This is the title of the thesis

Martin Fleischmann

A thesis presented for the degree of
Doctor of Philosophy

Supervised by:
Dr Ombretta Romice
Professor Sergio Porta

Urban Design Studies Unit
Department of Architecture
University of Strathclyde, UK
Month 2020

This thesis is the result of the author's original research. It has been composed by the author and has not been previously submitted for examination which has led to the award of a degree.

Signed:

Date:

Abstract

Lorem ipsum dolor sit amet, consectetur adipiscing elit. Nam et turpis gravida, lacinia ante sit amet, sollicitudin erat. Aliquam efficitur vehicula leo sed condimentum. Phasellus lobortis eros vitae rutrum egestas. Vestibulum ante ipsum primis in faucibus orci luctus et ultrices posuere cubilia Curae; Donec at urna imperdiet, vulputate orci eu, sollicitudin leo. Donec nec dui sagittis, malesuada erat eget, vulputate tellus. Nam ullamcorper efficitur iaculis. Mauris eu vehicula nibh. In lectus turpis, tempor at felis a, egestas fermentum massa.

Acknowledgements

Interdum et malesuada fames ac ante ipsum primis in faucibus. Aliquam congue fermentum ante, semper porta nisl consectetur ut. Duis ornare sit amet dui ac faucibus. Phasellus ullamcorper leo vitae arcu ultricies cursus. Duis tristique lacus eget metus bibendum, at dapibus ante malesuada. In dictum nulla nec porta varius. Fusce et elit eget sapien fringilla maximus in sit amet dui.

Mauris eget blandit nisi, faucibus imperdiet odio. Suspendisse blandit dolor sed tellus venenatis, venenatis fringilla turpis pretium. Donec pharetra arcu vitae euismod tincidunt. Morbi ut turpis volutpat, ultrices felis non, finibus justo. Proin convallis accumsan sem ac vulputate. Sed rhoncus ipsum eu urna placerat, sed rhoncus erat facilisis. Praesent vitae vestibulum dui. Proin interdum tellus ac velit varius, sed finibus turpis placerat.

Table of Contents

List of Figures

List of tables

Abbreviations

API	**A**pplication **P**rogramming **I**nterface
JSON	**J**ava**S**cript **O**bject **N**otation

Chapter 1

Introduction

Chapter 2

Existing approaches to classification of urban form

6 000 words (if less, better)

2.1 Introduction

- Explain prior focus on quantitative morphology (link to introduction), but say that the chapters gives overview of all, with the focus on quantitative.

2.2 The need for the classification

- *Why is classification important, what can it bring to the table, why should we bother doing it.*
- What is classification
  - a bit of definitions
  - different ways of making classification
    - typology/taxonomy distinction **important**
- Why is classification useful in general

- Why is classification useful in urban morphology

2.3 Existing methods of classification of urban form

- *Literature review of existing methods of classification and its analysis and description of patterns within the field.*
- Introduction

2.3.1 The history of classification attempts

- *A brief overview of the history of classification of urban form focusing on its origins and early attempts. People like Lynch, Kostof.*
- **Research TO DO**
- link between history and qualitative

2.3.2 Qualitative

- Traditional schools of urban morphology
  - Conzen
  - Muratori
  - Duany
- City-based approaches (Portland, Berlin, Prague)
- Spatial typology
  - Kohout, a+t
- The qualities of such approaches, their limits.
  - **Research TO DO (a bit)**
  - expert knowledge needed
  - concepts based
  - might be biased (not necessarily)

  - good in interpretation, could be detailed
  - time consuming, information demanding
  - limited applicability

2.3.3 Mixed (predominantly non-morphological)

- Socio-demography as a main branch
- Additional (energy)
- The qualities of such approaches, their limits
  - capturing non-morphological classes
  - good for specific purposes
  - good source for link between form and soft data
- **Research TO DO (a bit)**

2.3.4 Quantitative

- introduction
  - what does it mean quantitative method
  - two major groups divided by the data source
    * remote sensing — raster data
    * morphometrics — vector data
      · *morphometrics can in theory be done on remote sensing as well, so it might be better to use another term*

2.3.4.1 Remote sensing

- Introduce RS
  - satellite or aerial data, automatic (multi-spectral) image recognition, supervised ML
- Units of analysis

  - patch
  - block
  - grid
  - *add some figures as an illustration*

- Number of categories
  - 1 - 10

- The qualities of such approaches, their limits
  - possible extent
  - only "visible" spectrum - roofs can make a lot of difference in RS but minimal in reality
  - mostly supervised nature - you have to predefine ground truth
  - the aspect of resolution and data availability
  - number of categories is generally low related to low number of actual indicators (like Copernicus)

2.3.4.2 Urban Morphology (quantitative)

- *This is the key focus of the whole chapter, and the majority of scrutinised works fall into this category. The rest mentioned above and below is to draw a full picture, but it does not aim to provide an in-depth understanding, unlike this part.*
- **Research TO DO - check recent papers, some might be included**
- Introduce quantitative morphology
- units of classification
  - *Assessment based on the unit of classification and its placement on the scale.*
  - gradient of scales
  - from city scale to building and plot
  - *do some quantitative assessment of the db*

- number of classes

  - generally low, in few cases higher
  - *do some quantitative assessment of the db*

- mention number of characters used for classification (scrutinised in the next chapter)
- Synthesis of the corpus of works
  - *taxonomic relations between types?*
  - The qualities of such approaches, their limits

2.4 The gap in the systematic classification

- lack of systematic classification based on the small-scale unit
- gap in unsupervised classification
- gap in detailed classification (i.e. number of classes)
- gap in exploration of relationships between classes (*check before writing*)

2.5 Conclusion

- *conclusion: the existing approaches and methods have gaps: the lack of systematic classification based on the small-scale units ~~using an extensive, inclusive set of indicators~~ enabling detailed classification into larger number of types/taxa/classes. That should help position my work within the field and say what I am bringing new in later stages. BE CAREFUL TO CONCLUDE ONLY BASED ON THE CONTENT OF THE CHAPTER NOT MORE. FIND A GAP WHICH MAKES SENSE. THIS TEXT IS MIXING TOGETHER RESULTS OF THIS AND THE NEXT CHAPTER. THIS LOOKS AT THE UNIT AND NUMBER OF CLASSES MOSTLY. NUMBER OF CHARACTERS SHOULD BE LEFT TO THE NEXT CHAPTER.*

Chapter 3

Measuring of urban form

- The need for measuring
- Based mostly on my MSc
- How others measured form?
- Where is the gap?

Chapter 4

Evolution and urban form

- Evolutionary perspective in the context of urban design
  - Biological as well as cultural
  - Taxonomy
- Explain the principles of evolution in the context of urban morphology
- deifne viable analogies
- use cultural evolution alongside with biological
- introduce evolutionary approach to classification - cladistics, taxonomy
- current views of evolution and cities (e.g. Marshall)

Chapter 5

Propositions

Chapter 6

Morphometric elements of urban form

- ~~Identification of "individual" (OTU)~~
  - ~~Urban Tissue~~
- ~~Exploration of concept of DHC~~
- What is urban tissue
- Why is it worth studying
- How others approached it
- City as an ecosystem
- What is an individual within this ecosystem
- Principles of identification of individuals
- Introduction of DHC (as a theoretical concept)

6.1 What is the *individual* in the urban form?

- We are trying to identify and describe distinct kinds of urban form. *Individuals* forming the population - the city.
- Dibble et al. used Sanctuary Areas (SA), I argue that the concept of SA is limited. By qualitative definition of SA, and by possible heterogeneity of

it.

  - The problem with SAs is that their definition and identification is *phylogenic* process, it is based on the process of development of the settlement. The rest of the taxonomic systematisation is, however, based on purely *phenetic* attributes.
  - Use of SA as OTU assumes that whole cities are in fact ideal according to 'Emergent Neighbourhood Model' (Mehaffy et al). Even the authors states that they are not (e.g. the three pathologies). While the concept of SA in this model perfectly works, in case of unrestricted taxonomy it doesn't.

