.

The way around is to estimate the ideal number of components based on the goodness of fit of the model for each of them. That means that the GMM is trained multiple times based on the range of feasible options of the number of components and each of the models is then assessed against the whole dataset (to determine how well clusters are distinguished). The assessment is of a quantitative statistical nature, keeping the method relatively unsupervised. The only input researcher needs to make at this stage is an interpretation of the resulting values and the curve of the goodness of fit to specify the number of components for the final clustering.

###### Goodness of fit

The *goodness of fit* measures a fit of a trained model to a set of observations (e.g., the original dataset)[@d1986goodness]. It describes how consistent is the distribution of clustered model to the distribution of the whole dataset. With K-means clustering is often used silhouette method, which could, in theory, be used with GMM as well. Another option is measuring the average log-likelihood score. However, the optimal method for GMM is the Bayesian information criterion (BIC), a model based partly on the likelihood function [@schwarz1978estimating]. Unlike similar Akaike information criterion [@akaike1973information], BIC implements penalisation for a high number of clusters trying to mitigate possible overfitting of the model.

In practice, BIC is measured for each *n* within the tested range. The lowest the BIC is, the better the model represent original data.

The interpretation of the goodness of fit score is not a question of comparing the numbers only, but understanding the resulting curve. In theory, the lower the BIC score is, the better the model fits the original data. However, it has to be kept in mind that there is a certain confidence interval and that BIC itself penalises a higher number of clusters. The optimal number is not always the one which reaches the lowest BIC score, especially if the score is within the confidence interval of other options. The clustering aims to simplify the whole dataset into the smallest number of meaningful clusters, but not too small. Hence in the situation with multiple options within the same confidence interval, we should select the first significant minimum, i.e., the smallest number of components which has its mean score within the confidence interval of the numerically best fit.

In the ideal case, the BIC curve would reach the minimum for an optimal number of components and then start growing again, making the interpretation relatively straightforward. However, due to the possibility of overfitting, the curve may not culminate but only change the gradient. In such cases, the gradient itself should be analysed and as optimum should be selected a number of components before the flattening of the gradient.

###### Stability of procedure

Non-deterministic nature of GMM means that each of the trials should be repeated multiple times to understand what is the confidence interval of possible outcomes. Testing each number of components only once might lead to incorrect interpretation of results. The ideal situation is to compute multiple runs (the higher the number, the better the result) of each option and plot the confidence interval to help with the interpretation later. To better understand the magnitude of the effect, the model should be trained multiple times and resulting BIC score should be reported for each of them. The same should happen during the final clustering based on the selected number of components - the model should be initialised repeatedly, and the best of the resulting models should be kept and used.

The result of clustering is never the same, especially with the amount of the data this research is using. There is an inevitable variability, but that is mostly represented by unstable boundaries between clusters rather that significant results in clusters themselves. The boundaries should never be interpreted as a fixed line. There is always a certain degree of fuzziness, which could be captured by an overlay of resulting clusters form multiple models of same parameters.

##### Sample-based clustering

As the dataset grows, it may become increasingly impossible to perform clustering on the whole dataset, especially if we want our data with a meaningful confidence interval. The calculation of dimensions between components of the model in the hyperspace of 296 dimensions is a demanding task requiring time and computational power. While data for Prague (~140 000 features) could be processed on a desktop with modern multi-core processors within days (multiple options with a confidence interval, not a single run), that is not true for larger metropolitan areas where features count can reach millions. The data like this can be run in the same way on cloud-based services providing significantly more computational power and servers tailored to data analysis, but this solution can be costly.

For that reason, it might be worth training the method on sampled data before classifying the whole dataset. Instead of using all features to train the model, randomly samples subset could be used as a training set for GMM, which, once fitted, could be used to classify the whole dataset. This solution lowers computational demands as the number of features used in the learning process is smaller, but there are also issues with it. The random sample should reflect the structure of the whole dataset to provide results comparable with GMM trained on the whole dataset. However, that is never entirely true. The larger the sample is, the more similar to the complete data is, but at the same time, the effect of sampling on computation is becoming less significant. Even larger samples may, in some cases, miss smaller clusters present in the full-data clustering as features composing these cluster would not be present in the sample (the smaller the cluster, the higher the probability than it will be missed in the sample).

The decision whether to train GMM on the full or sampled data should reflect the balance between what is ideal (full) and what is possible in certain conditions. The different options of sample-based clustering are tested and compared to the default clustering in the following section, to assess the behaviour of sample-based clustering in the case of Prague. The behaviour will be likely different at different places as the real structure and distribution of values affects the sampling-effect. Places with more diverse structure and several smaller tissues will be probably affected more than places with a homogenous structure where the likelihood of proper sampling of all clusters is higher.

##### Sub-clustering

There are situations when resulting clustering is not refined enough for the specific analysis. The components are too big, and one may want a better resolution of clustering. One way to do it is to iteratively cluster individual already identified clusters, i.e. to do sub-clustering of existing clusters.

The morphometric dataset is rich in information, so if there is an assumption that a cluster should be divided, it is expected that the difference will be reflected in the data. The reason why it did not split the cluster in two initially is that such a difference is not significant from the perspective of the whole datasets, but it may be significant on a local scale. So when it is appropriate, the same data used for initial cluster recognition can be used again only on the sample belonging to one of the clusters.

The relation of sub-clusters to other than parental cluster is different from the relation between initial clusters themselves, and the difference has to be retained throughout the analysis and has to be correctly interpreted. Doing selective sub-clustering and then approaching initial clusters and sub-clusters as equal is not recommended even though there might be a particular situation when this approach might be viable. However, it has to be done consciously after an assessment of possible consequences.

The other way, aggregating clusters together based on their similarly will be discussed in the next chapter 8.

Either way, it is crucial to acknowledge that clustering is always based on the actual structure of the used data. That means that the result of clustering is always local. Clusters identified in Prague using solely Prague-based data would not be equal to clusters identified in Amsterdam using Amsterdam-based data only. The structure of both datasets determines what the optimal division is and as both structures are different, the optimal division is done along different lines. It is expected that results will be comparable as the optimal cluster should reflect optimal urban tissues. Chapter 8 will test whether the misalignment is significant or not to further explore the link between two local clustering models.

Selection of the clustering model and its parameters affects the results of identification of urban tissues. The decision has to be made based on detail theoretical considerations of what the behaviour of morphometric datasets likely is. While many of its properties are still unknown, based on the assumptions outlined in this sections, it is believed that GMM, in combination with BIC for determination of the number of GMM components, can identify distinct homogenous clusters as a proxy of urban tissue types.

