The Urban Atlas: Foundation of a Morphometric Taxonomy of Urban Form

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:

TODO: ADD UNITS TO ALL

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 (Bourdic et al. 2012).

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 Schirmer & Axhausen (2015). 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 (Steiniger et al. 2008) to exclude non-corner-like vertices.

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

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

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 Basaraner & Cetinkaya (2017) 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 (Schirmer & Axhausen 2015).

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

$$(13)\ CCD_{blg} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}\left(ccd_i - \bar{ccd}\right)^2}$$

where ccd_i is a distance between centroid and corner i and $\bar{ccd}$ is mean of all distances. It captures the variety of shape. As corner is considered vertex with angle $< 170^{o}$ to reflect potential circularity of object and topological imprecision of building polygon.

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

$$(14)\ CCM_{blg} = \frac{1}{n}\left(\sum_{i=1}^{n} ccd_i\right)$$

where ccd_i is a distance between centroid and corner i. It is a character measuring dimension of the object dependent on its shape (Schirmer & Axhausen 2015).

Spatial distribution of a single building is captured by following three characters:

15. **Solar orientation of building** is denoted as

$$(15)\ Ori_{blg} = |o_{blgB} - 45|$$

where o_{blgB} is an orientation of the longest axis of bounding rectangle in a range 0 - 45. It captures the deviation of orientation from cardinal directions. There are multiple ways of capturing orientation of a polygon. As reported by Yan et al. (2007), Duchêne et al. (2003) assessed five different options (longest edge, weighted bisector, wall average, statistical weighting, bounding rectangle) and concluded bounding rectangle as the most appropriate. Deviation from cardinal directions is used to avoid sudden changes between square-like objects.

16. **Street alignment of building** is denoted as

$$(16)\ SAl_{blg} = |Ori_{blg} - Ori_{edg}|$$

where Ori_{blg} is a solar orientation of building and Ori_{edg} is a solar orientation of street edge. It reflects the relationship between building and its street, whether it is facing the street directly or indirectly (Schirmer & Axhausen 2015).

17. **Cell alignment of building** is denoted as

$$(17)\quad CAl_{blg} = |Ori_{blg} - Ori_{cell}|$$

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

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

18. **Longest axis length of tessellation cell** is denoted as

$$(18)\quad LAL_{cell} = d_{cellC}$$

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

19. **Area of tessellation cell** is denoted as

$$(19)\quad a_{cell}$$

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

20. **Circular compactness of tessellation cell** is denoted as

$$(20)\quad CCo_{cell} = \frac{a_{cell}}{a_{cellC}}$$

where a_{cellC} is area of minimal enclosing circle. It captures the relation of tessellation cell footprint shape to its minimal enclosing circle, illustrating the similarity of a shape and circle.

21. **Equivalent rectangular index of tessellation cell** is denoted as

$$(21)\quad ERI_{cell} = \sqrt{\frac{a_{cell}}{a_{cellB}}} * \frac{p_{cellB}}{p_{cell}}$$

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

22. **Solar orientation of tessellation cell** is denoted as

$$(22)\quad Ori_{cell} = |o_{cellB} - 45|$$

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

23. **Street alignment of building** is denoted as

$$(23)\quad SAl_{cell} = |Ori_{cell} - Ori_{edg}|$$

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

24. **Coverage area ratio of tessellation cell** is denoted as

$$(24)\quad CAR_{cell} = a_{blg}/a_{cell}$$

where a_{blg} is an area of building and a_{cell} is an area of related tessellation cell (Schirmer & Axhausen 2015). Coverage area ratio (CAR) is one of the commonly used characters capturing *intensity* of development. However, the definitions vary based on the spatial unit.

25. **Floor area ratio of tessellation cell** is denoted as

$$FAR_{cell} = fa_{blg}/a_{cell} \tag{25}$$

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

Following characters measure aspect related to street segment.

26. **Length of street segment** is denoted as

$$l_{edg} \tag{26}$$

and defined as a length of a `LineString` geometry in metres (Dibble et al. 2017; Gil et al. 2012).

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

$$w_{sp} = \frac{1}{n}\left(\sum_{i=1}^{n} w_i\right) \tag{27}$$

where w_i is width of a street section i. Algorithm generates street sections every 3 meters alongside the street segment and measures mean value. In case of open-ended street, 50 metres is used as a perception-based proximity limit (Araldi & Fusco 2019).

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

$$h_{sp} = \frac{1}{n}\left(\sum_{i=1}^{n} h_i\right) \tag{28}$$

where h_I is mean height of a street section i. Algorithm generates street sections every 3 meters alongside the street segment and measures mean value (Araldi & Fusco 2019).

29. **Height/width ratio of a street profile** is denoted as

$$(29)\quad HWR_{sp} = \frac{1}{n}\left(\sum_{i=1}^{n}\frac{h_i}{w_i}\right)$$

where h_I is mean height of a street section i and w_i is width of a street section i. Algorithm generates street sections every 3 meters alongside the street segment and measures mean value (Araldi & Fusco 2019).

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

$$(30)\quad Ope_{sp} = 1 - \frac{\sum hit}{2\sum sec}$$

where $\sum hit$ is sum of section lines (left and right sides separately) intersecting buildings and $\sum sec$ total number of street sections. Algorithm generates street sections every 3 meters alongside the street segment.

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

$$(31)\quad wd_{sp} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}\left(w_i - w_{sp}\right)^2}$$

where w_i is width of a street section i and w_{sp} is mean width. Algorithm generates street sections every 3 meters alongside the street segment.

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

$$(32)\quad hd_{sp} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}\left(h_i - h_{sp}\right)^2}$$

where h_i is height of a street section i and h_{sp} is mean height. Algorithm generates street sections every 3 meters alongside the street segment.

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

(33) $Lin_{edg} = \frac{l_{eucl}}{l_{edg}}$

where l_{eucl} is Euclidean distance between end points of a street segment and l_{edg} is a street segment length. It captures the deviation of a segment shape from a straight line. Adapted from Araldi & Fusco (2019).

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

(34) $a_{edg} = \sum_{i=1}^{n} a_{cell_i}$

where a_{cell_i} is area of tessellation cell i belonging to the street segment. It captures the area which is likely served by each segment.

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

(35) $BpM_{edg} = \frac{\sum blg}{l_{edg}}$

where $\sum blg$ is a number of buildings belonging to a street segment and l_{edg} is a length of a street segment. It reflects the granularity of development along each segment.

The last character measured on topological distance 0 is based on street node.

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

(36) $a_{node} = \sum_{i=1}^{n} a_{cell_i}$

where a_{cell_i} is area of tessellation cell i belonging to the street node. It captures the area which is likely served by each node.

Characters 37 and above take into account more than a single element, reflecting the relationship between them in a set context.

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

$$SWR_{blg} = \frac{p_{blg_{shared}}}{p_{blg}} \tag{37}$$

where $p_{blg_{shared}}$ is a length of a perimeter shared with adjacent buildings and p_{blg} is a perimeter of a buildings. It captures the amount of wall space facing the open space (Hamaina et al. 2012).

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

$$Ali_{blg} = \frac{1}{n} \sum_{i=1}^{n} |Ori_{blg} - Ori_{blg_i}| \tag{38}$$

where Ori_{blg} is the solar orientation of a building and Ori_{blg_i} is the solar orientation of building i on a neighbouring tessellation cell. It calculates the mean deviation of solar orientation of buildings on adjacent cells from a building. Adapted from Hijazi et al. (2016).

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

$$NDi_{blg} = \frac{1}{n} \sum_{i=1}^{n} d_{blg,blg_i} \tag{39}$$

where d_{blg,blg_i} is a distance between building and building i on a neighbouring tessellation cell. Adapted from Hijazi et al. (2016). It captures the average proximity to other buildings.

40. **Weighted neighbours of tessellation cell** is denoted as

$$WNe_{cell} = \frac{\sum cell_n}{p_{cell}} \tag{40}$$

where $\sum cell_n$ is a number of cell neighbours and p_{cell} is a perimeter of a cell. It reflects granularity of morphological tessellation.

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

(41) $a_{cell_n} = \sum_{i=1}^{n} a_{cell_i}$

where a_{cell_i} is area of tessellation cell i within topological distance 1. It captures the scale of morphological tessellation.

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

(42) $RC_{edg_n} = \sum_{i=1}^{n} cells_{edg_i}$

where $cells_{edg_i}$ is number of tessellation cells on segment i within topological distance 1. It captures accessible granularity.

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

(43) $a_{edg_n} = \sum_{i=1}^{n} a_{edg_i}$

where a_{edg_i} is an area covered by a street segment i within topological distance 1. It captures accessible area.

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

(44) $deg_{node_i} = \sum_{j} edg_{ij}$

where edg_{ij} is an edge of a street network between node i and node j. It reflects the basic degree centrality.

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

(45) $MDi_{node} = \frac{1}{n} \sum_{i=1}^{n} d_{node,node_i}$

where $d_{node,node_i}$ is a distance between node and node i within topological distance 1. It captures the average proximity to other nodes.

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

(46) $RC_{node_n} = \sum_{i=1}^{n} cells_{node_i}$

where $cells_{node_i}$ is number of tessellation cells on node i within topological distance 1. It captures accessible granularity.

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

(47) $a_{node_n} = \sum_{i=1}^{n} a_{node_i}$

where a_{node_i} is an area covered by a street node i within topological distance 1. It captures accessible area.

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

(48) $NCo_{blg_{adj}}$

where $NCo_{blg_{adj}}$ is a number of interior rings of a polygon composed of footprints of adjacent buildings (Schirmer & Axhausen 2015).

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

(49) $p_{blg_{adj}}$

where $p_{blg_{adj}}$ is a length of exterior ring of a polygon composed of footprints of adjacent buildings.

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

$$IBD_{blg} = \frac{1}{n} \sum_{i=1}^{n} d_{blg,blg_i} \tag{50}$$

where d_{blg,blg_i} is a distance between building and building i on a tessellation cell within topological distance 3. Adapted from Caruso et al. (2017). It captures the average proximity between buildings.

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

$$BuA_{blg} = \frac{\sum blg_{adj}}{\sum blg} \tag{51}$$

where $\sum blg_{adj}$ is a number of joined built-up structures within topological distance 3 and $\sum blg$ is a number of building within topological distance 3. Adapted from Vanderhaegen & Canters (2017).

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

$$GFAR_{cell} = \frac{\sum_{i=1}^{n} FAR_{cell_i}}{\sum_{i=1}^{n} a_{cell_i}} \tag{52}$$

where FAR_{cell_i} is a floor area ratio of tessellation cell i and a_{cell_i} is an area of tessellation cell i within topological distance 3. Based on Dibble et al. (2017).

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

$$WRB_{cell} = \frac{\sum blk}{\sum_{i=1}^{n} a_{cell_i}} \tag{53}$$

where $\sum blk$ is a number of blocks within topological distance 3 and a_{cell_i} is an area of tessellation cell i within topological distance 3.

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

(54) a_{blk}

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

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

(55) p_{blk}

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

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

(56) $CCo_{blk} = \frac{a_{blk}}{a_{blkC}}$

where a_{blkC} is area of minimal enclosing circle. It captures the relation of block footprint shape to its minimal enclosing circle, illustrating the similarity of a shape and circle.

57. **Equivalent rectangular index of a block** is denoted as

(57) $ERI_{blk} = \sqrt{\frac{a_{blk}}{a_{blkB}}} * \frac{p_{blkB}}{p_{blk}}$

where a_{blkB} is area of minimal rotated bounding rectangle of a block (MBR) footprint and p_{blkB} its perimeter of MBR.

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

(58) $CWA_{blk} = d_{blkC} \times \left(\frac{4}{\pi} - \frac{16(a_{blk})}{p_{blk}^2}\right)$

where d_{blkC} is a diameter of minimal circumscribed circle around the block polygon, a_{blk} is area of block and p_{blk} is a perimeter of a block. It is a proxy of permeability of an area.(Feliciotti 2018)

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

(59) $Ori_{blk} = |o_{blkB} - 45|$

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

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

(60) $wN_{blk} = \frac{\sum blk_n}{p_{blk}}$

where $\sum blk_n$ is a number of block neighbours and p_{blk} is a perimeter of a block. It reflects granularity of a mesh of blocks.

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

(61) $wC_{blk} = \frac{\sum cell}{a_{blk}}$

where $\sum cell$ is number of cells composing a block and a_{blk} is an area of a block. It captures granularity of each block.

62. **Meshedness of a street network** is denoted as

(62) $Mes_{node} = \frac{e-v+1}{2v-5}$

where e is a number of edges in a subgraph and v is the number of nodes in a subgraph (Feliciotti 2018). A subgraph is defined as a network within topological distance 5 around a node.

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

$$MSL_{edg} = \frac{1}{n} \sum_{i=1}^{n} l_{edg_i} \tag{63}$$

where l_{edg_i} is a length of a street segment i within a topological distance 3 around a segment.

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

$$CDL_{node} = \sum_{i=1}^{n} l_{edg_i}, \text{ if } edg_i \text{ is cul-de-sac} \tag{64}$$

where l_{edg_i} is a length of a street segment i within a topological distance 3 around a node.

65. **Reached cells by a street network segments** is denoted as

$$RC_{edg} = \sum_{i=1}^{n} cells_{edg_i} \tag{65}$$

where $cells_{edg_i}$ is number of tessellation cells on segment i within topological distance 3. It captures accessible granularity.

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

$$D_{node} = \frac{\sum node}{\sum_{i=1}^{n} l_{edg_i}} \tag{66}$$

where $\sum node$ is a number of nodes within a subgraph and l_{edg_i} is a lengths of a segment i within a subgraph. A subgraph is defined as a network within topological distance 5 around a node.

67. **Reached cells by a street network nodes** is denoted as

$$RC_{node} = \sum_{i=1}^{n} cells_{node_i} \tag{67}$$

where $cells_{node_i}$ is number of tessellation cells on node i within topological distance 3. It captures accessible granularity.

68. **Reached area by a street network nodes** is denoted as

$$a_{node_{net}} = \sum_{i=1}^{n} a_{node_i} \tag{68}$$

where a_{node_i} is an area covered by a street node i within topological distance 3. It captures accessible area.

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

$$pCD_{node} = \frac{\sum_{i=1}^{n} node_i, \text{ if } deg_{node_i}=1}{\sum_{i=1}^{n} node_i} \tag{69}$$

where $node_i$ is a node whiting topological distance 5 around a node. Adapted from (Boeing 2017).

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

$$p3W_{node} = \frac{\sum_{i=1}^{n} node_i, \text{ if } deg_{node_i}=3}{\sum_{i=1}^{n} node_i} \tag{70}$$

where $node_i$ is a node whiting topological distance 5 around a node. Adapted from (Boeing 2017).

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

$$p4W_{node} = \frac{\sum_{i=1}^{n} node_i, \text{ if } deg_{node_i}=4}{\sum_{i=1}^{n} node_i} \tag{71}$$

where $node_i$ is a node whiting topological distance 5 around a node. Adapted from (Boeing 2017).

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

$$(72)\quad wD_{node} = \frac{\sum_{i=1}^{n} deg_{node_i} - 1}{\sum_{i=1}^{n} l_{edg_i}}$$

where deg_{node_i} is a degree of a node i within a subgraph and l_{edg_i} is a length of a segment i within a subgraph. A subgraph is defined as a network within topological distance 5 around a node.

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

$$(73)\quad lCC_{node} = \frac{n-1}{\sum_{v=1}^{n-1} d(v,u)}$$

where $d(v,u)$ is the shortest-path distance between v and u, and n is the number of nodes within a subgraph. A subgraph is defined as a network within topological distance 5 around a node.

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

$$(74)\quad sCl_{node} = \frac{\sum_{u=1}^{k_v} \sum_{w=u+1}^{k_v} q_v(u,w)}{\sum_{u=1}^{k_v} \sum_{w=u+1}^{k_v} [a_v(u,w) + q_v(u,w)]}$$

where $q_v(u,w)$ are the number of common neighbours of u and w other than v (ie squares), and $a_v(u,w) = (k_u - (1 + q_v(u,w) + \theta_{uv}))(k_w - (1 + q_v(u,w) + \theta_{uw}))$, where $\theta_{uw} = 1$ if u and w are connected and 0 otherwise (Lind et al. 2005).

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.

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

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.

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.

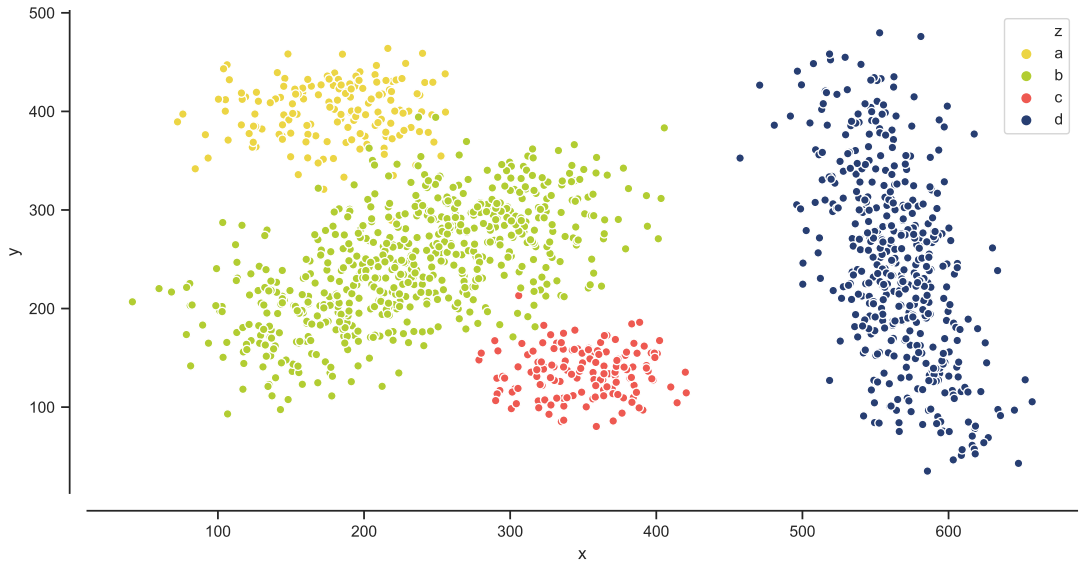

Figure 7.3: Artificial two dimensional (x, y) dataset containing 4 known cluster.

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.

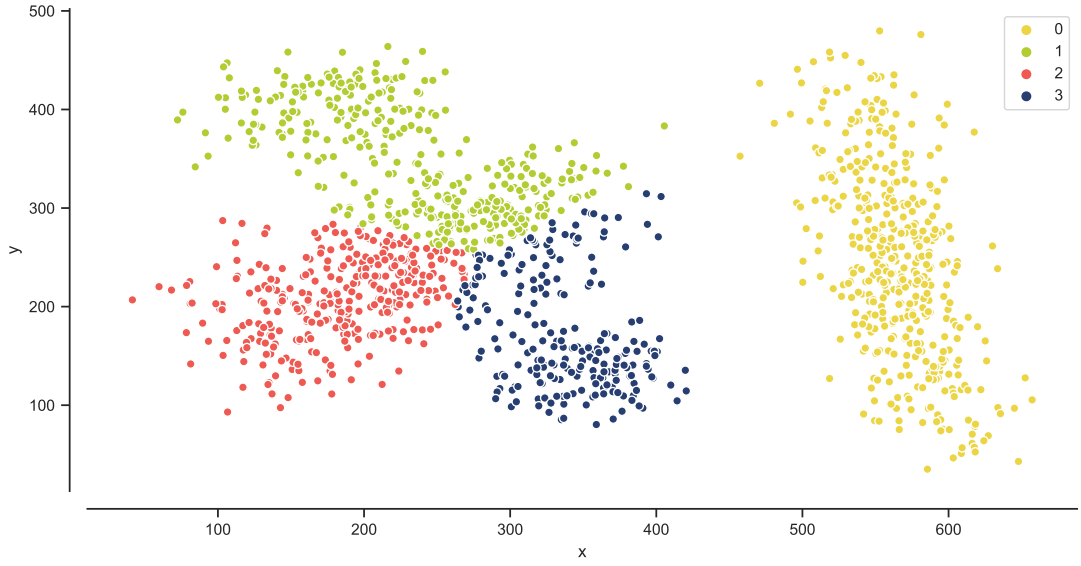

Figure 7.4: K-means clustering (k=4) of the artificial two dimensional (x, y) dataset containing 4 known cluster. Apart from the one cluster (0) which is clearly separated, none was correctly distinguished.

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.

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.

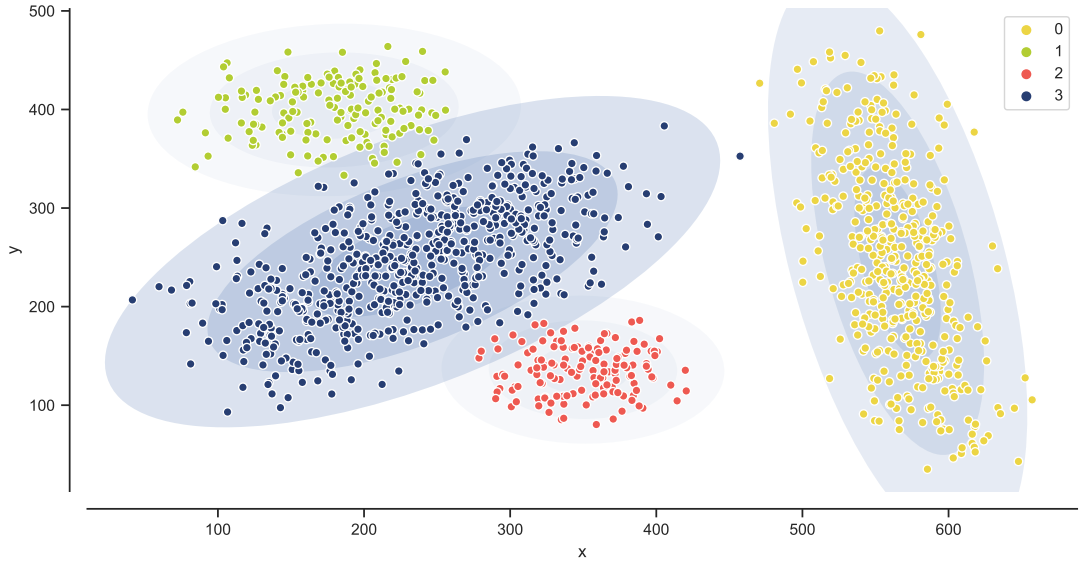

Figure 7.5: GMM clustering (4 components) of the artificial two dimensional (x, y) dataset containing 4 known cluster. All clusters are fairly successfully distinguished. Figure also shows underlying Gaussian distributions as ellipses reflecting the probability by change of the shade.

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.

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.

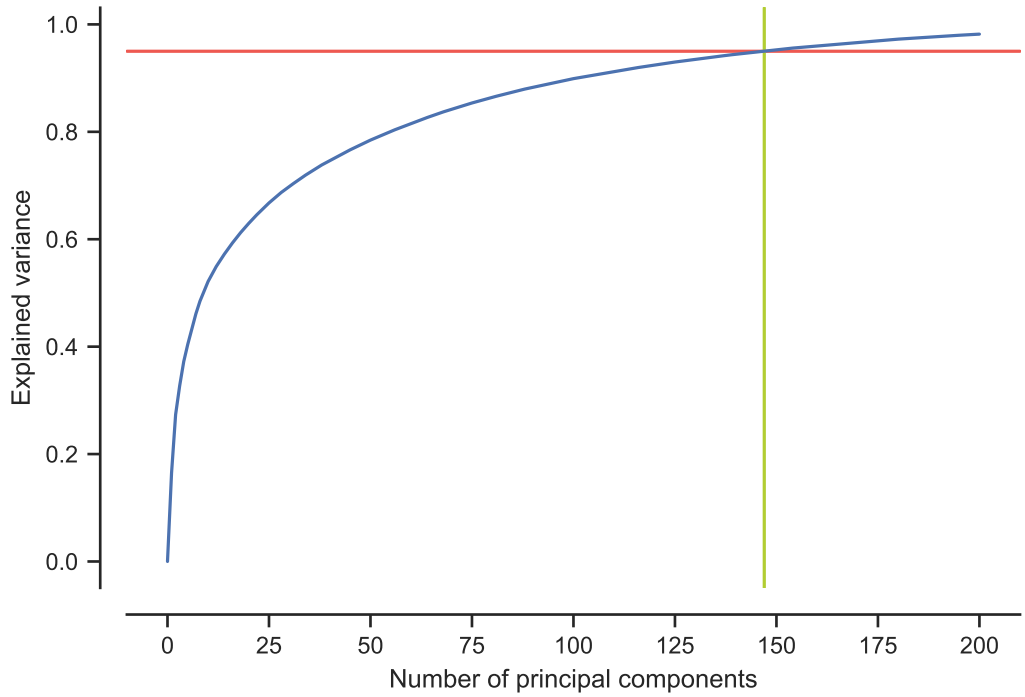

Figure 7.6: Principal Component Analysis results on the contextual characters (n=296) on Prague data. Red line marks 0.95 explained variance, green line denotes 147 principal components as a first value reaching 0.95.

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.

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			
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.

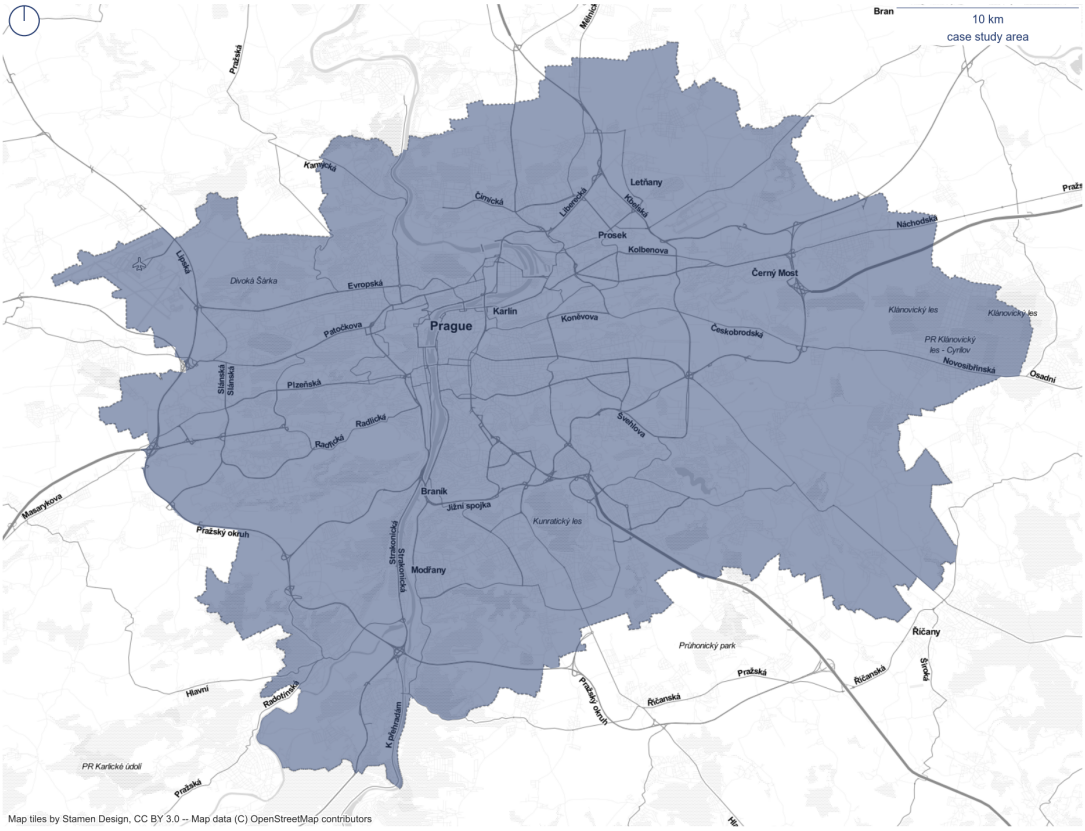

Figure 7.7: Prague case study area, which matches the administrative boundaries. Data source © IPR Praha, CC BY-SA 4.0

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

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 according to the data model. All morphometric characters are measured using `momepy` classes using reproducible Jupyter notebook `XX_XXXXX` presented as an Appendix XXX.

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

7.3.1.1 Spatial distribution

The spatial distribution of resulting values, i.e., spatial morphometric patterns, could be projected on maps and assessed visually to determine the character of such a pattern or statistically. Since the aim of measuring is, eventually, to identify homogenous areas defined by distinct patterns of spatial configuration, it is crucial that each of the characters capture local patterns. Statistically speaking, each of the characters needs to be spatially autocorrelated, which can be assessed using Moran's I (REF)[4].

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

[4]The same method has been used during the selection of primary characters to ensure that all do capture spatial patterns. See Annex 2 for details.

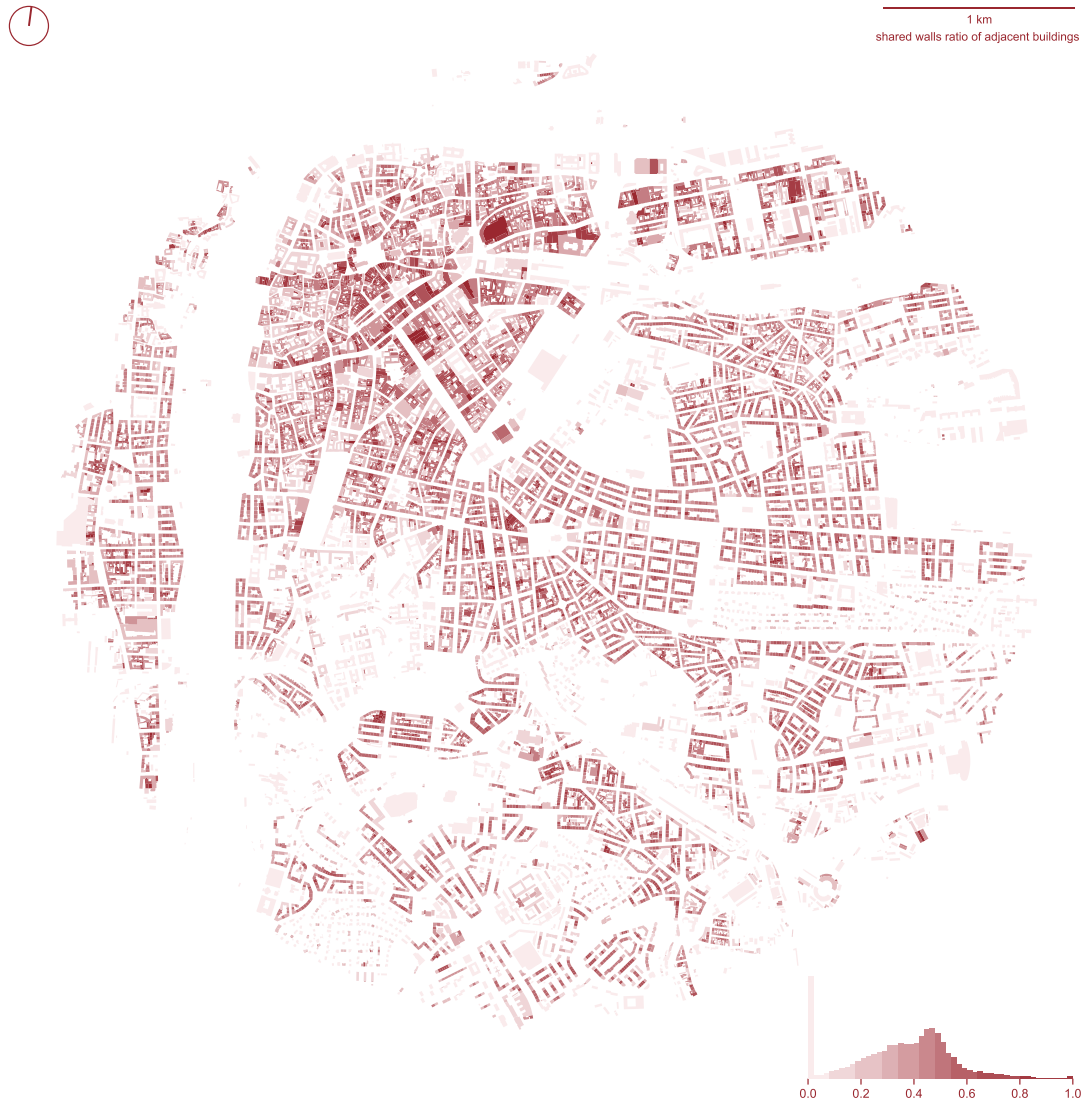

Figure 7.8: Spatial distribution of shared walls ratio of adjacent buildings in the area of Prague's city centre and its surroundings. Figure illustrates clear spatial patterns with the presence of sudden changes.

Figure 7.8 shows *shared walls ratio of adjacent buildings* in the part of Prague's city centre. There is a clear distinction between buildings having shared walls and those standing independently. The values show a relative homogeneity in the centre of the figure (Vinohrady), but high variability in some other places, especially in the Old Prague neighbourhood (top left). There are sudden changes in values on neighbouring tessellation cells. This pattern is not unique and it is

somewhat expected for characters based on individual elements as these do not have a notion of contiguity.

Figure 7.9: Spatial distribution of proportion of 4-way intersections of street network in the area of Prague's city centre and its surroundings. Figure illustrates clear continuous spatial patterns with unclear boundaries between low and high values.

Second example on figure 7.9 shows *proportion of 4-way intersections of street network* within the same area. This is purely network-based character measuring properties of subgraphs around each network node (i.e. a junction). Subgraphs, by definition, overlap causing the smooth transition of values across the study

area. It is relatively simple to describe resulting patterns visually, with high values in more grid-like areas (Vinohrady - centre, Smíchov - left). However, definition of boundaries between high and low values would be relatively complicated procedure, due to the inherent spatial smoothing. Characters based on a larger topological context tend all to have continuous patterns like this.

Figure 7.10: Spatial distribution of equivalent rectangular index of tessellation cell in the area of Prague's city centre and its surroundings. Figure illustrates visually unclear spatial patterns.

The last, not very frequent though, is the example on figure 7.10 showing visu-

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

Figure 7.11 below shows the value of Moran's I compared to reference distribution and a Moran scatterplot based on the contiguity of morphological tessellation. The I value for the whole Prague is 0.07, showing significant autocorrelation. It is not a high value, for a reference two previous example have I 0.387 and 0.912 respectively. However, it is still significant (the value itself is likely not within a reference distribution), meaning that even not visually clear, spatial pattern is still present. The whole set of character contains a couple of other examples similar to this one, but overall this situation is not a frequent one.

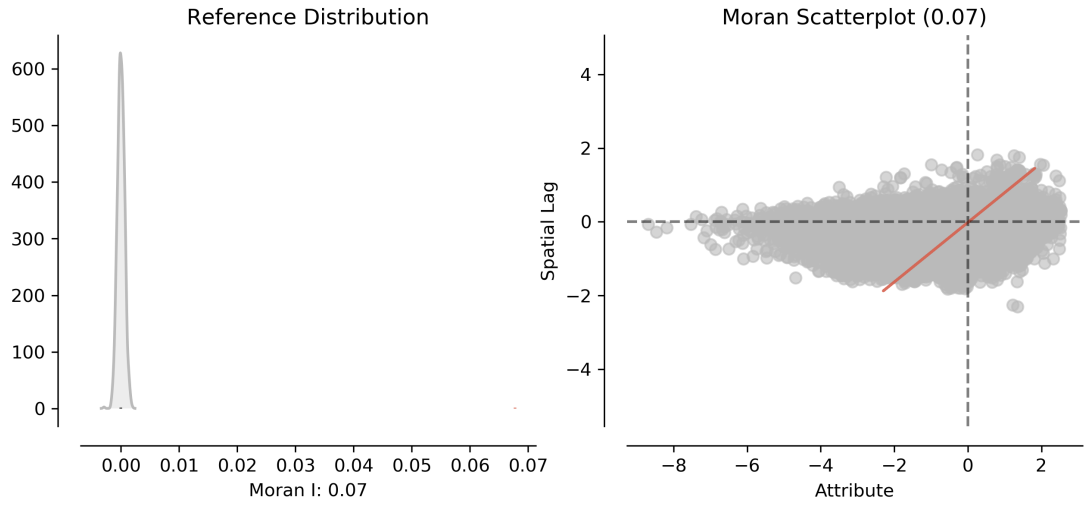

Figure 7.11: Reference distribution in relation to actual Moran's I value and Moran scatterplot of equivalent rectangular index of tessellation cell based on the whole Prague. The results indicate significant, however weak spatial autocorrelation.