- We are looking for the structure indicating the smallest distinct kind (Sneath and Sokal) of urban form, which urban morphologists define as urban tissue (Kropf)

6.2 Urban Tissue and similar concepts

- Urban tissue has several definitions
  - principal unit of growth
  - the ensemble of aggregated buildings, spaces and access routes (Cannigian analysis)(Samuels, 1982, p.3)
  - a distinct area of a settlement in all three dimensions, characterised by a unique combination of streets, blocks/plot series, plots, buildings, structures and materials and usually the result of a distinct process of formation at a particular time or period (Kropf, 2017)
- Urban morphologists are using a few concepts which are very similar. Those are **urban tissue**, conzenian **plan unit**, cannigian **tessuto urbano**, and **urban structural unit**. However, there are minor differences.
  - We can say, that **urban tissue** is a broad theoretical concept, which is defined above.
  - All the other, are methodological terms capturing urban tissue in different ways. Conzenian plan unit is based on qualitative analysis of

two-dimensional town plans, 'tessuto urbano' by Muratori, Cannigia and Maffei uses the principle of aggregation of smaller hierarchical elements (also qualitative approach), urban structural unit, originating in studies of metabolism of urban systems (Pauliet and Duhme, 1998 + some others) is mixed-use method incorporating, beside the built form, also structure of open spaces (Osmond).

- I am proposing another concept of capturing the urban tissue.
  * **the smallest distinct physiognomically homogenous cluster** (DHC)
    · In short, DHC is formed by clustering method based on measurable characters.
  * Unlike methods described above, DHC is purely quantitative one

One of the result of the research should therefore be the taxonomy of urban tissues (defined as DHC).

~~###### 05.x.x Complexity (of urban form)~~ ##### Generating blocks Blocks are generated based on the street network and morphological tessellation. Because the street network obtained from open data portal is capturing car-based network, it sometimes does not connect where it should. This should be fixed.

In the case of Prague, using original street network I have generated 9428 blocks, out of which 1839 were "unusual". (19.5%)

```
bdkSec > 500 or bskCom < 0.2 or bskCon <0.76 or bskERI < 0.7 or bskShI <0.5
```

After that I fixed the street network so it snapped to itself and closed gaps in street network - if the 20m extension of line intersects street network - snap. If the 70 extension of line intersects boundary of built-up area (defined by tessellation), snap.

The result gave me 9800 blocks and 1092 unusual (11%). 10% of unusual blocks are randomly selected and assessed whether they are correct blocks or incorrect. Based on that, the approximate error is estimated.

Out of 109 randomly selected blocks, 76 were marked as correct representation of block, 33 as incorrect. Based on that, the **estimated error is 3.4%**. That includes blocks which were incorrect before the network snapping as well as blocks which were falsely identified by the snapping.

Additionally, there should be a subchapter talking about exceptions which momepy is not able to capture (Krizikova, Karlin, Nabrezi Karluv most).

6.2.0.0.0.1 Problem with blocks in modernist structure As block is defined by street, it is expecting that street is major divider of space and was there first. In modernist structures, street is often designed as a way though the area, in the middle of what we could see as morphological piece (or block). We are effectively trying to define something which does not exist. **WHAT IS THE CONSEQUENCE OF THIS???**

Chapter 7

Identification of urban tissues through urban morphometrics

The concept of urban tissue introduced 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 actually 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 (REF), one still remains. How to consistently identify urban tissues across metropolitan areas in an automatised, algorithmic way.

The aim of this chapter is to provide theoretical and practical grounds to the novel method allowing automatic detection of distinct types of urban tissues. While similar research has been done before (REF), it was never linked to the coherent theory of morphometrics and numerical taxonomy, nor it was both rich in terms of number of characters used within a model and the spatial extent (see Chapter 3). Following pages present a method which aims to be both inclusive as per morphometric characters and at the same time automatised and efficient to allow for examination of large datasets spanning across metropolitan regions.

Following chapter will introduce key principles of systematic morphometric description, which will be later applied to the methodology. Then it will outline the basis for the recognition of distinct homogenous clusters (DHC), from the se-

lection and definition of morphometric characters to unsupervised classification using Gaussian Mixture Model clustering. Methodological proposition will be later tested on the case of Prague, Czechia.

7.1 Principles of systematic morphometric description

In the context of the whole research, theory of numerical taxonomy is applied twice - in the DHC recognition and then in development of a taxonomy (Chapter 8). This chapter builds on the idea of morphometrics, the idea stating that it is possible to classify *individuals* based on the measured feature of their form. However, the hypothesis of this chapter cannot follow this statement *per se*, due to the complicated nature of the term *individual* in the urban morphology. In turn, it is hence assumed that we are able to identify individuals (in a sense of urban tissue) based on the similarity of morphometric characterisation of their fundamental parts. Initial morphometric assessment then focuses on the description of the form of individual elements, which is later used to identify distinct physiognomically homogenous clusters of urban form, i.e., urban tissues.

To develop a robust model, the methodology of description needs to be both systematic, i.e., being methodological and replicable, and comprehensive, i.e., being inclusive, capturing the wide scope of features. This method is trying to be both by proposing clear rules of character selection and providing tools to measure them using `momepy` and by unbiased inclusion of wide range of morphometric characters based on relational model of urban form and characters' classification system, hence providing both scalar and structural complexity of urban form.

7.2 Methodological proposition

The detection of DHCs within their spatial context is not simple nor straightforward process. The design of the method consist of several steps outlined in the following section. The first step is definition of principles of DHC recognition which are then followed as subsequent steps through the rest of the method

design, and consequently reflected in the structure of the section.

7.2.1 Principle of DHC recognition

Recognition of DHCs is based on the principles we know from numerical taxonomy, but is a slightly specific way. In biology, the issue of individual delimitation is non-existent. Single individual of selected species is usually well defined in space (e.g., a bird), however in urban form this distinction is not so simple. Hence, the methodology which is used in biology needs to be adapted, while keeping the fundamental principles in place. The specificity is in the shift of the scale. While previous chapters identified urban tissue as *individual* of urban form, at this stage we pretend that this role holds duality building-tessellation cell as the smallest entity of urban form. The whole DHC recognition is then based on the assumption that entities 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).

Another difference between traditional method outlined by numerical taxonomy and the one adapted for the purpose of DHC recognition is the nature of morphometric characters. While in biology, each individual is usually measured independently of the rest (REF), that is not viable for urban form. The overall aim is to identify built-up patterns within urban fabric. However, the urban form itself is full of exceptions form the pattern. Individual plots follow different development process and are in some cases amalgamated or split. That does not happen tot he 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. To overcome the issue of exceptions, proposed method is working with two kinds of characters - primary and contextual.

The primary characters are those focusing on the individual elements and their relationships as identified in a relational model (Chapter 6). Typical example could be building height or area. Both are specific to each individual building and in the context of plots with internal construction, buildings in the head and the back of the plot will have significantly different values.

As primary characters by definition do not describe the pattern but rather its individual elements, they should not be used within pattern detection algorithms. The second kind of characters, contextual, has been designed specifically to turn values captured by primary characters into values describing the central tendency in the area - describing the pattern. As such, they can be used as an input for clustering aiming to distinguish DHCs.

Finally, 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.

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

7.2.2 Morphometric characters

The main scope of this research is not to develop new morphometric characters (even though there are some), but to use existing knowledge in urban morphometrics and combine it in a systematic framework providing a complex description of urban form. The chapter 3 mapped in detail the existing characters used across the field and the resulting database and classification is the basis for selection and definition of primary characters and to some extent even contextual characters.

7.2.2.1 Primary characters

As briefly outlined above, primary characters describe different elements and their relationships as are identified within the relational model of urban form. Building on the definition of the term *primary* from Oxford English Dictionary (REF), we can define primary characters witting the context of DHC recognition as *characters occurring first in a sequence of methodological steps capturing individual features of urban form elements and their basic relations.* The link to the relational model is crucial here as it defines which relations are meant and later reflected in the whole recognition model.

Chapter 3 shows that there is a large number of characters which could be, in theory, used within the model. However, the selected set of characters needs to have specific nature. The information captured should be non-overlapping, each of them should describe different unrelated feature of urban form to avoid clustering result distortion towards features occurring multiple times. For that reason, specific principles of characters selection were defined.

