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

Table of Contents

Table of Contents

List of figures

List of tables

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

intro as a link back to theory in chapter 4 and 6 reintroducing morphometrics and numerical taxonomy the aim of this chapter is to develop a morphometric method able to distinguish distinct types of urban tissues

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. Similar research has been done before (REF), but 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 selection 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

Restate principles of numerical taxonomy in biology 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).

Link to the urban form and this specific methodology Systematic methodology means that it is… Comprehensive methodology means that it is… This method is trying to be both by…

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 contextualised.

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, contextualised, 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 contextualised 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, contextualised 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 complex coherent framework. 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 contextualised 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. Adapting 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 *To select set of primary characters, following principles are followed. rules based on relational model and characters classification rules based on Sneath and Sokal (check with Annex 2) Initial selection and then expansion cleaning of the selection (check with rules above) full details of the steps are in Annex 2*

The process of selection itself starting form the database retrieved from chapter

3 and details of each decision which characters should be part of the final set and why is available as Annex 2. 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 detailed description of each character and its implementation please refer to the original reference and to the documentation and code of momepy.

LARGE TABLE OF CHARACTERS WITH FORMULAS AND REFERENCES

The final set is 74 morphometric characters spanning across the subsets of relational model and covering all categories, even though not equally.[^The balance across categories within the specific set is not required as different categories offer different information relevant for different purposes.] 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 Contextualised 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 contextualised 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). Contextualised character is then capturing

a central tendency or a distribution of a primary character within a set context.

Within this method, four types of contextualised 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 contextualised 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 contextualised characters.

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 contextualised 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 trun-

cated 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 reductionalist.

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 contextualised 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 7.1 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 TABLE WITH FORMULAS

Table 7.1: This is the table caption. Suspendisse blandit dolor sed tellus venenatis, venenatis fringilla turpis pretium.

Column 1	Column 2	Column 3
Row 1	0.1	0.2
Row 2	0.3	0.3
Row 3	0.4	0.4
Row 4	0.5	0.6

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 7.1 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 [^Complete results of the analysis are available as an Appendix XXX.] 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 contextualised characters to obtain rich description of underlying distribution of values.

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 contextualised 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 contextualised 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.

Each of the primary characters is represented by its local central tendency and local diversity (using all 3 characters) Conclude contextualised characters conclude all characters

7.2.3 Gaussian clustering

Once we have a description of individual elements, we have to cluster them to identify DHC General principle of clustering aka unsupervised machine learning Short overview of available methods and differences in their application

7.2.3.1 Gaussian Mixture Model clustering

introduction of Gaussian Mixture Model clustering and explanation of its selection and potential issues relation to k-means > k-means clustering tends to find clusters of comparable spatial extent, while the expectation-maximization mechanism allows clusters to have different shapes. (https://en.wikipedia.org/wiki/K-means_clustering) *definition based on gaussians probabilistic (soft clustering); given a data point x, what is the probability it came from Gaussian k Expectation — Maximization algorithm Scikit-learn implementation is used, for detail see REF.*

7.2.3.2 Dimensionality issue

As resulting morphometric description of each building/cell has ~300 values, we are facing 'dimensionality curse'. possible reduction of dimensionality (PCA, Factor analysis). PCA and how it works Tested PCA results - 95% and ~160, ~30 and 65% We'll have to deal with it and employ a bit more computational power, data are too cleaned to be reduced to PC.

7.2.3.3 Levels of DHC resolution and its scalability

Introduce discussion on the resolution of DHC (number of clusters) and scalability of method (exponential growth of resource needs as case study area grows).

7.2.3.3.1 Selection of number of components *we have to set number of components (clusters) trial of many options test the goodness of each number*

7.2.3.3.1.1 BIC, AIC, etc. *Introduce measures of goodness of clustering Silhouette, score, BIC, AIC, BIC gradient, TT distance intro We use BIC, BIC gradient and TT distance due to… Interpretation of scores is another question we can go with lower number of clusters to maximise stability of procedure (may incur under-fitting) or with the smallest BIC (might be overfitted). However,*

as the next step is hierarchical clustering, we can use its help in interpretation of smaller clusters. - [thoughts only] score is always only indicative, it will not give us a one final answer. There are generally two options - go for conservative clustering (elbow), which might be the best idea in this case, or go for the true minimum. However, there is a clear possibility of overfitting and the minimum can be influenced by the penalisation of BIC. Conservative clustering (15 clusters in this case) will likely need sub-clustering to get a better detail.

7.2.3.3.1.2 Stability of procedure *There is a certain effect of randomness in the process, so clustering comes with a confidence interval the answer of clustering is never fixed, there is certain variability confidence interval should help us in interpretation there is an issue of multiplication of computational demands*

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 confidence interval. For that reason, it might be worth training the method on sampled data before classifying the whole dataset. There are an issues linked to it. It is always a balance between what is ideal and what is possible.*

7.2.3.3.3 Subclustering *sometimes our cluster are too big and we want better resolution clusters defined by the lowest score can still be splitter as the dataset is rich, when appropriate iteration of the clustering method on a sample of one (stable) cluster relation to other clusters is different and has to be interpreted as such the other way, joining clusters to larger groups, will be discussed in the next chapter*

7.2.4 Data preprocessing

Before doing any of these steps, we have to make sure that our data are good enough to represent morphometric elements sometimes we need to preprocess data to have them in a correct shape

7.2.4.1 The common issues with input data

there are some common issues which are not unique to specific datasets, which needs to be resolved and some of them can be dealt with algorithmically

7.2.4.2 Preprocessing of buildings

to ensure precise results of tessellation and building-based characters topologically correct joined where joined non-overlapping the detail should be consistent - overly detailed shapes are bad, overly simplified as well buildings needs to come as one polygon - so either some way of splitting (complicated) or dissolving (depends on the data) needs to be employed missing height attribute it is not so complicated to find cases with ideal data, but these are not everywhere, esp. with height

7.2.4.3 Preprocessing of street network

to ensure topologically correct network representing streets in morphological terms correctly splitted representing morphology, not transport definition of what is street and what is not (lanes) transport-based network is fairly available, there are ways (not 100% though) how to generate morphological out of it conclude preprocessing

conclude methodology

7.3 DHC recognition | Case study Prague

Application of the whole methodology to the case study of Prague

7.3.1 Primary characters

illustration of primary characters on parts of Prague - Few examples, rest in Appendix

7.3.2 Contextualised characters

illustration of contextualised characters esp. in relation to primary ones - Few examples, rest in Appendix?

7.3.3 Clustering

introduction of clustering - abc will happen

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 Modersnist 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.

Citation examples:

`[@almazan2012]` (Almazán & Nakajima 2012) `[see @almazan2012]` (see Almazán & Nakajima 2012) `[@almazan2012, pp.33-34]` (Almazán & Nakajima 2012, pp.33–34) `@almazan2012` Almazán & Nakajima (2012) `@almazan2012 [p.2]` Almazán & Nakajima (2012, p.2) `[-@almazan2012]` (2012)

From Zotero export as Better BibTeX

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.