Due to the large variety of characters attempting to capture both structural

complexity and cross-scale complexity within a single set, the spatial distribution of resulting values may vary. However, all show significant spatial patterns.

7.3.1.2 Statistical distribution

Statistical distributions of resulting values are also different, based on the nature of each character. From the literature is known, that urban context is often described by exponential distributions like a power law (REF Salat), but that is far from being a rule for selected set of morphometric characters. Figure 7.12 shows four examples of distributions as captured in Prague.

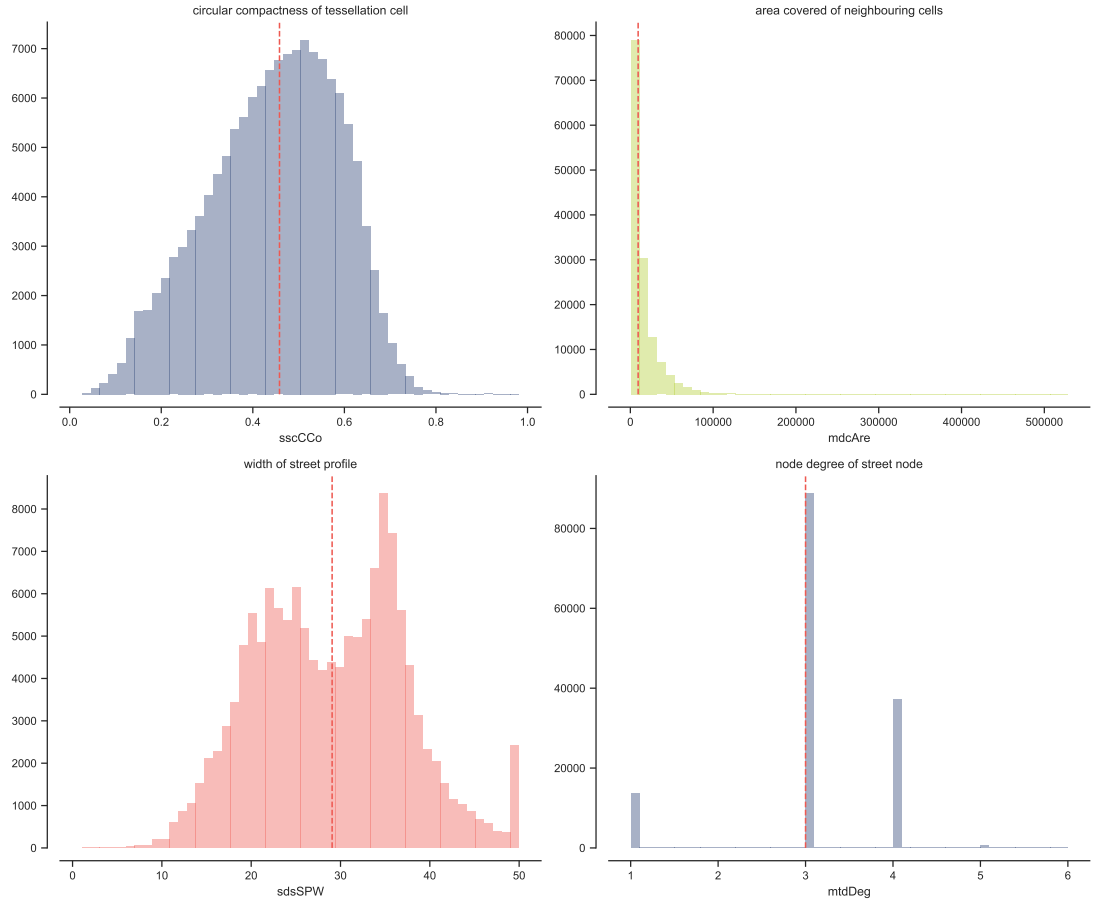

Figure 7.12: Histogram of four types of statistical distributions. Circular compactness of tessellation cell (top left), area covered by neighbouring cells (top right), width of a street profile (bottom left), and degree of a street node (bottom right).

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

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

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

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

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

Descriptive summary values of all character are presented in the table 7.2. Note that units for each character are available in section 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

id	mean	std	min	25%	50%	75%	max
ssbVFR	3	1.7	0.43	2.1	2.6	3.5	67
ssbCCo	0.53	0.11	0.026	0.47	0.56	0.61	1
ssbCor	8.8	7.4	0	4	8	10	390
ssbSqu	5.3	9.1	9.5e-09	0.48	1.1	5	85
ssbERI	0.94	0.086	0.25	0.91	0.96	1	1.1
ssbElo	0.71	0.2	0.026	0.56	0.74	0.87	1
ssbCCD	1.5	2.2	0	0.068	1	1.9	88
ssbCCM	9.4	6.6	3	6.3	7.6	10	210
stbOri	16	13	0	6.2	13	25	45
stbSAl	6.7	8.9	4.9e-10	0.61	2.5	9.5	45
stbCeA	6.9	9	8.9e-12	0.48	3	9.9	45
sdcLAL	67	42	7.9	40	52	79	970
sdcAre	2100	4100	31	540	940	1900	350000
sscCCo	0.45	0.14	0.027	0.35	0.46	0.55	0.98
sscERI	0.97	0.062	0.43	0.94	0.98	1	1.1
stcOri	18	13	0	7.1	16	29	45
stcSAl	9.2	9.7	1.9e-05	1.5	5.6	14	45
sicCAR	0.2	0.15	0.00092	0.092	0.16	0.26	1
sicFAR	0.67	0.92	0.00092	0.14	0.32	0.74	17
sdsLen	230	260	0.047	110	160	260	3300
sdsSPW	29	8.4	1	22	29	35	50
sdsSPH	10	6.1	0	6.4	8	13	57
sdsSPR	0.41	0.32	0	0.21	0.3	0.49	23
sdsSPO	0.58	0.21	0	0.44	0.58	0.71	1
sdsSWD	3.6	2.1	0	1.9	3.7	5.1	12
sdsSHD	2.3	2.3	0	0.94	1.5	2.7	24
sssLin	0.95	0.13	0	0.97	1	1	1
sdsAre	31000	56000	34	6900	13000	30000	740000
sisBpM	0.075	0.079	0.00056	0.046	0.068	0.095	21
sddAre	30000	46000	86	9400	16000	31000	660000
mtbSWR	0.18	0.2	0	0	0.15	0.32	1
mtbAli	4.8	5.1	1.4e-09	0.9	3	7	44

id	mean	std	min	25%	50%	75%	max
mtbNDi	25	18	0	13	20	30	200
mtcWNe	0.046	0.022	0.0012	0.03	0.045	0.059	0.26
mdcAre	16000	19000	390	5500	9400	19000	530000
misRea	44	25	1	27	40	55	290
mdsAre	86000	110000	770	33000	53000	94000	1.3e+06
mtdDeg	3.1	0.82	1	3	3	4	6
mtdMDi	170	150	0.047	99	130	190	3300
midRea	52	28	1	33	49	67	270
midAre	97000	110000	770	42000	65000	110000	1.3e+06
libNCo	0.6	3.3	0	0	0	0	58
ldbPWL	180	250	20	51	82	200	3400
ltbIBD	27	11	0	20	25	33	120
ltcBuA	0.65	0.24	0.043	0.49	0.7	0.84	1
licGDe	0.57	0.67	0.0022	0.18	0.35	0.66	5
ltcWRB	9e-05	6.7e-05	1.7e-06	3.9e-05	7.3e-05	0.00012	0.00072
ldkAre	120000	240000	710	15000	31000	110000	2e+06
ldkPer	1500	1800	100	550	830	1700	13000
lskCCo	0.43	0.13	0.11	0.33	0.44	0.53	0.98
lskERI	0.86	0.13	0.35	0.79	0.9	0.96	1.1
lskCWA	360	470	0.43	87	170	430	3100
ltkOri	18	13	0.00098	7	15	28	45
ltkWNB	0.0074	0.0043	0	0.004	0.0066	0.01	0.04
likWBB	0.00089	0.00066	8.3e-06	0.00037	0.00074	0.0013	0.006
lcdMes	0.15	0.06	-0.33	0.11	0.15	0.19	0.34
ldsMSL	150	76	45	110	130	170	1600
ldsCDL	280	390	0	13	160	380	4200
ldsRea	350000	310000	770	190000	260000	400000	4.2e+06
lddNDe	0.013	0.0055	0	0.0095	0.012	0.014	0.13
lddRea	190	86	1	130	190	240	680
lddARe	370000	310000	770	200000	280000	420000	4.2e+06
linPDE	0.13	0.087	0	0.067	0.11	0.17	1
linP3W	0.64	0.11	0	0.57	0.64	0.71	0.97

id	mean	std	min	25%	50%	75%	max
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

Without exploring the table 7.2 above in detail, it is worth pointing out two characters standing out - *courtyard area of building (sdbCoA)* and *number of courtyards of adjacent buildings (libNCo)*. Both are capturing similar concepts of closed courtyards (either in a single building or in a composite of adjacent buildings) and both are relatively invariant (min, 25%, 50% and 75% are all 0). While these might not be critical for identification of DHC in Prague, there are urban tissues, especially in warmer environments, characterised by these properties. This example illustrates the inclusiveness of a selected set of characters - it is not tailored to specific context.

Figures 7.13 - 7.17 show histograms capturing the (truncated) distribution of all measured characters. Not the differences outlined above and overall variety of distributions.

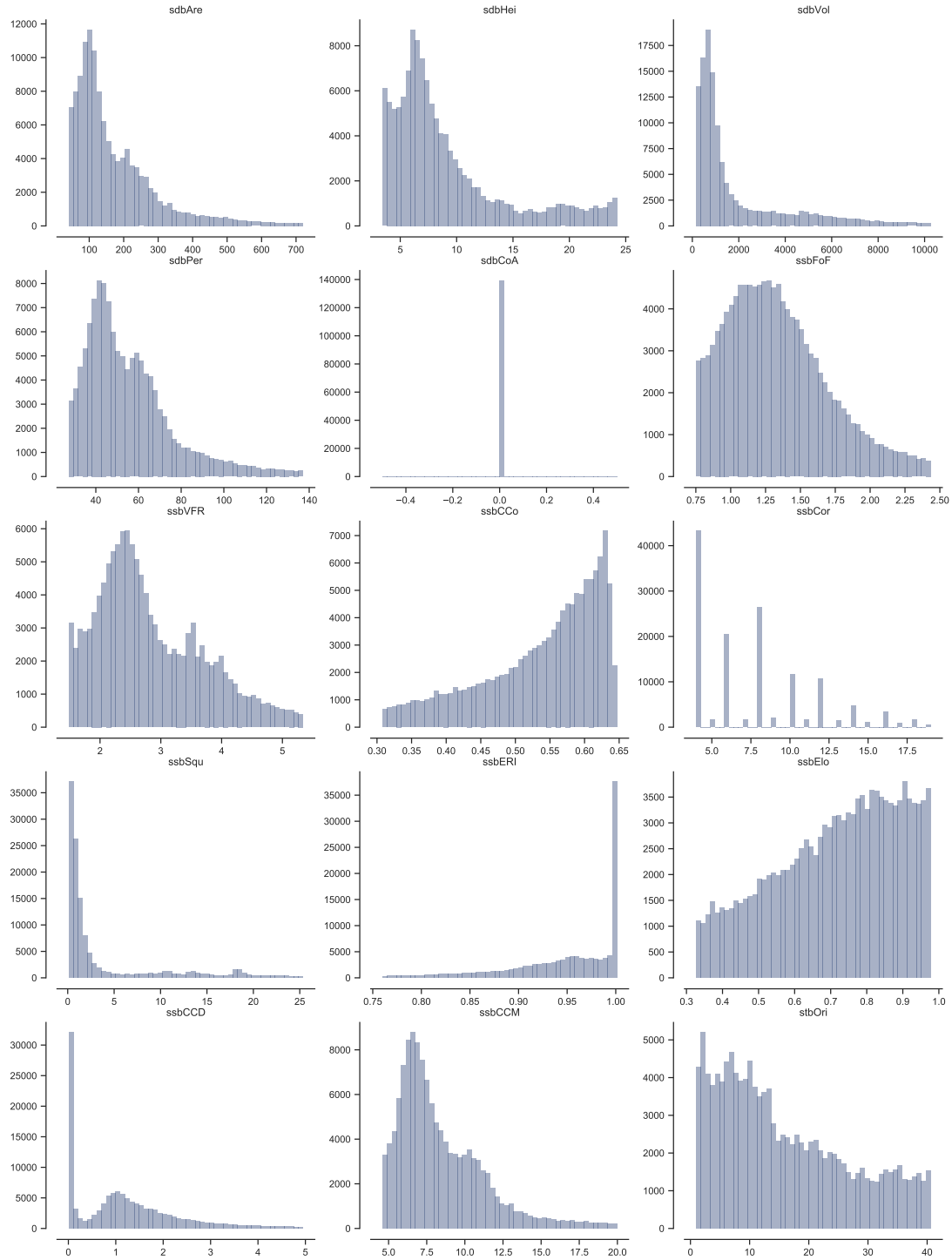

Figure 7.13: Histograms of characters 1-15 are showing the variety of distributions within the measured primary data. Histograms illustrate data within percentiles (5, 95) to avoid extreme skewing due to the presence of outliers. Data in table are presented complete for reference.

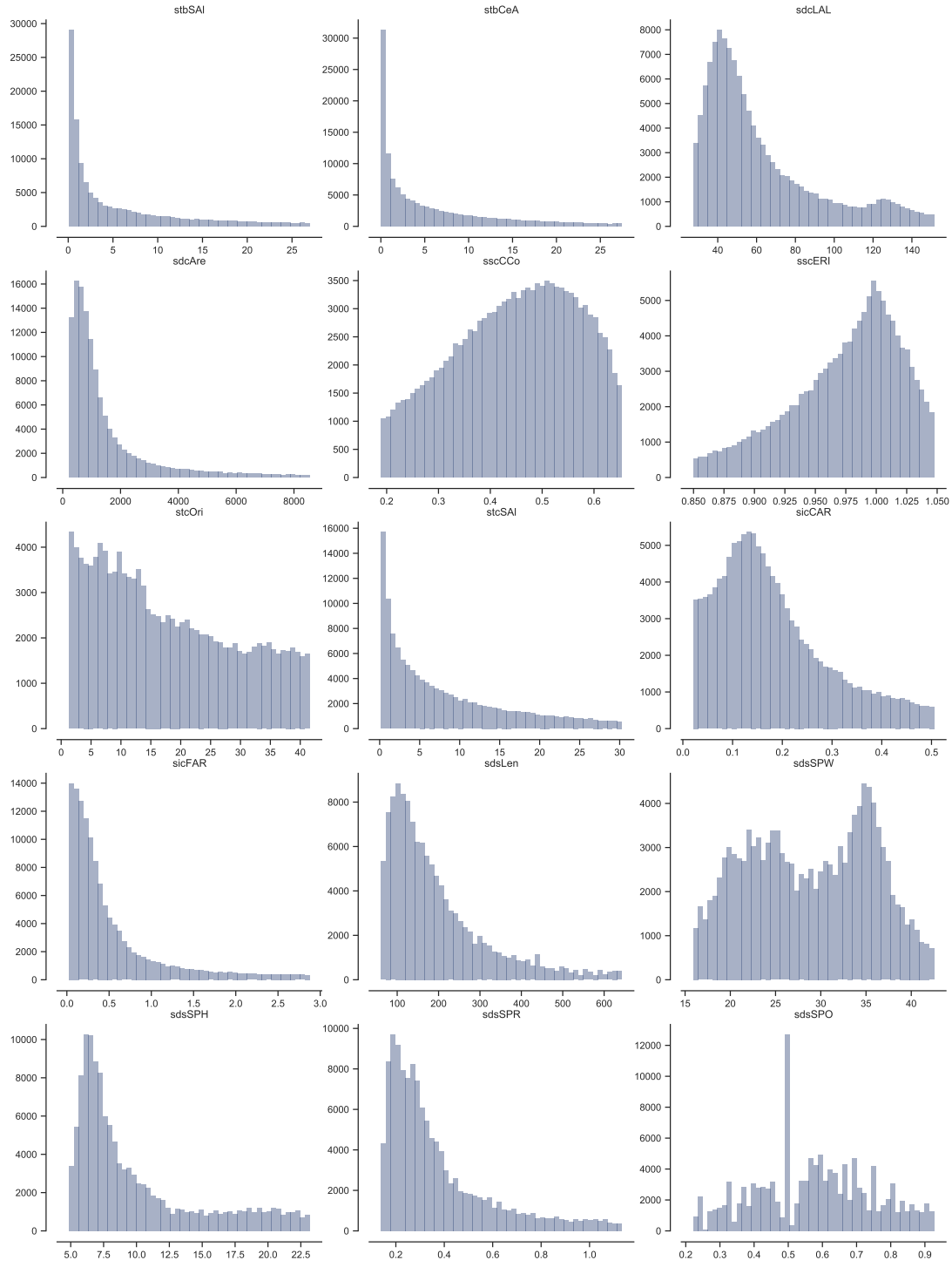

Figure 7.14: Histograms of characters 16-30 are showing the variety of distributions within the measured primary data. Histograms illustrate data within percentiles (5, 95) to avoid extreme skewing due to the presence of outliers. Data in table are presented complete for reference.

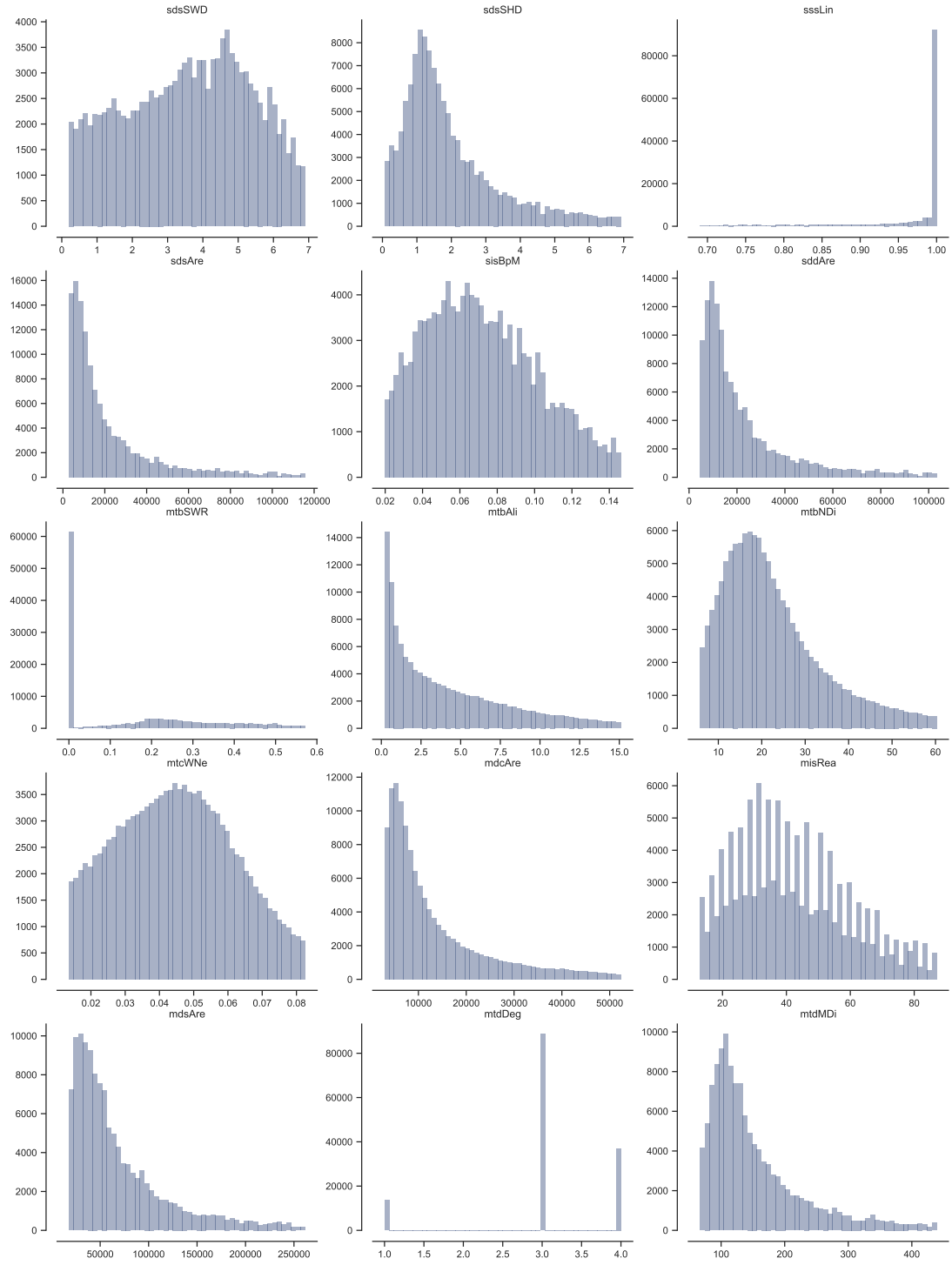

Figure 7.15: Histograms of characters 31-45 are showing the variety of distributions within the measured primary data. Histograms illustrate data within percentiles (5, 95) to avoid extreme skewing due to the presence of outliers. Data in table are presented complete for reference.

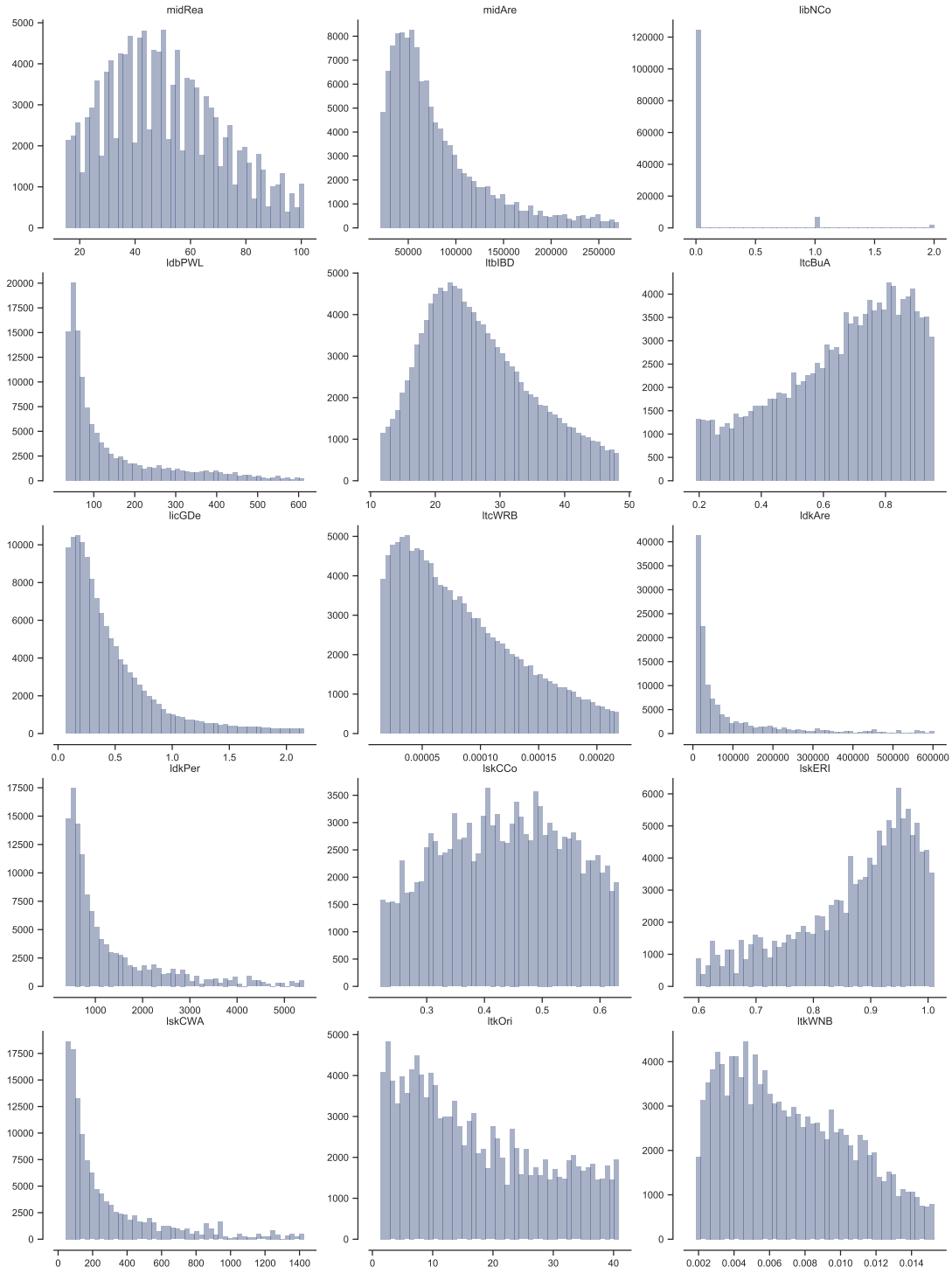

Figure 7.16: Histograms of characters 45-60 are showing the variety of distributions within the measured primary data. Histograms illustrate data within percentiles (5, 95) to avoid extreme skewing due to the presence of outliers. Data in table are presented complete for reference.

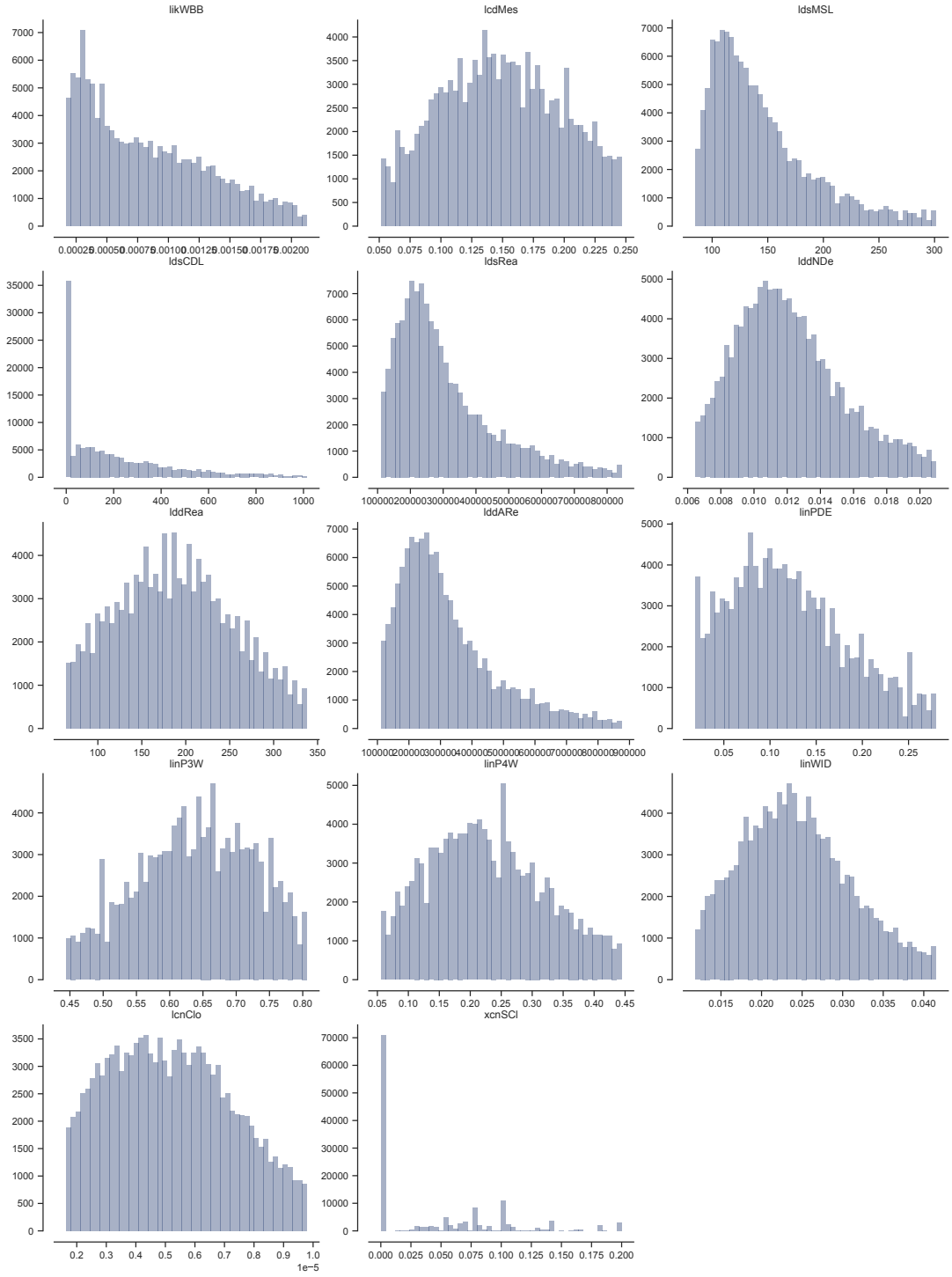

Figure 7.17: Histograms of characters 61-74 are showing the variety of distributions within the measured primary data. Histograms illustrate data within percentiles (5, 95) to avoid extreme skewing due to the presence of outliers. Data in table are presented complete for reference.

7.3.1.3 Statistical relationship of characters

Understanding the relationship between measured character is in important aspect of morphometric assessment. As specified in the section XXX (selection of chars), characters should not include many empirical correlations. Collinear characters (those being correlated and reflecting the same concept) should not be present in the resulting data as they might skew the hyperspace and adversely affect the result of clustering. As characters do not follow always follow normal distribution, Spearman's rank correlation is used to assess the relationship between characters. The results are illustrated in the correlation matrix on figure 7.18 below.

Figure 7.18: Correlation matrix of Spearman's rho values capturing statistical relationship between resulting morphometric values of primary characters. With a few exceptions, the relationship is none or very weak.

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

Primary morphometric characters are the core of the method of identification of distinct homogenous clusters. Their selection to capture different complexities aspects of urban form results in a very heterogenous set of measured data showing variable spatial patterns as well as statistical distributions. However, this data are the direct input of the clustering procedure, but they are an input of calculation

of contextual characters. The results of this consequent step are illustrated in the next section.

7.3.2 Contextual characters

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

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

7.3.2.1 Spatial distribution

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

Figure 7.19: Spatial distribution of interquartile mean of a width of a street profile measured within 3 topological steps on morphological tessellation in the area of Prague's city centre and its surroundings.

Figure 7.19 shows *interquartile mean of a width of a street profile* measured within 3 topological steps on morphological tessellation. As a version of truncated mean, this character directly reflect the actual values of primary characters and it is relatively simple to indicate areas with generally narrow streets (historical core) and those with wider profile (heterogenous areas on south and south-east). The overall distribution of values within shown area is very symmetrical with a peak

at 22 metres, which is a common street width in Prague.

Figure 7.20: Spatial distribution of interquartile range of a width of a street profile measured within 3 topological steps on morphological tessellation in the area of Prague's city centre and its surroundings.

Figure 7.20 illustrates *interquartile range of a width of a street profile* measured within 3 topological steps on morphological tessellation. That reflects range of values, so it is a first proxy of diversity. In the example above it does divide places with either major street or generally wider streets from predominantly homogenous areas. The distribution of values is balanced, but truncated at 0

(range could not be negative). We can identify certain similarity with the patterns on previous figure 7.19, because wider street (i.e., higher interquartile mean) causes bigger range, but the pattern are not identical.

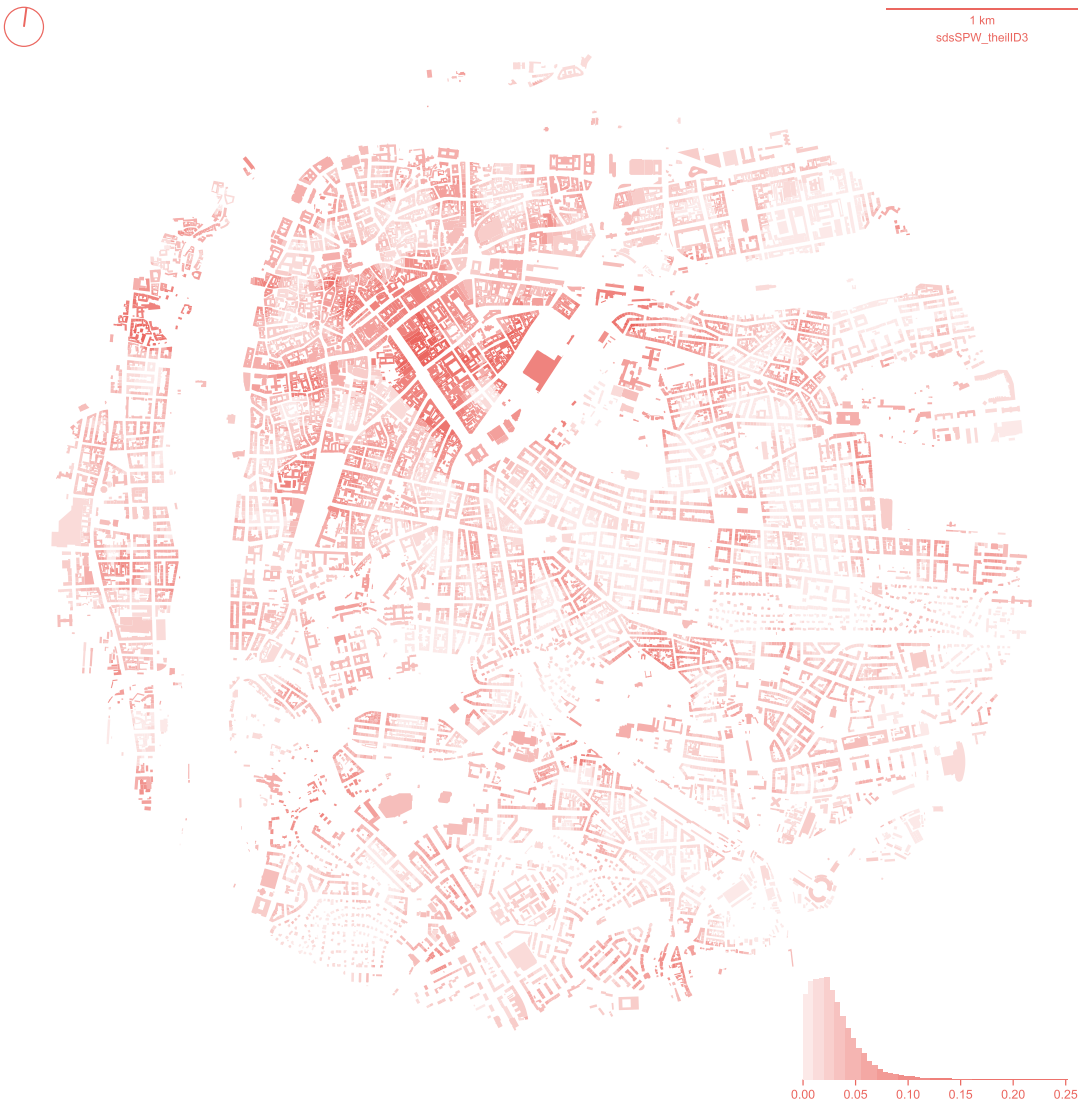

Figure 7.21: Spatial distribution of interdecile Theil index of a width of a street profile measured within 3 topological steps on morphological tessellation in the area of Prague's city centre and its surroundings.

Figure 7.21 illustrates *interdecile Theil index of a width of a street profile* measured within 3 topological steps on morphological tessellation. It is another character capturing diversity, this time based on inequality of distribution. The resulting

map then shows as the most *unequal* area around Vaclavske sq. (darker red) where one street (in this case elongated square) is significantly different from the other. Previously highlight areas of wider streets are not so from the perspective of Theil index. The distribution has a long tail, somewhat typical for Theil index applied to morphometric characters.