7.2.2.1.1 Principles of character selection and definition ***THIS SECTIONS NEEDS SIGNIFICANT CHANGES*** The idea of morphometric recognition of DHC is based on numerical taxonomy and the selection of morphometric characters then build on the principles used within selection of taxonomic characters in biology, as defined by Sneath & Sokal (1973). 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 provide the description of urban form suitable for a recognition of DHC. 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 3 and, more importantly, to the relational model of urban form. There are three top-level aims of the of the selected set 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. To generate complex description of urban form, all these aspects should be incorporated. However, as

different categories tend to focus on different scales and elements (REF Ch3), 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 DGC recognition model.

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

In this case in the context of the relational model, 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 model (coming from two data sources as MT is generated) hence this aim is focused on these only.

3. **Capture scalar complexity of urban form by covering all meaningful topological scales**

Relational model defines four topological scales:

- single/small
- medium
- large
- extralarge

For the purpose of DHC recognition, not all of them are equally meaningful, as the spatial extent of DHC is usually restricted and *extralarge* topological scale then

likely spans across multiple DHCs, rendering most of the characters occurring on that scale unhelpful. However, S, M and L are all relevant for the scale of DHC and should all be represented. The city and its urban form is composed of nested complexities (REF) occurring on different scales. Capturing them all together within the single model allows description of scalar complexity needed for complex and systematic morphometric characteristics of built-up patterns.

To fulfil the aims, relational model comes to help with defined subsets as a combinations of elements and scales, combining second and third aim into a single solution. Each of the subsets represent 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 Sneath & Sokal (1973) to determine a methodical approach to 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 3, so resulting database of morphometric characters can be directly used. This database works as the main source of characters. Due to its extent, it is expected that the majority of possible characters is included.

2. **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 model. Some can be adapted (e.g., morphological cell can be, in some cases, used instead of plot), but some are based on the different sources of data. Characters which could not be used within subsets of relational model are then excluded from the initial selection.

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

Those characters which are applicable, but are not readily available to be used within relational model 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 done with a sense of the meaning of each character which should not be significantly changed, otherwise is foundation in literature would be questionable and should be seen as a newly developed character.

4. **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.*" (Sneath & Sokal 1973, p.104)

5. **Eliminate ineffective characters**

Due to the nature of the analysis, working with large-scale data or even big data in some cases, the process of measuring has to be computationally effective. 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 axial maps or topological skeleton.

6. **Add characters where are clear gaps**

(diversity, plot-level Voronoi cell). Because I am using morphological cells the smallest scale spatial unit in a scope previously unused, there is a range of characters which had to be adapted from original plot-based to cell-based. The database of characters also showed imbalance of different categories and gaps in the measuring of diversity. New taxonomic characters have to be implemented to cover those gaps and provide coherent description of urban form. This part of the research is still ongoing.

7. **Exclude invariant characters**.

Some characters might be invariant over the entire sample of OTU's. Those should not be included as they are not bearing any taxonomic value. However, this exclusion is an ongoing process, because it depends on actual measured values.

8. **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 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.

9. **Exclude characters which does not have the ability to capture patterns**.

Test capability of each character to capture spatial patterns by measuring spatial autocorrelation as global Moran's I. Those without sufficient level of autocorrelation should be excluded as they do not bear any value in the process of identification of DHC.

10. **Balance scalarity 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 scales (street, block, vicinity). 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 street edge than block of vicinity 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 minimum.

The process of selection itself starting from the database retrieved from chapter 3 is available as Annex 2. It includes details of each decision on which characters should be part of the final set and why. Following section describes the final set only.

7.2.2.1.2 Identified set of primary characters Based on the principles described in the section above, following morphometric characters compose the final set of primary characters. For the implementation details please refer to the original referred work and to the documentation and code of momepy, which contains Python-based implementation of each character.

The most simple of the characters are those capturing *dimensions* of buildings:

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

(1) a_{blg}

and defined as an area covered by a building footprint in m^2.

2. **Height of building** is denoted as

(2) h_{blg}

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. The character based on the data provided by IPR Prague is illustrated on figure 7.1.

3. **Volume of building** is denoted as

(3) $v_{blg} = a_{blg} \times h_{blg}$

and defined as building footprint multiplied by its height in m^3.

4. **Perimeter of building** is denoted as

(4) p_{blg}

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

5. **Courtyard area of building** is denoted as

(5) a_{blgc}

and defined as the sum of areas of interior holes in footprint polygons in m^2.

Figure 7.1: Character height of building within central part of Prague as provided by IPR Prague. The distribution is truncated of extremes and captures only the visible area.

Further characters capture *shape* of buildings in both two and three dimensions (considering approximate building height as the third dimension):

6. **Form factor of building** is denoted as

$$(6)\ FoF_{blg} = \frac{a_{blg}}{v_{blg}^{\frac{2}{3}}}.$$

It captures three-dimensional shape characteristic of a building envelope unbiased by the building size (**???**).

7. **Volume to façade ratio of building** is denoted as

$$(7)\ VFR_{blg} = \frac{v_{blg}}{p_{blg} \times h_{blg}}.$$

It captures another aspect of three-dimensional shape of a building envelope distinguishing building types adapted from (**???**). It can be seen as a proxy of a volumetric compactness.

8. **Circular compactness of building** is denoted as

$$(8)\ CCo_{blg} = \frac{a_{blg}}{a_{blgC}}$$

where a_{blgC} is area of minimal enclosing circle. It captures the relation of building footprint shape to its minimal enclosing circle, illustrating the similarity of a shape and circle (Dibble et al. 2017).

9. **Corners count of building** is denoted as

$$(9)\ Cor_{blg} = \sum_{i=1}^{n} c_{blg}$$

where c_{blg} is defined as a vertex of building exterior shape with angle between adjacent line segments $\leq$ 170 degrees. Uses only external shape (`shapely.geometry.exterior`), courtyards are not included. Character is adapted from (**???**) to exclude non-corner-like vertices.

10. **Squareness of building** is denoted as

(10) $Squ_{blg} = \frac{\sum_{i=1}^{n} Dc_{blg_i}}{n}$

where D is the deviation of angle of corner c_{blg_i} from 90 degrees.

11. **Equivalent rectangular index of building** is denoted as

(11) $ERI_{blg} = \sqrt{\frac{a_{blg}}{a_{blgB}}} * \frac{p_{blgB}}{p_{blg}}$

where a_{blgB} is area of minimal rotated bounding rectangle of a building (MBR) footprint and p_{blgB} its perimeter of MBR. It is a measure of shape complexity identified by (**???**) as the shape characters with the best performance.

12. **Elongation of building** is denoted as

(12) $Elo_{blg} = \frac{l_{blgB}}{w_{blgB}}$

where l_{blgB} is length of MBR and w_{blgB} is width of MBR. It captures the ratio of shorter to longer dimension of MBR to indirectly capture the deviation of the shape from a square (**???**).

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

(13) $CCD_{blg} =

TODO

ADD KEY TO CHARACTERS IDS

The final set is 74 morphometric characters spanning across the subsets of relational model and covering all categories, even though not equally.[1] The set is non-overlapping and does not contain logically correlated characters. As such, it should provide unbiased and non-skewed description of each of the elements.

7.2.2.2 Contextual characters

Looking at the primary characters and their spatial distribution, they could be really abrupt and do not necessarily capture urban patterns as they are (even though all capture some patterns as per spatial autocorrelation). Two illustrations of such an abrupt change and the weak pattern description are XXX (fig) and YYY (fig). [TODO: ADD EXAMPLES AND THEIR DESCRIPTION]

To become useful for pattern detection within DHC recognition model, most of the characters defined above has to be expressed using their contextual versions. *Context* here is defined as neighbourhood of each tessellation cell within 3 topological steps on MT. That covers approximately 40 nearest neighbours (median 40, standard deviation ~13.4 based on Prague) providing 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 Annex XXX for sectional diagram analysis). Contextual character is then capturing a central tendency or a distribution of a primary character within a set context.

Within this method, four types of contextual characters are proposed. One capturing a local central tendency and three capturing the various kinds of diversity of values within the context. For each of the primary characters, each of the contextual is then calculated and then used within clustering algorithm itself. The resulting set of used characters is then composed of 4 times 74 characters, giving 296 individual contextual characters. Test Fig 7.2.

[1] The balance across categories within the specific set is not required as different categories offer different information relevant for different purposes.

Figure 7.2: Within this method, four types of contextual characters are proposed. One capturing a local central tendency and three capturing the various kinds of diversity of values within the context. For each of the primary characters, each of the contextual is then calculated and then used within clustering algorithm itself.

