Identifying Cities Using Clustering

submitted by

Iurie Tarlev

for the degree of BEng(Hons) Civil Engineering

of the

University of Bath

Department of Architecture and Civil Engineering

Supervisor: Dr Nick McCullen

11th April 2019

Abstract

Classifying urban areas can be a difficult and often subjective task, relying mostly on administrative data. However, administrative boundaries are not a very accurate representation of the delineation between rural and urban areas. In this thesis a classification of land use is proposed based on a number of geographical and socio-economic variables. Dimensionality reduction and clustering algorithms are applied to high-resolution geographical and census data, to identify urban clusters based on their similarities. Thereafter, a thorough comparison of various clustering algorithms and dimensionality reduction algorithms is provided. Bath and North East Somerset county (UK) is used as a training dataset, but a general comparison of the readily available dimensionality reduction and clustering methods is provided. The results of this work provide new understanding of various computational methods that can be applied to the urban data, which leads to a multi-disciplinary approach of defining urban/rural boundaries.

Acknowledgements

First and foremost I would like to express sincere gratitude to my supervisor Dr Nick McCullen for continuous guidance and motivation. His immense knowledge and encouragement always steered me in the right direction with my research and writing.

I am greatly indebted to my parents, for their unfailing support and continuous motivation throughout my education years and through the process of my work on this dissertation. Lastly, thanks go to my friends who are invariably helpful and continue to supplement my family's support.

Contents

List of Figures

List of Tables

List of Abbreviations

AGNES Agglomerative Nesting Clustering; a hierarchical clustering algorithm.

BATHNES Bath and North East Somerset; the area for which local clustering is performed.

CLARA Clustering Large Applications; partitional clustering algorithm.

DIANA Divisive Analysis Clustering; a heirarchical clustering algorithm.

GIS Geographical Information System; a system to manage, manipulate and present geographic information.

LSOA Lower layer Super Output Area; geographic census areas used by the ONS and covering between 1,000 and 3,000 people, and between 400 and 1,200 households.

ML Machine Learning; encompasses statistical models and algorithms that computer systems use to perform tasks without using explicit instructions.

MSOA Middle Layer Super Output Area; geographic census areas used by the ONS and covering between 5,000 and 15,000 people, and between 2,000 and 6,000 households.

OA Output Areas are geographic areas that are delineated by postal codes in the UK and are the base unit for census data releases.

ONS Office for National Statistics; UK's national statistical institute.

PAM Partitioning Around Medoids; a partitional clustering algorithm.

PC Principal Component, the rotated axis in multiple dimensions which explains a certain amount of data variance (based on transformation by PCA).

PCA Principal Component Analysis; a linear dimensionality reduction algorithm.

SPSS Statistical Package for the Social Sciences; statistical analysis software developed by IBM that allows for data exploration through cluster analysis.

t-SNE t-Distributed Stochastic Embedding; a non-linear dimensionality reduction technique.

TTWA Travel To Work Areas; geographical areas created to approximate labour market areas.

Chapter 1: Introduction

For much of the history, urban growth proceeded in an organic, unregulated manner (Pacione, 2009). However, beginning from the eighteenth century, a growing interest in definition of urban boundaries was observed among scientists (Masucci et al., 2015). The attention was given to this topic not solely for academic interest, but also in an attempt to guide rational urban development towards socially beneficial goals (Pacione, 2009).

In the 1930s in the UK, the concept known as "Green belt" has been introduced as a land use designation tool in urban planning, to contain urban growth within a demarcated boundary and protect rural areas (Elson et al., 1993). In Albania, the yellow line system has been used for several decades now, aiding the definition of "inhabitation centres" that delineate urban and rural areas (Turner, Hegedüs and Tosics, 1992). Today there are multiple interpretations of what constitutes urban areas, but they all vary across different countries. About 2/3 of countries worldwide define urban boundaries based on administrative data in combination with an additional parameter such as population size, density or economic output. The majority of the remaining countries utilise just population size or density as the defining criteria, in conjunction with another parameter in some instances (Moreno, 2017).