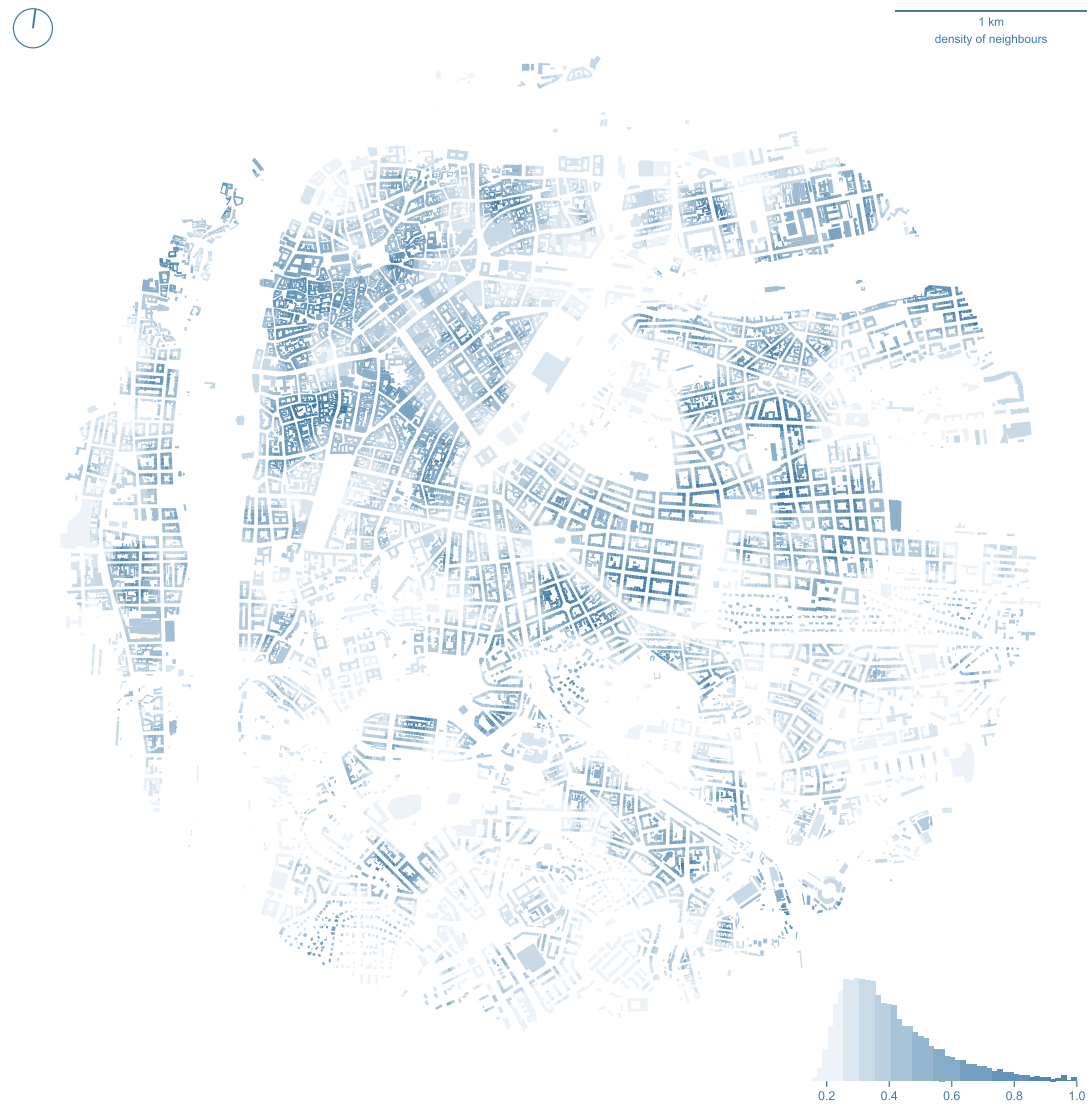

Figure 7.22: Spatial distribution of interdecile Simpson index of a width of a street profile measured within 3 topological steps on morphological tessellation in the area of Prague's city centre and its surroundings.

The last contextual character, shown on figure 7.22, is *Simpson diversity index of*

a width of a street profile measured within 3 topological steps on morphological tessellation. The values in this case are binned using Natural Breaks (k=7). It captures (inversely) similar information as Theil index, but that is not the rule. These two are, indeed, related as both capture diversity of values, but the relationship between them is not fixed as will be illustrated in the next section.

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

7.3.2.2 Statistical distribution

Figure 7.12 in previous section showed four types of distribution of primary characters. This section illustrates how each of them translates into distribution of contextual characters.

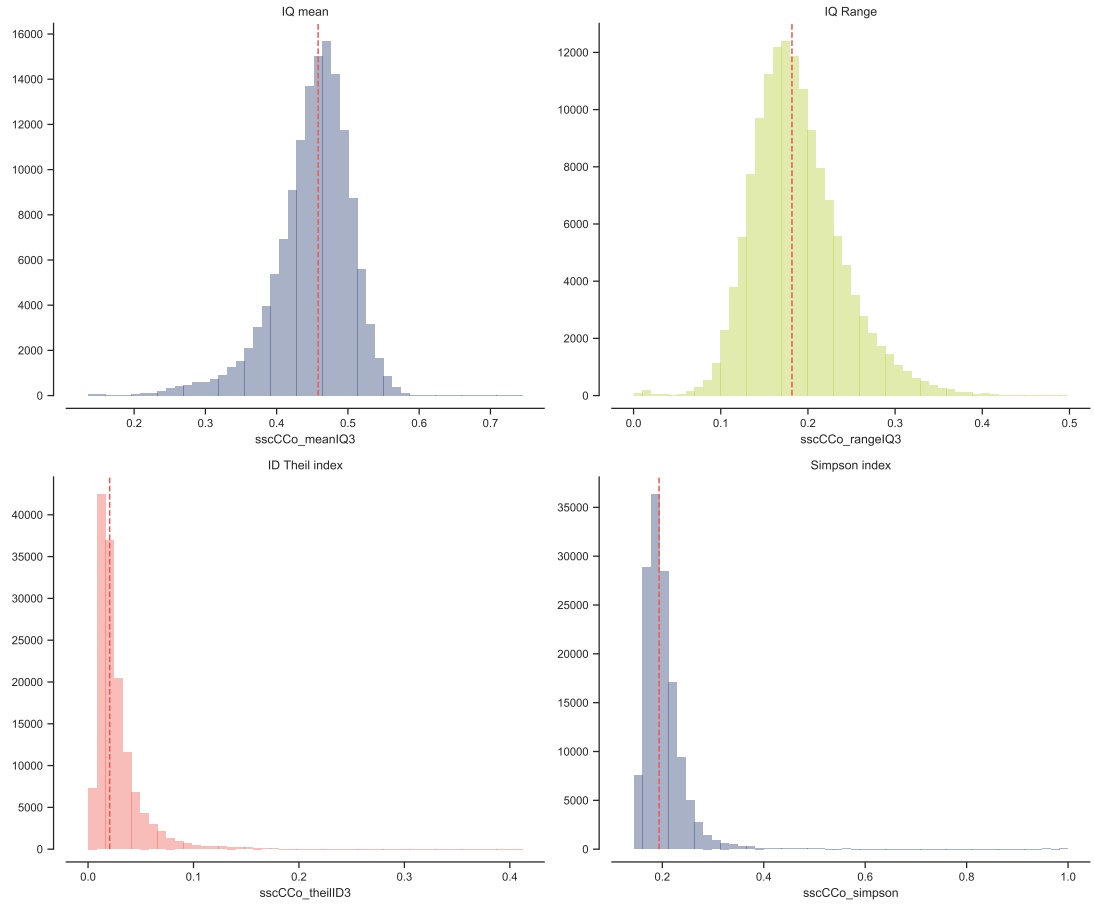

Figure 7.23: Histograms of statistical distribution of contextual versions of circular compactness of tessellation cell. Interquartile mean (top left), interquartile range (top right) interdecile Theil index (bottom left), Simpson index (bottom right).

The first example, *circular compactness of tessellation cell* on figure 7.23, was originally mildly skewed Gaussian-like distribution. In terms of IQ mean and IQ range, this property remains the same, both are relatively symmetrical distributions with small tail on one or the other side. On the other hand, the distribution of Theil index and Simpson diversity resembles exponential curve due to heavy tail in both.

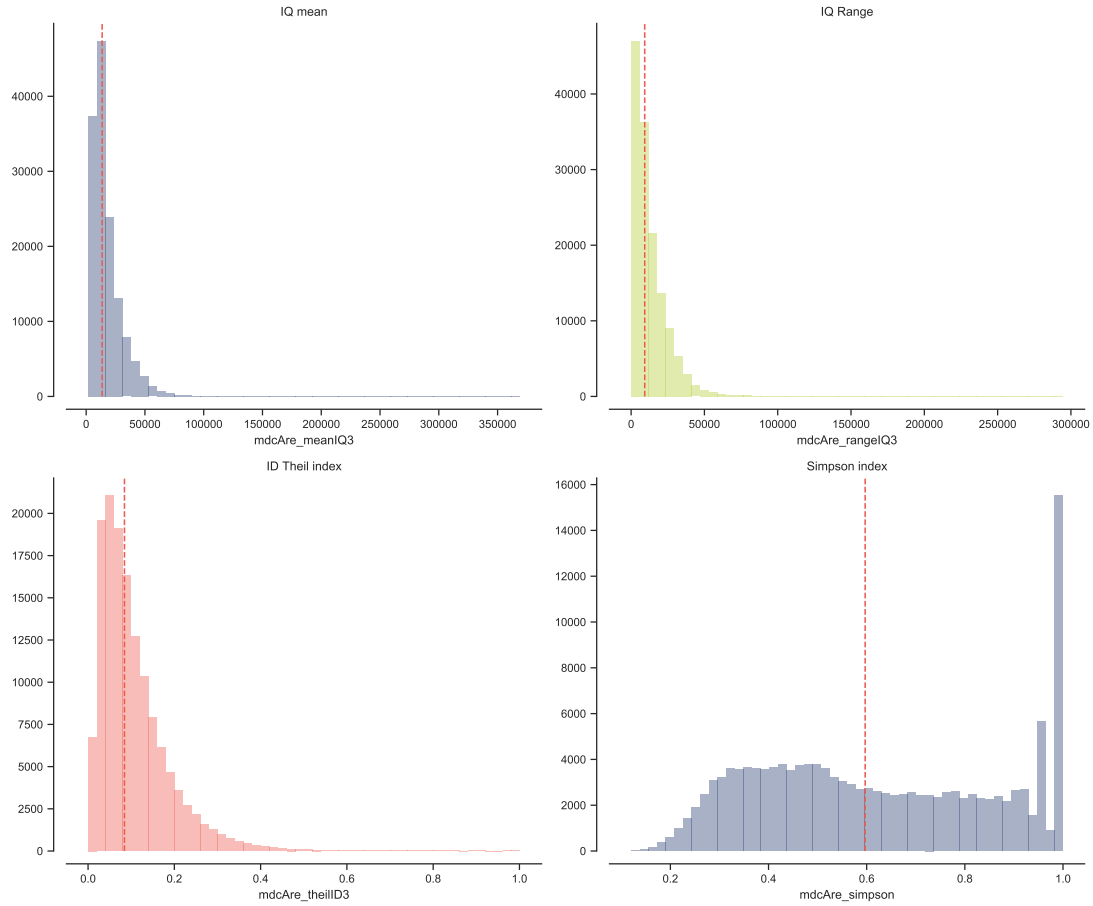

Figure 7.24: Histograms of statistical distribution of contextual versions of area covered by neighbouring cells. Interquartile mean (top left), interquartile range (top right) interdecile Theil index (bottom left), Simpson index (bottom right).

Initially exponential distribution of *area covered by neighbouring cells* remains exponential in both IQ mean and IQ range cases (figure 7.24). Theil index is also exponential, although the curve is not so unequal. Simpson index is significantly different from all three. The HeatTail binning used within the calculation is tailored to exponential distributions and resulting Simpson diversity is then relatively balanced across the values. The values 1 showing a big spike mean that the probability that any other value is within the same bin is 100%, hence no diversity in the area. It is typical for fairly homogenous compact urban tissues.

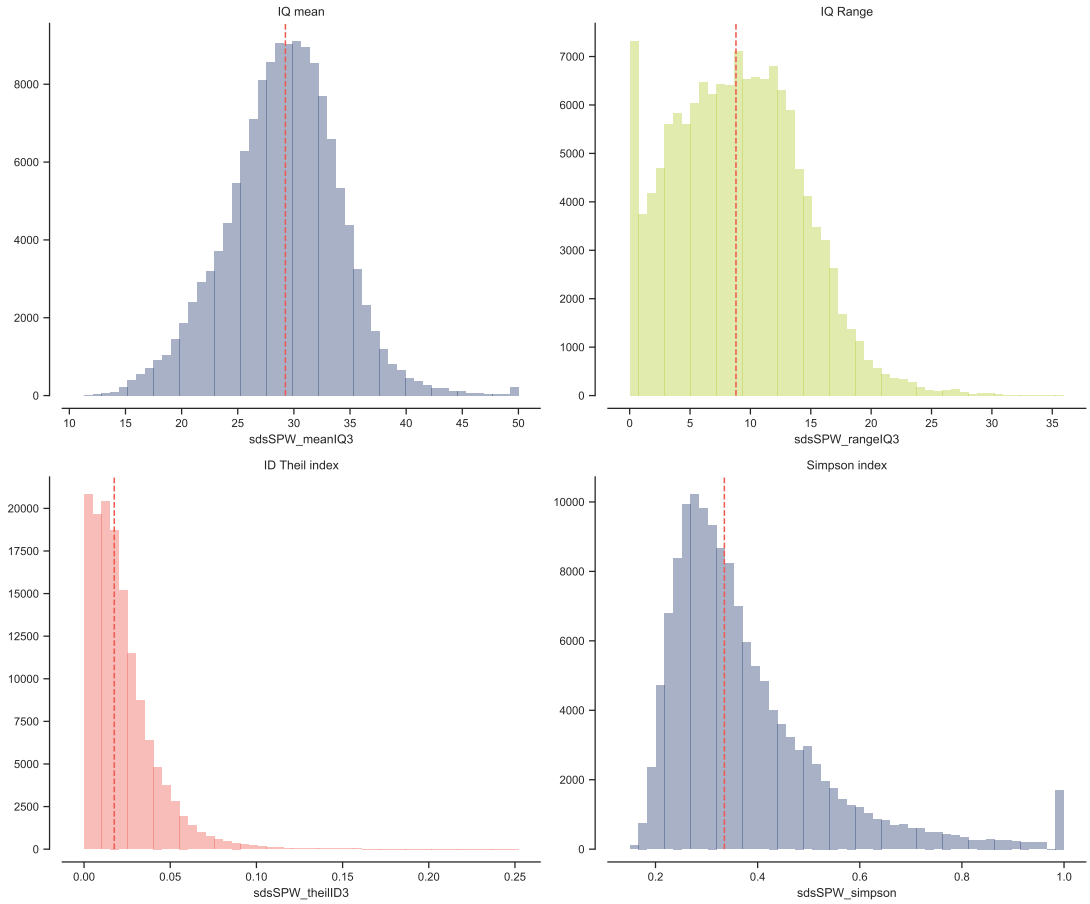

Figure 7.25: Histograms of statistical distribution of contextual versions of width of a street profile. Interquartile mean (top left), interquartile range (top right) interdecile Theil index (bottom left), Simpson index (bottom right).

The third example, *width of a street profile* which was illustrated on maps on previous pages, had initially specific distribution affected by rules on which streets are designed (there were spikes for narrower and wider streets). Figure 7.25 shows that none of the contextual character share this profile and, more importantly, all haver different distributions. IQ mean is almost ideal Gaussian distribution, IQ range is right-skewed and truncated with a spike on 0, Theil index is again exponential and Simpson diversity is right-skewed, but relatively symmetrical distribution. Even though figures 7.19 - 7.22 may seem similar, the difference in distributions on figure 7.25 indicates otherwise. What is important are numerical values, not visual perception.

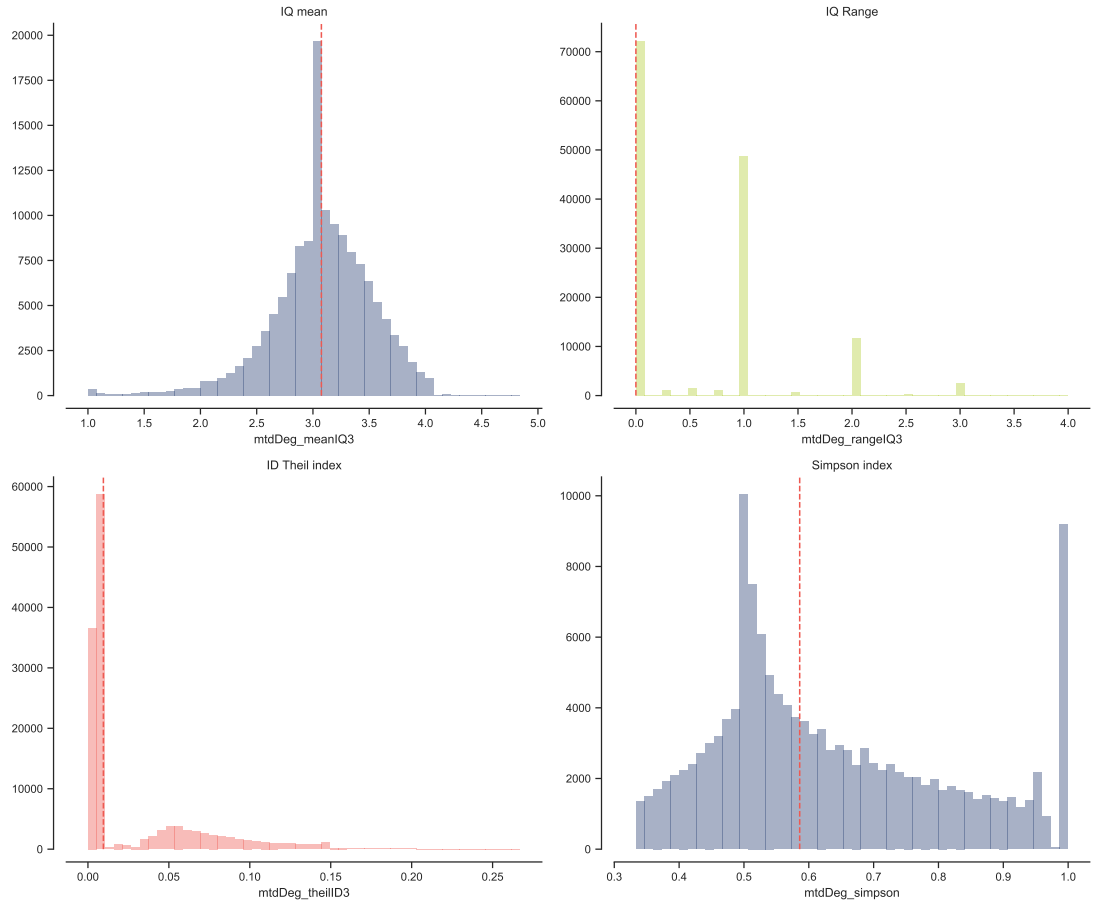

Figure 7.26: Histograms of statistical distribution of contextual versions of degree of a street node. Interquartile mean (top left), interquartile range (top right) interdecile Theil index (bottom left), Simpson index (bottom right).

Initially restricted values of *degree of a street node* remained present in IQ range (figure 7.26 but not in the other contextual characters. IQ mean is symmetrical with a large spike on 3, which is almost its median value. Theil index is very different from previous examples, and does not follow exponential distribution this time, while Simpson diversity index has two spikes on 0.5 and 1.0 and relative balance otherwise.

As the examples above indicates, the variety present in primary characters remained present, in a different way, in contextual characters as well.

7.3.2.3 Statistical relationship of characters

The statistical relationship between contextual characters will directly influence results of clustering in next steps. For that reason, we should aim for minimisation of such relationship in terms of Spearman's rank correlation. As illustrated above, we may expect some relations, however only selective, affected by the nature of primary characters. Below (figure 7.27) is a correlation matrix of contextual characters for illustration of the measured relationship[5].

[5]Due to the large number of characters, matrix is not optimal for presentation in this form. Its high quality version is available in Appendix.

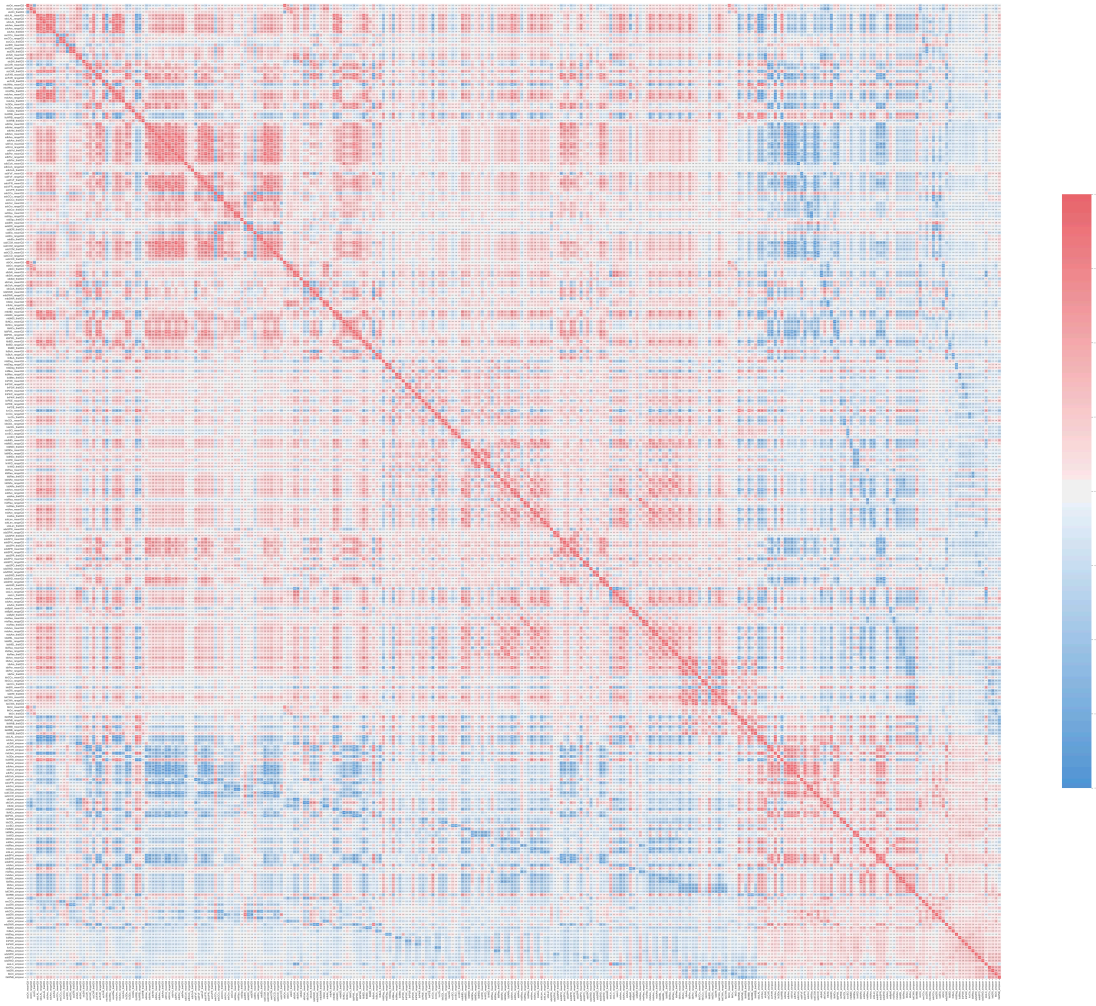

Figure 7.27: Correlation matrix of Spearman's values capturing statistical relationship between resulting morphometric values of contextual characters. With a few exceptions, the relationship is none or very weak. High-quality version of the matrix is in Appendix XXX.

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

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

7.3.3 Clustering

The critical point in the whole process if identification of distinct homogenous clusters is the clustering procedure itself. In the case of Prague it has to deal with ~140,000 features each with 296 dimensions, which is a challenging task. This section will explore the results of clustering on the complete set of data and then on sampled data, to understand the difference and possibility to lower computational demands. Both ways will start with assessment of an optimal number of components based on Bayesian Information Criterion (BIC) and follow with Gaussian Mixture Model (GMM) clustering itself. Furthermore, final part of this section explore the potential of sub-clustering, i.e., generating even more detailed distinction of urban tissues.

7.3.3.1 Complete data

Clustering based on complete data is likely the key result of the whole research. It will either support the main hypothesis or reject it depending on the resulting clusters. GMM clustering of a complete dataset means that all features (n=140,315) are used within a training set and GMM has to deal with 41,533,240 values. It is expected that the algorithm will be able to detect clusters, although there might be present some adverse effects of the dimensionality curse.

Before analysis itself, data are standardised by mean removal and variance scaling using `sklearn.preprocessing.StandardScaler` (Pedregosa et al. 2011):

$$z = \frac{x-\mu}{s} \quad (75)$$

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

7.3.3.1.1 Bayesian Information Criterion To perform GMM clustering, one needs to specify the number of components to look for. While this information is not a priori known, one have to determine the optimal number using other methods before GMM. In this research, Bayesian Information Criterion is used.

BIC analysis repeatedly generates GMM clusters for different number of components in a range (2, 39) and measures the goodness of fit of resulting clusters to the original dataset. The results in Prague are shown on figure 7.28. The lower the value, the better fit.

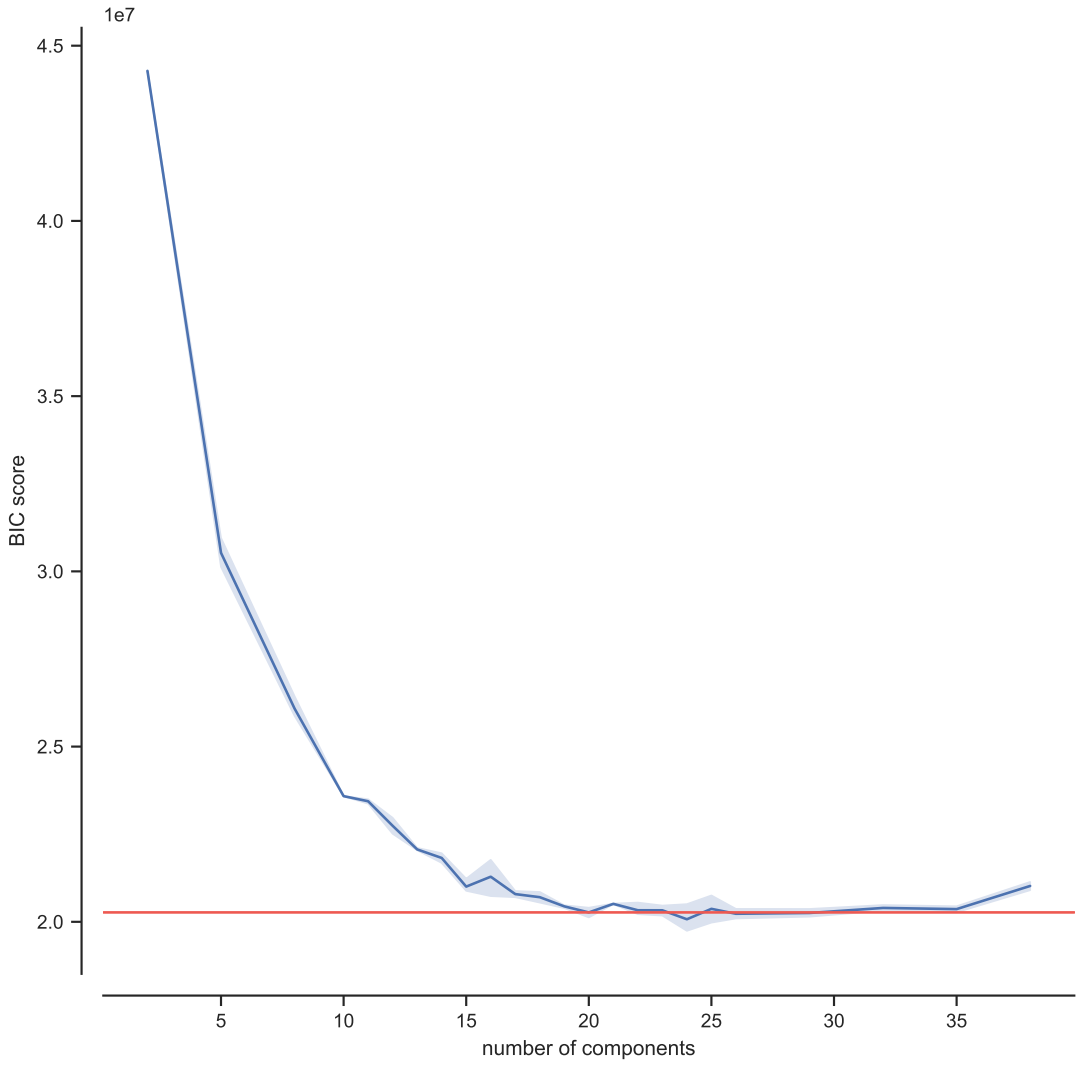

Figure 7.28: Bayesian Information Criterion score for changing number of components. Shaded area reflects .95 confidence interval, red line marks the first significant minimum.

The overall pattern shows steep decline from small number of components to approximately 15 components where it starts flattening. The results between 15 and 35 components are very similar and then the BIC starts growing again.

That suggests, that the optimal number of components for the final clustering is between these two values. The optimum is 20 as the value which is the first significant minimum. It is the smallest number after which no other is significantly (with the confidence interval) below the achieved score. The differentiation within the range in question is better recognisable in a zoomed figure 7.29 below.

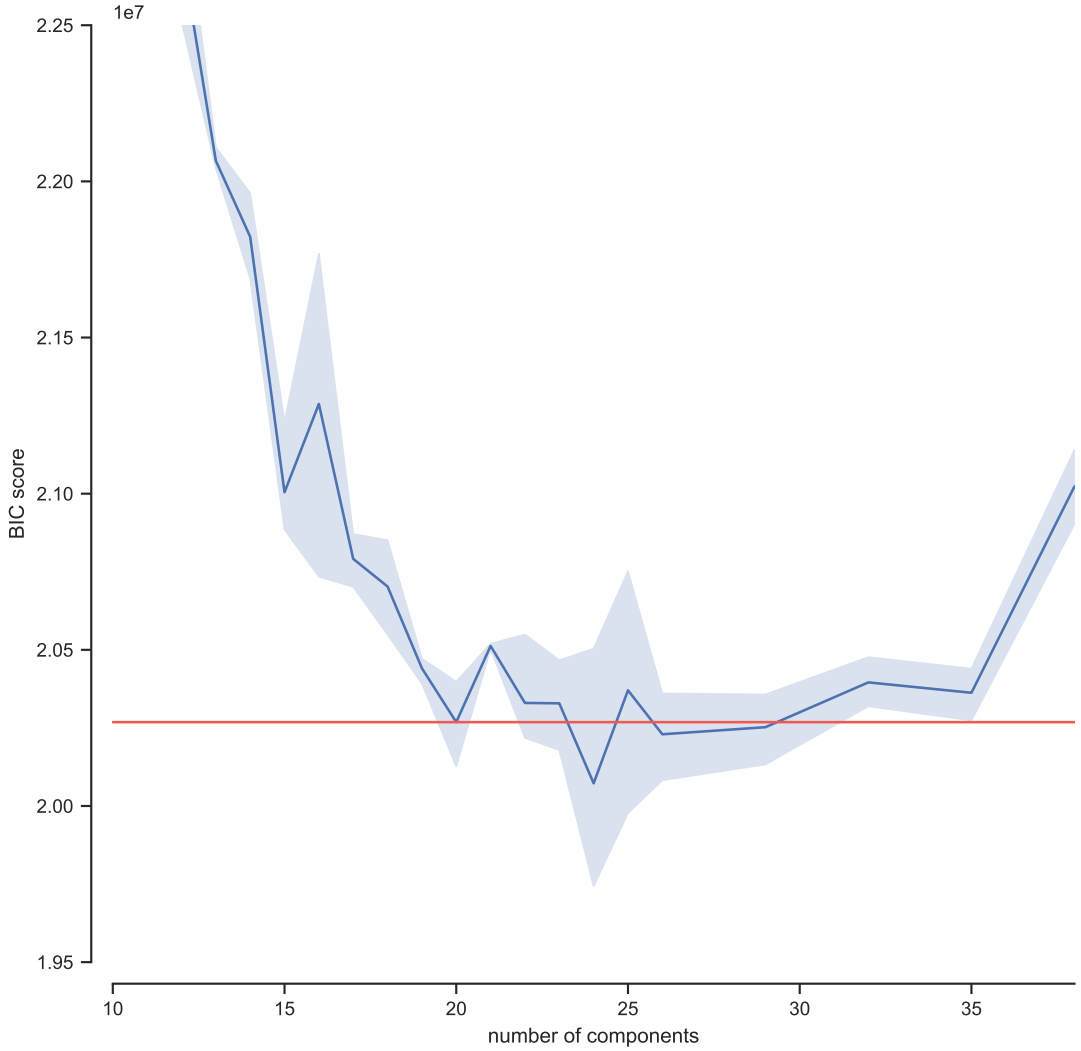

Figure 7.29: Trimmed plot of Bayesian Information Criterion score for changing number of components, to see the differentiation within values. Shaded area reflects .95 confidence interval, red line marks the first significant minimum.

Within the trimmed figure is more evident the difference, between the BIC score. The reason why 24 or 26 are not selected as optimum, while both being smaller

than 20 is the significance. The aim is to detect the smallest optimal number of clusters as larger numbers may have better due to overfitting. Hence we want the first significant minimum, which is 20. 24 is below, but its confidence interval goes above the score of 20. The same applies to 26.

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

7.3.3.1.2 Distinct homogenous clusters Gaussian Mixture Model clustering with 20 components is done using full covariance matrix and 5 initialisations, from which the best is selected based on the per-sample average log-likelihood. Following code snippet illustrates the exact implementation using `GaussianMixture` class from scikit-learn 0.22 (Pedregosa et al. 2011). The complete Jupyter notebook is available in Appendix XXX.

```
from sklearn.mixture import GaussianMixture
model = GaussianMixture(n_components=20, covariance_type="full",
                        max_iter=200, n_init=5).fit(data)
```

The resulting prediction of cluster membership is shown visually on the figure 7.30. Each feature (building/tessellation cell) is coloured according to cluster of the highest probability. The map shows the delineation of distinct homogenous clusters and their spatial distribution across the whole case study area. At this moment it is possible to say that the proposed method did identify certain type of proxy of urban tissues using purely quantitative method based on urban morphometrics. How well it did that will be assessed on following pages and later validated by other data in Chapter 8.

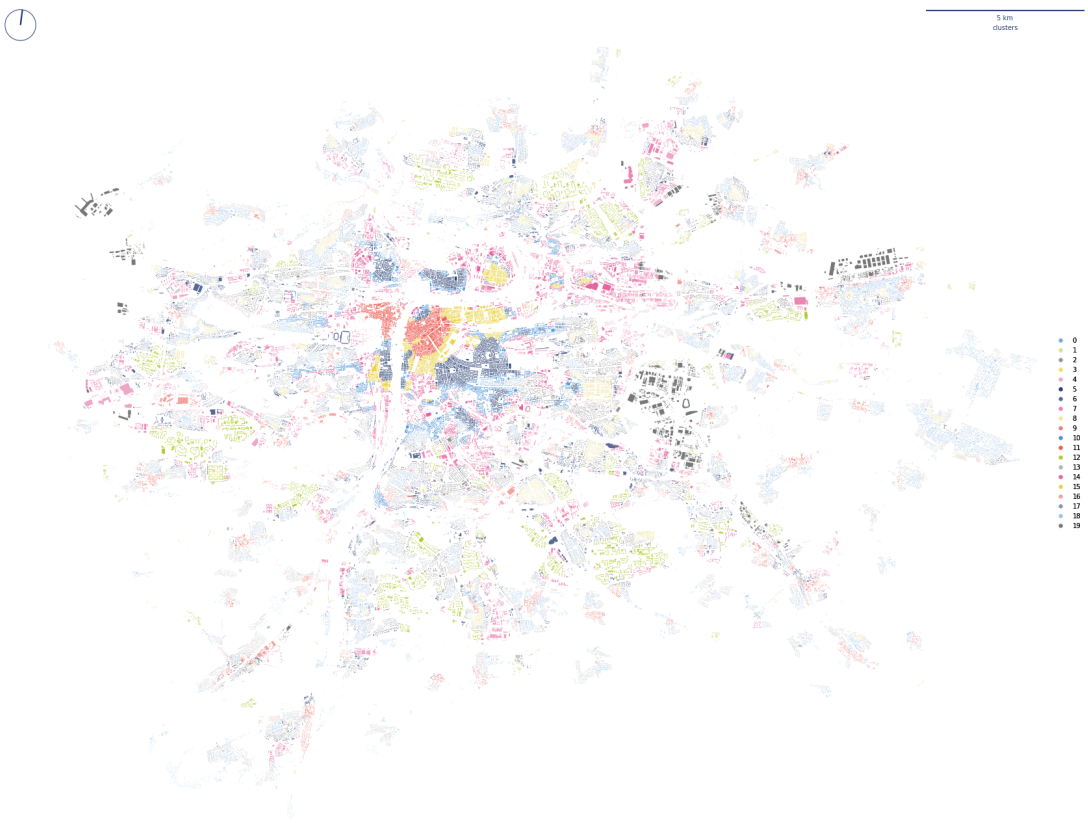

Figure 7.30: Spatial distribution of 20 clusters as identified by GMM based on complete data.