### Data model

The data model representing the elements of the urban form consists of two input and three generated layers, all linked together through the proxy of a building based on the system of unique identifiers according to the structure presented in a table .

Presence of different unique identifiers on different data layers. buildings contains all of them and are used as a connector.

layer

uID

nID

nodeID

bID

buildings

x

x

x

x

tessellation

x

street edges

x

street nodes

x

blocks

x

: Presence of different unique identifiers on different data layers. buildings contains all of them and are used as a connector.

Buildings are in the role of connecting elements and contain all identifiers. Morphological tessellation is based on the building layer, and cells hence inherit buildings’ uID. Street edges are linked to buildings based on the proximity of building centroid to street segment geometry (the nearest edge is linked using momepy). Street nodes are linked to buildings based on proximity either, but linked node has to be end node of linked nearest edge. Blocks are based on tessellation, and their id is linked to buildings using intersection-based spatial join during their creation.

Momepy uses unique identifiers to efficiently link elements together without the need for repeating costly spatial operations for every relevant character.

## Tissue type recognition | Case study Prague

The first trial of the proposed tissue type recognition method outlined above is the case study of Prague, a dataset which after pre-processing contains 140 315 individual buildings, 22 503 street edges, 16 207 street nodes and 7 395 tessellation-based blocks. Following section reports on each step of the method in terms of both results and interpretation. The overall discussion on the method itself, its relevance and applicability is in chapter 9 and includes results of the taxonomical analysis presented in chapter 8. The validation of results is included in chapter 8.

### Primary characters

The basis of the method lies with primary morphometric characters. These continuous variables describe individual aspects of fundamental elements and their combinations based on the relational framework. Following the method, all 74 of them are measured in Prague and then linked to the building-tessellation unit according to the data model. All morphometric characters are measured using momepy classes using reproducible Jupyter notebook presented in Appendix N.

The results of measured primary characters can be explored in two ways - 1) to assess a *spatial* distribution of values, and 2) to assess *statistical* distribution of values.

#### Spatial distribution

The spatial distribution of resulting values, i.e., spatial morphometric patterns, could be projected on maps and assessed visually, to determine the character of a pattern, or statistically. Since the aim of measuring is, eventually, to identify homogenous areas defined by distinct patterns of spatial configuration, each of the characters must capture local patterns. Statistically speaking, each of the characters needs to be spatially autocorrelated, which can be assessed using Moran’s I [@moran1950notes].

Based on the visual assessment, there are three types of characters within the measured set, represented by three examples below - 1) patterns with sudden changes, 2) smooth continuous patterns, 3) visually unclear patters.

Spatial distribution of shared walls ratio of adjacent buildings in the area of Prague’s city centre and its surroundings. The figure illustrates clear spatial patterns with the presence of sudden changes.

Figure shows *shared walls ratio of adjacent buildings* in the part of Prague’s city centre. There is a clear distinction between buildings having shared walls and those standing independently. The values show a relative homogeneity in the centre of the figure (Vinohrady), but high variability in some other places, especially in the Old Prague neighbourhood (top left). There are sudden changes in values on neighbouring tessellation cells. This pattern is not unique, and it is somewhat expected for characters based on individual elements as these do not have a notion of contiguity.

Spatial distribution of the proportion of 4-way intersections of the street network in the area of Prague’s city centre and its surroundings. The figure illustrates clear continuous spatial patterns with unclear boundaries between low and high values.

The second example on the figure shows *proportion of 4-way intersections of street network* within the same area. This is purely network-based character measuring properties of subgraphs around each network node (i.e. a junction). Subgraphs, by definition, overlap causing the smooth transition of values across the study area. It is relatively simple to describe resulting patterns visually, with high values in more grid-like areas (Vinohrady - centre, Smíchov - left). However, the definition of boundaries between high and low values would be a relatively complicated procedure due to the inherent spatial smoothing. Characters based on a broader topological context tend all to have continuous patterns like this.

Spatial distribution of equivalent rectangular index of tessellation cell in the area of Prague’s city centre and its surroundings. Figure illustrates visually unclear spatial patterns.

The last, not very frequent though, is the example on figure showing visually unclear spatial distribution. The figure shows *‌equivalent rectangular index of tessellation cell* in the same area. To determine spatial patterns visually require much effort, and still, the results are questionable. This is one of the examples where one might want to exclude such character for apparent randomness of resulting values. However, visual assessment should not be used for such a decision because it is naturally arbitrary and biased based on the ability of a researcher to detect patterns. For that reason, this work uses Moran’s I index of spatial autocorrelation to determine whether a character captures meaningful spatial pattern or not.

Figure below shows the value of Moran’s I compared to reference distribution and a Moran scatterplot based on the contiguity of morphological tessellation. The I value for the whole of Prague is 0.07, showing significant autocorrelation. It is not a high value, for a reference two previous example have I 0.387 and 0.912 respectively. However, it is still significant (the value itself is likely not within a reference distribution), meaning that even not visually apparent, spatial pattern is still present. The whole set of characters contains a couple of other examples similar to this one, but overall this situation is not a frequent one.

Reference distribution in relation to actual Moran’s I value and Moran scatterplot of the equivalent rectangular index of tessellation cell based on the whole Prague. The results indicate significant, however weak spatial autocorrelation. (Rey and Anselin, 2007)

Due to the large variety of characters attempting to capture both structural complexity and cross-scale complexity within a single set, the spatial distribution of resulting values may vary. However, all show significant spatial patterns.

#### Statistical distribution

Statistical distributions of resulting values are also different, based on the nature of each character. From the literature is known, that urban context is often described by exponential distributions like a power law [@salat2017], but that is far from being a rule for a selected set of morphometric characters. Figure shows four examples of distributions as captured in Prague.

Histogram of four types of statistical distributions. Circular compactness of tessellation cell (top left), ‌the area covered by neighbouring cells (top right), the width of a street profile (bottom left), and degree of a street node (bottom right).

The first case, *‌circular compactness of tessellation cell* (top left), is slightly skewed Gaussian distribution with a minimum of values being in one of the extremes. It illustrates the range of characters with a more or less distorted normal distribution. Median value tends to be in the middle of the range.

The second case, *‌area covered by neighbouring cells* (top right), tends to follow already mentioned exponential distribution, with a majority of values being in the lowest extreme and only a few in the highest. Median value tends to be close to the overall minimum.

The third case, *‌width of a street profile* (bottom left), reflects the specific rule of the spatial organisation of cities. In this bimodal case, we can see peaks around 22 and 35 metres, which are likely predominant street widths in the context of Prague. The minor peak at 50 meters is caused by the maximum value of the defaulting to 50 metres in case of open spaces. Median value tends to be in the middle of the distribution, but that is not the overall rule for all characters of similar type of distribution.