It can be deduced that in most cases, definition of urban areas is highly limited by a few population-related factors and the administratively established boundaries. Population-related indicators alone can be misleading, as they account for residential areas only, neglecting highly built-up commercial or industrial areas. The somewhat arbitrary administrative boundaries (Masucci et al., 2015) can also give a biased overview of urban extent, as throughout history jurisdictional boundaries were adopted to define rights and responsibilities on the land, which often ignored socio-economic and landscape aspects.

One would think that with current remote sensing technology and Geographical Information System (GIS) mapping capabilities, urban boundaries could be easily defined. This is hardly true, as to this day no broad agreement has been reached on demarcating urban areas and their boundaries (Masucci et al., 2015; Uchiyama and Mori, 2017; Liu et al., 2014). Researchers suggest that no uniform urban/rural classification system can be used for different geographies (Pateman, 2011; Seto et al., 2011) because the urban form can be highly correlated to local planning policies, cultural and other aspects specific to a location. This makes classification of urban areas an extremely complex (Chen et al., 2017; Moreno, 2017) and highly subjective task.

The real structure of modern urban systems is complex and the definition of urban boundaries is highly dependent on a large number of variables (Kasanko et al., 2006), which include landscape and socio-economic metrics related directly to urban form. Depending on the location, some of those variables might be more indicative of urban areas, than others. Researchers use different reasoning (Fang and Wang, 2018) when deciding which indicators are more important, but so far there has been limited attempt in quantifying the importance of used parameters when

defining urban areas.

In this study, dimensionality reduction algorithms (Section 4.4) are used as tools which will combine the metrics in a meaningful way, after which, clustering (Section 4.5) is performed to find similarities between geographical regions. Because a hard boundary between rural and urban areas does not exist, the scope of this study is identifying more than two different clusters, and therefore, outlining various levels of "urbanity". As a primary aim of this study, two dimensionality reduction techniques and five clustering algorithms are compared, overall totalling to ten different configurations (see Figure 3.1 for more details). Bath and North East Somerset (BATHNES) area has been taken as a case study to minimise the computational resources required, as it contains much less data than the global (England) dataset.

This paper is divided into eight chapters and is structured in the following way:

- Introduction - Chapter 1
- Literature Review - Chapter 2
- Aims and objectives - Chapter 3
- Methodology - Chapter 4
- Results and analysis - Chapter 5
- Discussion - Chapter 6
- Conclusion - Chapter 7
- Bibliography - Chapter 8

Chapter 2 provides a literature review which goes in-depth about what other researchers have discovered on this topic in the academic field. Chapter 3 introduces the aims and objectives of this study. Chapter 4 introduces the data used, dimensionality reduction and clustering methods evaluated as well as the method used for this evaluation. Chapter 5 shows all the results together with a thorough analysis. Chapter 6 deals with discussion of the results in this study, stating its advantages and limitations over other studies. Chapter 7 makes a summary of the findings, describing the contribution of this study and sets out future research directions in the field.

Chapter 2: Literature Review

The central theme of this dissertation is detection of similar geographical units by using clustering algorithms, thereby, identifying boundaries between different classes of urban/rural areas. Latest research into definition of urban/rural extent and methods used to accomplish that are presented in this chapter. The chapter allows to contextualise this study and identify gaps, some of which are covered in the current paper.

2.1 Importance of well-defined urban boundaries

This section will give a brief overview of the value that well-defined urban boundaries can bring to society in terms of sustainable development and economic benefit.

2.1.1 Sustainable development

Cities worldwide are undergoing constant transformation, due to new communication and transportation systems as well as changes in land use (Cao et al., 2013). This leads to an increasing amount of human activity and resource consumption to be focused in cities (Madlener and Sunak, 2011). Currently, around 75% of the world's resources are being expended in cities (Calcott and Bull, 2007), although urban areas cover less than 3% of earth's surface (Liu et al., 2014). This accelerating rate of urbanisation has the potential of impacting the ecosystems which regulate climate and air quality, maintain freshwater, natural resources and biodiversity (Shepherd, 2005; Foley et al., 2005; Kontgis et al., 2014). Nonetheless, cities could be the most environmentally friendly places to live in (Calcott and Bull, 2007), as compact urban settlements together with high employment and residential densities could lead to a reduction of energy consumption and carbon footprint (National Research Council, 2009). Therefore, society faces the challenge of developing long-term management strategies to reduce the environmental impact of changing land use, whilst maintaining the social and economic benefits (Foley et al., 2005; Kontgis et al., 2014; Madlener and Sunak, 2011).