7.2.2.2.1 Local central tendency Statistics knows central tendency as a measure of a typical value for a probabilistic distribution [Weisberg H.F (1992) Central Tendency and Variability, Sage University Paper Series on Quantitative Applications in the Social Sciences, ISBN 0-8039-4007-6 p.2]. Having 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 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. Each of them fits a different purposes. To use arithmetic mean to determine central values, underlying distribution should not be skewed, otherwise outliers may significantly affect the resulting value. 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 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 XXX 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 rather asymmetric, causing volatility of mean, which should not be used in such cases. The question is then limited to the distinction between 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 median, while being similarly robust to outliers, this research

settles on implementation of interquartile mean as a measure of local central tendency.

7.2.2.2.2 Diversity as a statistical dispersion Apart from local central tendency, which aims to capture representative value, it is fundamental to understand how the actual distribution of values within the context looks like. In other words, to capture the diversity of each of the characters. While discussion on importance of diversity has been central to urban discourse since the era of Jane Jacobs (REF), as shown in the chapter 3, there is not very wide range of characters actually measuring diversity and focus mostly on Simpson's diversity index, originally developed for categorical, not continuous variables and hence relies on pre-defined "bins" (classes of values). For example, Bobkova et al. (2017) 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. This appears to be a rather ubiquitous problem in applying the 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 morphology analysis to produce comparable outcomes, it is difficult to ensure specific descriptiveness to "universal" predefined bins. The use of the Simpson's diversity index in ecology is encouraged (Jost 2006) 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. The Simpson's diversity index and similar based on binning provide values specific to individual cases where binning was set and has to be interpreted as such.

Recent literature shows that we now have alternative ways to measure the diversity of morphological characters. Caruso et al. (2017) applied the Local Index of Spatial Autocorrelation (LISA) in a 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." (Caruso et al. 2017, 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 limited number of significant categories (4), its application is limited and rather reductionist.

Another approach grounds the diversity character on the statistical distribution of all measured values and compares it to the ideal distribution. One example is a test whether such distribution follows the principle of the Power Law used by Salat (2017), but that is a not straightforward measurement, especially if the distribution is of different shape. Another is an application of the Gini index initially used to measure inequality or entropy-based indices. In the case of diversity, the more unequal the distribution is, the more diverse. Since none of these measurements requires pre-defined grouping, they resolve the problem of binning highlighted above with reference to the Simpson's diversity index.

Moreover, diversity of continuous variables could be seen as a statistical dispersion, i.e., the ratio to which the distribution is stretched (wide distribution) or squeezed (narrow distribution). Together with central tendency, dispersion is often used to describe the distribution.

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 initial character) measures. Other options would be *dimensionless* (resulting values have no units) and to include Simpson's diversity index mentioned above, *binned* measures. To understand their properties and behaviour on the real morphometric data, wide selection of most relevant from each group is analysed as a way of selecting the most appropriate measures of dispersion/diversity to be used as contextual characters.

Dimensional measures of dispersion are the most common as they are generally easy to understand and interpret. Similarly to measure of central tendency, all van be measure on the full range of values or on limited, usually again as interquartile (IQ) or interdecile (ID) range. In the analysis are included *standard deviation (SD)*, *range*, and *absolute deviations (median - MAD, average - AAD)*. Both standard deviation and range is measured for IQ, ID and unrestricted range of values. Dimensionless measures are not expressed in the same units as original

characters, so while dimensional measure of dispersion for building area will be in meters, dimensionless will have no units (the values are relative). Included are *coefficient of variation (CoV)*, *quartile coefficient of dispersion (QCoD)*, *Gini index*, and *Theil index* (a special case of the generalised entropy index). In terms of binned measures, the key question is not which one should be used, either Simpson's diversity index as in Bobkova et al. (2017) or Gini-Simpson diversity index as in Feliciotti (REF), 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* REF (number of classes is based on the Goodness of Absolute Deviation Fit (GADF)), *Head Tail breaks* (Jiang 2013) Goodness of Absolute Deviation Fit and *quantiles* (5 and 10 bins). Details of the implementation of each are in table **??** below. The reason for inclusion of Simpson's diversity index, even though it may not be fully comparable across cases is the fact that DHC recognition is always local, always case-specific. However, using the values in further profiling and comparison of clusters across cases (identified separately) might lead to misleading results.

ADD Description of characters

Using four morphometric characters as test data - building area, building height, covered area ratio and floor area ratio, all potential measures of diversity listed in table **??** were measured on three topological steps around each building. Second steps was a visual assessment of resulting maps 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) (figure XXX). Then was built a correlation matrix of remaining measures for each of the characters and assessed to identify potential overlaps and uniqueness of values. Illustrative correlation matrix[2] based on building area (figure XXX) indicates that intra-group correlation is significant, while correlation between groups less so, suggesting that each of the groups capture different information. For that reason, it might be worth identifying the most suitable of each group and using all three of them as contextual characters to obtain rich description of underlying distribution of values.

[2]Complete results of the analysis are available as an Appendix XXX.

7.2.2.2.2.1 Selected diversity characters Complete analysis of selected measured is available in an appendix XXX. Within dimensional measures, IQ range and IQ SD are better in capturing boundaries between types of development and are robust to outliers. Interquartile range was used by Dibble et al. (2017) and is easier to interpret, hence has been chosen as a representative of the dimensional category to be used as contextual character.

Differences between tested dimensionless measures are very minor with selection from Theil index, Gini index and Coefficient of Variation, all based on ID or IQ values. Due to tis definition, CoV will tends to infinity when the mean value tends to zero, being very sensitive to changes of mean. Theil index and Gini index are both used to asses inequality, but Theil index, unlike Gini, is decomposable to within-group inequality and between-group differences, making it more suitable for spatial analysis than Gini index would be. ID values used within Theil index are better 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 a structures might be filtered out. In fact, it may help distinguishing between blocks with and without internal buildings, hence second contextual character will be *interdecile Theil index*.

I terms of Simpson's diversity index, due to the fact that most of the values follow power-law (or similar exponential) distribution within the whole dataset, binning method has to acknowledge that. For that reason, HeadTail Breaks are the ideal method as it is specifically tailored to exponential distributions (Jiang 2013). Those 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.

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. Then there are three characters describing the distribution of values within the local context. *Interquartile range (IQR)* as dimensional character of diversity captures the range of values around IQM, capturing where the values mostly lie. *Interdecile Theil index (IDT)*

describes the equality of distribution of values and *Simpson's diversity index (SDI)* captures the presence of various classes of values within the context. Together, these four characters have a potential to describe spatial distribution of morphometric values within a set context.

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 clustering method to identify distinct homogenous clusters. The fact that all input data for clustering are measured using this *cookie-cutter* method ensures that spatial clusters should be geographically coherent and mostly continuous. Such a nature of data allows use of spatially unconstrained clustering methods. That is important as spatially constrained clustering is less developed and mostly unfit for datasets of the size this research works with.

To sum up, after selection of primary morphometric characters from literature and their adaptation to fit *relational model of urban form*, the set of 74 characters is established to cover wide range of descriptive features capturing urban form configuration from dimensions of individual elements, through spatial distribution to diversity. To describe a central tendency in the area capturing morphological patterns, rather than description of individual elements, four contextual characters are introduced. These, combined, have a potential to capture the nature of each of the primary characters and its behaviour in the immediate spatial context. Thanks to their autocorrelated design, contextual characters can then be fed into the unsupervised machine learning procedure aiming to distinguish distinct homogenous clusters.

7.2.3 Identification of DHC

The actual identification of distinct homogenous clusters of urban form is in principle statistical clustering of buildings with similar information about itself and its context. Moreover, to derive DHC, such clusters needs to be contiguous and internally homogenous.

Contiguity is not easy to accomplish as spatially constrained clustering methods, which are designed to be contiguous and take into account spatial relationship of clustered elements, like Skater REF or Max-p Region Problem REF are computationally inefficient, which is multiplied by the size of the datasets used within this research. They essentially would not be able to crunch the amount of data. Second option how to include spatial dimension in clustering is actual inclusion of x and y coordinates of each object (in case of building likely x and y coordinates of building centroids). The geographical coordinates would then become another two dimensions in the dataset. This solution might work if the number of dimensions is low and two additional characters could make a significant effect. As the dataset of contextual characters is in composed of 296 dimensions, simple inclusion of two other might not make much of a difference and not ensure any spatial contiguity.