The last case, *degree of a street node* (bottom right), is specific as the results are always integer values with a limited range. Cases like this are a minority. Apart from this example, only those measuring number of corners show similar behaviour.

These are not the only types of distributions in the set, but they illustrate the variability of morphometric characters.

Descriptive summary values of all character are presented in table .

Overview of the primary morphometric values for the whole case study. The key to character IDs is available in table . Units, where applicable are in the section Identified set of primary characters

id

mean

std

min

25%

50%

75%

max

sdbAre

260

860

30

87

130

240

89000

sdbHei

9.9

6.7

3

5.5

7.4

12

110

sdbVol

3200

12000

90

550

960

3100

1.3e+06

sdbPer

64

56

20

40

51

67

3000

sdbCoA

2.1

64

0

0

0

0

11000

ssbFoF

1.4

0.57

0.23

1

1.3

1.6

11

ssbVFR

3

1.7

0.43

2.1

2.6

3.5

67

ssbCCo

0.53

0.11

0.026

0.47

0.56

0.61

1

ssbCor

8.8

7.4

0

4

8

10

390

ssbSqu

5.3

9.1

9.5e-09

0.48

1.1

5

85

ssbERI

0.94

0.086

0.25

0.91

0.96

1

1.1

ssbElo

0.71

0.2

0.026

0.56

0.74

0.87

1

ssbCCD

1.5

2.2

0

0.068

1

1.9

88

ssbCCM

9.4

6.6

3

6.3

7.6

10

210

stbOri

16

13

0

6.2

13

25

45

stbSAl

6.7

8.9

4.9e-10

0.61

2.5

9.5

45

stbCeA

6.9

9

8.9e-12

0.48

3

9.9

45

sdcLAL

67

42

7.9

40

52

79

970

sdcAre

2100

4100

31

540

940

1900

350000

sscCCo

0.45

0.14

0.027

0.35

0.46

0.55

0.98

sscERI

0.97

0.062

0.43

0.94

0.98

1

1.1

stcOri

18

13

0

7.1

16

29

45

stcSAl

9.2

9.7

1.9e-05

1.5

5.6

14

45

sicCAR

0.2

0.15

0.00092

0.092

0.16

0.26

1

sicFAR

0.67

0.92

0.00092

0.14

0.32

0.74

17

sdsLen

230

260

0.047

110

160

260

3300

sdsSPW

29

8.4

1

22

29

35

50

sdsSPH

10

6.1

0

6.4

8

13

57

sdsSPR

0.41

0.32

0

0.21

0.3

0.49

23

sdsSPO

0.58

0.21

0

0.44

0.58

0.71

1

sdsSWD

3.6

2.1

0

1.9

3.7

5.1

12

sdsSHD

2.3

2.3

0

0.94

1.5

2.7

24

sssLin

0.95

0.13

0

0.97

1

1

1

sdsAre

31000

56000

34

6900

13000

30000

740000

sisBpM

0.075

0.079

0.00056

0.046

0.068

0.095

21

sddAre

30000

46000

86

9400

16000

31000

660000

mtbSWR

0.18

0.2

0

0

0.15

0.32

1

mtbAli

4.8

5.1

1.4e-09

0.9

3

7

44

mtbNDi

25

18

0

13

20

30

200

mtcWNe

0.046

0.022

0.0012

0.03

0.045

0.059

0.26

mdcAre

16000

19000

390

5500

9400

19000

530000

misRea

44

25

1

27

40

55

290

mdsAre

86000

110000

770

33000

53000

94000

1.3e+06

mtdDeg

3.1

0.82

1

3

3

4

6

mtdMDi

170

150

0.047

99

130

190

3300

midRea

52

28

1

33

49

67

270

midAre

97000

110000

770

42000

65000

110000

1.3e+06

libNCo

0.6

3.3

0

0

0

0

58

ldbPWL

180

250

20

51

82

200

3400

ltbIBD

27

11

0

20

25

33

120

ltcBuA

0.65

0.24

0.043

0.49

0.7

0.84

1

licGDe

0.57

0.67

0.0022

0.18

0.35

0.66

5

ltcWRB

9e-05

6.7e-05

1.7e-06

3.9e-05

7.3e-05

0.00012

0.00072

ldkAre

120000

240000

710

15000

31000

110000

2e+06

ldkPer

1500

1800

100

550

830

1700

13000

lskCCo

0.43

0.13

0.11

0.33

0.44

0.53

0.98

lskERI

0.86

0.13

0.35

0.79

0.9

0.96

1.1

lskCWA

360

470

0.43

87

170

430

3100

ltkOri

18

13

0.00098

7

15

28

45

ltkWNB

0.0074

0.0043

0

0.004

0.0066

0.01

0.04

likWBB

0.00089

0.00066

8.3e-06

0.00037

0.00074

0.0013

0.006

lcdMes

0.15

0.06

-0.33

0.11

0.15

0.19

0.34

ldsMSL

150

76

45

110

130

170

1600

ldsCDL

280

390

0

13

160

380

4200

ldsRea

350000

310000

770

190000

260000

400000

4.2e+06

lddNDe

0.013

0.0055

0

0.0095

0.012

0.014

0.13

lddRea

190

86

1

130

190

240

680

lddARe

370000

310000

770

200000

280000

420000

4.2e+06

linPDE

0.13

0.087

0

0.067

0.11

0.17

1

linP3W

0.64

0.11

0

0.57

0.64

0.71

0.97

linP4W

0.23

0.12

0

0.15

0.22

0.3

0.73

linWID

0.025

0.01

0

0.019

0.024

0.029

0.18

lcnClo

5.3e-06

2.5e-06

0

3.4e-06

5.1e-06

6.9e-06

2e-05

xcnSCl

0.056

0.087

0

0

0

0.086

1

: Overview of the primary morphometric values for the whole case study. The key to character IDs is available in table . Units, where applicable are in the section [Identified set of primary characters](#identified-set-of-primary-characters)

Without exploring the table above in detail, it is worth pointing out two characters standing out - *courtyard area of a building (sdbCoA)* and *‌number of courtyards of adjacent buildings (libNCo)*. Both are capturing similar concepts of closed courtyards (either in a single building or in a composite of adjacent buildings), and both are relatively invariant (min, 25%, 50% and 75% are all 0). While these might not be critical for identification of clusters in Prague, there are urban tissues, especially in warmer environments, characterised by these properties. While the overall aim of this research is to be comparable, not tailored to a specific context, these characters are still included.

Figures - show histograms capturing the (truncated) distribution of all measured characters. Note the differences outlined above and overall variety of distributions.