2.1.2 Economic value

Managing the development of urban areas and extracting insights about it has become an increasingly valuable task in recent times since local and national governing bodies, as well as private-owned enterprises, invest millions of dollars each year obtaining important information about urban/suburban infrastructure (Jensen and Cowen, 1999). This highlights the importance of accurate representation of urban areas and their boundaries, as they can serve urban planners and businesses to develop deep insights into city development dynamics (Cao et al., 2013). Consequently, this can lead to better decision making, yielding higher economic output and efficiency levels.

2.2 Limitations of identifying urban areas using a single metric

Most countries in the world are divided into sub-national entities also referred to as counties, provinces or constituent units. These, in turn, are further subdivided into towns, cities and villages, which often have their own local government/administration. These subdivisions make management of land and other affairs an easier task. However, although the vast majority of statistics for any given country is reported for administratively divided regions, the idiosyncratic and somewhat arbitrary administrative boundaries (Masucci et al., 2015) are not a very good representation of urban/rural boundaries. This is mainly because the administrative division doesn't consider the socio-economic and landscape differences between regions. Henceforth, researchers attempted using various other indicators in order to establish boundaries within urban systems.

2.2.1 Population Density

Hasse and Lathrop (2003) suggested the use of population density for evaluation of land use and classification of areas as urban or rural. Majority of countries use this metric in combination with administrative urban extend delineation for boundaries definition (Moreno, 2017). However, on this basis, areas with high coverage of paved roads, concrete buildings and electricity are not necessarily classified as urban. Whereas areas with large amounts of green space, weak intensity of night light, but high population densities will be detected as urban. Furthermore, definition of urban areas based on population density implies determining a threshold, above which, areas will be considered as urban and this threshold is currently non-existent (Uchiyama and Mori, 2017). These factors make the sole use of population density for clustering urban areas a limited one.

2.2.2 Road Junctions

Long (2016) used road intersections in China, defining a city as a "spatial cluster with over 100 road junctions within a 300m distance". However, the limitation of using road junctions as the only indicators is the high subjectivity of the established threshold, which in Long's case is 100 junctions. While areas with 99 junctions and 101 junctions might be very similar in terms of their city characteristics, they are classed into two different clusters of city and non-city. Therefore, the utmost simplicity of using road junctions as the only indicator of urban areas is its own limitation. This also showcases the importance of identifying several different clusters, rather than splitting the geographical units into two groups.

2.2.3 Night-time lights data

Night-time lights data has been used by Zhou et al. (2014) to classify land use and cluster together similar urban areas. Zhou et al. argued that it is a cost and time effective method for clustering the data in urban form studies. However, using night-time lights as an indicator for identifying urban areas has its own limitations. It assumes that there is a strong correlation between the intensity of economic activity and night-time lights. This is an issue since it also classifies transportation infrastructure and oil mining sites (Uchiyama and Mori, 2017) as urban

areas because they have high-density night lights. This method can also be inaccurate due to seasonal variations caused by tourism. Many localities are busy during tourist season, while in other seasons the population is small (Chi et al., 2015). Being unable to distinguish between such tourist places and true urban areas is ,therefore, another limitation of using night-time lights.

2.2.4 Land Cover

Another type of indicator often used for detection of urban areas is landcover (El Garouani et al., 2017; Tayyebi, Pijanowski and Tayyebi, 2011). This method uses the built-up areas and low rates of vegetation as primary indicators for the purpose of defining urban form. As a consequence of this, large factories and empty developed districts away from cities will also be identified as "urban" (Uchiyama and Mori, 2017), thereby providing an inaccurate representation of urban areas.