The 20 cluster seems to be relatively well defined and based on the first observation tend to reflect homogenous form. Even though there is no spatial constraint in the clustering itself, results show apparent contiguity caused by the design of contextual characters. The figure 7.31 shows detail of section of Prague covering City Centre and area towards southern boundary for better understanding of results.

Figure 7.31: Long caption.

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

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

7.3.3.1.2.1 Individual clusters Each of the individual clusters is presented by one example (usually the largest contiguous area) and its surroundings within 1,5km buffer. Colour schema is the same as in figures 7.30 and 7.31 and will be kept throughout the chapter. Clusters are sorted according to their ID, which is randomly assigned.

The first cluster (figure 7.32), noted as 0, is composed of predominantly low-rise, single family housing. It has mostly residential character and tends to be located in the outer parts of the city, further away from the city centre. It is the largest of all clusters, with 15337 features, which is approximately 11% of all buildings in the study area.

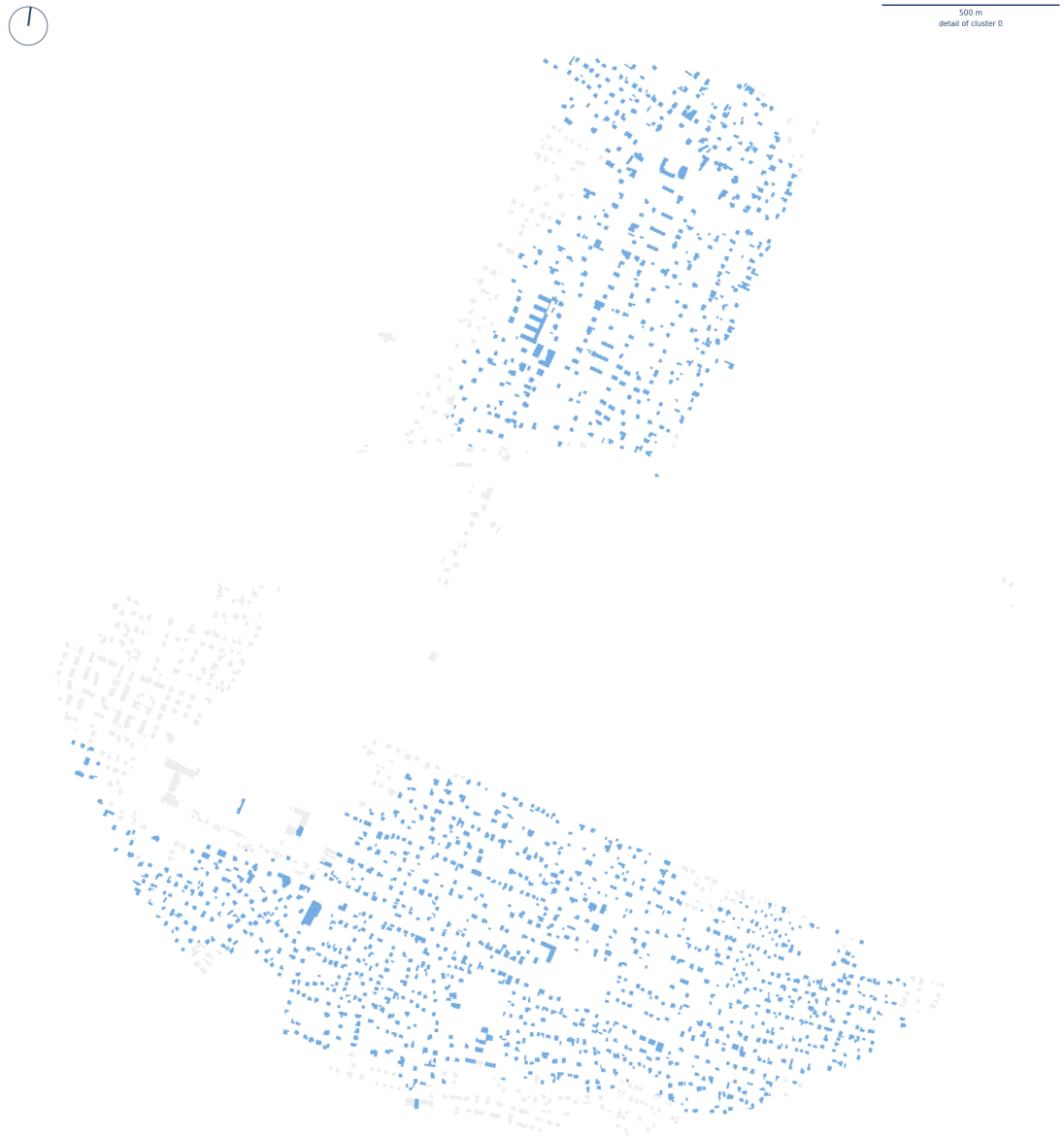

Figure 7.32: Example of cluster 0 and its surroundings within 1,5km buffer located at the eastern boundary of study area.

The cluster on figure 7.33, noted as 1, contains mostly small-scale industry areas with small coverage, relatively small buildings. Often is adjacent to other clusters. It tends to be located in outer rings of the city, but is overall very sparsely distributed. With only 2038 (less than 1.5%) features is one of the smallest clusters overall.

Figure 7.33: Example of cluster 1 and its surroundings within 1,5km buffer located at the north-east of study area.

Cluster 2, shown on figure 7.34 is one of the urban tissues following modernist principles of spatial configuration, with linear buildings, but still relatively con-

nected street network. These areas are mostly infills of the existing structure located within the city (except its central part) rather than on the periphery. It is relatively abundant with 12016 features (approximately 8.5%).

Figure 7.34: Example of cluster 2 and its surroundings within 1,5km buffer located at the north-west of study area.

Cluster 3 (figure 7.35) is one of the smaller ones. Its structure is defined by row-houses, a typology which is not very common in Prague. There are only 4133 features, less than 3% of all buildings scattered mostly in peripheral locations.

Figure 7.35: Example of cluster 3 and its surroundings within 1,5km buffer located at the east of study area.

Cluster 4 (figure 7.36) is one of the tissues with industrial character, in this case being distributed as sort of infill development in the fringe areas. It is largely adjacent to other urban tissues, relatively evenly distributed across the study area. It is composed of 5281, which is 3.8% of the total number.

Figure 7.36: Example of cluster 4 and its surroundings within 1,5km buffer located at the south of study area.

Cluster 5 (figure 7.37) can be best described as compact perimeter block-based residential area. This dense, grid-like development is located in the central areas of the city around the historical core and are one of the best defined urban tissues in Prague. There are 5930 of features belonging to this cluster, which is a bit more than 4.2% of total count.

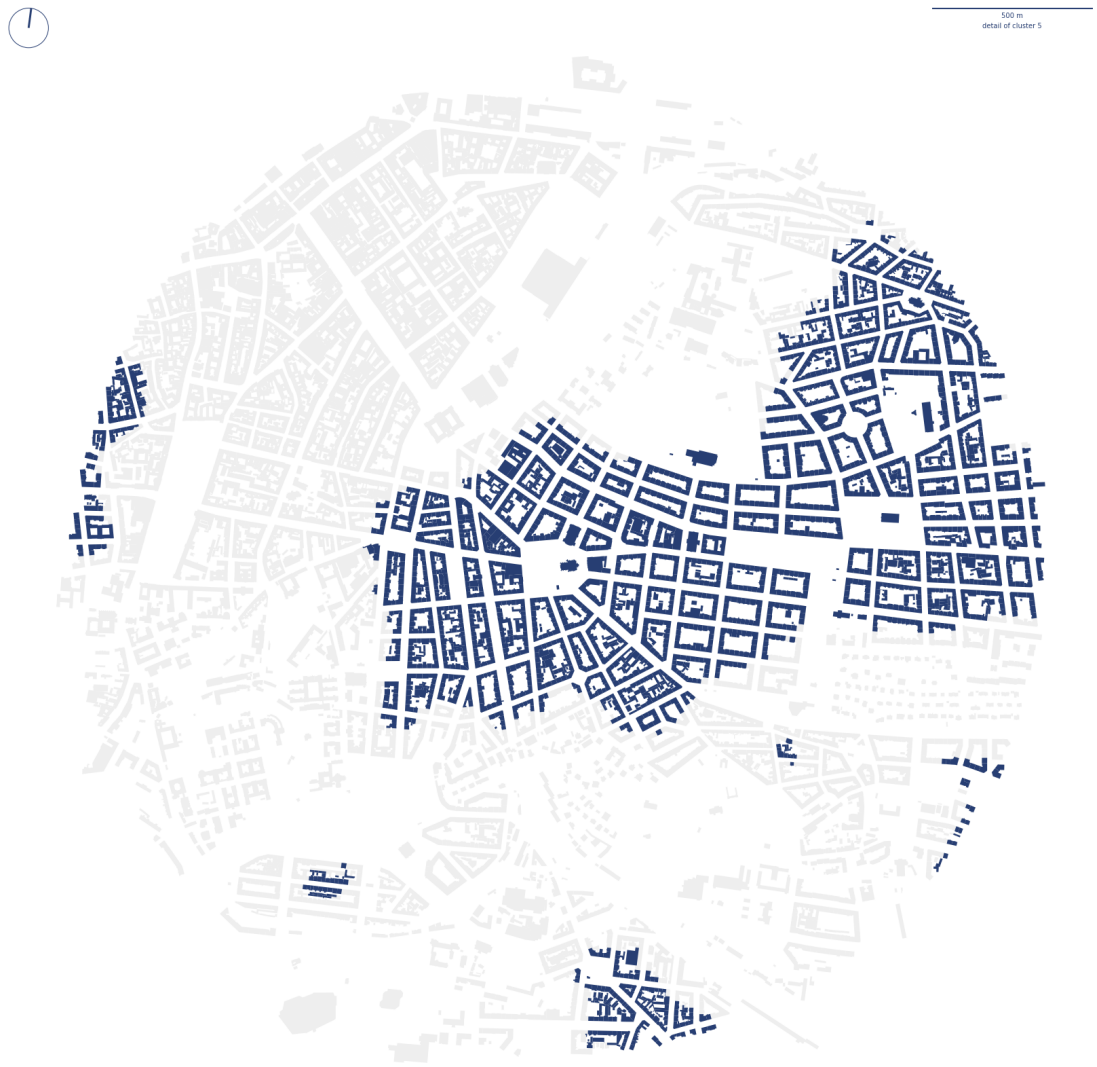

Figure 7.37: Example of cluster 5 and its surroundings within 1,5km buffer located at the centre of study area.

Cluster 6 (figure 7.38) is very different from the previous one as it contains fringe low-rise, not very well defined urban tissues. These are small-scale tissues scattered evenly around the study area, adjacent to other types of tissues, often filling topographically inconvenient areas. There are 10329 of these features, which is about 7.4%, so it is one of the more abundant clusters.

Figure 7.38: Example of cluster 6 and its surroundings within 1,5km buffer located the south-east direction from the city centre.

Cluster 7 (figure 7.39) is an example of more heterogenous area. It has a similar character as cluster 4, but unlike that it often contains other types of development with less defined structure, like contemporary housing or office parks which does not reflect traditional rules of spatial configuration. That leads to higher heterogeneity in the area making these tissues complicated to simply define. It consists of 4140 features, which is nearly 3% of total amount.

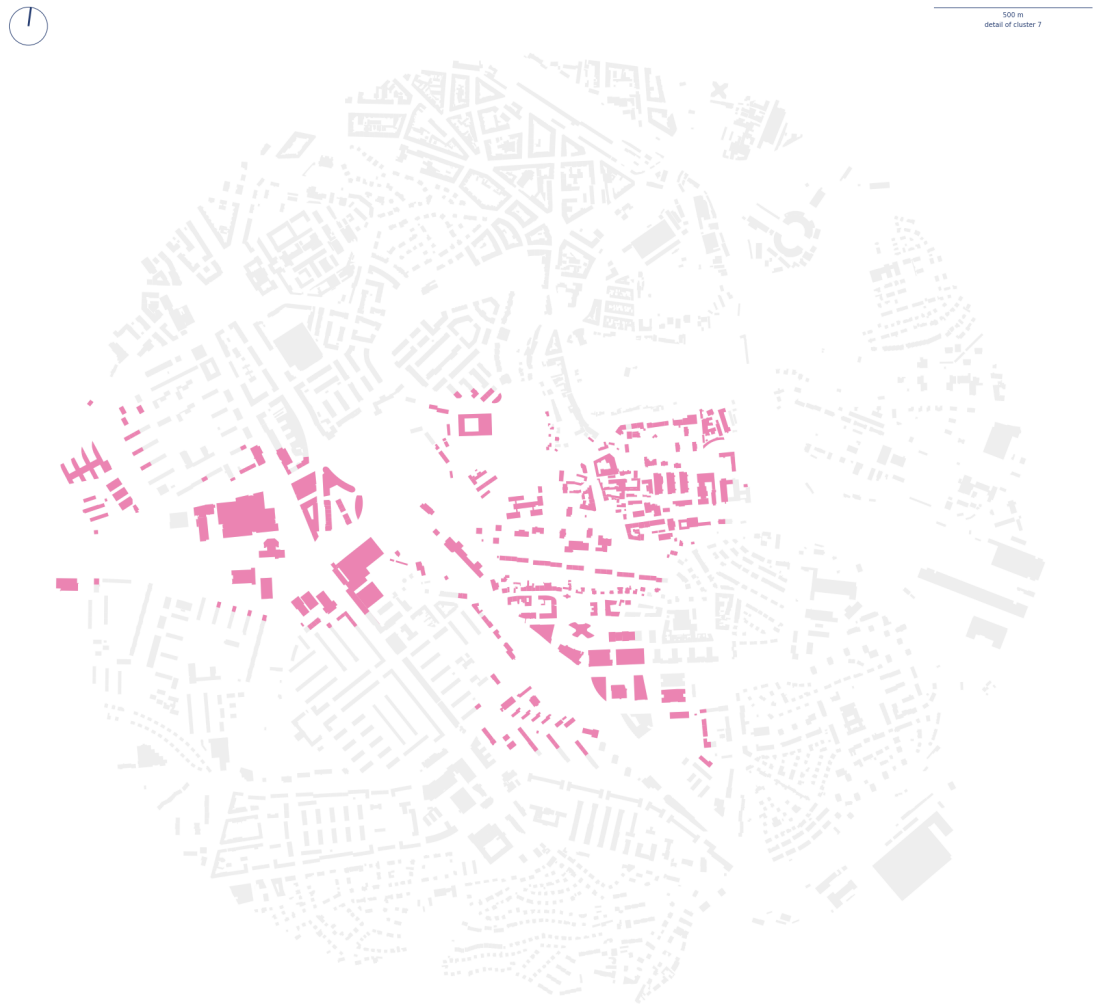

Figure 7.39: Example of cluster 7 and its surroundings within 1,5km buffer located the souther direction from the city centre.

Cluster 8 (figure 7.40) predominantly contains single family housing in relatively dense setting resembling garden city movement development. These places have interconnected network of relatively grid-like character, with buildings adjacent to each other either as row-house typology or similar. There are 7845 features within this cluster (5.6%).

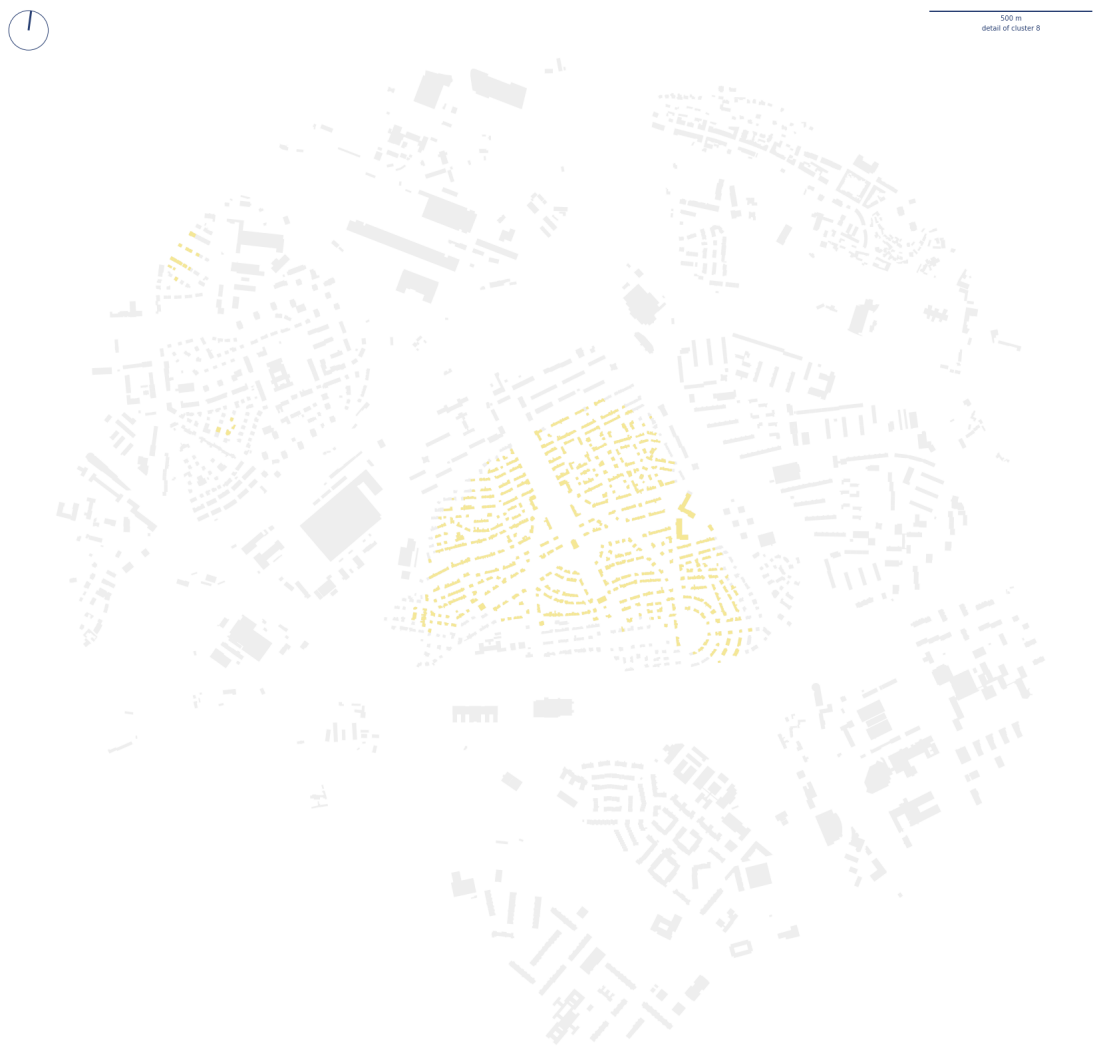

Figure 7.40: Example of cluster 8 and its surroundings within 1,5km buffer located the souther direction from the city centre.

Cluster 9 on figure 7.41 seems to identify low-rise areas of organic development, what seems to be cores of the historical villages around Prague. These are small-scale tissues evenly distributed in the outer ring of development, now mostly embedded in the other development. They compose 5.6% of total features (7862).

Figure 7.41: Example of cluster 9 and its surroundings within 1,5km buffer located at the south-west of the study area.

Cluster 10 (figure 7.42) is very often adjacent to cluster 5 (compact blocks) or composes its own areas of block-based development. However, unlike 5 these

blocks tend to be skewed or distorted in some other way. In some cases, this cluster could be seen as transitional area between homogenous compact blocks and other types of urban tissue. These are 7203 features within this group, making 5.1% of the total amount.

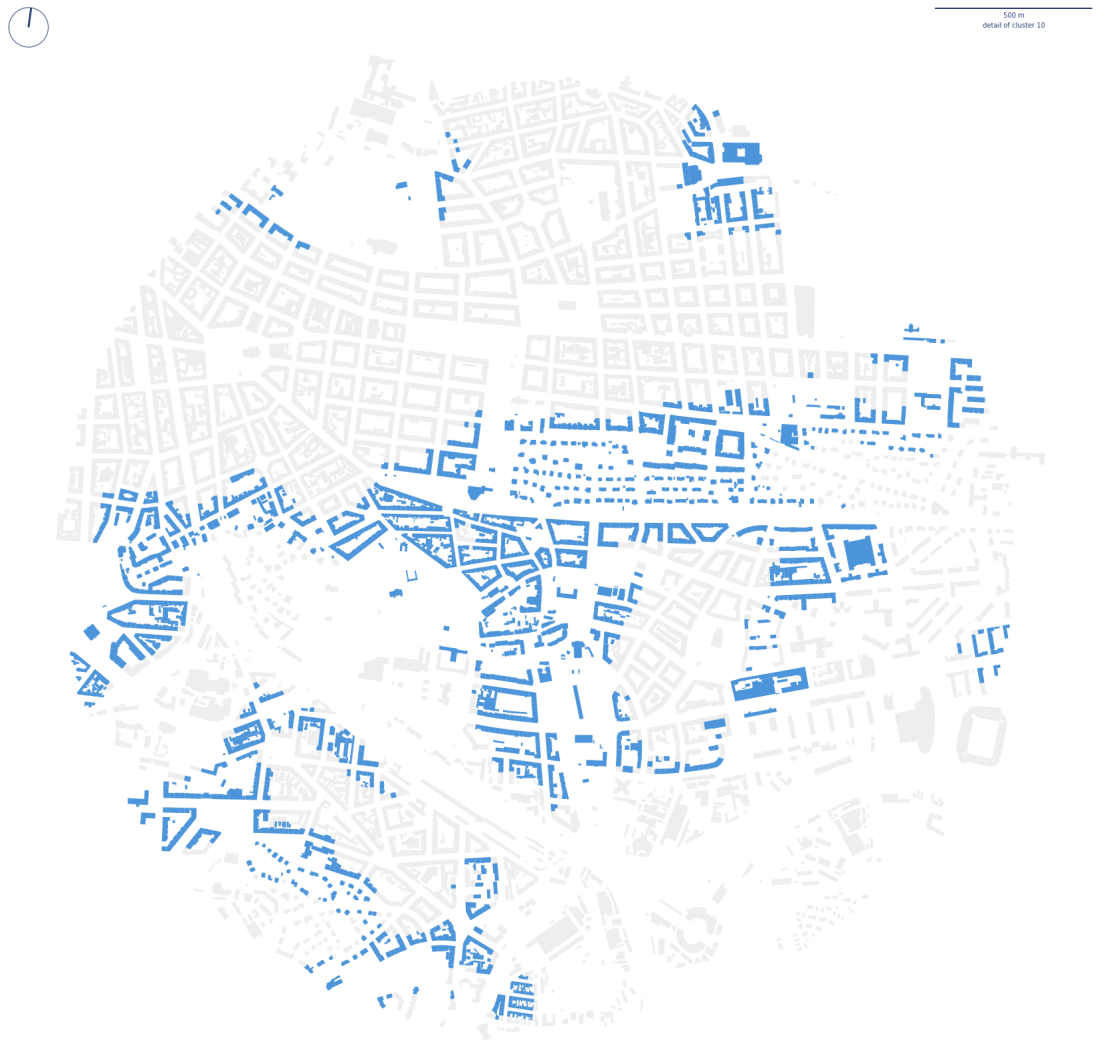

Figure 7.42: Example of cluster 10 and its surroundings within 1,5km buffer located next to the city centre.

Cluster 11 (figure 7.43) is very straightforward one to describe as it composes solely of historical medieval core of Prague. It includes areas on both sides of the river and correctly excludes the area cut-out of the Old Town, which has

been demolished in 19th century and rebuilt after that. There are 2167 features, making historical core one of the smallest clusters of all, composing only 1.5% of the total amount.

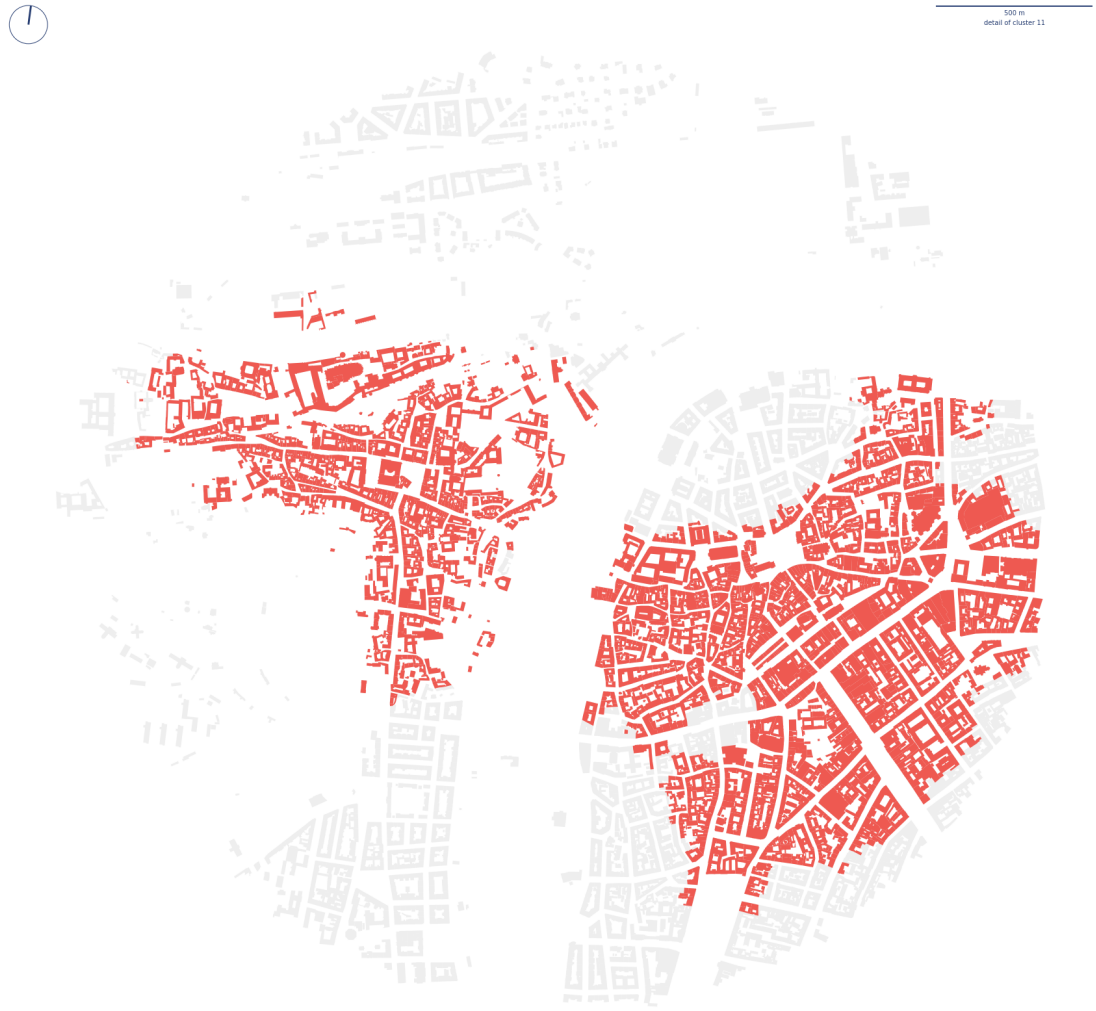

Figure 7.43: Example of cluster 11 and its surroundings within 1,5km buffer located in the city centre.

Cluster 12 is another very distinct one. As illustrated on figure 7.44, the origin of the development is modernist, covering large-scale modernist housing estates. These are typical with slab buildings, incoherent relationship between buildings, plots and streets and large amounts of open spaces, among other characteristics. In Prague they are almost exclusively on the peripheral ring of the city, forming

so-called modernist belt of Prague. They consist of 6885 features, which is 4.9% of the total amount.

Figure 7.44: Example of cluster 12 and its surroundings within 1,5km buffer located on the southern edge of the city.

Cluster 13 (figure 7.45) is another example consisting of single family housing. This time it is low-density development with predominantly detached buildings. It is typical with elongated blocks which in part is a reaction to the underlying topography. It is very abundant cluster with 14992 features, making more than 10.6% of the total amount, distributed along the periphery of the city.

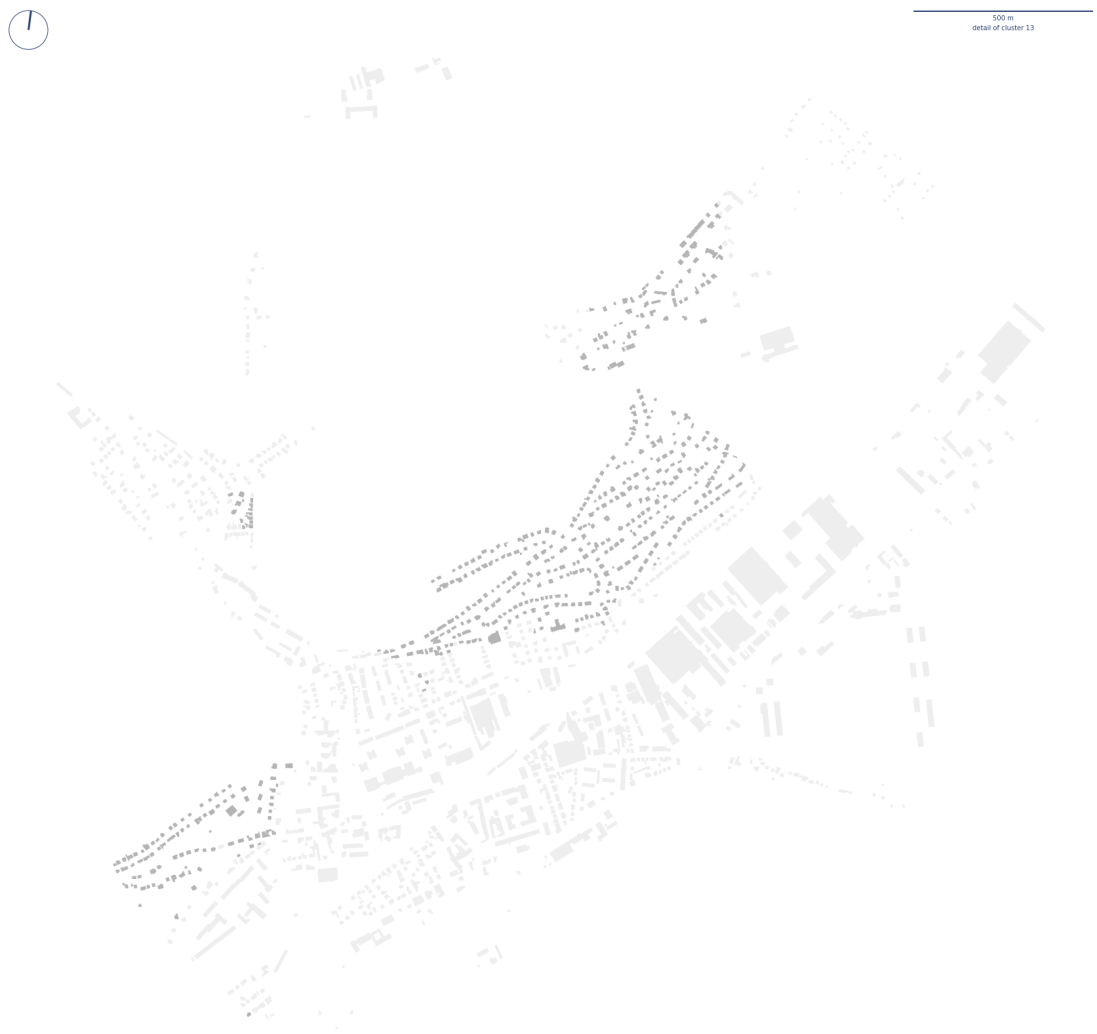

Figure 7.45: Example of cluster 13 and its surroundings within 1,5km buffer located on the south-western edge of the city.

Cluster 14 is distributed almost exclusively within the wider centre of Prague, often adjacent to homogenous compact city as is illustrated on the figure 7.46. The cluster could be defined as an inner fringe composed of heterogenous developments on the edge of existing homogenous one. There are 4984 features within it, making 3.6% of the data.

Figure 7.46: Example of cluster 14 and its surroundings within 1,5km buffer located north of the city centre.

Cluster 15 (7.47) is perimeter-block based tissue with very heterogenous development in the block interiors. It has very high coverage area ratio located in in the city centre either as a transitional area between medieval core and compact city or as industrial development. There are only 3060 features within the cluster (2.2%).

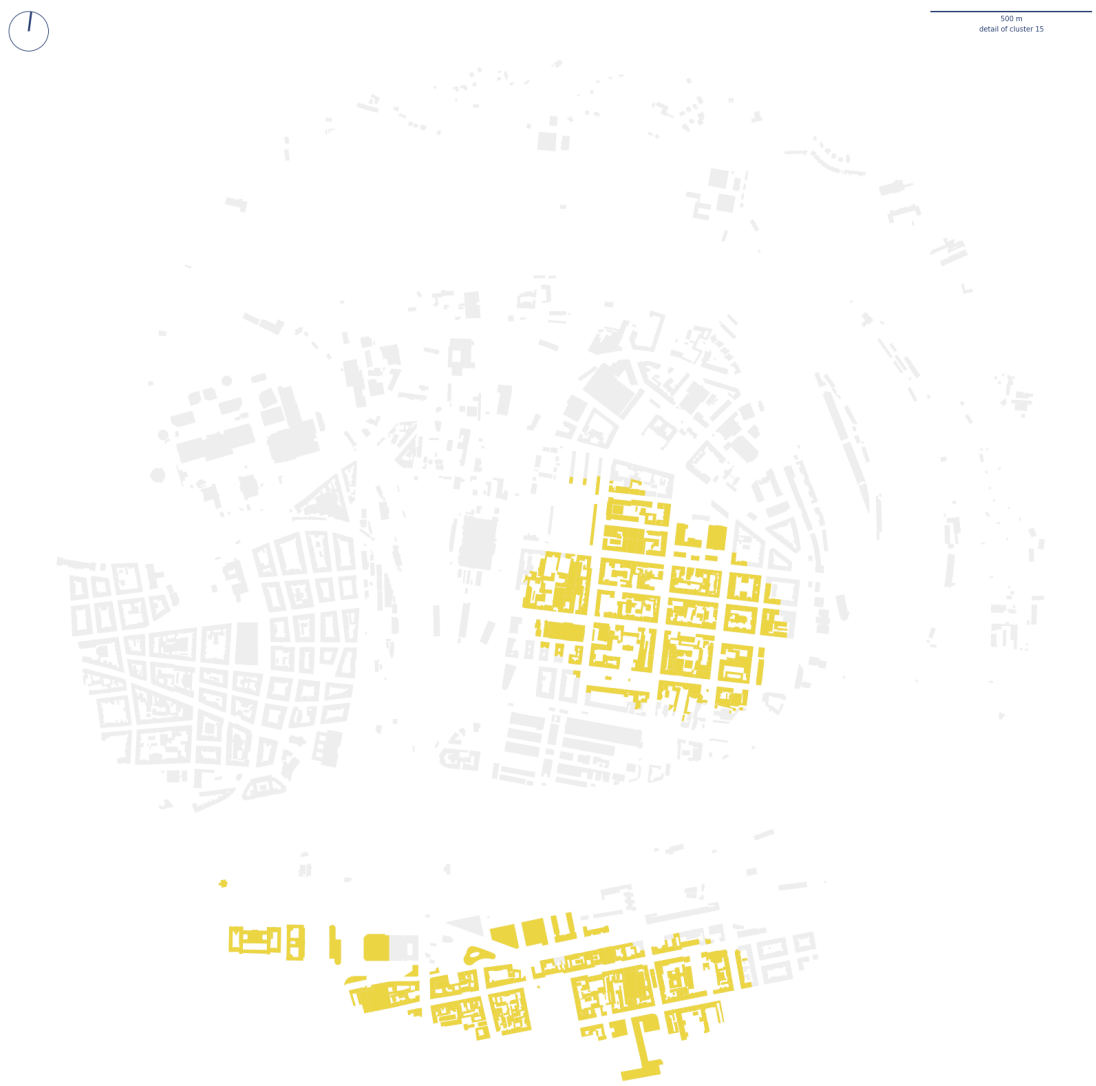

Figure 7.47: Example of cluster 15 and its surroundings within 1,5km buffer located north of the city centre.