#### Statistical relationship of characters

Understanding the relationship between measured characters is an essential aspect of the morphometric assessment. As specified in section 7.1.2.1.1, characters should not include many empirical correlations. Collinear characters (those being correlated and reflecting the same concept) should not be present in the resulting data as they might skew the hyperspace and adversely affect the result of clustering. As characters do not follow always follow a normal distribution, Spearman’s rank correlation [@spearman1961proof] is used to assess the relationship between characters. The results are illustrated in the correlation matrix on figure below.

Correlation matrix of Spearman’s rho values capturing the statistical relationship between resulting morphometric values of primary characters. With a few exceptions, the relationship is none or very weak.

As expected (the set is designed in such a way), characters generally show minimal correlations, with a few exceptions. These are reflecting different concepts and are capturing different phenomena, which makes them admissible.

Primary morphometric characters are the core of the method of identification of tissue types. Their selection, to capture different complexities aspects of urban form, results in a very heterogeneous set of measured data showing variable spatial patterns as well as statistical distributions. However, this data are the direct input of the clustering procedure, but they are an input of calculation of contextual characters. The results of this following step are illustrated in the next section.

### Contextual characters

While the importance of primary morphometric characters is that they bring the fundamental information about the spatial order of elements of urban form, the values which are used for the identification of clusters itself are based on the contextualisation. Resulting contextual characters are of four types (interquartile mean, interquartile range, interdecile Theil index, Simpson diversity index), where each describes the same primary character from a different perspective. Together, they reflect the context of each tessellation cell, defined as three topological steps, comprehensively and inclusively.

The actual values measured in Prague could, once again, be explored visually to assess the spatial distribution of resulting values and numerically to assess resulting statistical distributions.

#### Spatial distribution

Unlike in the case of primary characters, contextual characters are always capturing spatially consistent patterns. The reason is the inclusion of the topological context in each of them. However, the actual distribution of values differ. Following four figures show contextual characters based on *width of a street profile* to illustrate the differences and similarities between contextual characters. Note that this is an only illustrative example, and spatial distribution would differ for other characters.

Spatial distribution of the interquartile mean of a width of a street profile measured within three topological steps on morphological tessellation in the area of Prague’s city centre and its surroundings.

Figure shows *interquartile mean of a width of a street profile* measured within three topological steps on morphological tessellation. As a version of truncated mean, this character directly reflects the actual values of primary characters, and it is relatively simple to indicate areas with generally narrow streets (historical core) and those with a wider profile (heterogenous areas on south and south-east). The overall distribution of values within the shown area is very symmetrical with a peak at 22 metres, which seems to be a common street width in Prague.

Spatial distribution of the interquartile range of a width of a street profile measured within three topological steps on morphological tessellation in the area of Prague’s city centre and its surroundings.

Figure illustrates *interquartile range of a width of a street profile* measured within three topological steps on morphological tessellation. That reflects the range of values, so it is a proxy of statistical dispersion. In the example above it does divide places with either major street or generally wider streets from predominantly homogenous areas. The distribution of values is balanced but truncated at 0 (range could not be negative). We can identify certain similarity with the patterns on previous figure , because wider street (i.e., higher interquartile mean) causes a bigger range, but the patterns are not identical.

Spatial distribution of interdecile Theil index of a width of a street profile measured within three topological steps on morphological tessellation in the area of Prague’s city centre and its surroundings.

Figure illustrates *interdecile Theil index of a width of a street profile* measured within three topological steps on morphological tessellation. The resulting map shows as the most *unequal* area around Vaclavske sq. (darker red) where one street (in this case, elongated square) is significantly different from the other. Previously highlighted areas of wider streets are not so from the perspective of Theil index. The distribution has a long tail, somewhat typical for Theil index applied to morphometric characters.

Spatial distribution of interdecile Simpson index of a width of a street profile measured within three topological steps on morphological tessellation in the area of Prague’s city centre and its surroundings.

The last contextual character, shown on figure , is *Simpson diversity index of a width of a street profile* measured within three topological steps on morphological tessellation. The values, in this case, are binned using Natural Breaks (*k=7*). It captures (inversely) similar information as Theil index, but that is not the rule. These two are somewhat related as both capture dispersion of values, but the relationship between them is not fixed as will be illustrated in the next section.

Depending on the spatial distribution of primary characters, the contextual pattern may be more similar to each other (like in the case above) or less similar. However, as illustrated in the chapter on primary characters, the visual assessment is not enough.

#### Statistical distribution

Figure in previous section showed four types of distribution of primary characters. This section illustrates how each of them translates into the distribution of contextual characters.

Histograms of statistical distribution of contextual versions of circular compactness of tessellation cell. Interquartile mean (top left), interquartile range (top right) interdecile Theil index (bottom left), Simpson index (bottom right).

The first example, *circular compactness of tessellation cell* on figure , was initially mildly skewed Gaussian-like distribution. In terms of IQ mean and IQ range, this property remains the same. Both are relatively symmetrical distributions with small tail on one or the other side. On the other hand, the distribution of the Theil index and Simpson diversity resembles exponential curve due to heavy tail in both.

Histograms of statistical distribution of contextual versions of area covered by neighbouring cells. Interquartile mean (top left), interquartile range (top right) interdecile Theil index (bottom left), Simpson index (bottom right).

Initially, exponential distribution of *area covered by neighbouring cells* remains exponential in both IQ mean and IQ range cases (figure ). Theil index is also exponential, although the curve is not so unequal. Simpson index is significantly different from all three. The HeadTail binning used within the calculation is tailored to exponential distributions and resulting Simpson diversity is then relatively balanced across the values. The values 1 showing a significant spike mean that the probability that any other value is within the same bin is 100%, hence no diversity in the area. It is typical for relatively homogenous compact urban tissues.

Histograms of statistical distribution of contextual versions of ‌width of a street profile. Interquartile mean (top left), interquartile range (top right) interdecile Theil index (bottom left), Simpson index (bottom right).

The third example, *‌width of a street profile* which was illustrated on maps on previous pages, had initially specific distribution affected by rules on which streets are designed (there were spikes for narrower and wider streets). Figure shows that none of the contextual character share this profile and, more importantly, all have different distributions. IQ mean is almost ideal Gaussian distribution, IQ range is right-skewed and truncated with a spike on 0, Theil index is again exponential, and Simpson diversity is right-skewed, but relatively symmetrical distribution. Even though figures - may seem similar, the difference in distributions on figure indicates otherwise. What is important are numerical values, not visual perception.

Histograms of statistical distribution of contextual versions of ‌ degree of a street node. Interquartile mean (top left), interquartile range (top right) interdecile Theil index (bottom left), Simpson index (bottom right).