The solution of the contiguity issue is, in fact, built in the design of contextual characters. As their measuring follows location-based manner, so called *cookie-cutter* method of spatial aggregation, all characters are actually significantly spatially autocorrelated by design[3]. There is a significant overlap between areas used for computation of contextual characters of two neighbouring cells that indirectly ensured 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.

The general principle of clustering, i.e. unsupervised machine learning is using the learning data (which in case of clustering is often the whole dataset, but

[3]Median I is 0.77, St.Dev 0.12, with values ranging between 0.42 (Square Clustering of Street Network Theil Index) and 0.98 (Gross Density Interquartile Mean) all with *p < 0.001*. Complete Spatial Autocorrelation analysis is available as Appendix XX

can be sampled) to iteratively determine the optimal division of observed data into homogenous clusters and then applying learned model to the whole data to predict to which cluster each element belongs. In terms of probabilistic methods, this prediction can have associated probability that 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 most likely k-means clustering REF and its derivatives (k-medoid, k-median or Gaussian mixture models). 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. 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 unfit 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 heterogenous, meaning that algorithms expecting and identifying noise (in this case buildings which do not belong to any cluster) in the data like DBSCAN REF, HDBSCAN REF or OPTICS REF are not ideal either.

7.2.3.1 Gaussian Mixture Model clustering

Clustering method which does reflect the nature of the problem is Gaussian Mixture Model (GMM), which is a probabilistic derivative of k-means, but unlike the k-means itself it does not rely on squared Euclidean distances only. GMM is based on an assumption that each dimension of each cluster is represented by a Gaussian distribution, hence the cluster itself is defined by a mixture of Gaussians.

To illustrate the behaviour in a visual way, take the following example (figure 7.3)

of a two dimensional dataset with 4 known clusters. The clusters are of unequal size, density and shape. Because we do not know what properties will have DHC in hyperspace, it is safe to assume that they could be similarly complicated.

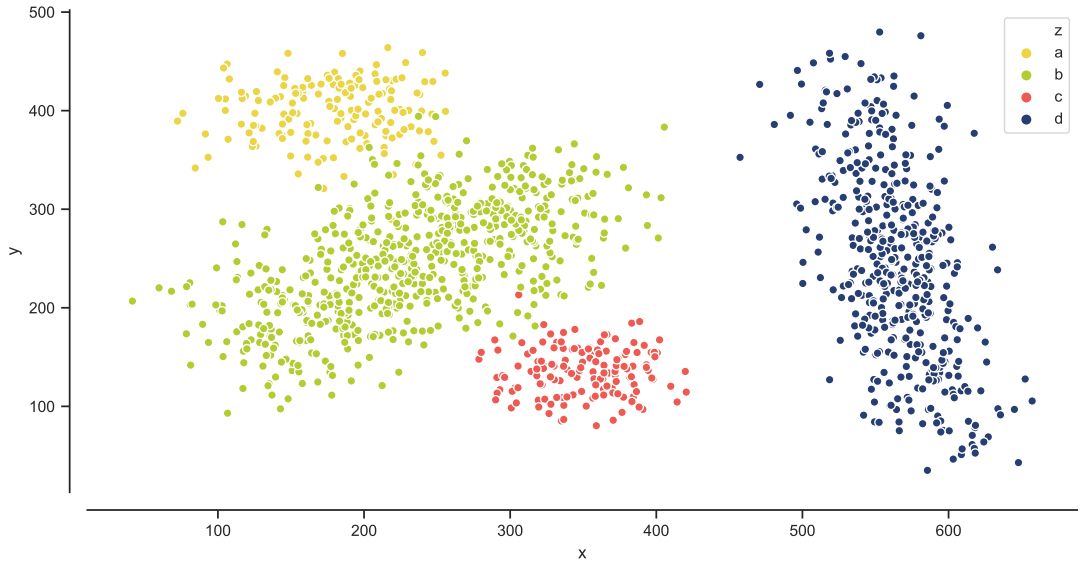

Figure 7.3: Artificial two dimensional (x, y) dataset containing 4 known cluster.

Let's first check what would be the result of clustering using k-means algorithm with k=4. The figure 7.4 shows 4 clusters, but only one (0) being correctly labeled. The variable shape and density of other three cluster together with the close proximity unveils the weak points of k-means algorithm. We can see that the purely distance-based definition does not provide the appropriate results.

Where k-means looks for clusters of similar extent, GMMs embedded expectation-maximization (EM) algorithm allows identification of different shapes. EM is an iterative method which starts from random points (like k-means) but is able to find 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.

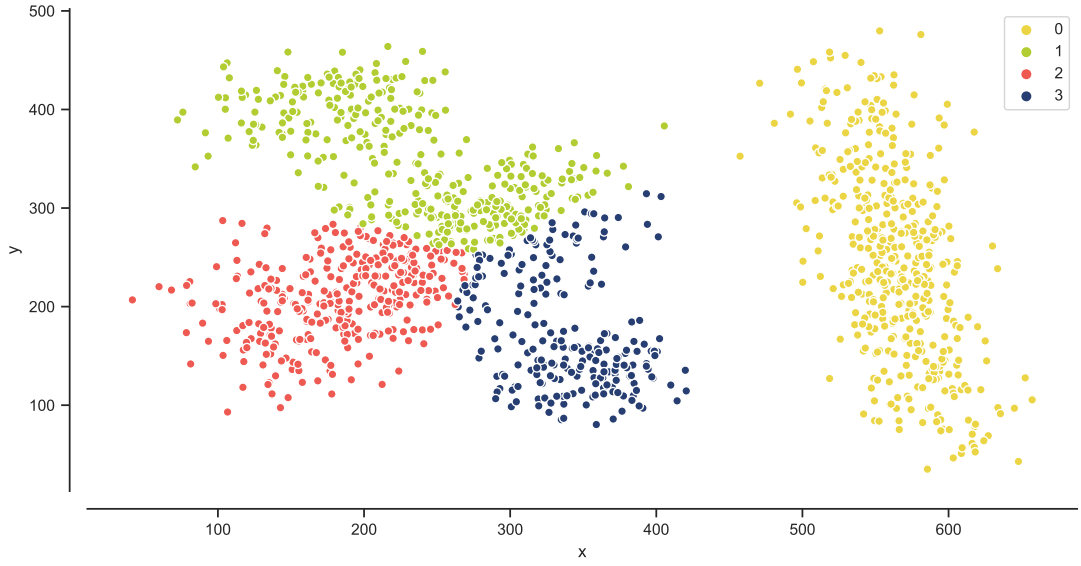

Figure 7.4: K-means clustering (k=4) of the artificial two dimensional (x, y) dataset containing 4 known cluster. Apart from the one cluster (0) which is clearly separated, none was correctly distinguished.

The result of GMM applied to the artificial dataset, as shown on figure 7.5, 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 belongs to the selected cluster.

Due to the fact that 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. To ensure the stability of the clustering, it has to be done repeatedly in several initialisations of which the best should be used.

Within this research, an `sklearn.mixture.GaussianMixture()` implementation of GMM within open-source python package scikit-learn v.0.22 (REF) is used. Further details on the exact algorithm are available in scikit-learn documentation and code.