Cluster 16 (7.48) is not very straightforward to define as it is a heterogenous one. It mostly consists of small patches of not very well defined tissues with predominant role of small-scale buildings but not exclusively. It may be seen as *other*, combining parts of the dataset which do not fit elsewhere, but at the same time all places have similar character of being *out of sight*. It is evenly distributed, but not very abundant one with 3548 making approximately 2.5% of the dataset.

Figure 7.48: Example of cluster 16 and its surroundings within 1,5km buffer located north of the city centre.

Cluster 17 (7.49) is another of the low-density single-family tissue types. It has less defined and rigid structure, it is often adjacent to open space. It does have a certain inner heterogeneity expressed as various kinds of buildings from detached to row houses. As the other similar clusters, this is also relatively abundant with 12145 features (8.7%).

Figure 7.49: Example of cluster 17 and its surroundings within 1,5km buffer located west of the city centre.

Cluster 18 (7.50) consists of relatively independent detached areas of low-density village-like development. DHCs can be only a strip along the road or other open-space facing tissues. It is located mostly on the periphery of the city and entails 8764 features (6.2%)

Figure 7.50: Example of cluster 18 and its surroundings within 1,5km buffer located on the north of the city.

The last cluster, 19 illustrated on figure 7.51 is industrial urban tissue, consisting of mixture of large-scale and small-scale buildings, convoluted street network and minimum of residential use. It is only at a few places, but of a large-scale, mostly towards the edge of the city. There are only 1656 features within the cluster making 1.2% of the total amount.

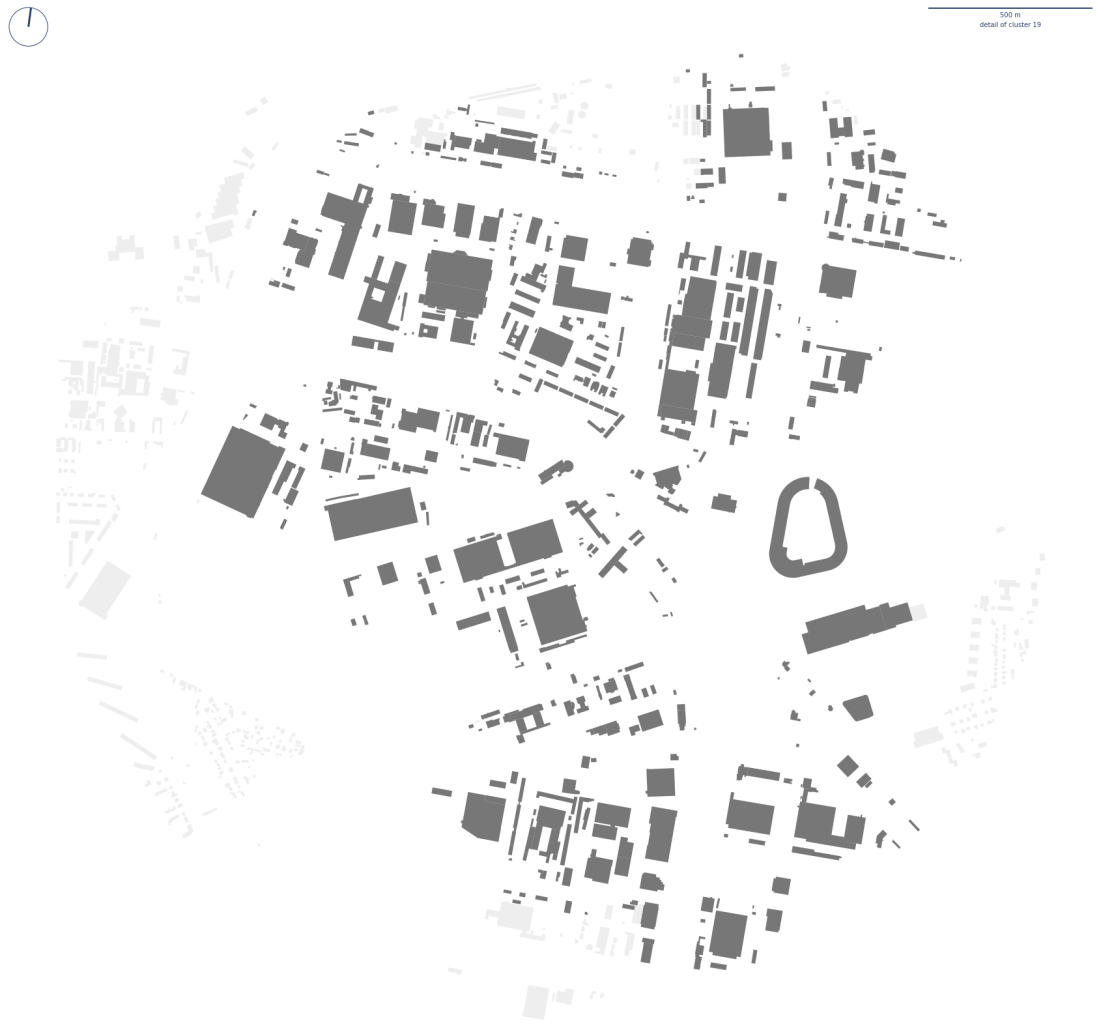

Figure 7.51: Example of cluster 19 and its surroundings within 1,5km buffer located on the eastern edge of the city.

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

7.3.3.2 Sampled data

Gaussian Mixture Model is an unsupervised machine learning procedure which means that it uses a training data on input to estimate the optimal clustering and then predicts the probability that each feature belongs to any of the components. That means that training data do not have to equal the data we want to classify. The GMM and especially the estimation of number of components, which does GMM repeatedly, could have relatively high computational demands as the size of the dataset grows. The Prague example, with 140,000 features took approximately 100 hours to measure all BIC values and do the final clustering on a desktop computer with 12-core Intel Xeon processor. Running larger areas at once may get unfeasible, it is hence critical to understand if the method can work with sampled data.

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

7.3.3.2.1 Bayesian Information Criterion Three versions of sampling are assessed - 10%, 25%, and 50% of the dataset. Because random sampling results in different samples each time, which could affected BIC, each option is sampled three times and GMM is run three times on each (in total 9 runs of GMM per option). To asses number of components, values from range 2 - 40 were tested. BIC is measured using the complete dataset. Simplified code below illustrates the principle.

```
for s in range(3):
    sample = data.sample(sample_size)
    for n_components in range(2, 40):
```

```
for i in range(3):
    gmm = GaussianMixture(n_components=n_components,
                          covariance_type="full",
                          max_iter=200)
    fitted = gmm.fit(sample)
    bic = gmm.bic(data)
```

The resulting values are shown on figure 7.52 below.

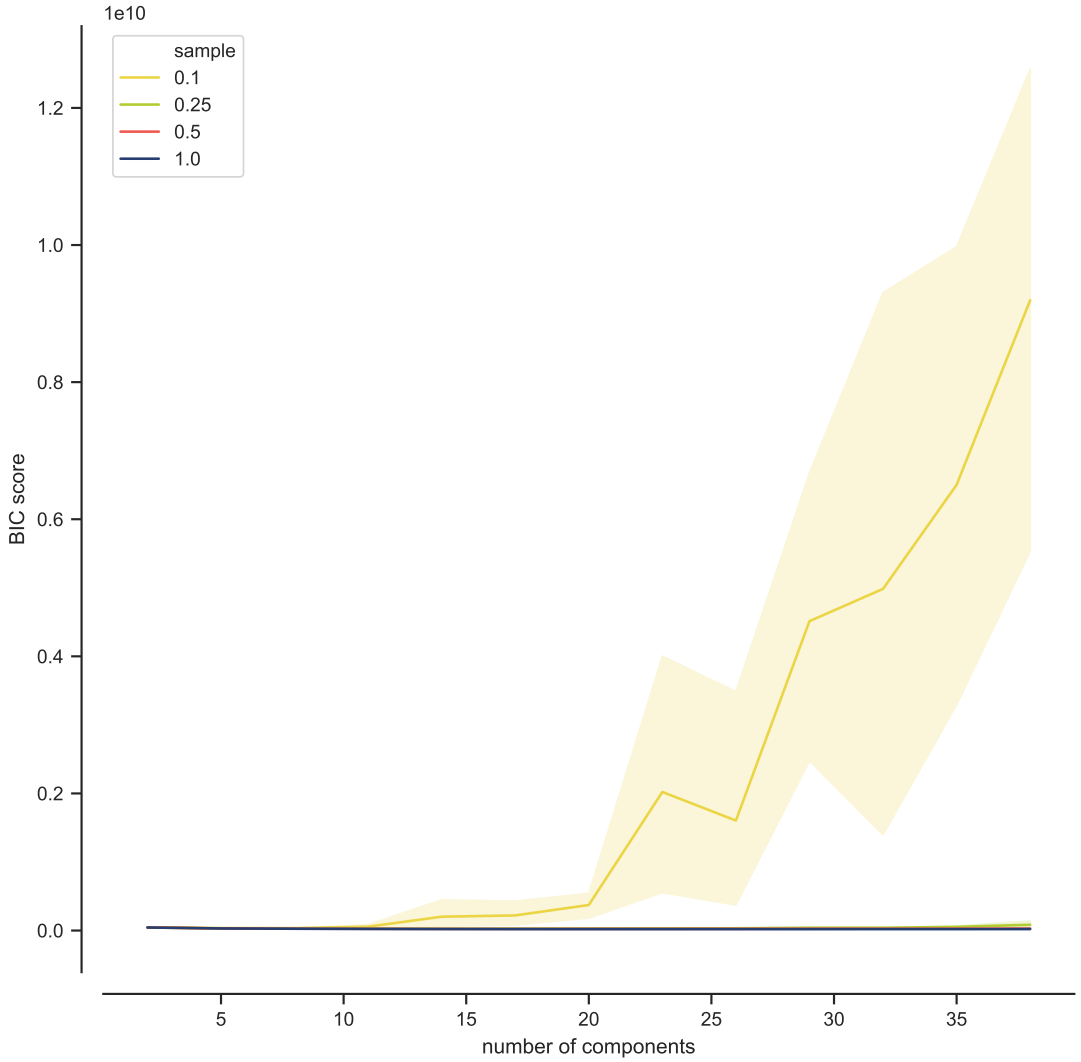

Figure 7.52: Bayesian Information Criterion score for sampled clustering. Shaded area reflects .95 confidence interval.

Figure 7.52 shows a striking difference between results of 0.1 (10%) sampling and the rest. The BIC score for this option is significantly higher that for the rest indicating that the sample is way too restricted to capture the structure of the dataset and generate meaningful clustering. Due to this difference, any differences between the other options are not recognisable. For that reason, figure 7.53 shows the same data without 0.1 option.

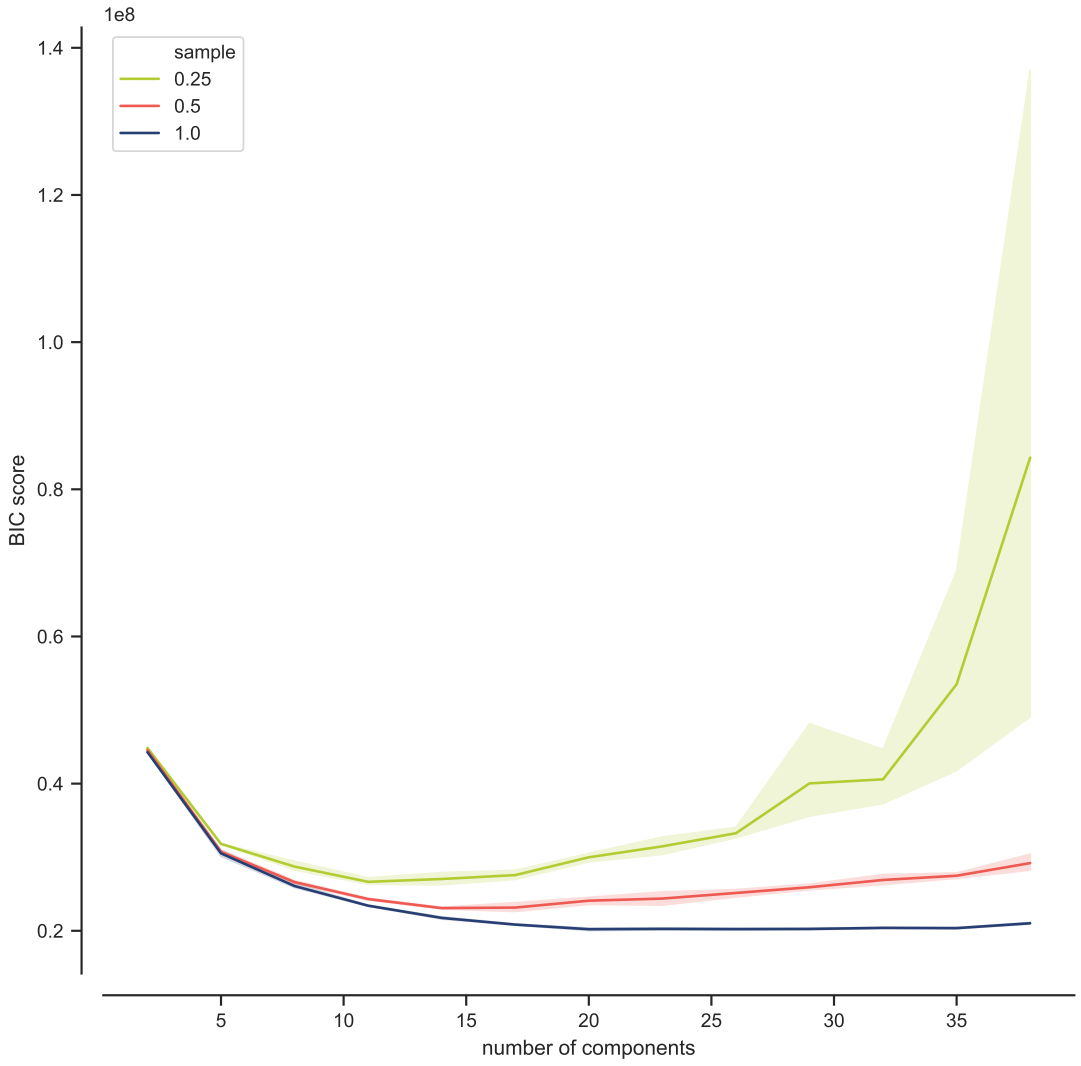

Figure 7.53: Bayesian Information Criterion score for sampled clustering excluding 0.1 sampled results. Shaded area reflects .95 confidence interval.

All two remaining options show similar curves as was already seen in the complete

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

[Bayesian Information Criterion score for sampled clustering excluding 0.1 and 0.25 sampled results. Shaded area reflects .95 confidence interval.](source/figures/ch7/PRG_sampled_BIC_05_10.pdf "BIC score for sampled clustering without 0.1, 0.25""){#fig:PRG_sampled_BIC_05_10 width=100%}

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

7.3.3.2.2 Distinct homogenous clusters Identification of distinct homogenous clusters based on sampled data (random sample 50%) using 15 components results in a spatial distribution of clusters illustrated on figure 7.54.

Figure 7.54: Spatial distribution of clusters based on sampled data (0.5) and 15 components within the whole study area.

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

7.3.3.2.2.1 Comparison of sampled and complete clustering To understand what are the actual on-ground differences between two versions of clustering apart from the different number of components, three easy-to-interpret clusters are compared 1:1 and their composition and shape is assessed.

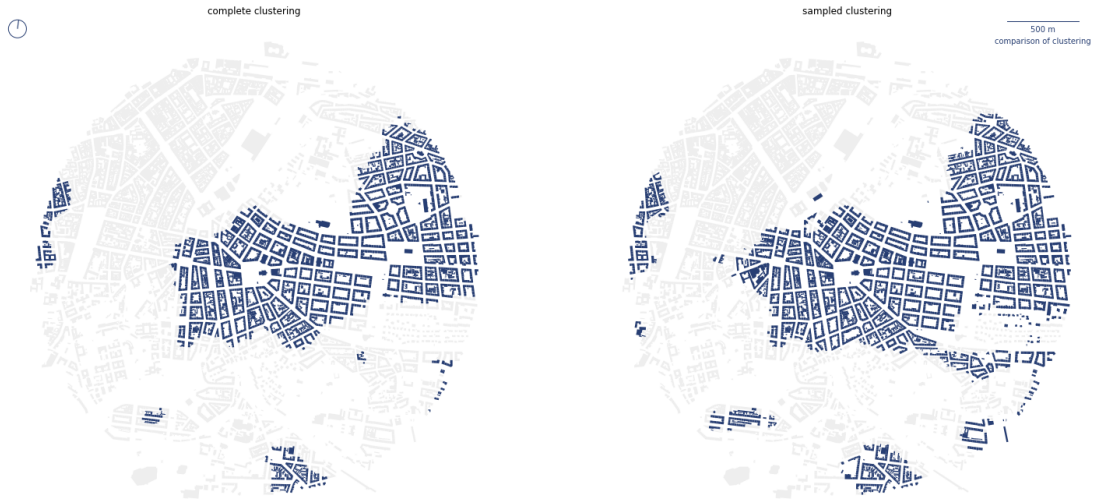

Figure 7.55: Comparison of spatial distribution of cluster 5 and sampled cluster 4 in the city centre.

Original cluster 5, capturing compact perimeter blocks has its counterpart in sampled cluster 4. The example of their spatial distribution is illustrated on figure 7.55. Both versions capture essentially the same type of urban tissue with very similar footprint. The only apparent difference is that sampled cluster is more inclusive (covering larger area) than complete cluster, likely due to smaller overall number of clusters (clusters needs to be naturally larger).

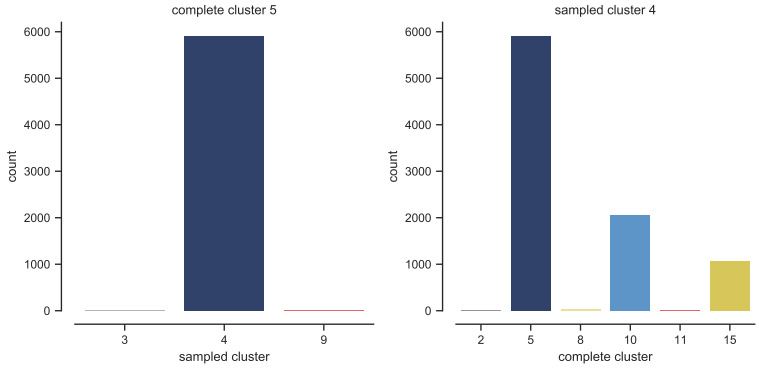

Figure 7.56: Composition of cluster 5 and sampled cluster 4 in relation to each other. Shows number of features labeled as studied cluster and their labels in the other clustering variant.

The comparison of composition of both clustering versions in relation to each other on figure 7.56 shows that features originally marked as being in the complete

cluster 5 are almost entirely within sampled cluster 4. On the other hand, features labeled as sampled cluster 4 are predominantly located in complete cluster 5, but due to higher inclusiveness also in clusters 10 and 15.

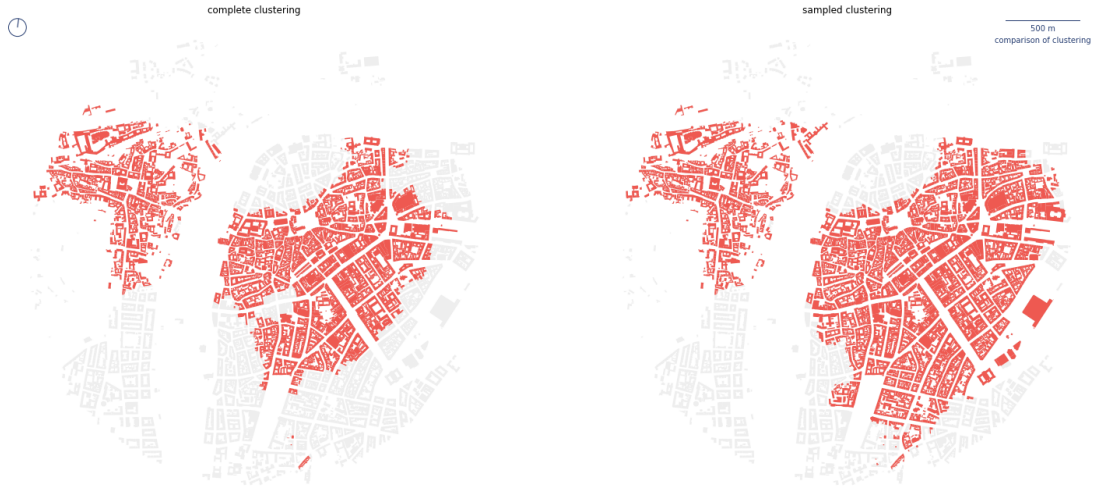

Figure 7.57: Comparison of spatial distribution of cluster 11 and sampled cluster 9 in the historical core.

Second example based on complete cluster 11 (figure 7.57) representing historical core of Prague shows the similar story as the first one. The both versions correctly delineate medieval urban tissue, and avoid newer redevelopment of former Jewish quarter in the North. The difference is in inclusiveness, where sampled cluster covers larger areas, which are in complete clustering seen as cluster 15, the transitional one.

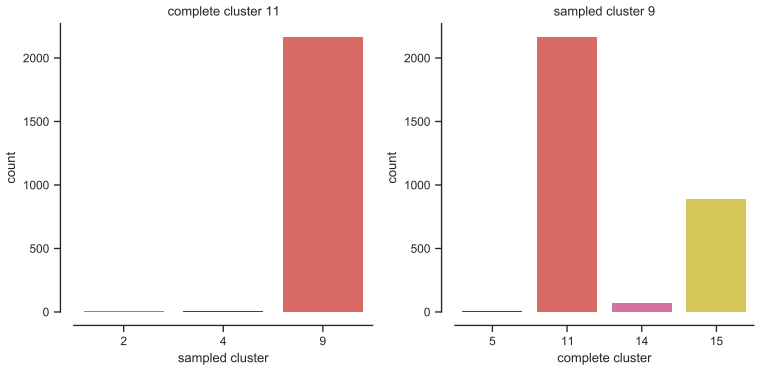

Figure 7.58: Composition of cluster 11 and sampled cluster 9 in relation to each other. Shows number of features labeled as studied cluster and their labels in the other clustering variant.

The assumption derived from visual assessment is correct based on the numerical data on the actual composition (figure 7.58). Features classified as cluster 11 in complete clustering, are classified as cluster 9 in sampled clustering. Features classified as cluster 9 in sampled clustering, however, include parts of complete cluster 15 (less than 1/3 of cluster 9 comes from 15).

Figure 7.59: Comparison of spatial distribution of cluster 12 and sampled cluster 5 in the west of the city.

Even more similarity shows a comparison of large-scale modernist urban tissues (figure 7.59) with a very few differences which could be derived from visual observation. Even the complicated case of modernist area mimicking perimeter blocks in the South of the example shows the same pattern, with middle part being excluded from the rest (likely an effect of a configuration of street network).

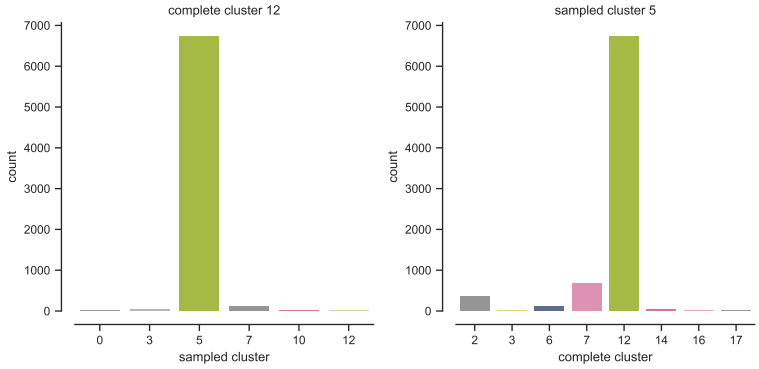

Figure 7.60: Composition of cluster 12 and sampled cluster 5 in relation to each other. Shows number of features labeled as studied cluster and their labels in the other clustering variant.

The composition of clusters is, with a few exceptions equal (figure 7.60). The same features which belong to cluster 12 in complete clustering are labelled as cluster 5 in the sampled clustering and vice versa.

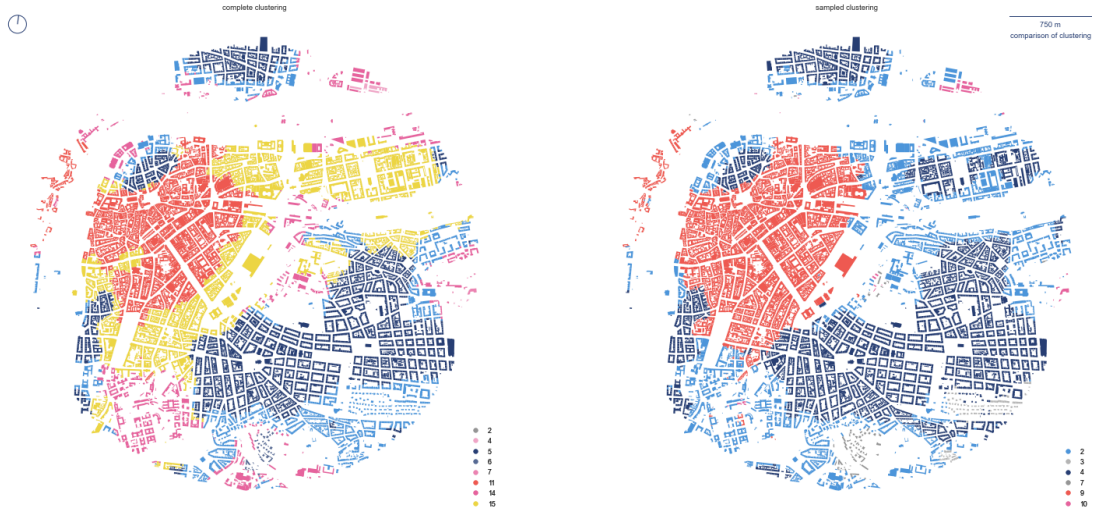

Figure 7.61: Comparison of the city centre focusing on cluster 15, which is not present in the sampled clustering.

Three examples above shows that there is a striking similarity between both results. However, there is a different number of clusters, so where is the difference? The example on figure 7.61 shows cluster 15 based on complete data, which does not have its counterpart in sampled clustering. Instead, it is split into three

almost equal parts (figure 7.62) each linked to another cluster. What was so-called transitional area between medieval core and historical compact city is no longer present. That by itself is likely not a big issue, but it illustrates the behaviour of sampled clustering with smaller number of components. It does not necessarily merge two similar clusters into one, but as some places split clusters into multiple pieces. GMM in this case sees different data and hence might exclude some smaller clusters. Because these might have been *in between* other, parts are now closer to one and other part closer to another cluster. The resulting clustering should then be seen as different perspective using different resolution, rather than coarser version of complete clustering.

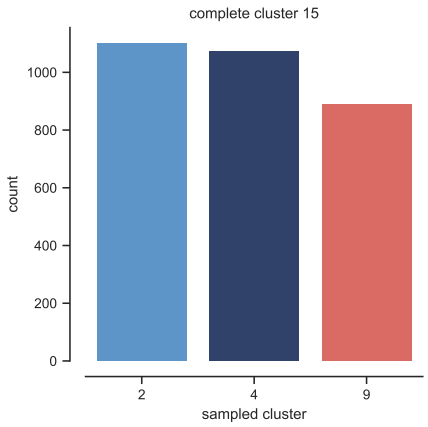

Figure 7.62: Composition of cluster 15 in relation to sampled clustering. Shows number of features labeled as studied cluster and their labels in the other clustering variant.

Looking onto other clusters which do not have counterpart in sampled clustering (apart from 15, 2, 6, 9, and 10), none of them is *swallowed* by one larger cluster, all are split into two sampled clusters. Sometimes more equally (e.g., cluster 2 is equally split between 10 and 14), sometimes less equally (e.g., cluster 6 is more present in sampled 0 than sampled 8). This illustrates the probabilistic rather than hierarchical nature of GMM. Full comparison is available in Appendix XXX.*TODO: add full data to appendix (relation_sampled-complete, relation_complete-sampled)*

Depends on the aim of the study, sampled clustering could likely be used instead of complete clustering, considering the fact that results based on samples smaller

than 50% are not precise enough. However, for the optimal, detailed identification of distinct homogenous clusters, sampled clustering might provide sub-optimal results.

7.3.3.3 Note on probability

The Gaussian Mixture Model clustering is probabilistic, which means that each feature has predicted probability that it belongs to any of the components. What is shown on all maps and data above, is the cluster with the highest probability. In theory, it should be possible to work with secondary or tertiary labels for each feature, but the actual data on probability tell otherwise. The probability that features belong to any other than primary cluster tends to be insignificant. Only 89 out of 140,315 features have probability that they belong to any other than primary cluster bigger than 0.1. The reason behind it is likely related to the richness of the data and especially related dimensionality causing big differences in Euclidean distance between clusters. So while GMM is in theory probabilistic, in practice it provides a single primary label only.

7.3.3.4 Sub-clustering

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

7.3.3.4.1 Compact Prague The first case the cluster 5, which could be interpreted as urban tissue of compact rigid perimeter blocks. The reason for its selection is that due to the varied topography, these blocks has to react to steeper

surface at some places and the perceptional character of such areas is different from those laying on the flat grounds.

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

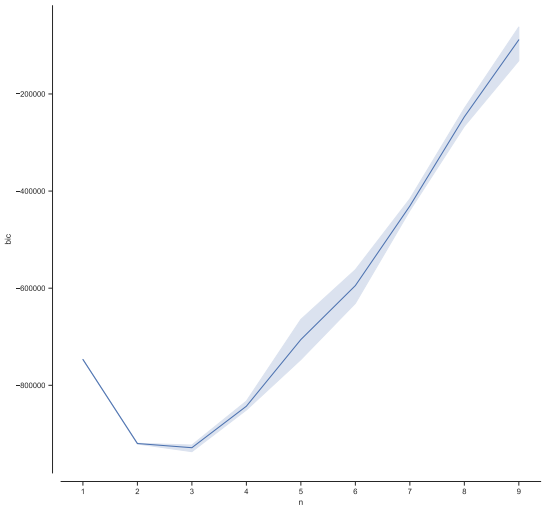

Figure 7.63: TITLE

Bic shows that there's potential either 2 or even 3

Bayesian Information Criterion illustrated on figure 7.63 indicates that there is a scope for sub-clustering as both 2 and 3 components have better score than a single component. If the situation would be otherwise, and a single component would have the lowest BIC score, there would be no significant sub-clusters in the data and results of forced clustering would likely suffer from discontinuity. Following the rule of the first significant minimum, this trial works with 2 components.

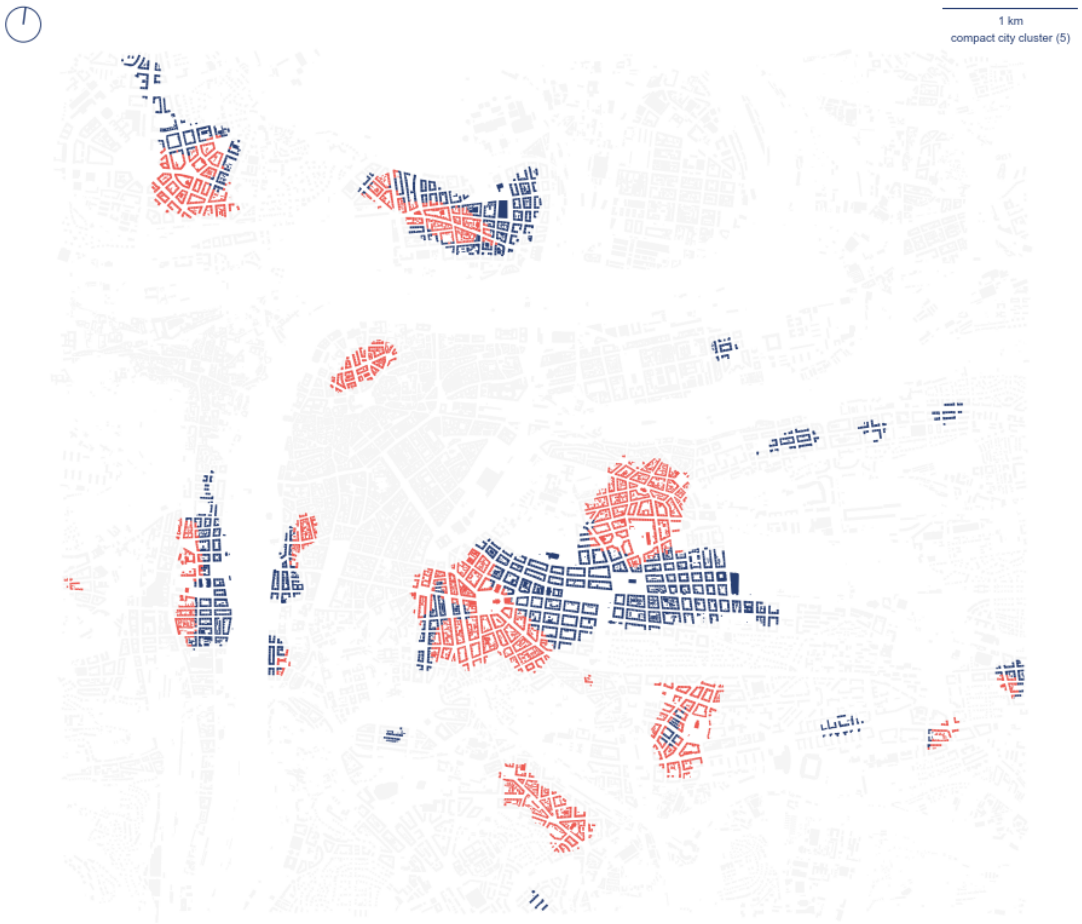

Figure 7.64: TITLE

The result of sub-clustering of cluster 5 is shown on figure 7.64. It feels fair to conclude, that newly identified sub-clusters have a meaning and distinguish between areas which are more rigidly gridded and those which tend to have grid distorted.

7.3.3.4.2 Modernist Prague -second example is modernism because

Second sub-clustering trial focus on large-scale modernist housing estates on the periphery of Prague. There is an assumption of inner differentiation of the relevant cluster 12 because each of these neighbourhoods has been designed and there were different authors and approaches in different places and periods of

development. It is assumed that morphometric data should be able to reflect this difference.

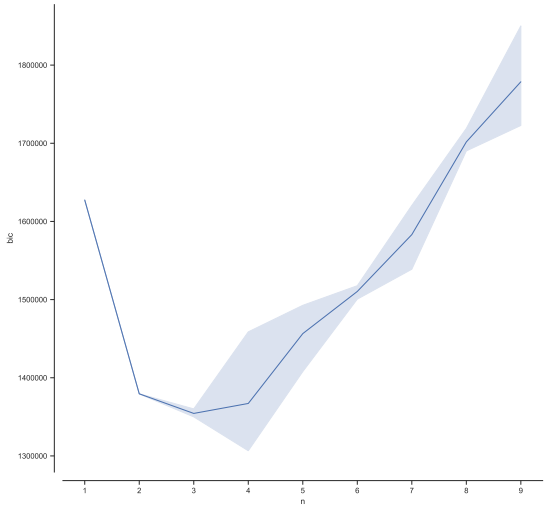

Figure 7.65: TITLE

BIC results on figure 7.65 indicates that subdivision of the cluster is significantly better than a single group with all the options between 2 and 7 having lower score that one component. The first significant minimum in this case are 3 components.