2.2.5 Travel to Work Areas

In the UK, Travel To Work Areas (TTWA), provides an alternative view on urban form, as the boundaries are more reflective of socio-economic characteristics (Coombes, 2016), rather than the somewhat limited definition of administrative urban extent and boundaries based on population-related metrics. However, TTWA boundaries are based solely on indicators of where the population would generally commute to a larger town or city, ignoring other important socio-economic and landscape aspects.

2.2.6 Disadvantages of using single metrics

Uchiyama and Mori (2017) suggested that from the indicators discussed above, one may be preferred depending on the country of interest: more developed, less developed or least developed. According to their study, for less developed countries more urban areas are identified when using population density and night-time lights as the indicator. On the other hand, for least developed countries there were significant differences between the size of identified urban areas, depending on which one of the aforementioned indicators was used. Therefore, it can be concluded that the use of a single metric is highly dependable on a given location and it doesn't consider all the characteristics related to urban form. Consequently, the approach of combining multiple metrics taken in the current paper can be described as more unified.

2.3 Limitations of other urban definition studies

A number of studies have utilised a collection of metrics in identifying urban areas. Schirmer and Axhausen (2015); Tayyebi, Pijanowski and Tayyebi (2011); Cao et al. (2013); Huang, Lu and Sellers (2007) have all used various landscape indicators to define urban morphology, where they were usually derived from satellite imagery. Although several metrics have been used, the primary limitation of such studies is that social and demographic characteristics have been totally ignored.

Other studies rely solely on a combination of census or socio-economic attributes. Kendig (1976) study was aimed at classifying residential areas based on a range of social factors alone. Zhikharevich, Rusetskay and Mladenović (2015) study identified similarities between 120 different cities based on socio-economic attributes. However, this is a very small scale task, with no output defining urban morphology. Although there is a large number of similar studies, they are all very small in scale, location specific and only utilising a limited number of social characteristics which authors were interested in exploring.

Other studies (Fang and Wang, 2018; Tsai, 2005) have used both socio-economic and landscape metrics. However, the choice of the metrics is not justified, since there is no guided procedure of selecting those indicators. The indicators are usually chosen by the authors, who decide what is deemed as relevant. Furthermore, only a small number of indicators was chosen, due to the limitation of the methods used in determining similarities between various regions in those studies.

Osorio (2017) study is a comprehensive analysis of a large number of physical and socio-economic indicators. PCA dimensionality reduction algorithm (further introduced in Section 4.4) is used to mathematically calculate which attributes are the most relevant. Then a clustering algorithm called K-means was used to determine similarities between various regions within England and a resulting cluster map plot was presented. The limitation of this study is the use of dimensionality reduction and clustering algorithms based on widespread use and familiarity. No assessment of methods has been done to decide on which algorithms to be used for increased accuracy of results.

The current paper addresses the limitations of the studies described above, by analysing 32 different metrics from both landscape and socio-economic data. It also compares two dimensionality reduction algorithms and five different clustering methods, assessing their performance and output.

2.4 Clustering techniques application

Clustering algorithms have been used for numerous purposes in identifying certain patterns of urban form but were not always related to identification of urban boundaries. Chi et al. (2015) study explored which cities in China can be classified as Ghost Cities, identifying those that have a highly developed infrastructure but low population rates. Smartphone live geographical data was combined with people's home locations on record, after which clustering was used to identify cities with low population. Other studies have used clustering for identifying cities with similar pollution rates (Gao et al., 2011; Baxter and Sacks, 2014), entrepreneurship development (Wei, 2010) and comparability of other specific metrics.

Statistical Package for the Social Sciences (SPSS) (IBM, 2014) is commonly used in social sciences to cluster geographical areas based on some parameters of interest. However, this standard software is limited by the methods used and the number of variables it can analyse. IBM's SPSS software can only perform two of the most commonly used clustering algorithms. Furthermore, there is no flexibility in choosing a method for dimensionality reduction. Thus

standardised software like this is not deemed as a suitable approach for identifying urban boundaries and providing a comparison of different techniques of doing it.