Initially, restricted values of *degree of a street node* remained present in IQ range (figure but not in the other contextual characters. IQ mean is symmetrical with a large spike on 3, which is almost its median value. Theil index is very different from previous examples and does not follow exponential distribution this time, while the Simpson diversity index has two spikes on 0.5 and 1.0 and relative balance otherwise.

As the examples above indicate, the variety present in primary characters remained present, in a different way, in contextual characters as well. Complete results for all contextual characters are available as Appendix 7.7.

#### Statistical relationship of characters

The statistical relationship between contextual characters will directly influence the results of clustering in the next steps. For that reason, we should aim for minimisation of such relationship in terms of Spearman’s correlation. As illustrated above, we may expect some relations, however only selective, affected by the nature of primary characters. Below (figure ) is a correlation matrix of contextual characters for illustration of the measured relationship.

Correlation matrix of contextual capturing the statistical relationship between resulting morphometric values of contextual characters. With a few exceptions, the relationship is none or very weak. High-quality version of the matrix is in Appendix 7.8.

Even though there, once again, are some characters which tend to be correlated, the overall correlation is minimal. Such data then have the potential to provide the meaningful, unskewed result of clustering.

The exploration of measured primary and consequent contextual character shows that the data comply with requirements set by the method and show a high variety of information. It is assumed, that they describe urban form in its structural complexity as well as cross-scale complexity.

### Cluster analysis

The critical point in the whole process if the identification of distinct clusters is the clustering procedure itself. This section will explore the results of clustering on the complete set of data and then on sampled data, to understand the difference and possibility to lower computational demands. Both ways will start with an assessment of an optimal number of components based on the Bayesian Information Criterion (BIC) and follow with Gaussian Mixture Model (GMM) clustering itself. Furthermore, the final part of this section explores the potential of sub-clustering, i.e., generating even more detailed distinction of urban tissues.

#### Complete data

Clustering based on complete data is likely the key result of the whole research. It will either support the main hypothesis or reject it depending on the resulting clusters. GMM clustering of a complete dataset means that all features (n=140315) are used within a training set. It is expected that the algorithm will be able to detect clusters, although there might be present some adverse effects of the dimensionality curse.

Before analysis itself, data are standardised by mean removal and variance scaling:

where is the mean of the training values, and is the standard deviation.

##### Bayesian Information Criterion

To perform GMM clustering, one needs to specify the number of components to look for. While this information is not a priori known, one has to determine the optimal number using other methods before GMM. In this research, the Bayesian Information Criterion is used. BIC analysis repeatedly generates GMM clusters for a different number of components in a range (2, 40) and measures the goodness of fit of resulting clusters to the original dataset. The results in Prague are shown in figure . The lower the value, the better fit.

Bayesian Information Criterion score for the variable number of components. Shaded area reflects .95 confidence interval, red line marks the first significant minimum.

The pattern shows a steep decline from a small number of components to approximately 15 components where it starts flattening. The results between 15 and 35 components are very similar, and then the BIC starts growing again. That suggests that the optimal number of components for the final clustering is between these two values. The optimum is 20 as the value, which is the first significant minimum. It is the smallest number after which no other is significantly (with the confidence interval) below the achieved score. The differentiation within the range in question is better recognisable in a zoomed figure below.

Trimmed plot of Bayesian Information Criterion score for the variable number of components, to see the differentiation within values. Shaded area reflects .95 confidence interval, red line marks the first significant minimum.

Within the trimmed figure, it is more evident the difference between the BIC score. The reason why 24 or 26 are not selected as optimum, while both being smaller than 20 is the significance. The aim is to detect the smallest optimal number of clusters as larger numbers may have better due to overfitting. Hence we want the first significant minimum, which is 20. 24 is below, but its confidence interval goes above the score of 20. The same applies to 26.

Based on the BIC results, GMM clustering of complete data will focus on the identification of 20 components (clusters).

##### Distinct clusters as urban tissue types

Gaussian Mixture Model clustering with 20 components is done using full covariance matrix and five initialisations, from which the best is selected based on the per-sample average log-likelihood. The complete Jupyter notebook is available in Appendix N.

The resulting prediction of cluster membership is shown visually on the figure . Each feature (building/tessellation cell) is coloured according to a cluster of the highest probability. The map shows the delineation of distinct homogenous clusters and their spatial distribution across the whole case study area. At this moment, it is possible to say that the proposed method did identify the particular type of proxy of urban tissues using the purely quantitative method based on urban morphometrics. How well it did that will be assessed on the following pages and later validated by other data in Chapter 8.

Spatial distribution of 20 clusters as identified by GMM based on complete data.

The 20 cluster seems to be relatively well defined and based on the first observation tend to reflect homogenous form. Even though there is no spatial constraint in the clustering itself, results show apparent contiguity caused by the design of contextual characters. The figure shows detail of the section of Prague covering City Centre and the area towards the southern boundary for better understanding of results.

Detail of spatial distribution of 20 clusters as identified by GMM based on complete data.

Starting from the top left corner, where the historical core of Prague is, we can see (id 11 in red) delineation of what could be seen as medieval urban form, transitioning to compact perimeter blocks of Vinohrady neighbourhood (id 5 in dark blue). That is surrounded by less rigid heterogenous perimeter block-like tissues (id 10 in light blue) and then fringe areas (id 7 in pink). Towards south and east are present low-rise tissues (id 8 and 3 in lighter yellow) and modernist developments (ids 2 and 12 in grey and green). Drawing from the pure observation, clusters seems to be very precise and detailed and, most importantly, meaningful in terms of their link to the concept of the urban tissue.

The following section describes each of the identified clusters to give a detailed overview and understanding of what each cluster is composed.

###### Individual clusters

Each of the individual clusters is presented by one example (usually the largest contiguous area) and its surroundings within 1,5km buffer. Colour schema is the same as in figures and and will be kept throughout the chapter. Clusters are sorted according to their ID, which is randomly assigned.

The first cluster (figure ), noted as 0, is composed of predominantly low-rise, single-family housing. It has mostly residential character and tends to be located in the outer parts of the city, further away from the city centre. It is the largest of all clusters, with 15337 features, which is approximately 11% of all buildings in the study area.

Example of cluster 0 and its surroundings within 1,5km buffer located at the eastern boundary of study area. Aerial image courtesy of mapy.cz

The cluster on figure , noted as 1, contains mostly small-scale industry areas with small coverage, relatively small buildings. Often is adjacent to other clusters. It tends to be located in outer rings of the city but is overall very sparsely distributed. With only 2038 (less than 1.5%) features is one of the smallest clusters overall.