Figure 7.66: TITLE

The map on the figure 7.66 shows the whole cluster 12 divided into three sub-clusters. Interesting case is the green group, located exclusively on the western edge of the study area. The fact that it is not present anywhere else indicates that sub-clustering indicates that results are not affected by randomness. Closer look at the differences as illustrated on the figure 7.67 shows why these tissues are split in such a way. The green sub-cluster have large blocks and circular character, the red one tends to be a large-scale orthogonal configuration, while blue is smaller-scale more compact urban tissue.

Figure 7.67: TITLE

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

7.4 DHC as an urban tissue

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

While the term *urban tissue* is used in paragraphs above interchangeably with clusters, this link still needs to be studied especially through qualitative assessment of DHCs. Morphometric characters certainly help in description of urban tissues and clustering helps make sense out of it, but one should be aware that DHC is a numerical, morphometric statistical **proxy** of urban tissue, not its definition and replacement. GMM clustering is non-deterministic, so boundaries are not fixed, but rather indicative. It is not a ground truth (there is no ground truth

at all in fact) and the meaning of clusters and relation between them has to be determined and interpreted before any further steps. The one approach how to do so is proposed in the next chapter. # Taxonomy of urban tissues

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

Chapter 8

Synthesis

Appendix 1: Contextual characters

8.1 Interquartile mean

Table 8.1: Overview of the contextual morphometric values of interquartile mean for the whole case study. Key to character IDs is available in table XXX.

	mean	std	min	25%	50%	75%	max
stcOri	18	6.8	0.12	13	18	23	42
sdcLAL	69	25	25	51	64	82	310
sdcAre	2400	2100	180	1100	1700	2900	50000
sscCCo	0.45	0.055	0.14	0.42	0.46	0.49	0.74
sscERI	0.97	0.018	0.86	0.96	0.97	0.98	1.1
stcSAl	9.3	3.7	0.12	6.7	9	12	36
sicCAR	0.19	0.1	0.0022	0.13	0.16	0.21	0.73
sicFAR	0.66	0.7	0.0022	0.25	0.39	0.73	4.4
mtcWNe	0.045	0.013	0.0016	0.036	0.045	0.054	0.15
mdcAre	18000	14000	1600	8700	14000	22000	370000
licGDe	0.57	0.64	0.0022	0.2	0.36	0.65	4.1
ltcWRB	8.8e-05	5.6e-05	2.7e-06	4.6e-05	7.8e-05	0.00012	0.00048
sdbHei	9.9	4.7	3	6.3	8.1	12	37
sdbAre	310	330	33	130	200	360	10000
sdbVol	3800	4600	130	910	2000	5100	150000
sdbPer	68	26	24	49	60	79	460

	mean	std	min	25%	50%	75%	max
sdbCoA	3.3	15	0	0	0	0	340
ssbFoF	1.4	0.3	0.73	1.2	1.4	1.6	4.8
ssbVFR	3.1	0.9	1.4	2.5	2.9	3.7	19
ssbCCo	0.53	0.044	0.27	0.5	0.53	0.56	0.72
ssbCor	9.2	2.7	4	7.5	8.6	10	54
ssbSqu	5.4	3.5	0.016	2.9	4.5	7.1	39
ssbERI	0.93	0.03	0.7	0.92	0.94	0.95	1.1
ssbElo	0.7	0.085	0.23	0.65	0.72	0.76	0.96
ssbCCM	9.8	3.5	3.8	7.3	8.9	11	59
ssbCCD	1.7	0.81	0.00036	1.1	1.5	2	16
stbOri	16	8.8	0.026	9.8	15	22	44
stbSAl	6.8	4.5	0.038	3.3	6.1	9.3	39
stbCeA	7	3	0.058	4.8	6.8	9	27
mtbSWR	0.17	0.12	0	0.077	0.14	0.25	0.75
mtbAli	4.9	2.7	0.005	2.8	4.7	6.7	38
mtbNDi	26	10	0	19	25	32	120
libNCo	0.59	2.7	0	0	0	0.049	51
ldbPWL	180	190	24	69	110	210	2100
ltbIBD	27	9.7	0	21	26	33	120
ltcBuA	0.65	0.22	0.083	0.5	0.69	0.82	1
mtdDeg	3.1	0.44	1	2.9	3.1	3.4	4.8
lcdMes	0.15	0.054	-0.23	0.11	0.15	0.19	0.32
linP3W	0.64	0.1	0	0.57	0.64	0.71	0.93
linP4W	0.23	0.11	0	0.15	0.22	0.3	0.72
linPDE	0.13	0.077	0	0.072	0.11	0.17	1
lcnClo	5.3e-06	2.3e-06	6.8e-08	3.6e-06	5.1e-06	6.7e-06	1.7e-05
ldsCDL	280	310	0	78	190	370	3600
xcnSCl	0.056	0.055	0	0.012	0.046	0.083	0.75
mtdMDi	170	120	36	110	140	190	3300
lddNDe	0.013	0.004	0.0028	0.01	0.012	0.014	0.063
linWID	0.025	0.0079	0	0.02	0.024	0.029	0.11
lddRea	190	71	2	140	190	230	630

	mean	std	min	25%	50%	75%	max
lddARe	370000	270000	39000	230000	300000	410000	3.8e+06
sddAre	30000	39000	3000	12000	19000	32000	660000
midRea	52	22	2	38	49	62	270
midAre	97000	96000	12000	51000	70000	110000	1.2e+06
sdsLen	230	200	36	140	180	260	3300
sdsSPW	29	5	11	26	29	32	50
sdsSPH	10	5.1	0	6.6	8.5	13	38
sdsSPR	0.41	0.27	0	0.24	0.31	0.47	2
sdsSPO	0.58	0.14	0.027	0.5	0.58	0.66	1
sdsSWD	3.6	1.2	0	2.8	3.6	4.4	9.8
sdsSHD	2.3	1.5	0	1.2	1.8	2.9	14
sssLin	0.95	0.077	0	0.93	0.97	0.99	1
sdsAre	31000	46000	2100	10000	17000	32000	740000
sisBpM	0.074	0.031	0.0013	0.056	0.07	0.086	0.79
misRea	44	19	2	32	40	52	230
mdsAre	86000	92000	8600	41000	59000	95000	1.1e+06
ldsMSL	150	67	57	110	140	170	1200
ldsRea	350000	270000	39000	220000	280000	390000	3.8e+06
ldkAre	120000	200000	3200	26000	56000	130000	2e+06
ldkPer	1500	1400	250	720	1100	1800	13000
lskCCo	0.43	0.09	0.13	0.37	0.43	0.5	0.98
lskERI	0.86	0.096	0.36	0.81	0.88	0.93	1.1
lskCWA	370	370	0.43	140	240	450	3100
ltkOri	18	9.4	0.034	10	17	25	45
ltkWNB	0.0074	0.0035	0	0.0047	0.0071	0.0097	0.025
likWBB	0.00088	0.00051	3e-05	0.00051	0.0008	0.0012	0.004

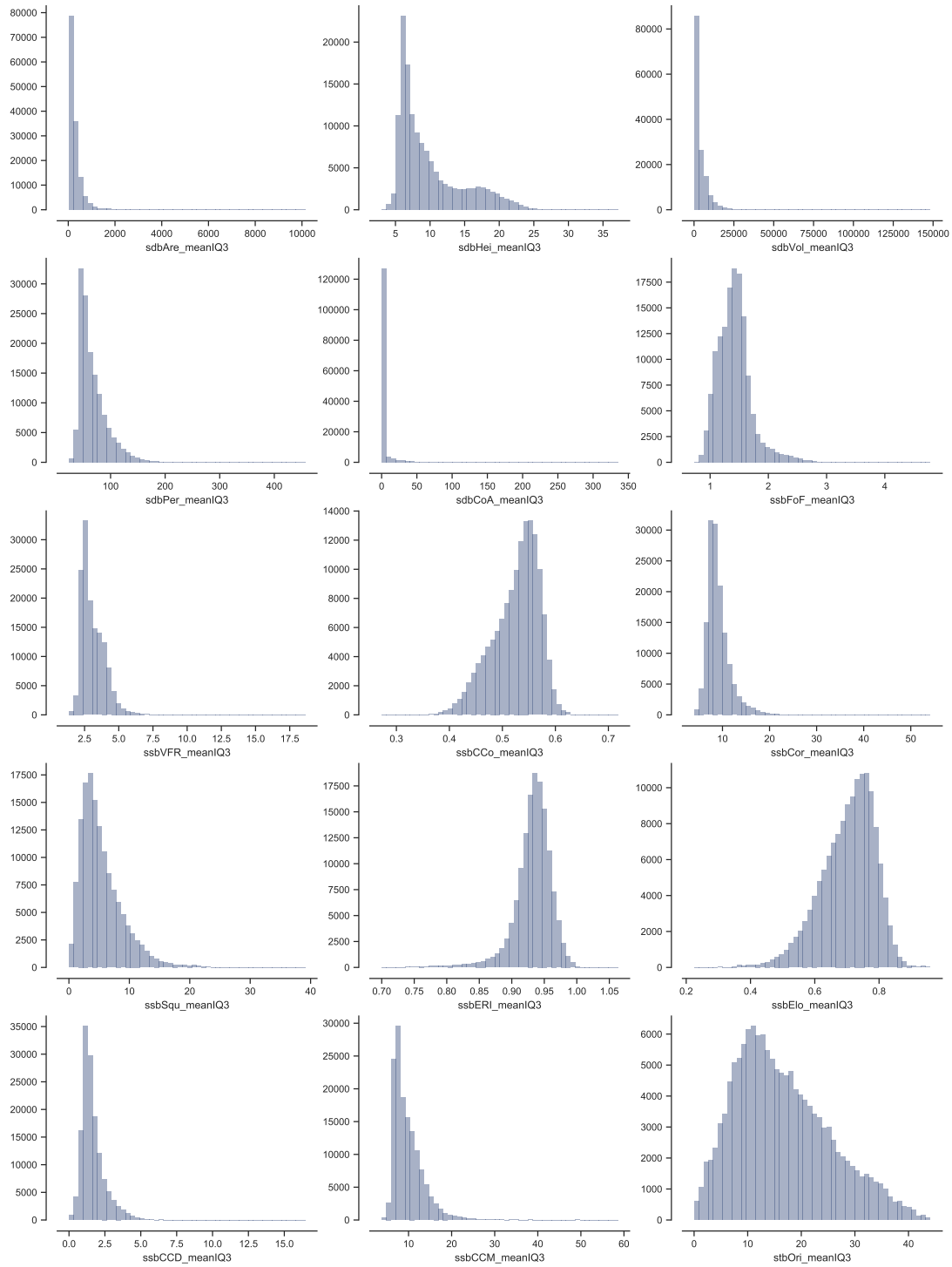

Figure 8.1: Histograms of interquartile mean for characters 1-15 are showing the variety of distributions within the measured contextual data.

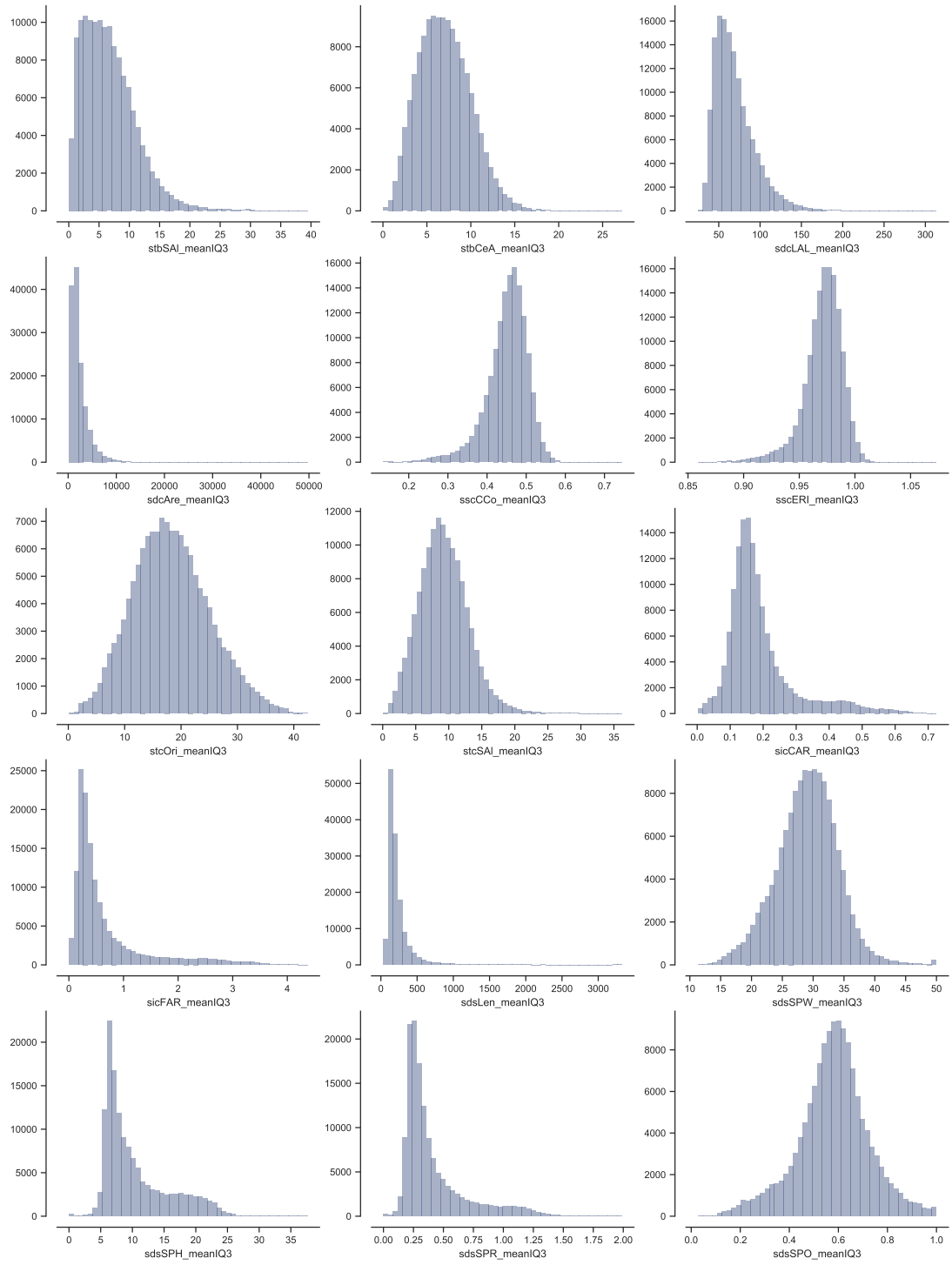

Figure 8.2: Histograms of interquartile mean for characters 16-30 are showing the variety of distributions within the measured contextual data.

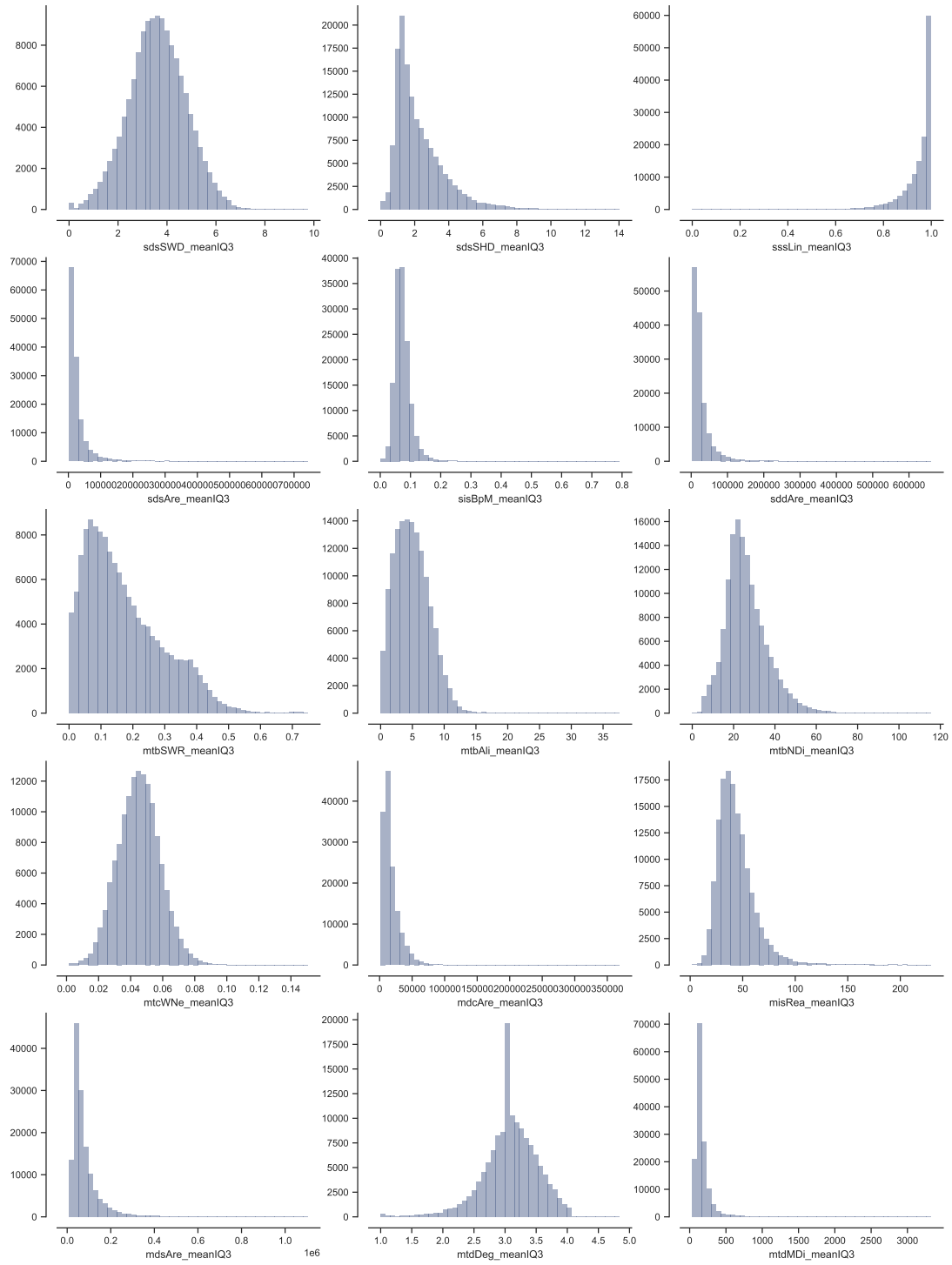

Figure 8.3: Histograms of interquartile mean for characters 31-45 are showing the variety of distributions within the measured contextual data.

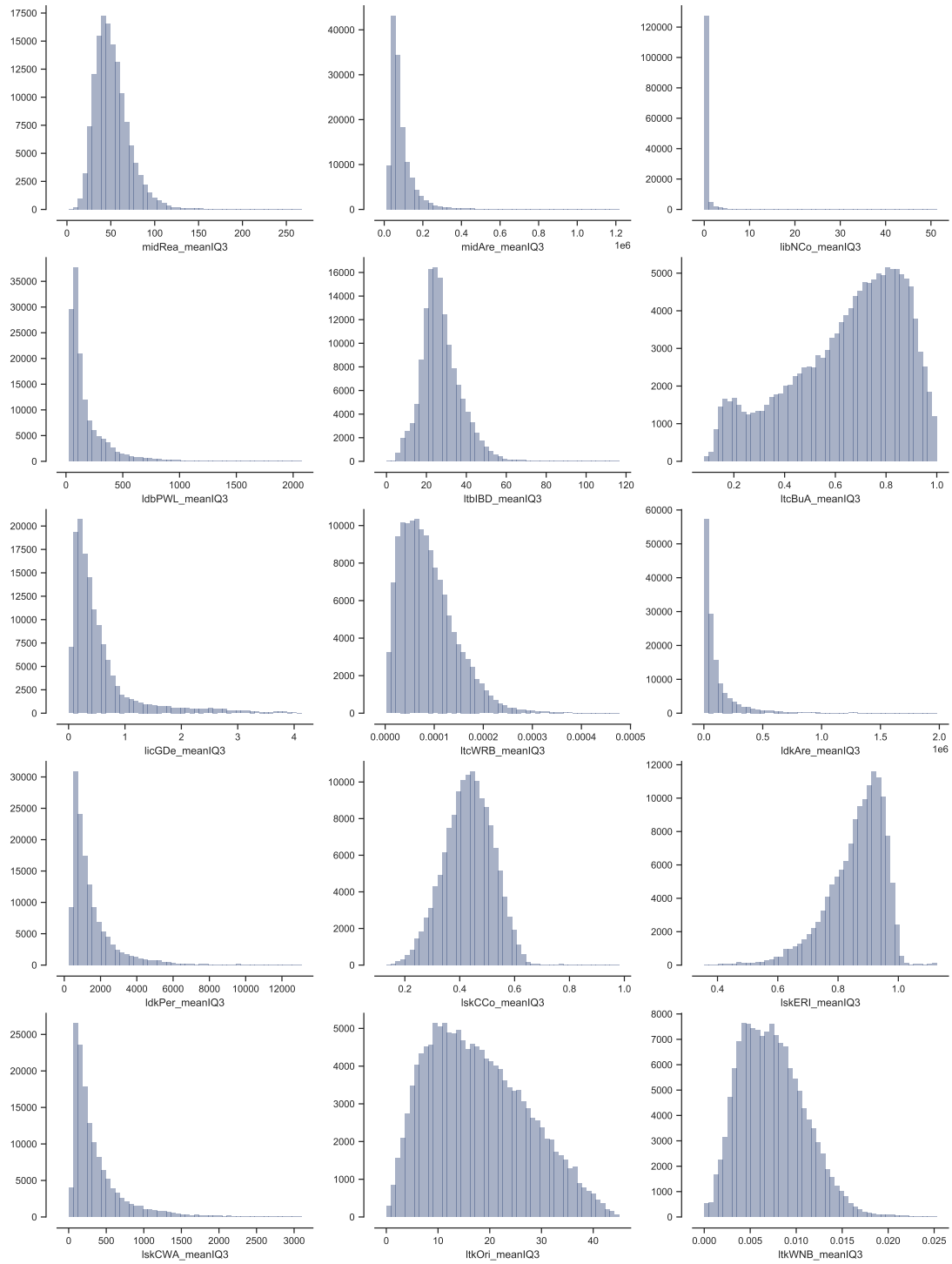

Figure 8.4: Histograms of interquartile mean for characters 46-60 are showing the variety of distributions within the measured contextual data.

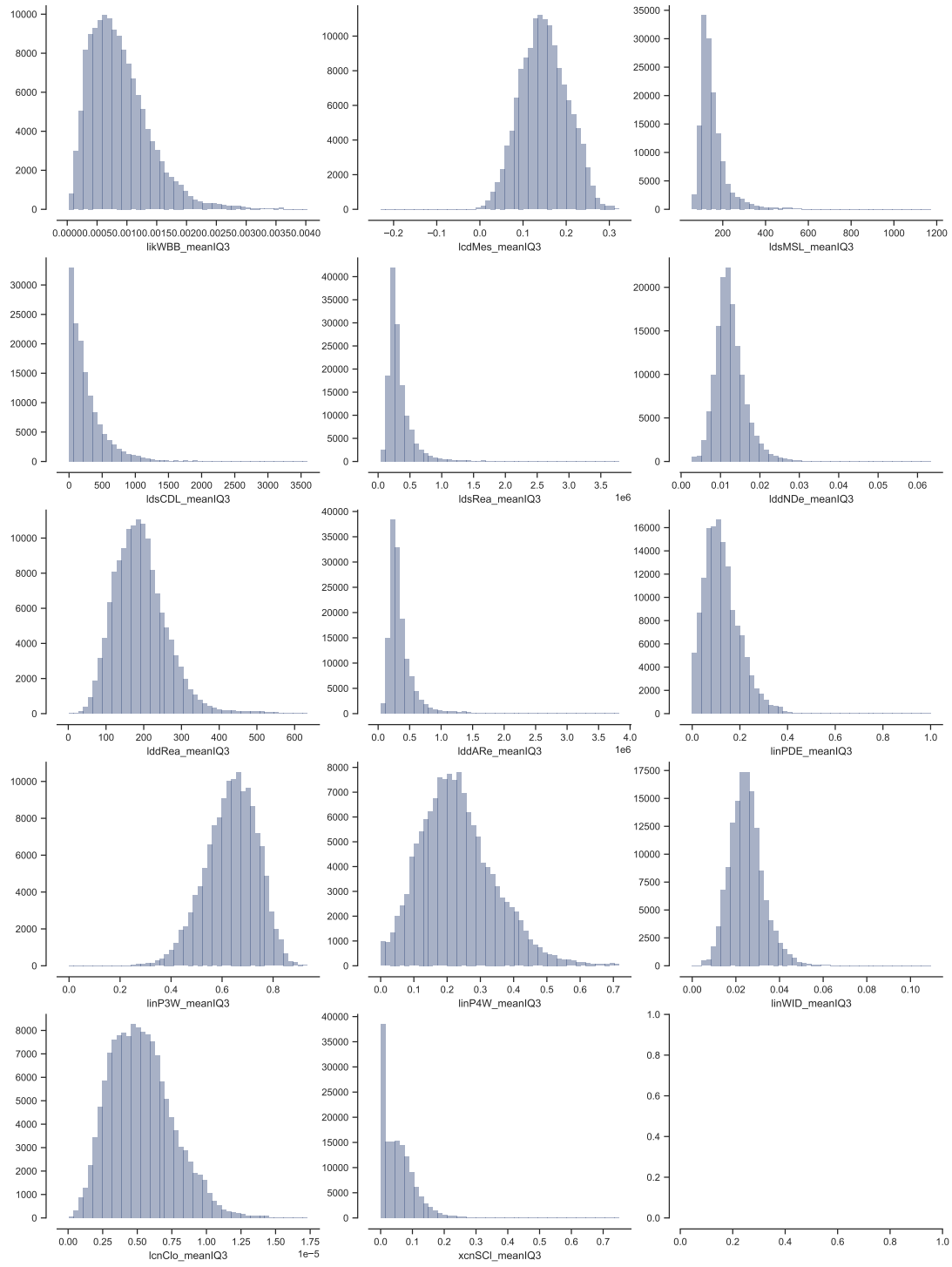

Figure 8.5: Histograms of interquartile mean for characters 61-74 are showing the variety of distributions within the measured contextual data.

8.2 Interquartile range

Table 8.2: Overview of the contextual morphometric values of interquartile range for the whole case study. Key to character IDs is available in table XXX.

	mean	std	min	25%	50%	75%	max
stcOri	14	7.4	0.013	8.8	14	19	45
sdcLAL	36	23	0.045	17	30	51	160
sdcAre	1800	2000	3.7	590	1100	2200	47000
sscCCo	0.19	0.052	0.00011	0.15	0.18	0.22	0.5
sscERI	0.069	0.022	0.00024	0.054	0.067	0.081	0.27
stcSAl	11	5.4	0.013	7.2	10	14	41
sicCAR	0.13	0.062	2e-05	0.083	0.11	0.15	0.62
sicFAR	0.56	0.56	2e-05	0.19	0.32	0.73	3.8
mtcWNe	0.021	0.0089	7.8e-07	0.015	0.02	0.025	0.15
mdcAre	13000	13000	0	4500	9300	18000	290000
licGDe	0.19	0.22	0	0.062	0.12	0.24	2.4
ltcWRB	4e-05	3.1e-05	0	1.8e-05	3.3e-05	5.4e-05	0.00038
sdbHei	5	4.9	0	2	3.1	6.1	42
sdbAre	170	250	0.097	57	96	180	17000
sdbVol	2400	3300	2.1	480	1100	3500	170000
sdbPer	28	26	0.018	13	20	32	420
sdbCoA	0.096	2	0	0	0	0	160
ssbFoF	0.47	0.29	0.00044	0.28	0.39	0.58	4.2
ssbVFR	1.1	0.84	0.0029	0.58	0.86	1.4	21
ssbCCo	0.13	0.054	0.00027	0.086	0.12	0.16	0.37
ssbCor	5.6	2.9	0	4	5.2	6.8	76
ssbSqu	5.7	6	0.001	1.1	2.4	9.9	45
ssbERI	0.087	0.046	1.6e-05	0.059	0.079	0.1	0.44
ssbElo	0.26	0.09	0.00025	0.2	0.26	0.32	0.69
ssbCCM	3.7	3.7	0.0029	1.6	2.5	4.2	57
ssbCCD	1.7	1.1	0.00014	1	1.5	2.1	22
stbOri	9.8	8.8	7.3e-09	2.3	7.6	15	45
stbSAl	7.9	6.7	2.5e-08	2.5	6.4	11	42

	mean	std	min	25%	50%	75%	max
stbCeA	8.8	4.7	0.015	5.3	8.5	12	43
mtbSWR	0.2	0.13	0	0.12	0.21	0.28	0.9
mtbAli	4.7	2.8	0	2.6	4.6	6.5	24
mtbNDi	15	9.4	0	8.6	13	20	82
libNCo	0.62	3.2	0	0	0	0	52
ldbPWL	130	220	0	28	62	140	3200
ltbIBD	6.6	4.2	0	3.5	5.7	8.7	49
ltcBuA	0.1	0.061	0	0.057	0.085	0.13	0.62
mtdDeg	0.6	0.73	0	0	0	1	4
lcdMes	0.029	0.024	0	0.013	0.024	0.038	0.47
linP3W	0.051	0.046	0	0.024	0.043	0.067	0.8
linP4W	0.044	0.037	0	0.02	0.036	0.059	0.46
linPDE	0.039	0.045	0	0.016	0.03	0.05	0.97
lcnClo	1.2e-06	9.1e-07	0	5.5e-07	1e-06	1.7e-06	7e-06
ldsCDL	200	290	0	11	110	260	4100
xcnSCl	0.056	0.074	0	0	0.04	0.08	1
mtdMDi	74	98	0	21	43	87	1500
lddNDe	0.0028	0.0038	0	0.00098	0.0019	0.0035	0.096
linWID	0.0055	0.0065	0	0.002	0.0039	0.0069	0.15
lddRea	56	44	0	25	46	76	480
lddARe	140000	180000	0	42000	89000	170000	3.7e+06
sddAre	20000	31000	0	4400	9700	22000	590000
midRea	22	17	0	10	18	29	220
midAre	49000	66000	0	14000	28000	56000	900000
sdsLen	140	190	0	44	85	160	2900
sdsSPW	8.9	5.3	0	4.8	8.8	13	36
sdsSPH	2.9	3.5	0	0.77	1.6	3.7	38
sdsSPR	0.18	0.17	0	0.063	0.12	0.23	2.2
sdsSPO	0.19	0.11	0	0.11	0.18	0.26	0.91
sdsSWD	2.3	1.3	0	1.3	2.2	3.1	8.7
sdsSHD	1.7	2	0	0.53	1	2.1	20
sssLin	0.073	0.13	0	0	0.011	0.09	1

	mean	std	min	25%	50%	75%	max
sdsAre	23000	40000	0	4900	11000	25000	710000
sisBpM	0.039	0.029	0	0.021	0.034	0.051	0.92
misRea	19	16	0	9	16	25	180
mdsAre	46000	67000	0	12000	25000	53000	970000
ldsMSL	31	43	0	8	17	36	1100
ldsRea	130000	170000	0	40000	80000	160000	2.9e+06
ldkAre	98000	200000	0	5300	21000	96000	2e+06
ldkPer	940	1500	0	98	340	1100	13000
lskCCo	0.12	0.095	0	0.042	0.1	0.18	0.56
lskERI	0.11	0.11	0	0.031	0.08	0.16	0.65
lskCWA	270	410	0	30	100	340	3000
ltkOri	9.9	9.5	0	2.3	6.9	15	45
ltkWNB	0.0028	0.0024	0	0.00085	0.0023	0.0041	0.019
likWBB	0.00049	0.00048	0	0.00014	0.00037	0.0007	0.0053

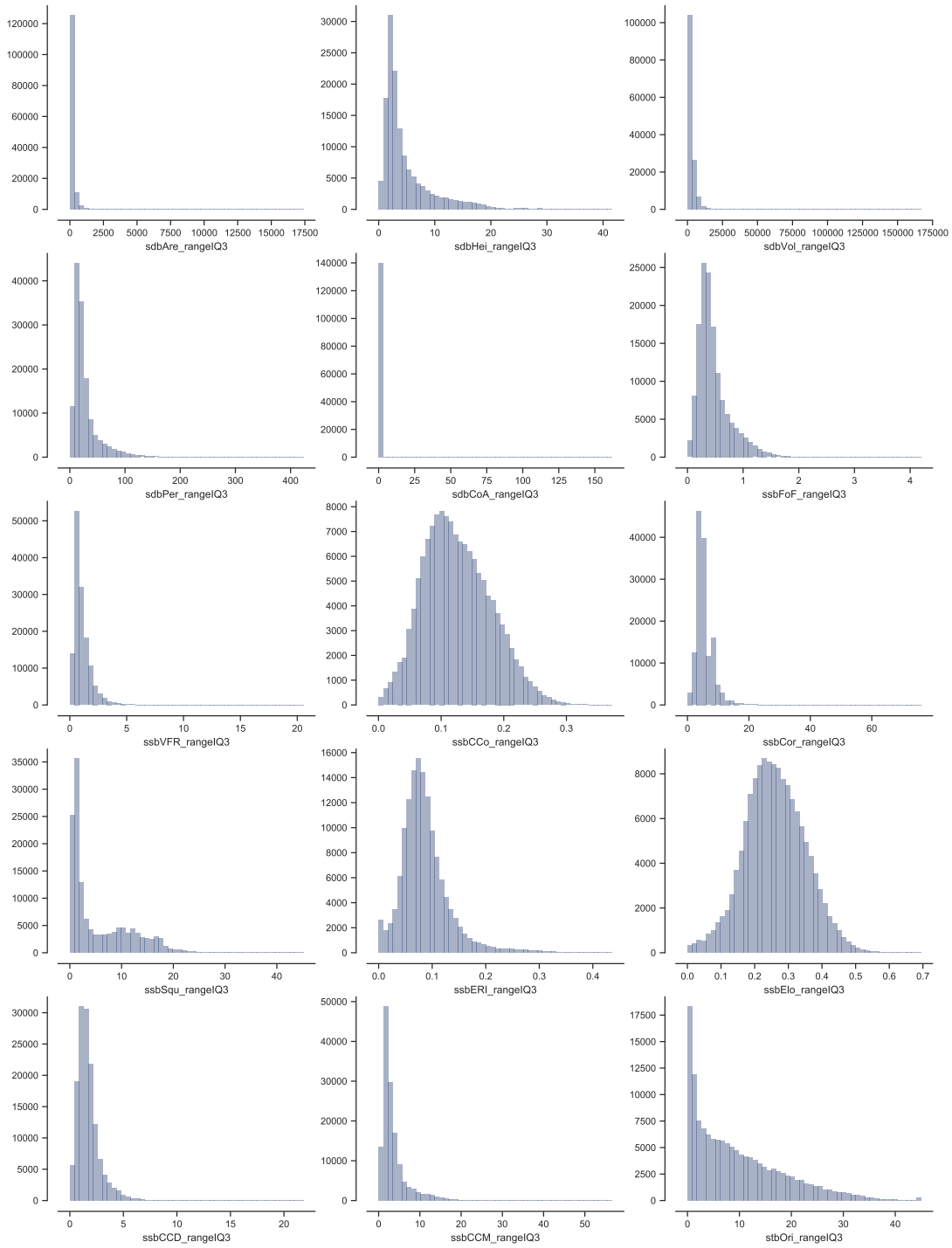

Figure 8.6: Histograms of interquartile range for characters 1-15 are showing the variety of distributions within the measured contextual data.