Clustering techniques have been used for a variety of applications as outlined above, but there is a limited exploration of readily available dimensionality reduction and clustering techniques which would outline the advantages and disadvantages of using them. This study deals primarily with this issue.

2.5 Studies comparing clustering algorithms

A comparison of 6 different clustering methods has been performed on ecological background and urban evolution data in China. The comparison included: grey clustering, coefficient of mode ratio, comprehensive index, systematical clustering and k-means clustering. Through qualitative and quantitative judgement, it was identified that grey clustering is the best clustering method for the eco-city classification. This rating is due to grey clustering performing better in interpreting of regional ecological background and urban evolution (Zhang et al., 2013).

Murphy and Maggioni (2018) explored various clustering algorithms used for hyperspectral imaging as Machine Learning (ML) is applied to remote sensing. In hyperspectral imaging, the clustering algorithms are aimed at identifying the material type and label it by assigning a colour. This paper presented a thorough comparison of various advanced clustering methods.

In literature, there is a number of studies similar to those described above, which give a comparison between relatively advanced and modified clustering algorithms. However, the scope of the current paper is to assess readily available algorithms which can be used not only in academia but for wider use as well.

2.6 Summary

Because clustering algorithms produce easily interpretable results, they are used across different fields for various studies. In the big data era, an increasing amount of data that describes various characteristics and behaviours within cities (Fang and Wang, 2018) is allowing researches in the field to emerge with advanced algorithms to the data in order to define boundaries between urban and rural areas. To ensure a sustainable, smooth and efficient urbanisation process, there is a growing need to develop urban computing and analysis tools to guide urban planning at different scales (Tsompanoglou and Photis, 2013; Cao et al., 2013). In this study, the aim is to assess some of the readily available computational algorithms for urban studies and provide a thorough comparison of them.

Chapter 3: Aims and Objectives

This research is a methodology assessment study, where the main aim is to compare readily available dimensionality reduction and clustering techniques as tools for determining boundaries between urban/rural areas. These techniques will be applied to a range of socio-economic and landscape metrics. The following objectives specifically will be met:

1. Pre-processing of BATHNES dataset.
2. A thorough comparison of different configurations of clustering and dimensionality reduction techniques applied to the chosen dataset.

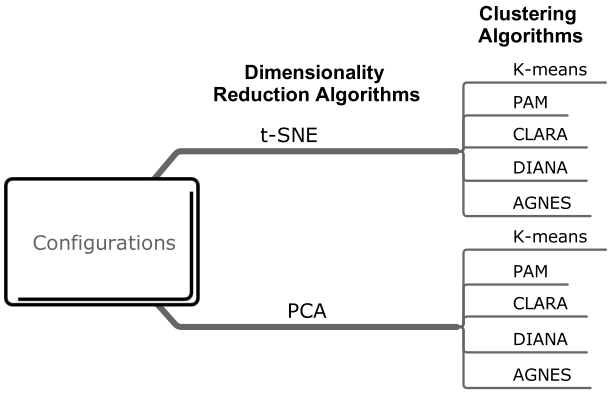

Figure 3.1: Dim. reduction and clustering configurations

3. A comparison of classification of areas being clustered locally as opposed to being clustered globally, as part of the countrywide dataset (see Figure 3.2 for more details).