Example of cluster 1 and its surroundings within 1,5km buffer located at the north-east of study area. Aerial image courtesy of mapy.cz

Cluster 2, shown in figure is one of the urban tissue types following modernist principles of spatial configuration, with linear buildings, but still relatively connected street network. These areas are mostly infills of the existing structure located within the city (except its central part) rather than on the periphery. It is relatively abundant with 12016 features (approximately 8.5%).

Example of cluster 2 and its surroundings within 1,5km buffer located at the north-west of study area. Aerial image courtesy of mapy.cz

Cluster 3 (figure ) is one of the smaller ones. Its structure is defined by row-houses, a typology which is not very common in Prague. There are only 4133 features, less than 3% of all buildings scattered mostly in peripheral locations.

Example of cluster 3 and its surroundings within 1,5km buffer located at the east of study area. Aerial image courtesy of mapy.cz

Cluster 4 (figure ) is one of the types with an industrial character, in this case being distributed as sort of infill development in the fringe areas. It is mostly adjacent to other urban tissues, relatively evenly distributed across the study area. It is composed of 5281, which is 3.8% of the total number.

Example of cluster 4 and its surroundings within 1,5km buffer located at the south of study area. Aerial image courtesy of mapy.cz

Cluster 5 (figure ) can be best described as compact perimeter block-based residential area. This dense, grid-like development is located in the central areas of the city around the historical core and is one of the best defined urban tissues in Prague. There are 5930 of features belonging to this cluster, which is a bit more than 4.2% of the total count.

Example of cluster 5 and its surroundings within 1,5km buffer located at the centre of study area. Aerial image courtesy of mapy.cz

Cluster 6 (figure ) is very different from the previous one as it contains fringe low-rise, not very well defined urban tissues. These are small-scale tissues scattered evenly around the study area, adjacent to other types of tissues, often filling topographically inconvenient areas. There are 10329 of these features, which is about 7.4%, so it is one of the more abundant clusters.

Example of cluster 6 and its surroundings within 1,5km buffer located the south-east direction from the city centre. Aerial image courtesy of mapy.cz

Cluster 7 (figure ) is an example of more heterogeneous area. It has a similar character as cluster 4, but unlike that, it often contains other types of development with a less defined structure, like contemporary housing or office parks which do not reflect traditional rules of spatial configuration. That leads to higher heterogeneity in the area, making these tissues complicated to define. It consists of 4140 features, which is nearly 3% of the total amount.

Example of cluster 7 and its surroundings within 1,5km buffer located the souther direction from the city centre. Aerial image courtesy of mapy.cz

Cluster 8 (figure ) predominantly contains single-family housing in a relatively dense setting resembling garden city movement development. These places have interconnected network of relatively grid-like character, with buildings adjacent to each other either as row-house typology or similar. There are 7845 features within this cluster (5.6%).

Example of cluster 8 and its surroundings within 1,5km buffer located the souther direction from the city centre. Aerial image courtesy of mapy.cz

Cluster 9 on figure seems to identify low-rise areas of organic development, which seems to be cores of the historical villages around Prague. These are small-scale tissues evenly distributed in the outer ring of development, now mostly embedded in the other development. They compose 5.6% of total features (7862).

Example of cluster 9 and its surroundings within 1,5km buffer located at the south-west of the study area. Aerial image courtesy of mapy.cz

Cluster 10 (figure ) is very often adjacent to cluster 5 (compact blocks) or composes its own areas of block-based development. However, unlike in cluster 5, these blocks tend to be skewed or distorted in some other way. In some cases, this cluster could be seen as a transitional area between homogenous compact blocks and other types of urban tissue. These are 7203 features within this group, making 5.1% of the total amount.

Example of cluster 10 and its surroundings within 1,5km buffer located next to the city centre. Aerial image courtesy of mapy.cz

Cluster 11 (figure ) is a very straightforward one to describe as it composes solely of historic medieval core of Prague. It includes areas on both sides of the river and correctly excludes the area cut-out of the Old Town, which has been demolished in 19th century and rebuilt after that [@hruuza2003urbanismus]ƒ. There are 2167 features, making historical core one of the smallest clusters of all, composing only 1.5% of the total amount.

Example of cluster 11 and its surroundings within 1,5km buffer located in the city centre. Aerial image courtesy of mapy.cz

Cluster 12 is another very distinct one. As illustrated in figure , the origin of the development is modernist, covering large-scale modernist housing estates. These are typical with slab buildings, the incoherent relationship between buildings, plots and streets and large amounts of open spaces, among other characteristics. In Prague they are almost exclusively on the peripheral ring of the city, forming a so-called modernist belt of Prague. They consist of 6885 features, which is 4.9% of the total amount.

Example of cluster 12 and its surroundings within 1,5km buffer located on the southern edge of the city. Aerial image courtesy of mapy.cz

Cluster 13 (figure ) is another example consisting of single-family housing. This time it is low-density development with predominantly detached buildings. It is typical with elongated blocks which in part is a reaction to the underlying topography. It is a very abundant cluster with 14992 features, making more than 10.6% of the total amount, distributed along the periphery of the city.

Example of cluster 13 and its surroundings within 1,5km buffer located on the south-western edge of the city. Aerial image courtesy of mapy.cz

Cluster 14 is distributed almost exclusively within the wider centre of Prague, often adjacent to the homogenous compact city as is illustrated on the figure . The cluster could be defined as an inner fringe composed of heterogeneous developments on the edge of existing homogenous one. There are 4984 features within it, making 3.6% of the data.

Example of cluster 14 and its surroundings within 1,5km buffer located north of the city centre. Aerial image courtesy of mapy.cz

Cluster 15 () is perimeter-block based tissue type with very heterogeneous development in the block interiors. It has a very high coverage area ratio located in in the city centre either as a transitional area between medieval core and compact city or as industrial development. There are only 3060 features within the cluster (2.2%).

Example of cluster 15 and its surroundings within 1,5km buffer located north of the city centre. Aerial image courtesy of mapy.cz

Cluster 16 () is not very straightforward to define as it is a heterogeneous one. It mostly consists of small patches of not very well defined tissues with the predominant role of small-scale buildings but not exclusively. It may be seen as *other*, combining parts of the dataset which do not fit elsewhere, but at the same time, all places have the similar character of being *out of sight*. It is evenly distributed, but not very abundant one with 3548 making approximately 2.5% of the dataset.

Example of cluster 16 and its surroundings within 1,5km buffer located north of the city centre. Aerial image courtesy of mapy.cz

Cluster 17 () is another of the low-density single-family tissue types. It has a less defined and rigid structure, and it is often adjacent to open space. It does have a certain inner heterogeneity expressed as various kinds of buildings from detached to row houses. Like the other similar clusters, this is also relatively abundant with 12145 features (8.7%).