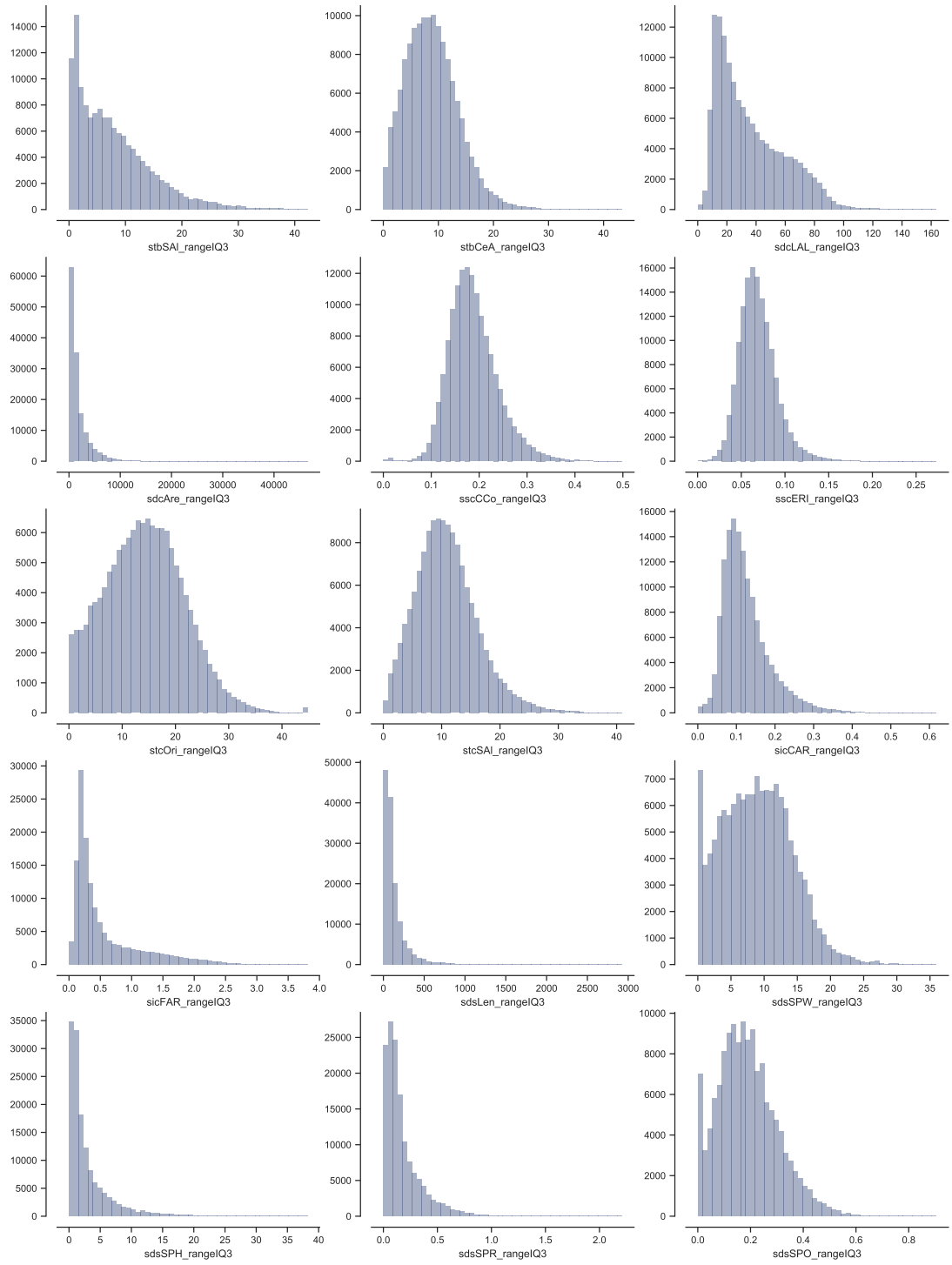

Figure 8.7: Histograms of interquartile range for characters 16-30 are showing the variety of distributions within the measured contextual data.

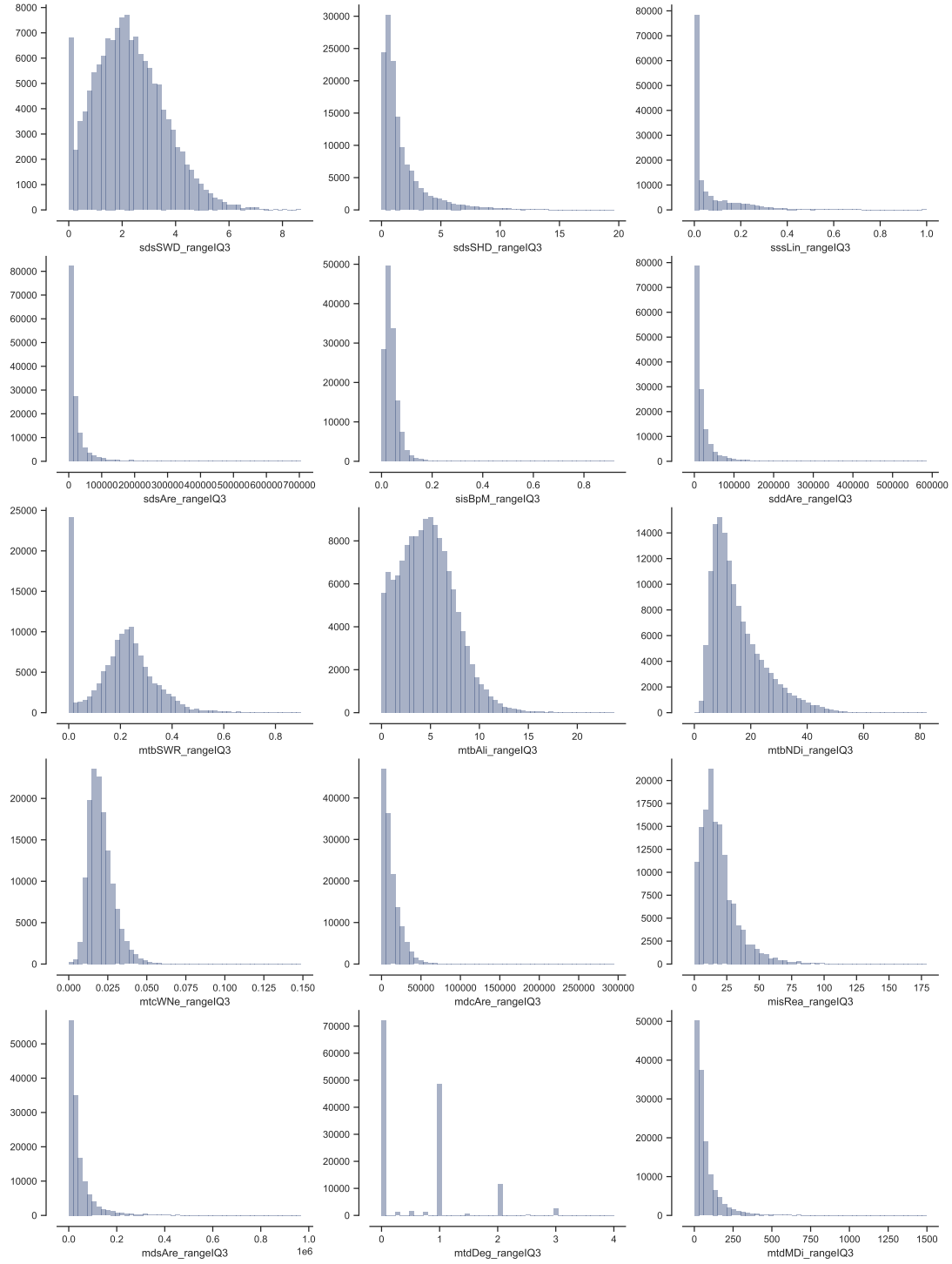

Figure 8.8: Histograms of interquartile range for characters 31-45 are showing the variety of distributions within the measured contextual data.

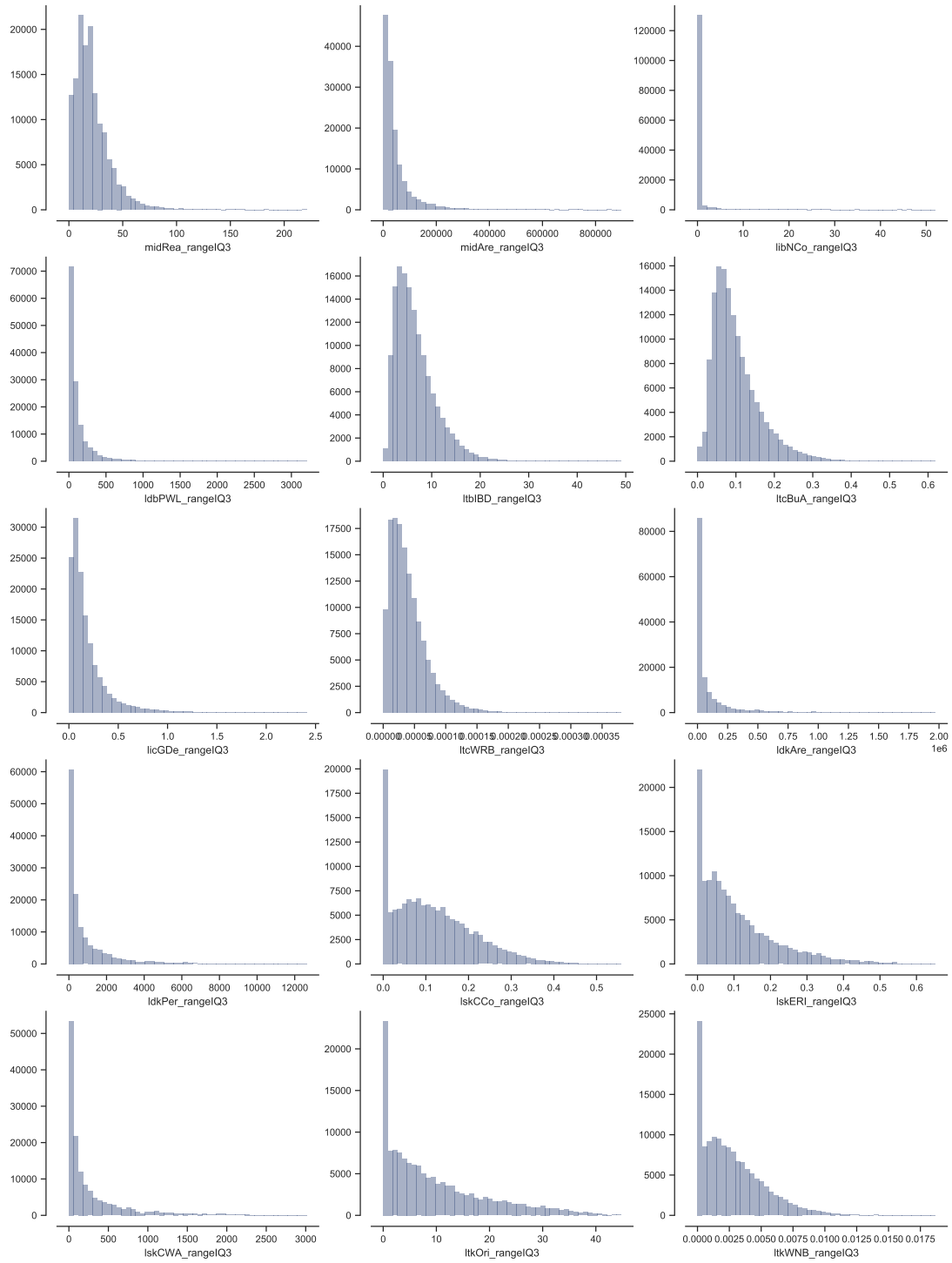

Figure 8.9: Histograms of interquartile range for characters 46-60 are showing the variety of distributions within the measured contextual data.

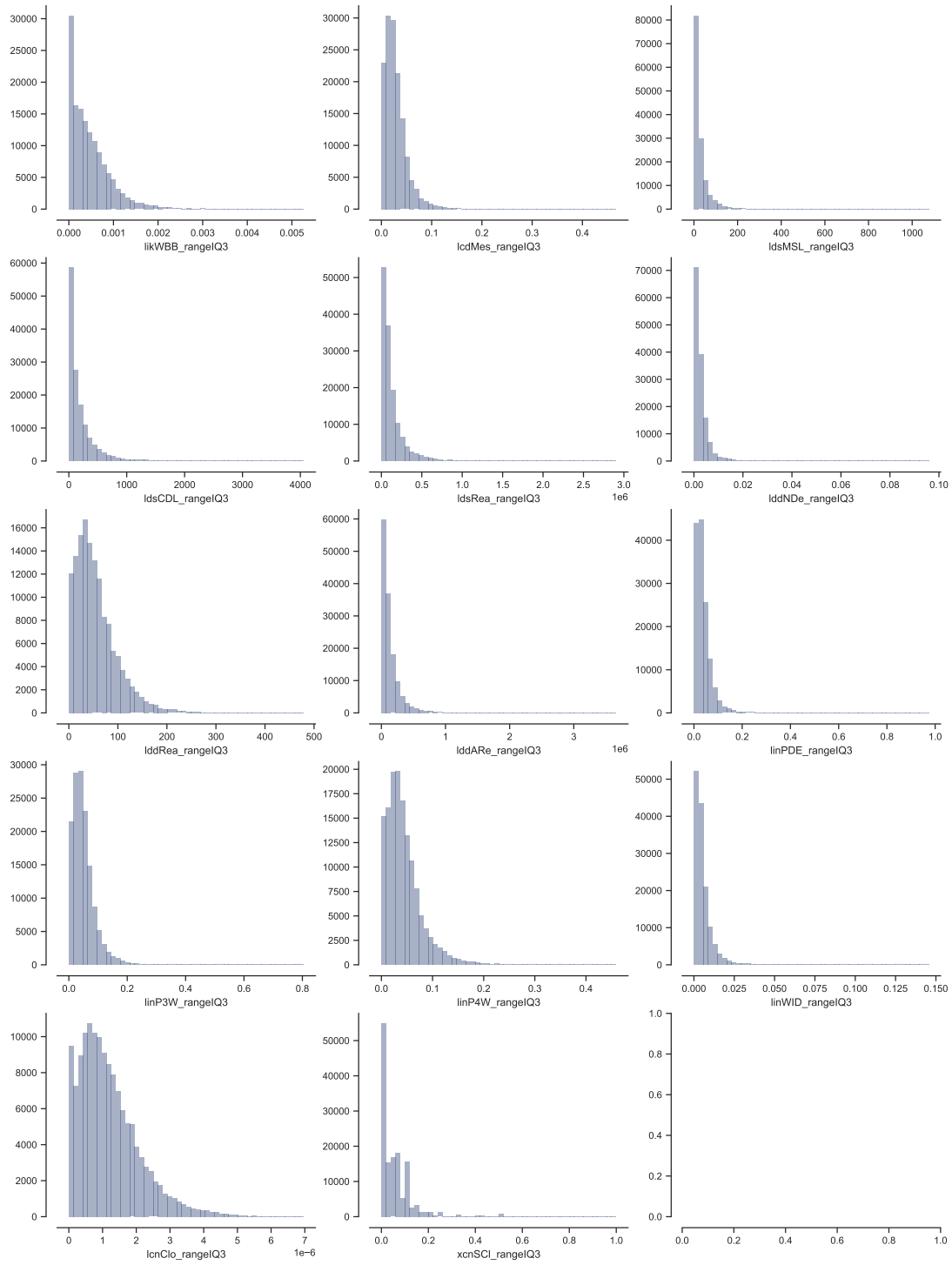

Figure 8.10: Histograms of interquartile range for characters 61-74 are showing the variety of distributions within the measured contextual data.

8.3 Interdecile Theil index

Table 8.3: Overview of the contextual morphometric values of interdecile Theil index for the whole case study. Key to character IDs is available in table XXX.

	mean	std	min	25%	50%	75%	max
stcOri	0.14	0.13	8.3e-08	0.059	0.11	0.19	2.2
sdcLAL	0.046	0.033	2e-08	0.02	0.039	0.066	0.27
sdcAre	0.18	0.12	1.7e-07	0.085	0.15	0.24	1.6
sscCCo	0.027	0.022	1.3e-08	0.014	0.021	0.032	0.41
sscERI	0.00078	0.0005	2.5e-08	0.00047	0.00067	0.00096	0.023
stcSAl	0.32	0.16	1e-06	0.21	0.29	0.4	2.4
sicCAR	0.093	0.072	1e-06	0.043	0.073	0.12	0.78
sicFAR	0.16	0.11	1e-06	0.084	0.14	0.22	1.2
mtcWNe	0.037	0.029	1.2e-07	0.017	0.029	0.049	0.39
mdcAre	0.11	0.082	0	0.048	0.084	0.14	1
licGDe	0.026	0.032	-1.1e-16	0.0073	0.017	0.034	0.75
ltcWRB	0.038	0.036	-2.2e-16	0.014	0.026	0.049	0.47
sdbHei	0.044	0.046	0	0.015	0.027	0.056	0.51
sdbAre	0.12	0.12	2.5e-07	0.04	0.073	0.15	1.5
sdbVol	0.18	0.17	2.5e-07	0.064	0.13	0.25	2.1
sdbPer	0.04	0.042	1.6e-08	0.014	0.024	0.05	0.47
sdbCoA	0.041	0.34	-1.1e-16	0	0	0	4.7
ssbFoF	0.023	0.025	2.8e-08	0.0075	0.013	0.029	0.34
ssbVFR	0.024	0.025	2.2e-07	0.0093	0.016	0.03	0.39
ssbCCo	0.01	0.0082	1.3e-07	0.004	0.0077	0.014	0.097
ssbCor	0.087	0.045	-1.1e-16	0.06	0.079	0.1	0.8
ssbSqu	0.54	0.27	0.00012	0.33	0.53	0.71	2.2
ssbERI	0.0017	0.0021	4.2e-11	0.00063	0.0011	0.0019	0.032
ssbElo	0.022	0.016	5.9e-07	0.01	0.018	0.03	0.19
ssbCCM	0.031	0.036	9.4e-09	0.0087	0.017	0.038	0.41
ssbCCD	0.38	0.28	3.4e-07	0.19	0.3	0.48	2.7
stbOri	0.13	0.18	4.4e-09	0.017	0.075	0.17	3.6
stbSAl	0.4	0.24	6.1e-07	0.23	0.35	0.52	3

	mean	std	min	25%	50%	75%	max
stbCeA	0.48	0.22	6.3e-07	0.33	0.45	0.6	2.4
mtbSWR	0.88	0.78	-1.1e-16	0.3	0.69	1.3	4.3
mtbAli	0.22	0.17	0	0.11	0.18	0.29	2.3
mtbNDi	0.065	0.051	0	0.031	0.052	0.085	0.67
libNCo	0.14	0.43	-1.1e-16	0	0	0	4.6
ldbPWL	0.11	0.11	-2.2e-16	0.03	0.07	0.15	1.2
ltbIBD	0.01	0.012	-2.2e-16	0.0031	0.0066	0.013	0.28
ltcBuA	0.0073	0.011	0	0.001	0.0029	0.0085	0.15
mtdDeg	0.029	0.04	-1.1e-16	0.005	0.0095	0.049	0.27
lcdMes	0.024	0.096	-5.6e-16	0.0021	0.0056	0.015	3.3
linP3W	0.0038	0.044	-6.7e-16	0.00042	0.00095	0.002	3.9
linP4W	0.038	0.16	-6.7e-16	0.0022	0.0058	0.016	3.9
linPDE	0.073	0.22	-4.4e-16	0.0065	0.017	0.046	3.9
lcnClo	0.02	0.041	-6.7e-16	0.0031	0.0088	0.023	1.6
ldsCDL	0.38	0.51	-6.7e-16	0.038	0.17	0.51	4.3
xcnSCl	0.54	0.65	-4.4e-16	0.008	0.31	0.84	4.7
mtdMDi	0.045	0.06	-5.6e-16	0.0099	0.025	0.057	0.96
lddNDe	0.017	0.044	-6.7e-16	0.0022	0.0058	0.015	1.5
linWID	0.017	0.056	-5.6e-16	0.0024	0.0061	0.015	3.9
lddRea	0.029	0.039	-1.1e-16	0.0052	0.014	0.037	0.94
lddARe	0.043	0.06	-5.6e-16	0.0079	0.022	0.053	0.97
sddAre	0.11	0.11	-6.7e-16	0.036	0.078	0.15	1.2
midRea	0.046	0.047	-1.1e-16	0.015	0.032	0.063	0.71
midAre	0.067	0.077	-6.7e-16	0.018	0.042	0.088	1.1
sdsLen	0.083	0.09	-6.7e-16	0.027	0.056	0.11	1.1
sdsSPW	0.022	0.019	-5.6e-16	0.0087	0.018	0.03	0.25
sdsSPH	0.03	0.082	-5.6e-16	0.0026	0.0076	0.025	3
sdsSPR	0.049	0.084	-6.7e-16	0.013	0.028	0.054	3.7
sdsSPO	0.033	0.049	-6.7e-16	0.0092	0.02	0.039	1.5
sdsSWD	0.13	0.15	-6.7e-16	0.034	0.084	0.18	3.7
sdsSHD	0.16	0.19	-5.6e-16	0.038	0.091	0.21	3.7
sssLin	0.009	0.044	-5.6e-16	0	0.00033	0.0043	2

	mean	std	min	25%	50%	75%	max
sdsAre	0.15	0.14	-6.7e-16	0.059	0.11	0.2	1.5
sisBpM	0.061	0.056	-6.7e-16	0.025	0.047	0.08	1.1
misRea	0.049	0.049	-1.1e-16	0.016	0.034	0.067	0.6
mdsAre	0.079	0.089	-5.6e-16	0.021	0.049	0.1	1
ldsMSL	0.012	0.025	-5.6e-16	0.0012	0.0039	0.011	0.35
ldsRea	0.042	0.061	-5.6e-16	0.0072	0.021	0.051	0.96
ldkAre	0.22	0.25	-6.7e-16	0.04	0.13	0.32	1.9
ldkPer	0.1	0.13	-6.7e-16	0.013	0.053	0.15	1
lskCCo	0.023	0.028	-6.7e-16	0.0048	0.014	0.032	0.31
lskERI	0.0063	0.011	-6.7e-16	0.00058	0.0022	0.0072	0.12
lskCWA	0.16	0.18	-6.7e-16	0.034	0.1	0.23	1.4
ltkOri	0.14	0.18	-6.7e-16	0.018	0.078	0.2	2.6
ltkWNB	0.041	0.047	-6.7e-16	0.0092	0.026	0.058	0.54
likWBB	0.081	0.082	-6.7e-16	0.019	0.057	0.12	0.84

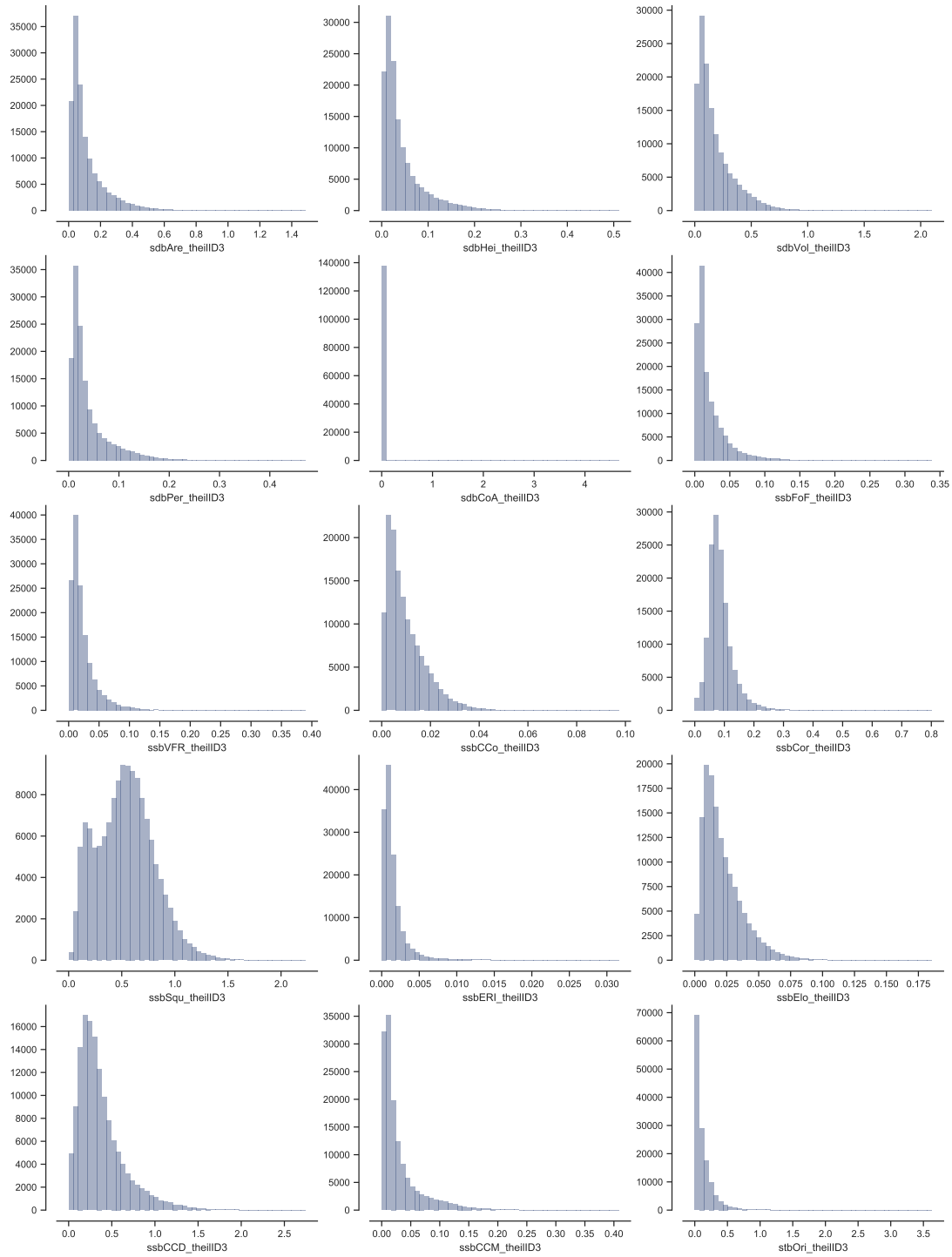

Figure 8.11: Histograms of interdecile Theil index for characters 1-15 are showing the variety of distributions within the measured contextual data.

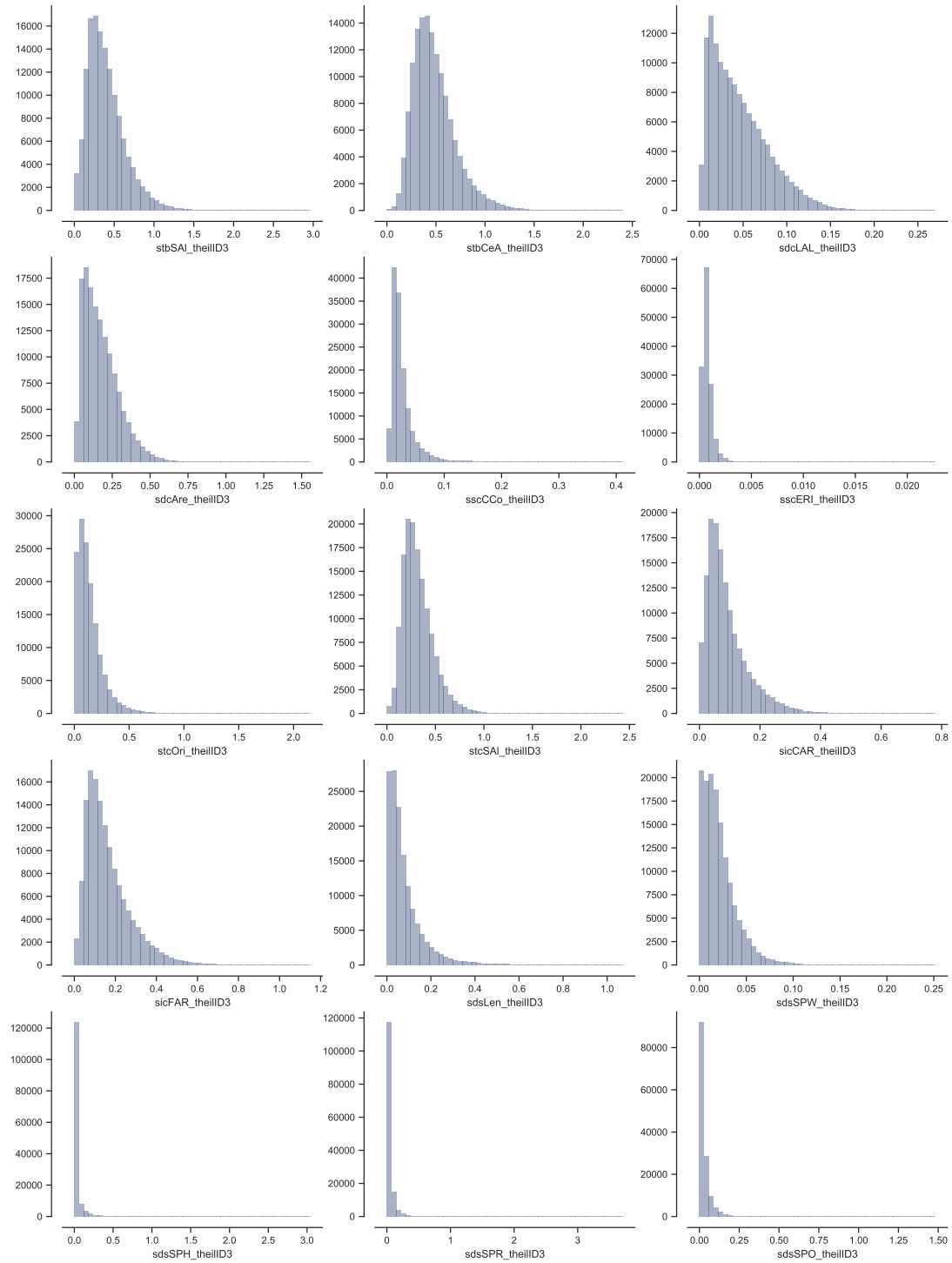

Figure 8.12: Histograms of interdecile Theil index for characters 16-30 are showing the variety of distributions within the measured contextual data.

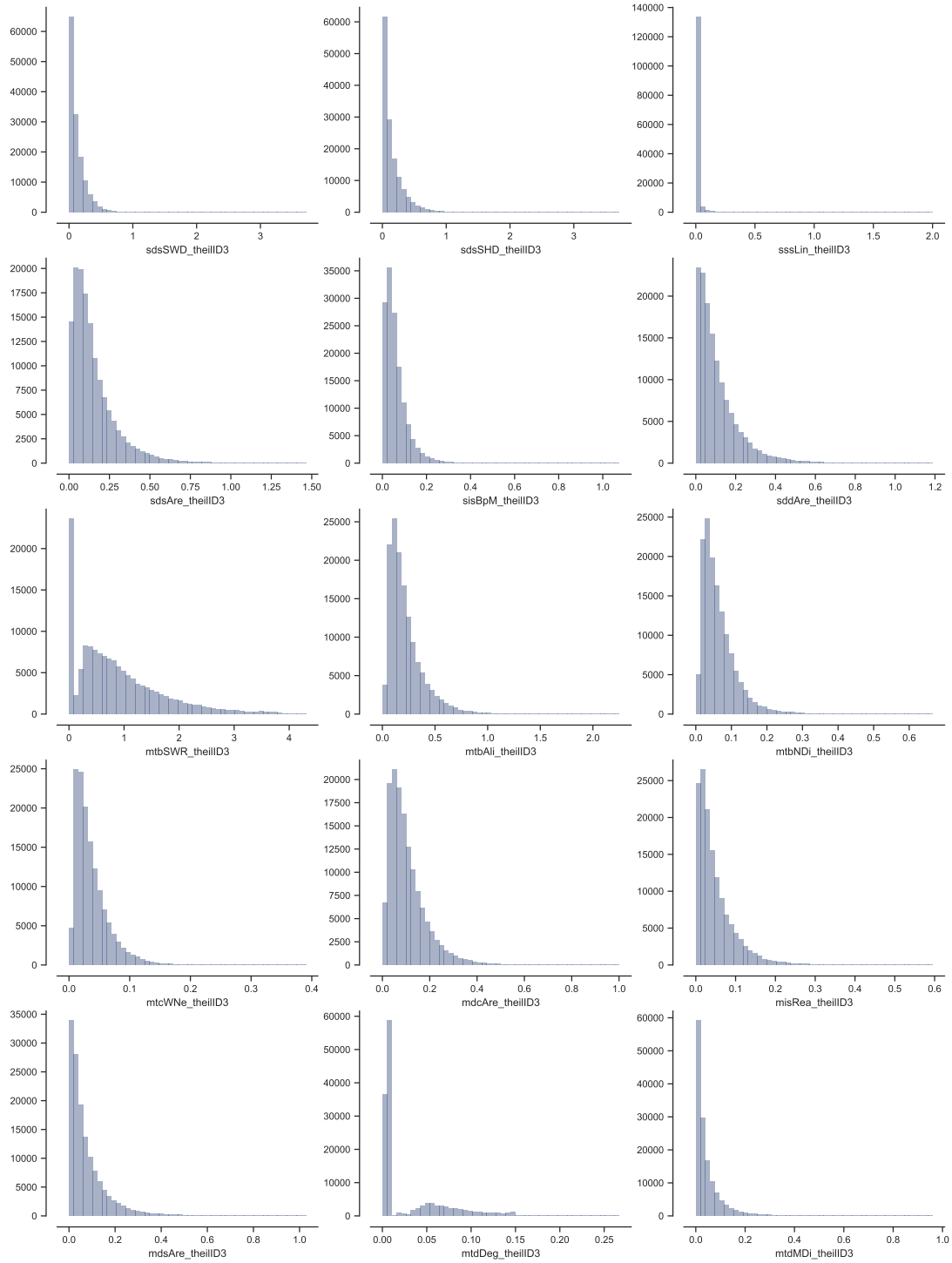

Figure 8.13: Histograms of interdecile Theil index for characters 31-45 are showing the variety of distributions within the measured contextual data.

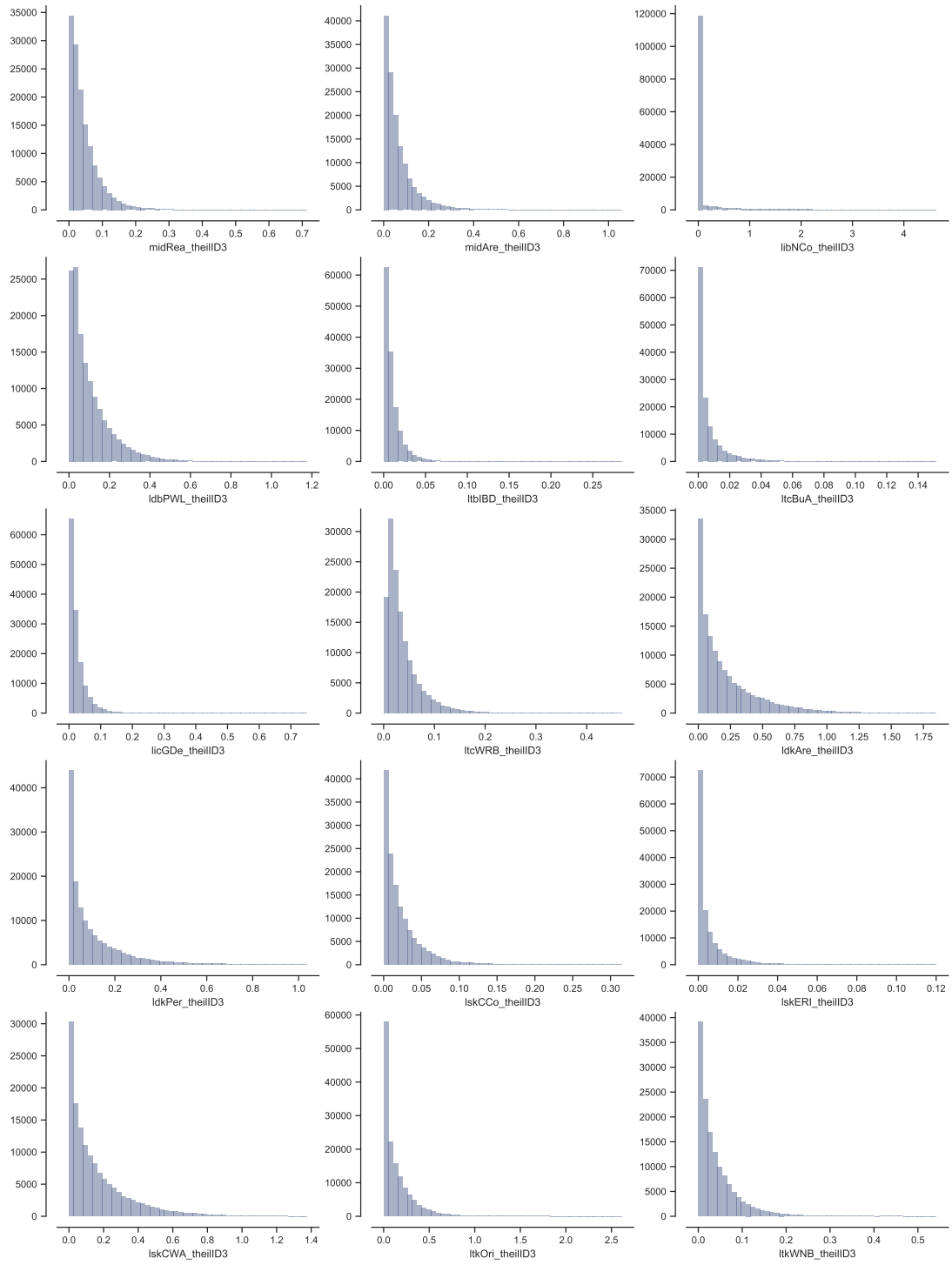

Figure 8.14: Histograms of interdecile Theil index for characters 46-60 are showing the variety of distributions within the measured contextual data.