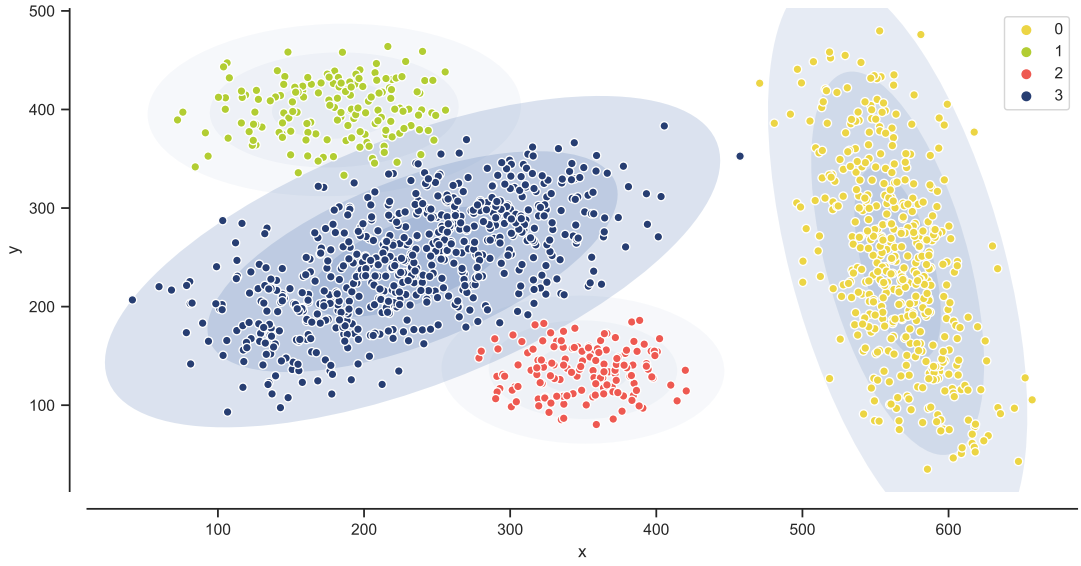

Figure 7.5: GMM clustering (4 components) of the artificial two dimensional (x, y) dataset containing 4 known cluster. All clusters are fairly successfully distinguished. Figure also shows underlying Gaussian distributions as ellipses reflecting the probability by change of the shade.

7.2.3.2 Dimensionality issue

The morphometric description of each building/cell has 296 values (each for each contextual character). In the case of Prague, composed of approximately 140,000 buildings, it means that clustering has to deal with more than 40,000,000 data points (140,000 buildings * 296 characters). That is a significant number, which is not only demanding in terms of computational power, but also tricky in terms of statistics itself. The high dimensionality of the dataset (each character is a dimension in a hyperspace) may come with a *curse of dimensionality*. That means that even though there is the value in additional data (additional dimensions), it may affect results in a negative way. The high-dimensional hyperspace tends to become inflated (bigger), which in turn may render clusters very sparse. Individual data points are further away and density-based, or distance-based clusterings (GMM is distance-based) may struggle to correctly identify them as Euclidean distances between pairs of points on sparse high-dimensional data would be of little difference, rendering clustering extremely unstable and insignificant. However, that is not always the case as it depends on the internal structure of the dataset and relations between dimensions.

One way how to deal with large number of characters is a reduction of dimensionality. Two of the most applied statistical methods to reduce number of dimensions of data are Factor analysis (FA) REF and Principal compoment analysis (PCA) REF. Both are aiming to describe the dataset using the smaller number of *factors* or *principal components* (essentially dimensionless variables hard to interpret). The key concept allowing the generation of meaningful clusters and effective reduction of dimension which is used in both is correlation of original variables. That causes an issue in reduction of used morphometric dataset as it is designed to limit empirical correlation, hence FA and PCA are expected to be not very effective in reduction.

The preliminary tests of PCA on the complete dataset of contextual characters shows that to retain at least 95% of variance, one need at least 147 principal components (Figure 7.6). That is a significant reduction, but the ideal number of dimensions is approximately 30-50, so the reduction is clearly not good enough. Using 30 principal components, the retained variance drops to 69%, for 59 components the value would be 78%. Because there is no set acceptable rate of explained variance needed, without validation data it is not possible to determine acceptable number of components. The results might or might not offer helpful reduction of dimension.

Difference between 296 dimensions of original dataset and 160 dimensions to keep at least 95% of variance might offer reduce computational demands, but at the same time complicates interpretation of clusters where each of the 147 components is a black box without a morphological meaning. It is expected that GMM will be able to handle 296 dimension, even though the computation might require more resources. The decision for the purpose of this research is to skip dimensionality reduction, unless GMM proves to struggle to identify clusters. In the further development of the method, it may be helpful to employ PCA, however that is left for future exploration.

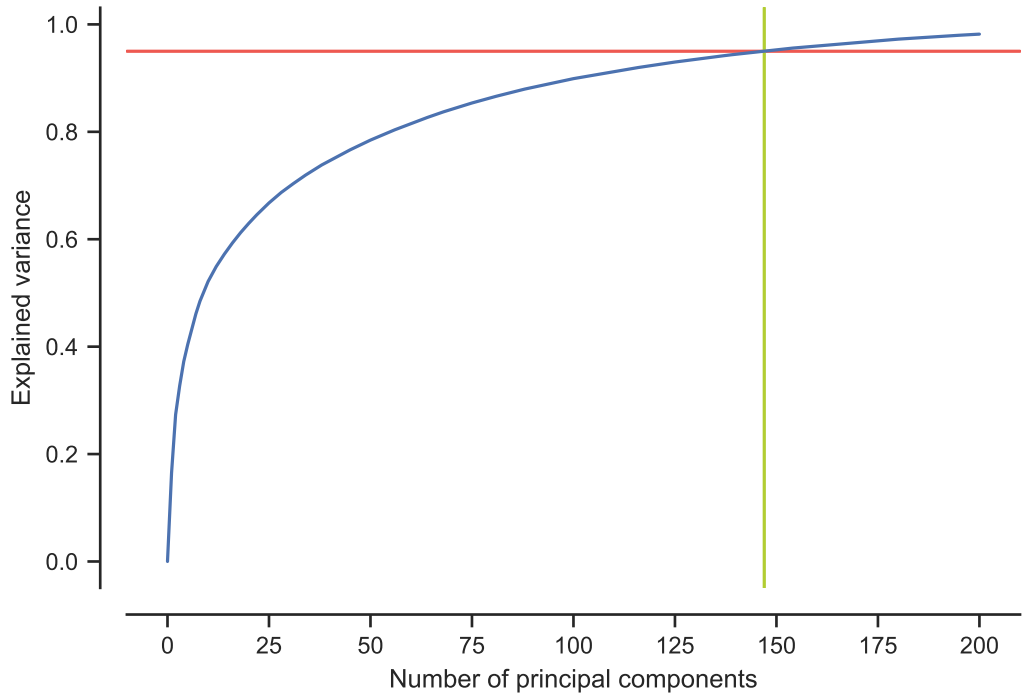

Figure 7.6: Principal Component Analysis results on the contextual characters (n=296) on Prague data. Red line marks 0.95 explained variance, green line denotes 147 principal components as a first value reaching 0.95.

7.2.3.3 Levels of DHC resolution and its scalability

The ideal outcome of DHC recognition is each cluster as a distinct urban tissue. 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 different one. This issue is actually 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 traditional sense, but their aggregations. On the other hand, detecting 100 would likely represent over-fitted model and each cluster would be only a part of a tissue. It is expected

that statistically optimal number of clusters should be close to what we would normally call urban tissue, however this link require further interpretative work, which should happen based on taxonomy of DHC to allow scalar flexibility. For that reason, this section focuses on the first part, i.e. detection of optimal number of clusters, and section XX in following chapter 8 discuss the relationship between tissue and DHC in detail.

7.2.3.3.1 Number of components Gaussian Mixture Model clustering requires, similarly to k-means, specification of number of components of the model (i.e., clusters) prior 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 essentially against the prepositions set in chapter 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 essentially means that the GMM is trained multiple times based on range of feasible options of number of components and each of the models is then assessed against the whole dataset (how well are clusters 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 goodness of fit to specify the number of components for the final clustering.

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

In practice, BIC is measured for each trained model based on the original data. The lowest the BIC is, the better the model represent original data.

The interpretation of the goodness of fit score is not 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 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 aim of the clustering is 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,REF 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 clusters 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 number of components before the flattening of the gradient.

7.2.3.3.1.2 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 plotting the confidence interval to help with the interpretation later. To better understand the magnitude of the effect, 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 selected number of components - model should be initialised repeatedly and the best of the resulting models should be kept and used.

The result of clustering is never exactly the same, especially with the amount of the data this research is using. There is a certain 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 certain degree of fuzziness, which could be captured by overlay of resulting clusters form multiple models of same parameters.

7.2.3.3.2 Sample-based clustering As the dataset grows, it may become increasingly impossible to run clustering on the whole dataset, especially if we want our data with meaningful confidence interval. The calculation of dimensions between components of the model in 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 number of feature can reach millions. The data like this can be run in a 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 fully true. The larger the sample is, the more similar to the whole 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 train GMM on the full of 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 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 number of smaller cases will be probably affected more than places with homogenous structure where the likelihood of proper sampling of all clusters is higher.