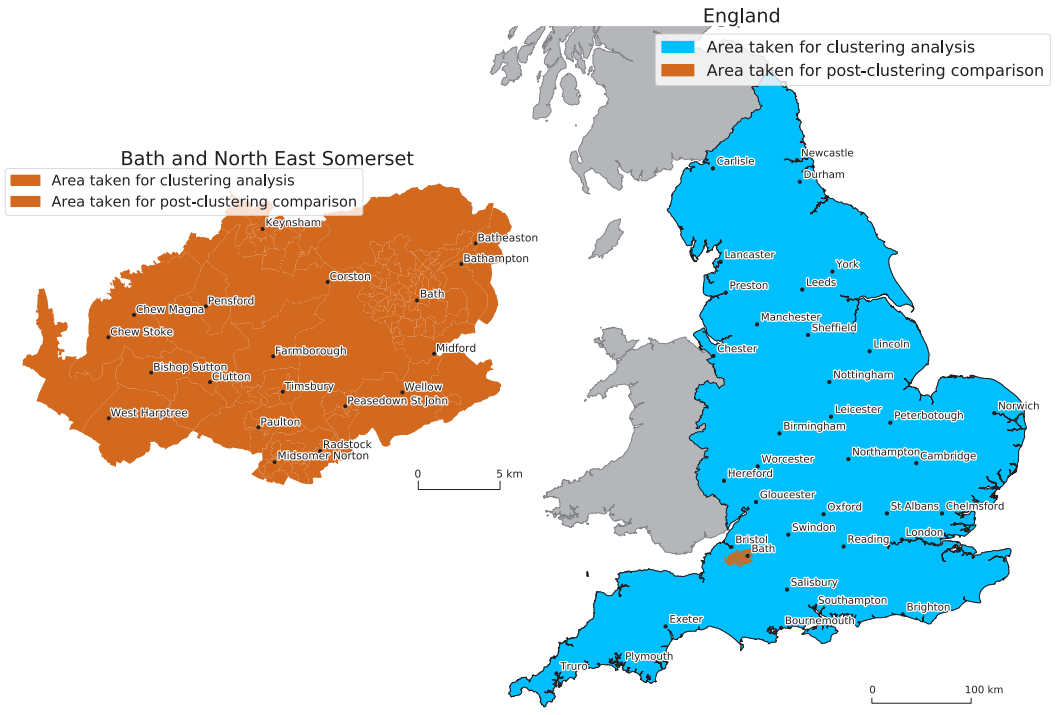

Figure 3.2: Local clustering and global clustering

Chapter 4: Methodology

4.1 Dataset resolution and size

To classify regions into urban, rural and other urban types it was desirable to acquire data for the smallest possible census unit. This ensured that the finest definition of boundaries could be defined by the clustering algorithm. In the UK, Office for National Statistics (ONS) publishes census data in two sizes of Output Areas (OA): Middle layer Super Output Areas (MSOA) and Lower layer Super Output Areas (LSOA). LSOA units have been chosen for this study, as they are the smallest geographical units, that ensure the highest resolution.

In the current paper BATHNES dataset has been selected as a case study for the assessment of clustering and dimensionality reduction techniques. This allowed for programming scripts to be run on a relatively small dataset of 115 LSOA regions, with 32 parameters each (see Section 4.2 for more details on parameters chosen). In contrast, the entire England dataset contains over 32,000 LSOA regions and each of them has 32 parameters, so running the analysis using this dataset would require much more computational power resources and time. England dataset is only used when comparing global clustering with local clustering as outlined in objective number 3 in Chapter 3.

4.2 Selection of urban form indicators

Urban form definition varies across literature (Uchiyama and Mori, 2017) and usually depends on landscape metrics (Huang, Lu and Sellers, 2007; Schneider and Woodcock, 2008; Long, 2016; Masucci et al., 2015) and/or socio-economic metrics (Kendig, 1976; Kasanko et al., 2006; Zhikharevich, Rusetskay and Mladenović, 2015). Landscape metrics tend to describe the physical structure of a particular area (e.g. length of roads), whereas socio-economic indicators rely on social characteristics (e.g. income per household).

There is no single parameter which could be used to portray full urban extent characteristics (Kasanko et al., 2006). Therefore, as mentioned in Chapter 2 it would be inadequate to rely on a single indicator for establishing boundaries between areas of different character. All parameters have their own strengths and weaknesses, but when used in conjunction, they can enable a comprehensive analysis (Kasanko et al., 2006).

For this study, a range of socio-economic and physical characteristics have been used in parallel for identification of boundaries between areas of different classes of "urbanity". Socio-economic indicators were obtained from Census statistics (Census Data, 2011) and landscape variables from land use (Ordnance Survey, 2017b) and road network (Ordnance Survey, 2017a) datasets. Additional parameters can be included in this assessment in the future if deemed to be relevant. Table 4.1 and Table 4.2 give a detailed overview of the parameters used for each LSOA geographical unit.