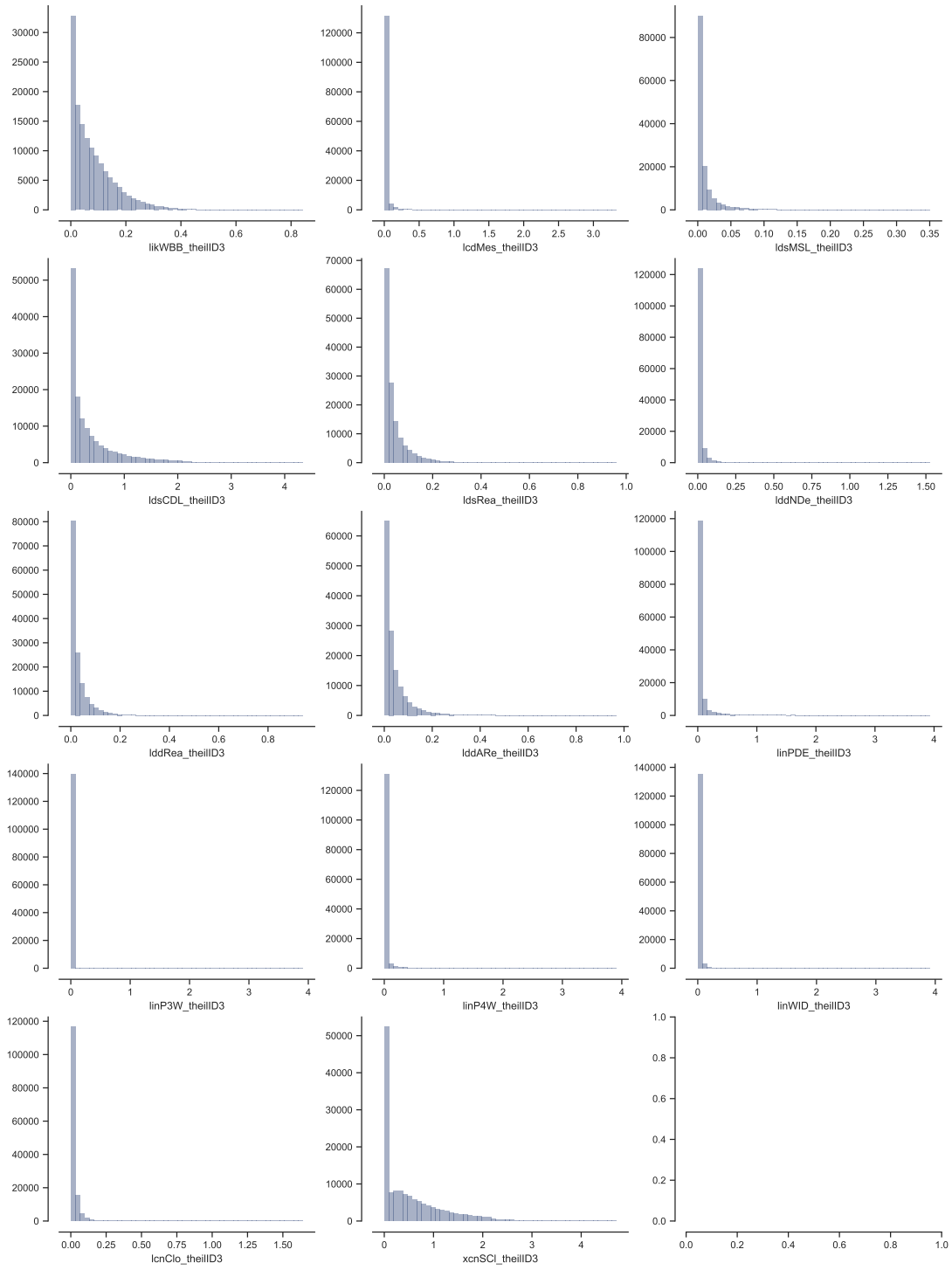

Figure 8.15: Histograms of interdecile Theil index for characters 61-74 are showing the variety of distributions within the measured contextual data.

8.4 Simpson index

Table 8.4: Overview of the contextual morphometric values of Simpson index for the whole case study. Key to character IDs is available in table XXX.

	mean	std	min	25%	50%	75%	max
sdcLAL	0.57	0.22	0.15	0.39	0.54	0.74	1
sdcAre	0.67	0.22	0.15	0.49	0.67	0.85	1
stcSAl	0.5	0.16	0.14	0.38	0.47	0.6	1
sicCAR	0.58	0.21	0.13	0.42	0.56	0.73	1
sicFAR	0.7	0.26	0.13	0.5	0.75	0.95	1
mdcAre	0.62	0.24	0.12	0.42	0.6	0.85	1
licGDe	0.83	0.22	0.12	0.62	1	1	1
ltcWRB	0.67	0.26	0.1	0.44	0.61	1	1
sdbHei	0.68	0.25	0.13	0.46	0.67	0.94	1
sdbAre	0.71	0.22	0.16	0.51	0.73	0.91	1
sdbVol	0.73	0.23	0.17	0.51	0.78	0.96	1
sdbPer	0.61	0.21	0.17	0.44	0.6	0.79	1
sdbCoA	0.99	0.048	0.43	1	1	1	1
ssbFoF	0.52	0.18	0.11	0.4	0.49	0.64	1
ssbVFR	0.59	0.21	0.14	0.41	0.56	0.78	1
ssbCor	0.53	0.14	0.16	0.43	0.52	0.62	1
ssbSqu	0.63	0.18	0.16	0.5	0.64	0.77	1
ssbCCM	0.62	0.23	0.14	0.44	0.6	0.83	1
ssbCCD	0.53	0.15	0.18	0.42	0.51	0.62	1
stbSAl	0.59	0.21	0.13	0.42	0.55	0.76	1
stbCeA	0.53	0.16	0.15	0.4	0.5	0.63	1
mtbAli	0.54	0.23	0.12	0.35	0.48	0.7	1
mtbNDi	0.53	0.21	0.13	0.37	0.49	0.66	1
libNCo	0.92	0.16	0.2	0.94	1	1	1
ldbPWL	0.74	0.23	0.15	0.53	0.78	1	1
linPDE	0.75	0.23	0.18	0.53	0.77	1	1
ldsCDL	0.75	0.23	0.18	0.53	0.77	1	1
xcnSCl	0.66	0.23	0.2	0.47	0.6	0.9	1

	mean	std	min	25%	50%	75%	max
mtdMDi	0.74	0.22	0.22	0.53	0.76	1	1
lddNDe	0.7	0.23	0.19	0.51	0.68	0.97	1
linWID	0.7	0.23	0.17	0.5	0.66	0.96	1
lddARe	0.76	0.22	0.21	0.54	0.8	1	1
sddAre	0.78	0.22	0.21	0.56	0.83	1	1
midRea	0.65	0.22	0.17	0.49	0.59	0.85	1
midAre	0.77	0.22	0.17	0.56	0.82	1	1
sdsLen	0.73	0.22	0.2	0.53	0.72	1	1
sdsSPH	0.75	0.25	0.15	0.52	0.82	1	1
sdsSPR	0.74	0.25	0.15	0.51	0.78	1	1
sdsSHD	0.69	0.25	0.15	0.48	0.66	1	1
sdsAre	0.78	0.22	0.2	0.57	0.83	1	1
sisBpM	0.6	0.21	0.2	0.45	0.54	0.74	1
misRea	0.65	0.22	0.16	0.49	0.59	0.85	1
mdsAre	0.78	0.22	0.17	0.56	0.83	1	1
ldsMSL	0.79	0.22	0.2	0.58	0.87	1	1
ldsRea	0.77	0.22	0.17	0.56	0.82	1	1
ldkAre	0.82	0.21	0	0.61	0.94	1	1
ldkPer	0.78	0.21	0	0.57	0.82	1	1
lskCWA	0.77	0.21	0.25	0.56	0.8	1	1
likWBB	0.67	0.24	0.16	0.5	0.61	0.95	1
stcOri	0.3	0.13	0.17	0.21	0.25	0.34	1
sscCCo	0.2	0.05	0.14	0.18	0.19	0.22	1
sscERI	0.2	0.041	0.13	0.18	0.2	0.22	1
mtcWNe	0.27	0.093	0.13	0.21	0.25	0.3	1
ssbCCo	0.24	0.081	0.14	0.19	0.22	0.28	1
ssbERI	0.32	0.11	0.14	0.25	0.29	0.36	1
ssbElo	0.24	0.075	0.17	0.2	0.23	0.27	1
stbOri	0.47	0.22	0.17	0.3	0.41	0.59	1
mtbSWR	0.49	0.17	0.25	0.36	0.44	0.58	1
ltbIBD	0.45	0.19	0.15	0.31	0.4	0.52	1
ltcBuA	0.56	0.21	0.17	0.41	0.51	0.67	1

	mean	std	min	25%	50%	75%	max
mtdDeg	0.63	0.18	0.33	0.5	0.59	0.76	1
lcdMes	0.55	0.19	0.16	0.41	0.51	0.66	1
linP3W	0.54	0.19	0.13	0.4	0.5	0.63	1
linP4W	0.59	0.2	0.16	0.45	0.54	0.72	1
lcnClo	0.52	0.2	0.16	0.37	0.48	0.62	1
lddRea	0.47	0.18	0.15	0.34	0.43	0.55	1
sdsSPW	0.38	0.15	0.15	0.27	0.33	0.43	1
sdsSPO	0.38	0.16	0.15	0.27	0.34	0.43	1
sdsSWD	0.37	0.15	0.17	0.27	0.33	0.42	1
sssLin	0.74	0.22	0.2	0.54	0.75	1	1
lskCCo	0.47	0.19	0.15	0.33	0.42	0.54	1
lskERI	0.49	0.19	0.16	0.36	0.45	0.57	1
ltkOri	0.53	0.21	0.17	0.37	0.48	0.63	1
ltkWNB	0.53	0.22	0.15	0.36	0.47	0.64	1

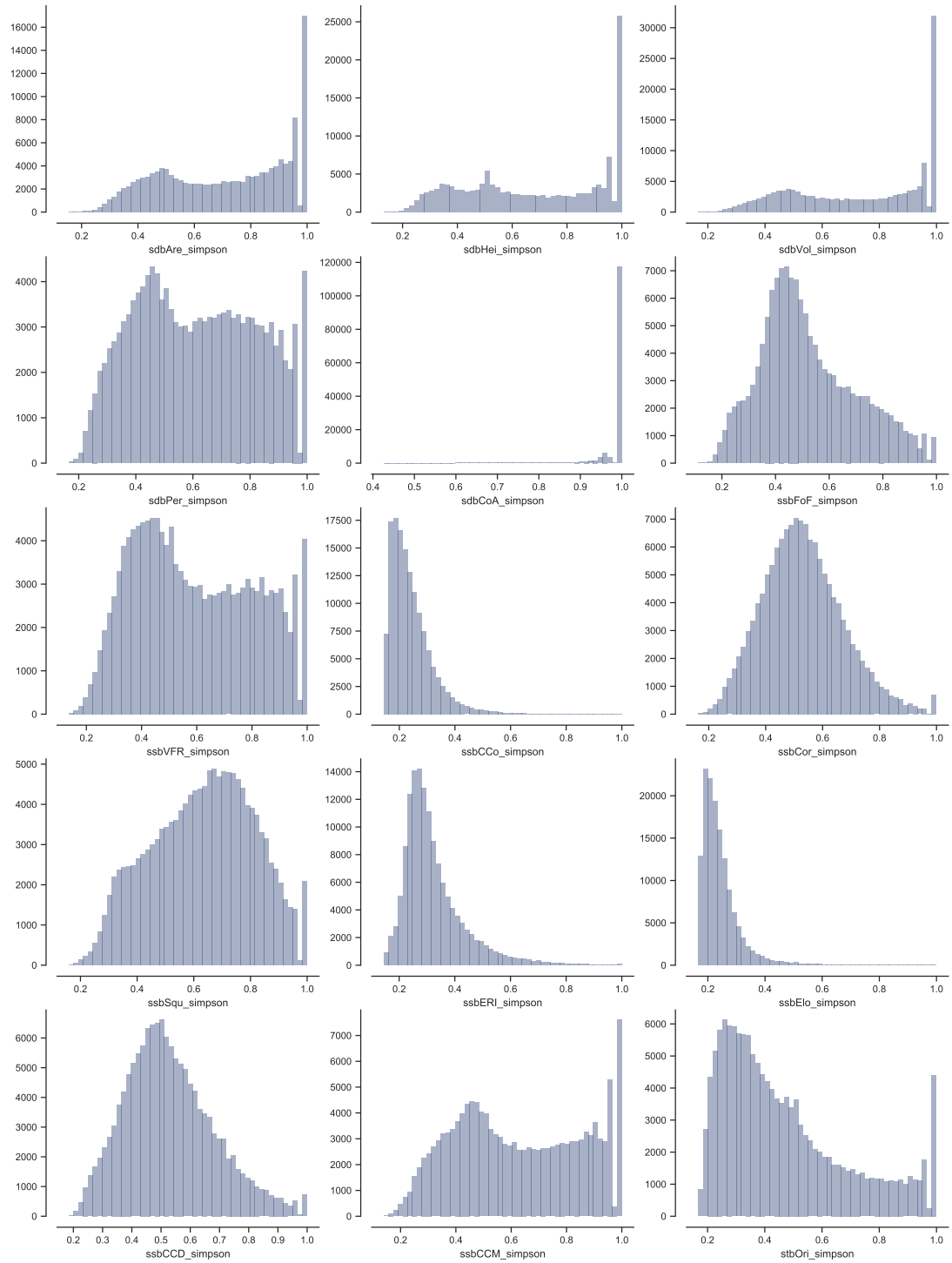

Figure 8.16: Histograms of Simpson index for characters 1-15 are showing the variety of distributions within the measured contextual data.

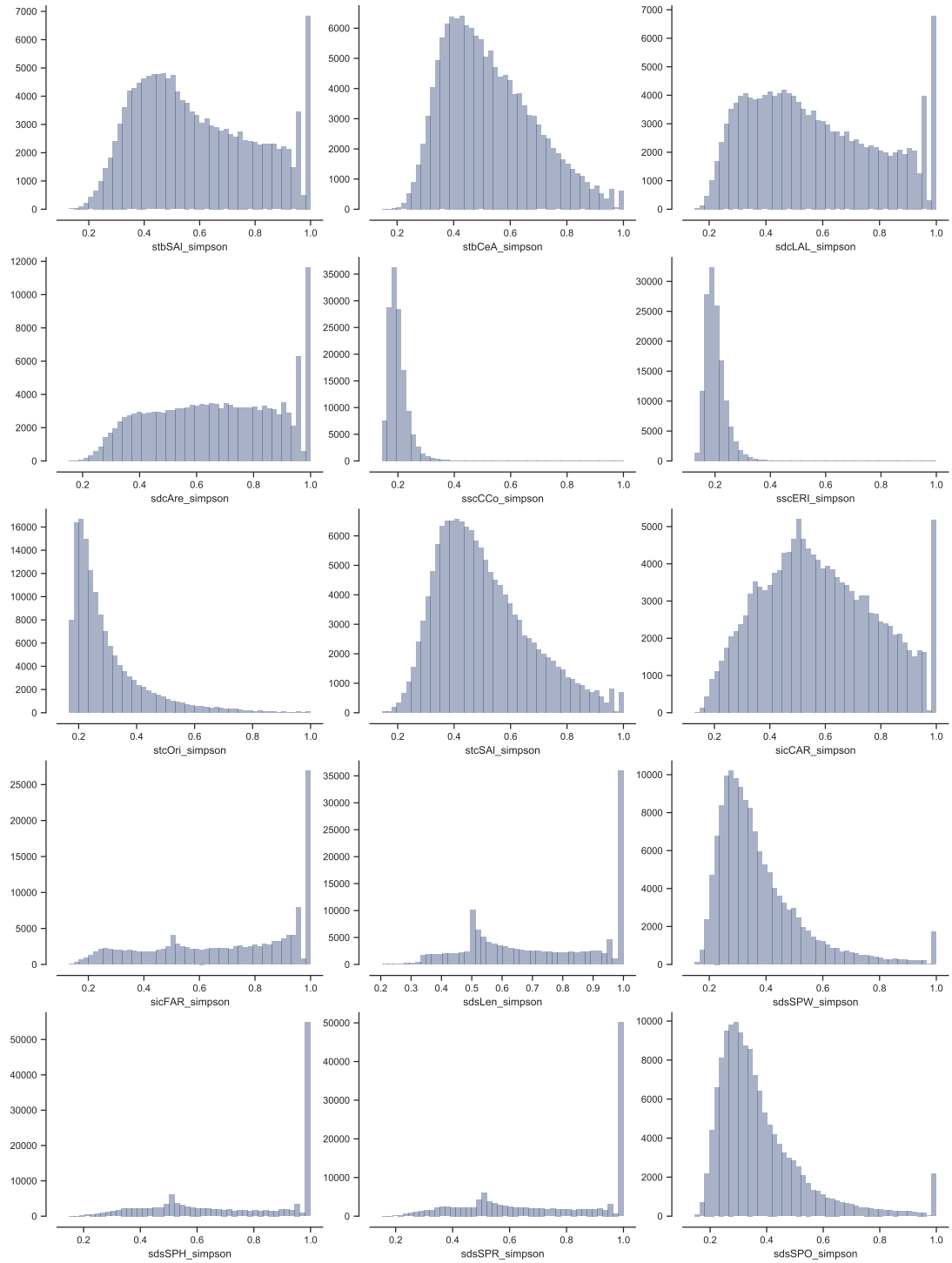

Figure 8.17: Histograms of Simpson index for characters 16-30 are showing the variety of distributions within the measured contextual data.

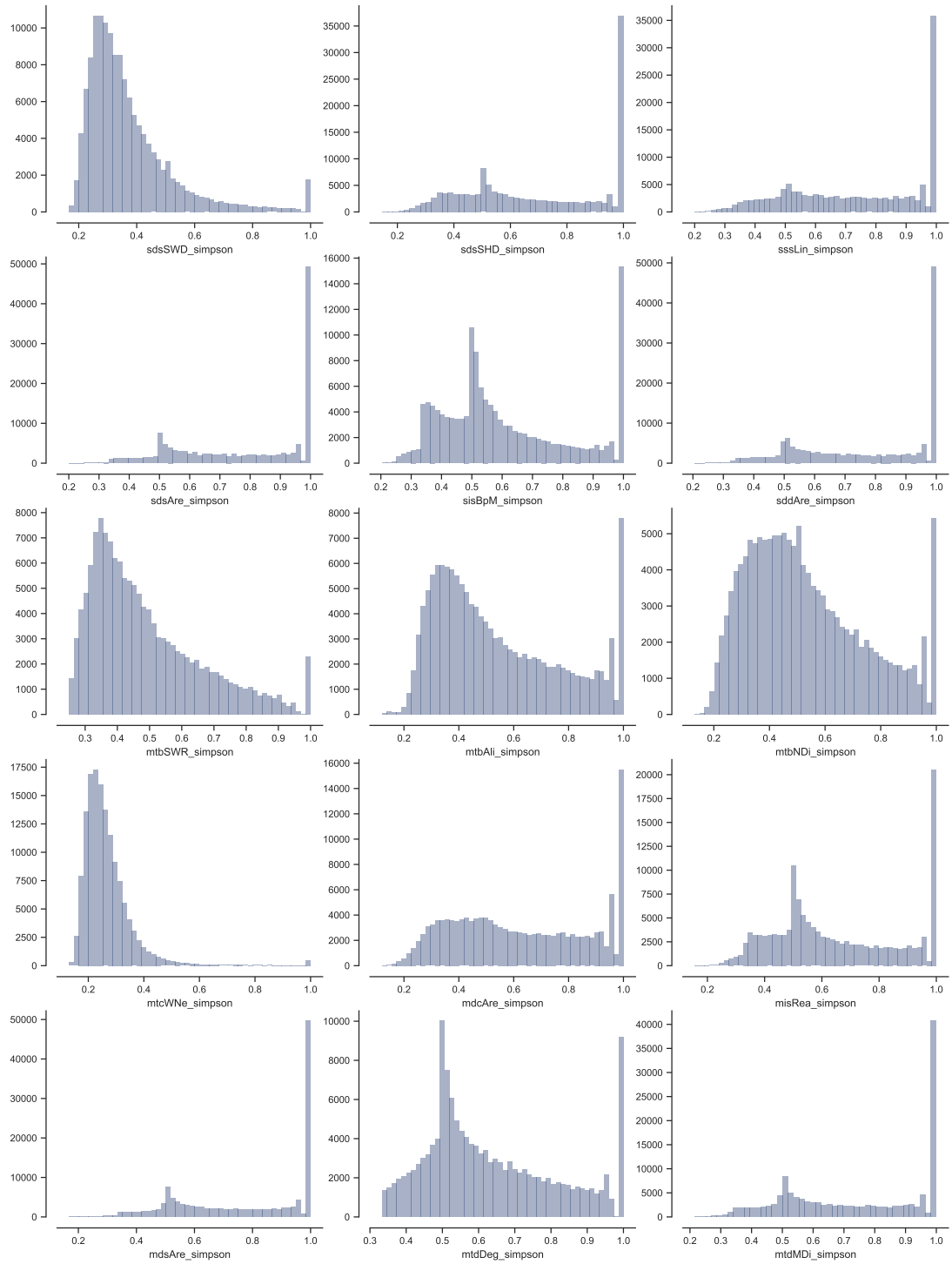

Figure 8.18: Histograms of Simpson index for characters 31-45 are showing the variety of distributions within the measured contextual data.

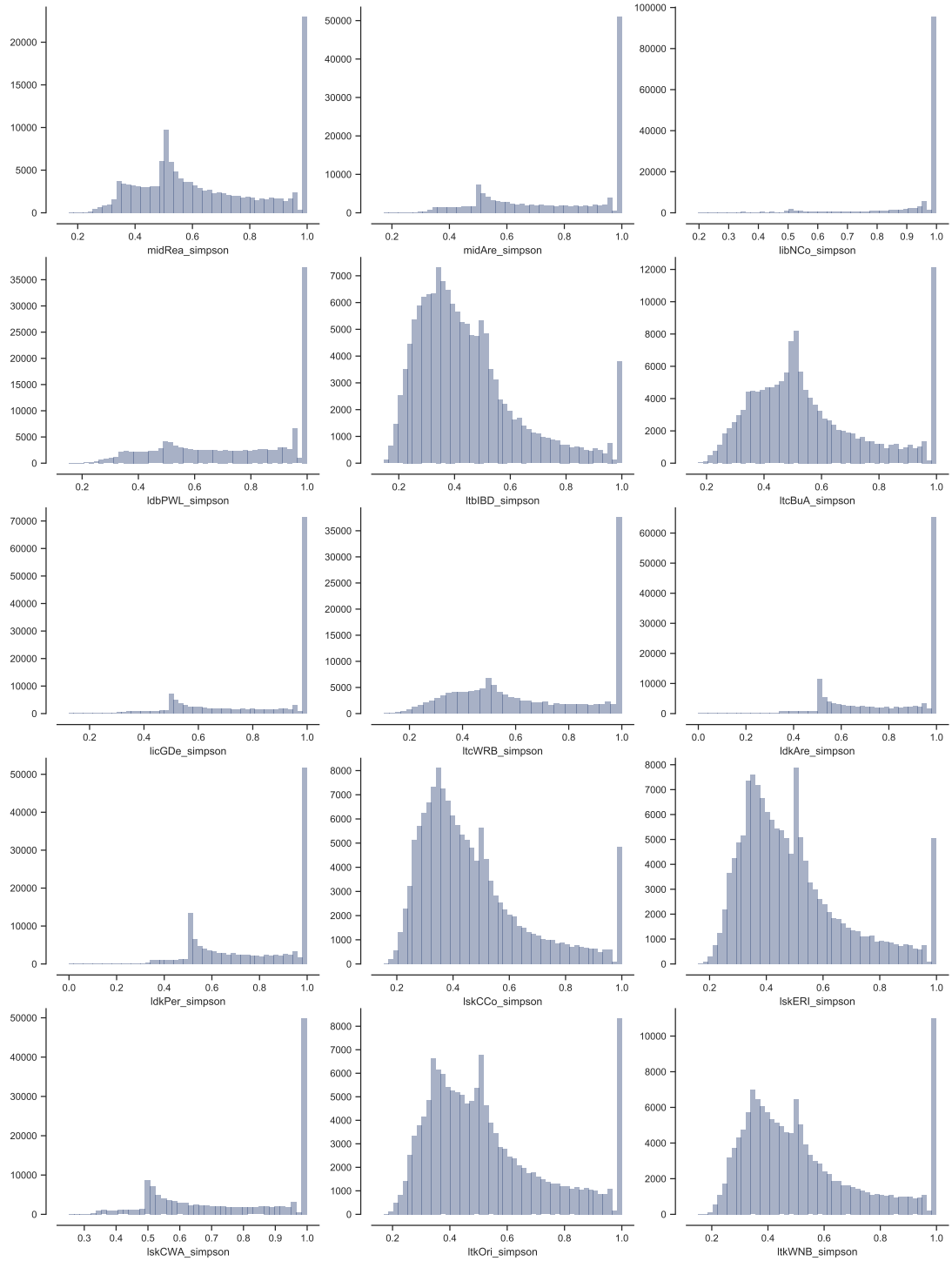

Figure 8.19: Histograms of Simpson index for characters 46-60 are showing the variety of distributions within the measured contextual data.

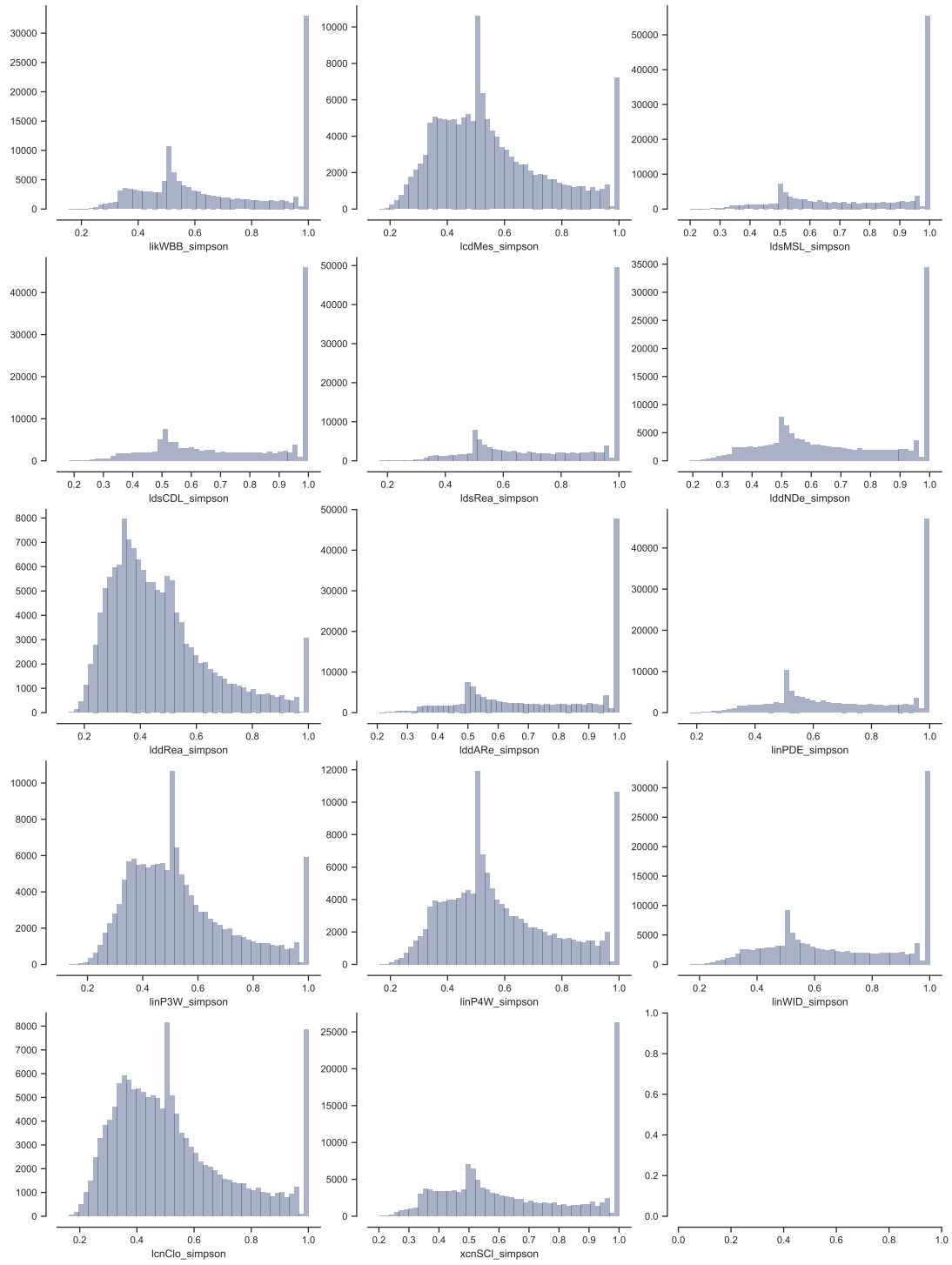

Figure 8.20: Histograms of Simpson index for characters 61-74 are showing the variety of distributions within the measured contextual data.

<!–

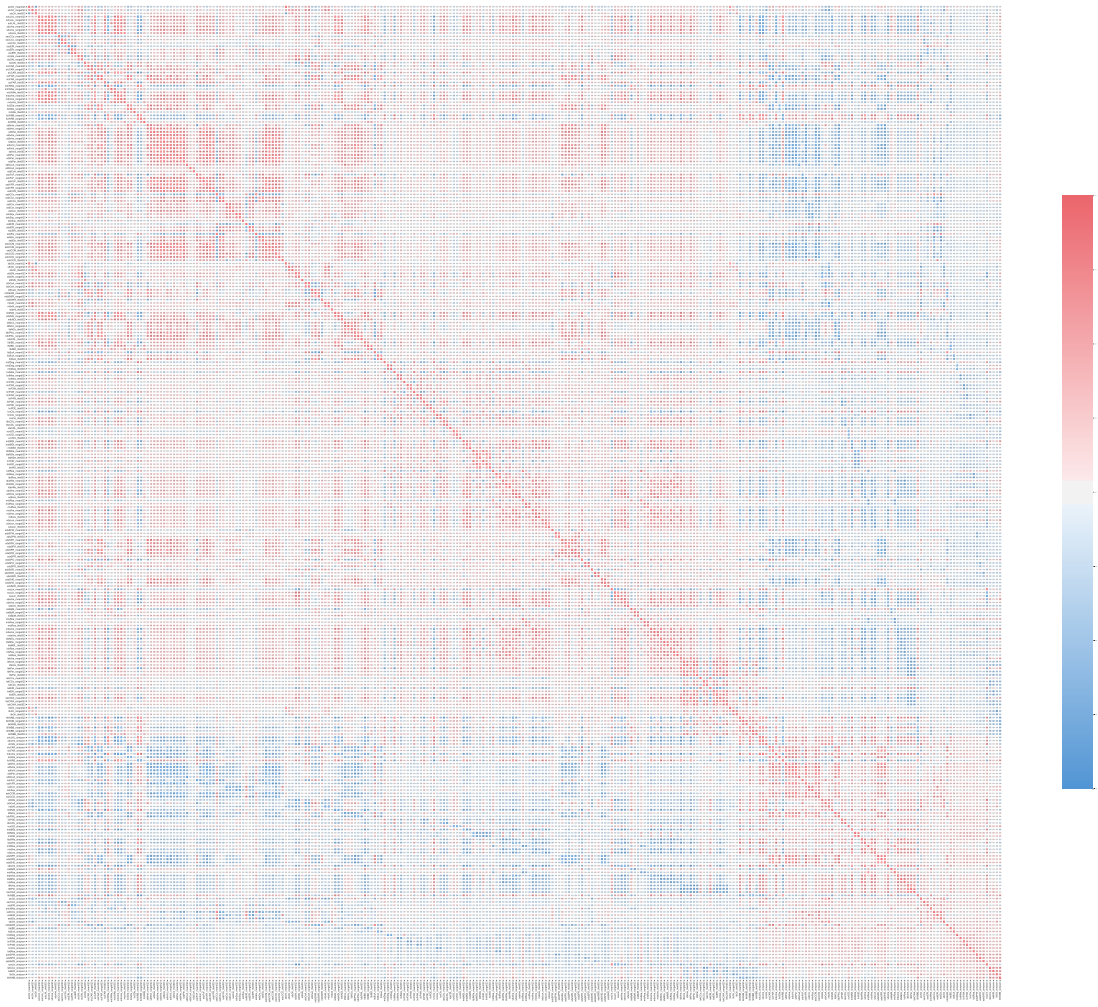

Figure 8.21: TITLE.

–!>

References

Araldi, A. & Fusco, G., 2019. From the street to the metropolitan region: Pedestrian perspective in urban fabric analysis: *Environment and Planning B: Urban Analytics and City Science*, 46(7), pp.1243–1263.

Basaraner, M. & Cetinkaya, S., 2017. Performance of shape indices and classification schemes for characterising perceptual shape complexity of building footprints in GIS. *International Journal of Geographical Information Science*, 31(10), pp.1952–1977.

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.

Boeing, G., 2017. OSMnx: New methods for acquiring, constructing, analyzing, and visualizing complex street networks. *Computers, Environment and Urban Systems*, 65, pp.126–139.

Bourdic, L., Salat, S. & Nowacki, C., 2012. Assessing cities: A new system of cross-scale spatial indicators. *Building Research & Information*, 40(5), pp.592–605.

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.

Duchêne, C., Bard, S. & Barillot, X., 2003. Quantitative and qualitative description of building orientation. In *He 5th ICA workshop on progress in automated map generalization*. Paris.

Feliciotti, A., 2018. *RESILIENCE AND URBAN DESIGN: A SYSTEMS APPROACH TO THE STUDY OF RESILIENCE IN URBAN FORM*. PhD thesis. Glasgow.

Gil, J. et al., 2012. On the Discovery of Urban Typologies: Data Mining the Multi-dimensional Character of Neighbourhoods. *Urban Morphology*, 16(1), pp.27–40.

Hamaina, R., Leduc, T. & Moreau, G., 2012. Towards Urban Fabrics Characterization Based on Buildings Footprints. In *Bridging the Geographic Information Sciences*. Berlin, Heidelberg: Springer, Berlin, Heidelberg, pp. 327–346.

Hijazi, I. et al., 2016. Measuring the homogeneity of urban fabric using 2D geometry data. *Environment and Planning B: Planning and Design*, pp.1–25.

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.

Lind, P.G., González, M.C. & Herrmann, H.J., 2005. Cycles and clustering in bipartite networks. *Physical Review E*, 72(5), p.056127.

Pedregosa, F. et al., 2011. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12, pp.2825–2830.

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.

Schirmer, P.M. & Axhausen, K.W., 2015. A multiscale classification of urban morphology. *Journal of Transport and Land Use*, 9(1), pp.101–130.

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

Steiniger, S. et al., 2008. An Approach for the Classification of Urban Building Structures Based on Discriminant Analysis Techniques. *Transactions in GIS*, 12(1), pp.31–59.

Vanderhaegen, S. & Canters, F., 2017. Mapping urban form and function at city block level using spatial metrics. *Landscape and Urban Planning*, 167, pp.399–409.

Yan, H., Weibel, R. & Yang, B., 2007. A Multi-parameter Approach to Automated Building Grouping and Generalization. *GeoInformatica*, 12(1), pp.73–89.