7.2.3.3.3 Sub-clustering There are situations when resulting clustering is not refined enough for the purpose of the specific analysis. There are simply 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 don 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 DHC 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 than 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 certain 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. DHCs identified in Prague using solely Prague-based data would not be equal to DHCs identified in Amsterdam using Amsterdam-based data only. The structure of both datasets determines what is the optimal division and as both structures are different, the optimal division is done along different lines. It is expected that results will be comparable as optimal DHC should reflect optimal urban tissues, but there will always be certain misalignment of clusters. Chapter 8 will test whether the misalignment is significant or not to further explore the link between two local clustering models.

conclude clustering

7.2.4 Data preprocessing

Before doing any of these steps, it is fundamental to ensure that data are good enough to represent morphological/morphometric elements. That could be an issue for both building and street network layers, so there are cases when the data needs to be prepared for morphometric analysis. The preprocessing can be in some cases automatised, in other, unfortunately, manual or at least semi-manual to have the data in the correct shape in the end.

While each dataset coming from different source is specific hence the cleaning procedure needs to be tailored to each source, there are some common issues which are not unique to specific datasets. Following section outlines these common issues and how to resolve them or at least minimise the error under the significant level. As the method described above is error-prone due to the design of contextual characters, the data do not have to be perfect all the time. However there are cases where even contextual character can be significantly affected and skew the result of clustering.

7.2.4.1 Preprocessing of buildings

ADD ILLUSTRATIVE FIGURES HERE AND BELOW

Having data layer correctly representing building footprints is crucial from two reasons as it not only affect morphometric characters based on buildings, but also morphological tessellation and consequently characters based on morphological cells, which in the end are all contextual characters. There are several aspects which needs to be fulfilled - topological correctness, consistency in detail, representation of individual buildings and building height attribute presence. Overall, it is expected to have a building data representing Level of Detail 1 (LoD1) REF Bilejcki.

Topological correctness ensures that geometry represent the actual relationship between buildings on the ground. There are characters measuring continuity of a perceived wall in a joined buildings or shared walls ratio which require building polygons to be correctly snapped together when two buildings touch. In that case it is expected that neighbouring polygons will share vertices and boundary segments. There should not be a gap between polygons when there is none in reality and vice versa. Also, polygons must not overlap at any case as that would cause significant disruption of tessellation geometry.

The building detail should be consistent across the dataset and represent optimal approximation of building shape based on LOD specification as proposed by REF Bilejcki. The approximation should represent LOD1.1 (no details, but shape is kept) or LOD 1.2 (minor details), building shapes should not be overly detailed nor overly simplified. In the case of inconsistency, simplification of more detailed shapes needs to be done before morphometric assessment.

Each polygon has to represent a single building. There are datasets (often of remote sensing origin) capturing all structures which are joined by any means as a single polygon. Such a data do not represent the morphological truth on the grounds. Their preprocessing is complicated as it requires splitting of existing geometries according to additional dataset. The second extreme is the opposite situation, when a single building is represented by multiple polygons. These usually represent different height levels, through routes or similar features. If these polygons, representing parts of buildings, have a common ID which allows joining them together to get a single polygon representing a single buildings, the preprocessing of such a data is only a simple dissolution. However, there

are many cases when this ID is missing and correct pre-processing require either clever algorithms understand which polygon belongs to which or laborious manual work.

Certain number of primary and subsequently contextual characters uses building height attribute, which has to be present in original input dataset. The resolution should be able to capture the distinction between levels, further detailing is not significant. The input can be either in meters (optimal) or in number of storeys, which should then be represented as a metric approximation as characters expect height to be in meters.

7.2.4.2 Preprocessing of street network

Similar situation as with building layer is with street network. Incorrectly drawn street network may cause significant errors in morphometric results and consequently in clustering. There are three most important cases which needs to be checked before the analysis - topological correctness, morphological correctness and consistency in classification.

Topological correctness ensures that each street segments is represented by a single `LineString` geometry, that neighbouring segments share end vertex and that geometry is not split if the segments intersects only on projected plane and not in reality (typically multilevel communications, when one is on the bridge across the other so that projected intersection is not real intersection).

Moreover, street networks have to be morphologically correct, which means that geometries represent morphological connections, not other, usually transport-focused. That often mean simplifications of networks to eliminate transport geometries like roundabouts or similar types of junctions, or dual lines representing dual carriageways. In certain cases networks have to be snapped together, because due to traffic calming measures some junctions might not be connected when they should be.

Finally, network needs to be consistently drawn in terms of inclusion of different levels of network hierarchy. The definition of what is street and should be included

and what is minor connection and should not is crucial for comparability of results. *DEFINE PROPERLY WHAT IS STREET*

As per data availability, networks are widely availably. However, geometries mostly represent transport network and often do not follow ideal topological rules. The preprocessing to ensure that all three points above are fulfilled is hence necessary and can be partially automatised either using `momepy.network_false_nodes` or using methodology outlined by Krenz (REF pp.), using conventional GIS tools. However, there might be cases when more complicated procedures should be employed, either to ensure that algorithm is more precise or to include manual steps.

It is not complicated to find case studies offering the data in a required quality and detail, but it is true that data of this level of precision are not available everywhere around the world. That is true especially for building height parameters. Having all data as outlined above is the ideal situation, which will be tested in this research. In the real world, situation might be less ideal, so preprocessing procedures has to be employed before performing the analysis itself. The case analysis using extremely sub-optimal data is available as Annexe X, outlining the work done on Grand Rapids, Michigan using building footprints not representing individual buildings and missing any height attributes.

7.2.5 Data model

The data model representing the elements of 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 7.1.

Table 7.1: 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			

layer	uID	nID	nodeID	bID
street edges		x		
street nodes			x	
blocks				x

Buildings are in the role of a connecting elements and contain all identifiers. Morphological tessellation is based on the building layer, 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.get_network_id`). Street nodes are linked to buildings based on proximity either, but linked node has to be end node of linked nearest edge (`momepy.get_node_id`). Blocks are based on tessellation and their id is linked to buildings using intersection-based spatial join during their creation (`momepy.Blocks`).

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

conclude methodology

7.3 DHC recognition | Case study Prague

The first trial of DHC recognition method outlined above is the case study of Prague, limited to its administrative boundary, which in the case of Prague extends the continuous built-up area and ensures that the edge effect cause by street network cutting is minimised (figure 7.7). Following section reports on each step of 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 taxonomical analysis presented in chapter 8.

Prague dataset after pre-processing contains 140 315 individual buildings, 22 503 street edges, 16 207 street nodes and 7 395 tessellation-based blocks.