Example of cluster 17 and its surroundings within 1,5km buffer located west of the city centre. Aerial image courtesy of mapy.cz

Cluster 18 () consists of relatively independent detached areas of low-density village-like development. Clusters can be only a strip along the road or other open-space facing tissues. It is located mostly on the periphery of the city and entails 8764 features (6.2%)

Example of cluster 18 and its surroundings within 1,5km buffer located on the north of the city. Aerial image courtesy of mapy.cz

The last cluster, 19 illustrated on figure is industrial urban tissue type, consisting of a mixture of large-scale and small-scale buildings, convoluted street network and a minimum of residential use. It is only at a few places, but of a large-scale, mostly towards the edge of the city. There are only 1656 features within the cluster, making 1.2% of the total amount.

Example of cluster 19 and its surroundings within 1,5km buffer located on the eastern edge of the city. Aerial image courtesy of mapy.cz

From the overview is clear that some clusters are very distinct like the historical core (11) or modernist estates (12), while others resemble each other as is the case of low-density single-family clusters (0, 8, 13, 17). However, even between these seemingly similar clusters are recognisable differences. Numerical assessment of differences between clusters is part of Chapter 8, to determine which characters are causing the distinction and understand the clusters based on their morphometric profiles.

#### Sampled data

Gaussian Mixture Model uses training data on the input to estimate the optimal clustering and then predicts the probability that each feature belongs to any of the components. That means that training data do not have to equal the data we want to classify. The GMM and especially the estimation of the number of components, which does GMM repeatedly, could have relatively high computational demands as the size of the dataset grows. The Prague example, with 140,000 features took approximately 60 hours to measure all BIC values and do the final clustering on a desktop computer with a 12-core Intel Xeon processor. Running larger areas at once may get unfeasible, it is hence critical to understand if the method can work with sampled data.

Sampled clustering would use a randomly selected fraction of the data as a training set and then use it for prediction on the complete data. That might significantly reduce computational demands because they rise exponentially with the growing dataset, but at the same time might not provide useful results. Sampling procedure might miss some clusters entirely (none or very few features are included in the sample) or affect the results in another way. The following section tries to answer some of the questions comparing the clustering based on the complete dataset with sampled one.

##### Sampled Bayesian Information Criterion

Three versions of sampling are assessed - 10%, 25%, and 50% of the dataset. Because random sampling results in different samples each time, which could affect BIC, each option is sampled three times and GMM is run three times on each (in total nine runs of GMM per option). To assess the number of components, values from range 2 - 40 were tested. BIC is measured using the complete dataset. The resulting values are shown in figure below.

Bayesian Information Criterion score for sampled clustering. Shaded area reflects .95 confidence interval.

Figure shows a striking difference between the results of 0.1 (10%) sampling and the rest. The BIC score for this option is significantly higher than for the rest, indicating that the sample is way too restricted to capture the structure of the dataset and generate meaningful clustering. Due to this difference, any differences between the other options are not recognisable. For that reason, figure shows the same data without 0.1 option.

Bayesian Information Criterion score for sampled clustering excluding 0.1 sampled results. Shaded area reflects .95 confidence interval.

All two remaining options show similar curves as was already seen in the complete clustering. The bigger the sample is, the better results can GMM provide. One key difference between the samples is the resulting optimal number of components. It seems that the smaller the sample is, the sooner BIC curve culminates, which results is a smaller number of optimal components. 0.25 sampling culminates at 11 components, 0.5 at 15 components and 1.0 (complete data) at already mentioned 20 components. The difference between 0.25 and 1.0 both in terms of BIC and an optimal number of components is big, so it is questionable if such a small sample can provide any similar results.

Bayesian Information Criterion score for sampled clustering excluding 0.1 and 0.25 sampled results. Shaded area reflects .95 confidence interval.

Figure compares 0.5 and 1.0 sampling only as the difference is not so dramatic. Unlike 1.0 sampling, where the point of culmination is not as clear, 0.5 culminates sooner, and the curve starts ascending quicker. That makes the decision of optimum easier. Following the same principle as in the previous case, the first significant minimum is 15 components. The BIC score overall is worse than in non-sampled case, but it is worth testing the similarity of the actual cluster recognition.

##### Sampled distinct homogenous clusters

Identification of distinct clusters based on sampled data (random sample 50%) using 15 components results in the spatial distribution of clusters illustrated on figure .

Spatial distribution of clusters based on sampled data (0.5) and 15 components within the whole study area.

Visual assessment of clustering indicates that the clusters seem to be meaningful and could be seen as a proxy of urban tissues. Due to the smaller number of components, some areas are showing less differentiation than in the complete clustering, but the difference does not seem to be in terms of correctness or wrongness of one or the other clustering, but only in terms of the change of the resolution of results.

###### Comparison of sampled and complete clustering

Three easy-to-interpret clusters are compared 1:1 and their composition and shape are assessed to understand what are the actual on-ground differences between two versions of clustering apart from the different number of components.

Comparison of spatial distribution of cluster 5 and sampled cluster 4 in the city centre.

Original cluster 5, capturing compact perimeter blocks, has its counterpart in sampled cluster 4. The example of their spatial distribution is illustrated in figure . Both versions capture mostly the same type of urban tissue with a very similar footprint. The only apparent difference is that sampled cluster is more inclusive (covering larger area) than a complete cluster, likely due to the smaller overall number of clusters (clusters needs to be naturally more abundant).

Composition of cluster 5 and sampled cluster 4 in relation to each other. Shows the number of features labelled as studied cluster and their labels in the other clustering variant.

The comparison of the composition of both clustering versions in relation to each other on figure shows that features initially marked as being in the complete cluster 5 are almost entirely within sampled cluster 4. On the other hand, features labelled as sampled cluster 4 are predominantly located in complete cluster 5, but due to higher inclusiveness also in clusters 10 and 15.

Comparison of spatial distribution of cluster 11 and sampled cluster 9 in the historical core.

The second example based on complete cluster 11 (figure ) representing the historical core of Prague shows a similar story as the first one. Both versions correctly delineate medieval urban tissue and avoid newer redevelopment of the former Jewish quarter in the North. The difference is in inclusiveness, where the sampled cluster covers more extensive areas, which are in the complete clustering seen as cluster 15, the transitional one.

Composition of cluster 11 and sampled cluster 9 in relation to each other. Shows the number of features labelled as studied cluster and their labels in the other clustering variant.

The assumption derived from the visual assessment is correct based on the numerical data on the actual composition (figure ). Features classified as cluster 11 in the complete clustering, are classified as cluster 9 in sampled clustering. Features classified as cluster 9 in sampled clustering, however, include parts of complete cluster 15 (less than 1/3 of cluster 9 comes from 15).