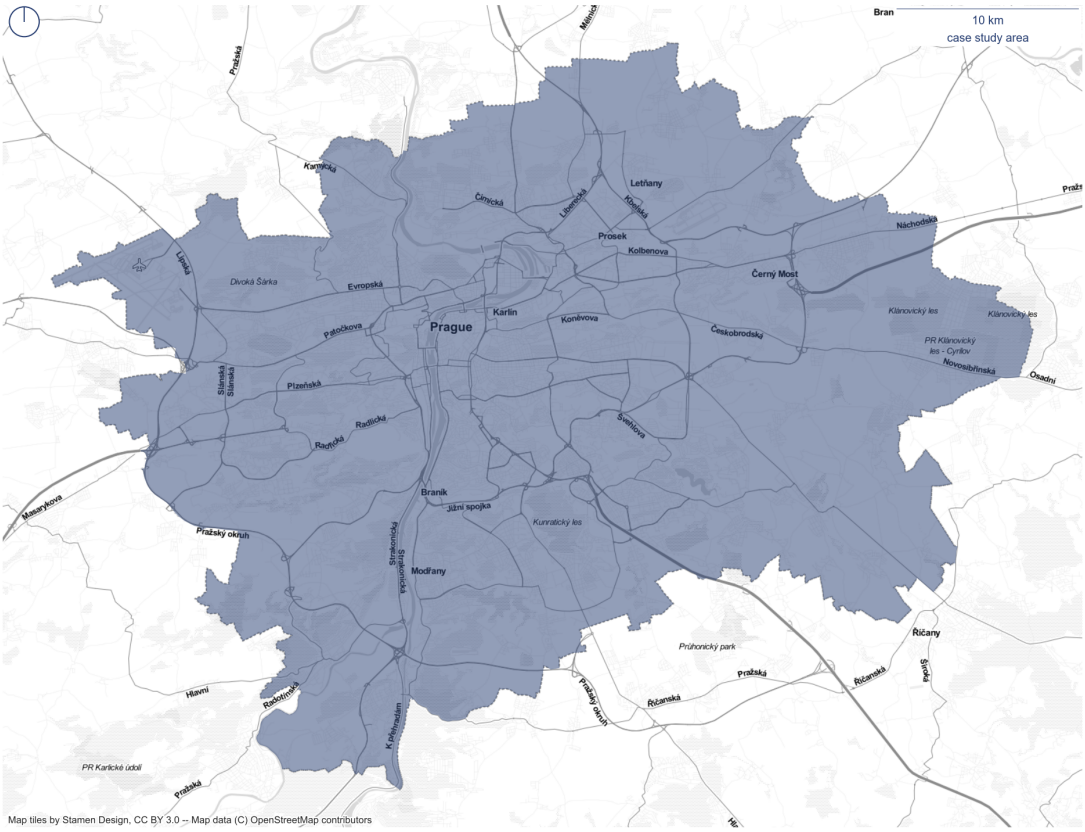

Figure 7.7: Prague case study area, which matches the administrative boundaries. Data source © IPR Praha, CC BY-SA 4.0

7.3.1 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 model. Following the method, all 74 of them are measured in Prague and then linked to the building-tessellation unit following the data model. Descriptive summary values are presented in the table 7.2. Note that units for each character are available in section XXX. All morphometric characters are measured using `momepy` classes using reproducible Jupyter notebook `XX_XXXXX` presented as an Appendix XXX.

Table 7.2: Overview of the primary morphometric values for the whole case study. Key to character IDs is available in table XXX.

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

id	mean	std	min	25%	50%	75%	max
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

id	mean	std	min	25%	50%	75%	max
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

The data shows the variety of statistical distributions and

- select few and show distributions and related maps
  - normal distributions
  - power-law distribution
  - other distribution (orient, …)
  - invariant
- show correlation matrix

Due to the large number of characters, maps visualising the data are not presented within this study.

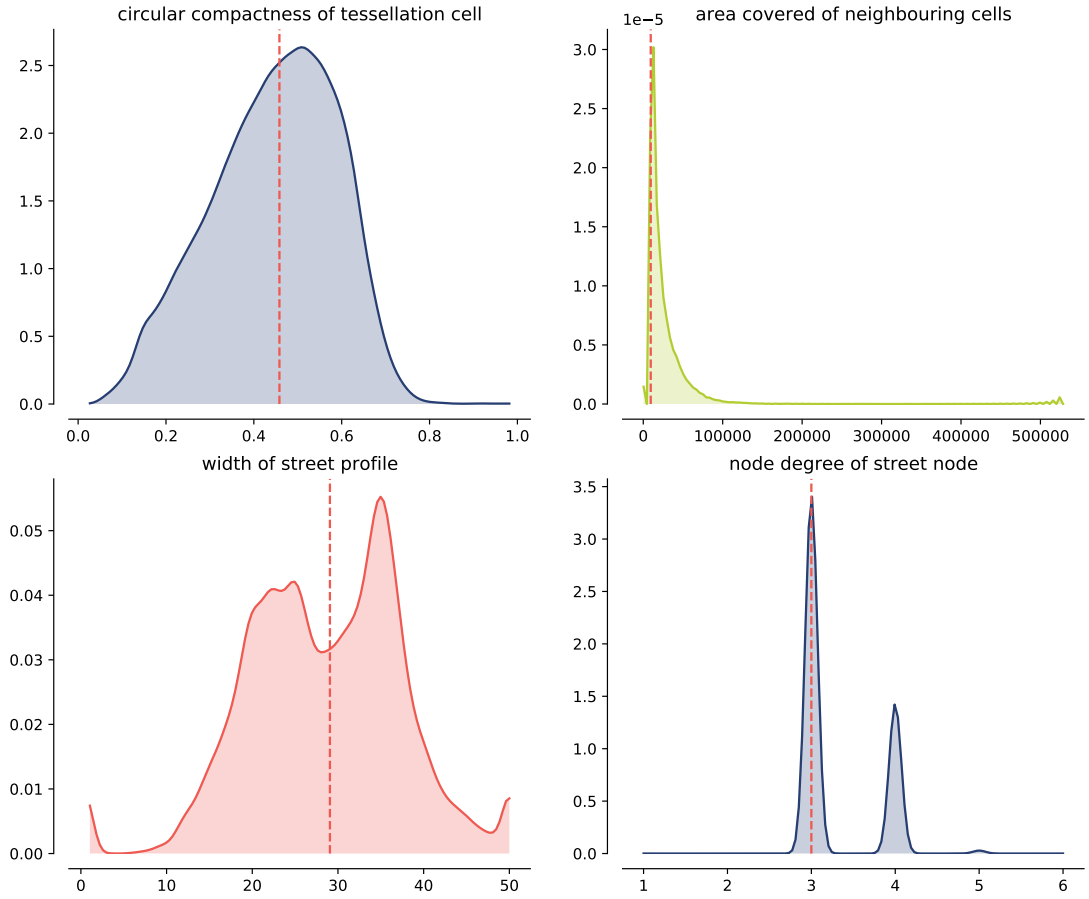

Figure 7.8: normals

7.3.2 Contextual characters

illustration of contextual characters esp. in relation to primary ones - Few examples, rest in Appendix? - Use the same examples as in Primary so illustrate smoothing

7.3.3 Clustering

introduction of clustering - abc will happen - major expectations

7.3.3.1 Complete data

BIC BIC gradient TT distance Interpretation of score map and its (basic, as detailed is in Ch8) interpretaion

7.3.3.2 Sampled data

Score BIC BIC gradient TT distance Interpretation of score Comparison of sampled and complete compared graphs and statistical values compared resulting clustermaps

7.3.3.3 Probability of cluster (change)

note on probability of cluster assignment due to the richness of data, clusters are very well defined, there is probability but they are insignificant

7.3.3.4 Subcluster illustration

Sub-clustering question test on compact urban form (perimeter blocks and modernism)

7.3.3.4.1 Compact Prague *BIC and others Map Interpretation*

7.3.3.4.2 Modernist Prague *BIC and others Map Interpretation*

7.4 DHC as an urban tissue

morphometric characters certainly help in description of urban tissues clustering helps make sense out of it DHC is a numerical, morphometric statistical proxy of urban tissue Clustering is non-deterministic, so boundaries are not fixed, rather

indicative. It is not a ground truth and the meaning and relation of clusters has to be interpreted before any further steps hierarchical clustering will help with that

Chapter 8

Taxonomy of urban tissues

- Forming a taxonomy from sample data (chosen UK cities?)

Chapter 9

Synthesis

Appendix 1: Some extra stuff

Add appendix 1 here. Vivamus hendrerit rhoncus interdum. Sed ullamcorper et augue at porta. Suspendisse facilisis imperdiet urna, eu pellentesque purus suscipit in. Integer dignissim mattis ex aliquam blandit. Curabitur lobortis quam varius turpis ultrices egestas.

References

Bobkova, E., Marcus, L.H. & Berghauser Pont, M., 2017. Plot systems and property rights: Morphological, juridical and economic aspects. In *XXIV International Seminar of Urban Form*. Valencia.

Caruso, G., Hilal, M. & Thomas, I., 2017. Measuring urban forms from inter-building distances: Combining MST graphs with a Local Index of Spatial Association. *Landscape and Urban Planning*, 163, pp.80–89.

Dibble, J. et al., 2017. On the origin of spaces: Morphometric foundations of urban form evolution. *Environment and Planning B: Urban Analytics and City Science*, 46(4), pp.707–730.

Jiang, B., 2013. Head/Tail Breaks: A New Classification Scheme for Data with a Heavy-Tailed Distribution. *The Professional Geographer*, 65(3), pp.482–494.

Jost, L., 2006. Entropy and diversity. *Oikos*, 113(2), pp.363–375.

Salat, S., 2017. A systemic approach of urban resilience: Power laws and urban growth patterns. *International Journal of Urban Sustainable Development*, 9(2), pp.107–135.

Sneath, P.H.A. & Sokal, R.R., 1973. *Numerical Taxonomy*, San Fransisco.