Comparison of spatial distribution of cluster 12 and sampled cluster 5 in the west of the city.

Even more, similarity shows a comparison of large-scale modernist urban tissues (figure ) with very few differences which could be derived from visual observation. Even the complicated case of modernist area mimicking perimeter blocks in the South of the example shows the same pattern, with middle part being excluded from the rest (likely an effect of a configuration of the street network).

Composition of cluster 12 and sampled cluster 5 in relation to each other. Shows the number of features labelled as studied cluster and their labels in the other clustering variant.

The composition of clusters is, with a few exceptions equal (figure ). The same features which belong to cluster 12 in the complete clustering are labelled as cluster 5 in the sampled clustering and vice versa.

Comparison of the city centre focusing on cluster 15, which is not present in the sampled clustering.

Three examples above show that there is a striking similarity between both results. However, there is a different number of clusters, so where is the difference? The example on figure shows cluster 15 based on complete data, which does not have its counterpart in sampled clustering. Instead, it is split into three almost equal parts (figure ) each linked to another cluster. What was the so-called transitional area between medieval core and historical compact city is no longer present. That by itself is likely not a big issue, but it illustrates the behaviour of sampled clustering with a smaller number of components. It does not necessarily merge two similar clusters into one, but at some places splits clusters into multiple pieces. GMM, in this case, sees different data and hence might exclude some smaller clusters. Because these might have been *in between* other, parts are now closer to one and other parts closer to another cluster. The resulting clustering should then be seen as a different perspective using different resolution, rather than a coarser version of complete clustering.

composition of cluster 15 in relation to sampled clustering. Shows a number of features labelled as studied cluster and their labels in the other clustering variant.

Looking onto other clusters which do not have a counterpart in sampled clustering (apart from 15, 2, 6, 9, and 10), none of them is *swallowed* by one larger cluster. All are split into two sampled clusters. Sometimes more equally (e.g., cluster 2 is equally split between 10 and 14), sometimes less equally (e.g., cluster 6 is more present in sampled cluster 0 than sampled cluster 8). This illustrates the probabilistic rather than hierarchical nature of GMM. The full comparison is available in Appendix 7.9.

Depending on the aim of the study, sampled clustering could likely be used instead of complete clustering, considering the fact that results based on samples smaller than 50% are not precise enough. However, for the ideal, detailed identification of urban tissues, sampled clustering might provide sub-optimal results.

#### Sub-clustering

The trial of sub-clustering, i.e., division of existing clusters, obtained using the complete dataset will be done on two of the original clusters, which are very different. The first example will focus on cluster 5, compact perimeter blocks, and the second on the modernist belt of Prague labelled as cluster 12. The assumption behind sub-clustering is that the richness of the data may allow us to determine differences within the cluster. These are not significant from the perspective of the whole dataset, that is why they were not picked initially as independent clusters, but they might be significant internally.

##### Compact Prague

The first case is the cluster 5, which could be interpreted as the urban type of compact, rigid perimeter blocks. The reason for its selection is that due to the varied topography. These blocks have to react to the steeper surface at some places, and the perceptional character of such areas is different from those laying on the flat grounds.

Sub-clustering uses contextual data of individual features within the cluster and performs the identification of types once again, starting from determination of the optimal number of components using BIC and consequent training of the model and prediction of labels.

Bayesian information criterion score for cluster 5 sub-clustering to determine optimal number of components.

Bayesian Information Criterion illustrated in figure indicates that there is a scope for sub-clustering as both 2 and 3 components have a better score than a single component. If the situation would be otherwise, and a single component would have the lowest BIC score, there would be no significant sub-clusters in the data and results of forced clustering would likely suffer from discontinuity. Following the rule of the first significant minimum, this trial works with two components.

Spatial distribution of sub-clusters of cluster 5 marks the distinction between rigidly and distorted grids.

The result of sub-clustering of cluster 5 is shown in figure . It feels fair to conclude, that newly identified sub-clusters have a meaning and distinguish between areas which are more rigidly gridded and those which tend to have grid distorted.

##### Modernist Prague

The second sub-clustering trial focuses on large-scale modernist housing estates on the periphery of Prague. There is an assumption of the inner differentiation of the relevant cluster 12 because each of these neighbourhoods has been designed, and there were different authors and approaches in different places and periods of development [@hruuza2003urbanismus]. It is assumed that morphometric data should be able to reflect this difference.

Bayesian information criterion score for cluster 12 sub-clustering to determine optimal number of components.

BIC results on figure indicates that subdivision of the cluster is significantly better than a single group with all the options between 2 and 7 having a lower score that one component. The first significant minimum, in this case, are three components.

Spatial distribution of sub-clusters of cluster 12 showing the different location of all three groups.

The map on the figure shows the whole cluster 12 divided into three sub-clusters. The interesting case is the green group, located exclusively on the western edge of the study area. The fact that it is not present anywhere else indicates that sub-clustering indicates that results are not affected by randomness. A closer look at the differences as illustrated on the figure shows why these tissues are split in such a way. The green sub-cluster has large blocks and a circular character. The red one tends to be a large-scale orthogonal configuration, while blue is smaller-scale more compact urban tissue.

Comparison of examples of sub-clusters of cluster 12 illustrating the structural differences.

Both examples above indicate that there is a scope for sub-clustering if the research using this method needs a more refined level of detail. As noted above, sub-clustering ability depends on the internal homogeneity of each cluster, and it may not be possible in some cases. However, in cases where this possibility is available, results show meaningful patterns, enabled by the richness of the morphometric dataset.

## Summary

This chapter took morphometric elements organised within the relational framework of urban form and defined the rich set of characters to be used within the rest of the study. Morphometric characters, divided into 74 primary and 296 contextual characters were then tested on the case of Prague. As the set followed specific rules driving its definition, it proved to provide a complex multi-scale characterisation of the local context of each individual building. That served as an input of cluster analysis using Gaussian Mixture Model method, which delineated 20 potential types of urban tissues within the fabric of the city.

While the validation is left for Chapter 8, results of clustering illustrated on previous pages indicate that the morphometric method of identification of urban tissues and their types has a potential. The outcome of the Gaussian Mixture Model learning procedure does match the expectations of what a tissue type should be. The question remains what the relation of these clusters to the actual concept of urban tissues is.

One should be aware that cluster is a numerical, morphometric statistical **proxy** of urban tissue, not its definition and replacement. GMM clustering is non-deterministic, so boundaries are not fixed, but rather indicative. It is not a ground truth (there is no ground truth at all in fact), and the meaning of clusters and relation between them has to be determined and interpreted before any further steps. The one approach on how to do so is proposed in the next chapter.