Table 4.1: Landscape metrics

Variable	Definition	Relevance	References
perimeter	total length of the border of each LSOA	measures complexity and size of an LSOA unit	(Schwarz, 2010)
area	total area of each LSOA	measures size of an LSOA unit	none
domestic buildings area	surface area covered by residential buildings	indicates density of residential areas	(Tratalos et al., 2007)
road area	surface area covered by road network	indicates the magnitude of complexity	(Osorio, 2017)
road length	total length of the road road network	indicates the magnitude of complexity and accessibility	(Osorio, 2017)
road length density	length of road network per resident	measures complexity and compactness	(Osorio, 2017)
rail area	surface area covered by railway	indicates how industrial an LSOA is	none
built-up area	surface area covered by physical man-made structures (e.g. roads, buildings, railways)	indicates urban area size	none
built-up proportion	percentage of surface in built-up use of the total individual LSOA area	indicates degree of urbanisation	(Osorio, 2017)
buildings area	surface area covered by buildings	indicates urban area size	none
buildings area proportion	percentage of surface covered by buildings of the total individual LSOA area	indicates degree of urbanisation	(Osorio, 2017)
non built-up area	surface area not covered by physical man-made structures (e.g. roads, buildings, railways)	indicates non-urban area size	none
non built-up area proportion	percentage of surface not in built-up use of the total individual LSOA area	indicates the inverse of the urbanisation degree	none
proportion of detached dwellings	percentage of detached dwellings out of total number of dwellings	indicates settlement structure	(Tratalos et al., 2007)
proportion of semidetached dwellings	percentage of semidetached dwellings out of total number of dwellings	indicates settlement structure	(Tratalos et al., 2007)
green space	area covered by green space (e.g. gardens)	indicates both compactness and heterogeneity	(Osorio, 2017)

Table 4.2: Socio-economic metrics

Variable	Definition	Relevance	References
resident population	no. of residents	indicates population size	(Schwarz, 2010)
male resident population	no. of male residents per female resident	indicates gender dynamics in cities	none
population density	no. of residents per unit surface area	indicates compactness	(Schwarz, 2010)
population density in built-up area	no. of residents per unit built-up surface area	indicates compactness	(Tratalos et al., 2007)
number of dwellings	total no. of dwellings	indicates degree of urbanisation	(Schwarz, 2010)
density of dwellings	no. of dwellings per unit surface area	measure of urbanisation density	(Tratalos et al., 2007)
number of household spaces	total no. of household spaces	indicates degree of urbanisation	(Schwarz, 2010)
density of household spaces	no. of household spaces per unit surface area	measure of urbanisation density	(Schwarz, 2010)
density of household spaces in built-up area	no. of household spaces per unit built-up surface area	measure of urbanisation density	(Osorio, 2017)
private car availability	no. of cars in households per 1000 residents	measures of welfare and transport structure	(Osorio, 2017)
proportion of population with higher education	population percentage that hold higher education degree	measures education impact on urban areas	(Tratalos et al., 2007)
proportion of population in employment	percentage of working age population in work	measures job structure impact on urban areas	(Osorio, 2017)
proportion of population employed in services	percentage of working age population employed in services	measures job structure impact on urban areas	none
proportion of flats in commercial buildings	percentage of flats located in non-residential buildings	indirect link to the degree of urbanisation	none
ratio of detached houses	no. of detached houses over total no. of houses	indirect link to the degree of urbanisation	(Tratalos et al., 2007)
yearly household income	yearly household earnings	indicates economic welfare	(Schwarz, 2010)

4.3 Pre-processing of data and resources used

The dataset with the data mentioned in Section 4.2 was stored in a CSV file, after which it was scaled to ensure variables measured in different units were not erroneously dominating the clustering. This scaled data was then used for comparison of two dimensionality reduction and five clustering algorithms. Technical details about the algorithms used are described in Section 4.4 and Section 4.5. All the scripts developed for clustering were written in R programming language (R Development Core Team, 2008); version 3.5.1. The plotting of clustering results was performed in Python programming language (Python Development Core Team, 2015); version 3.7.1. Original data used and all the developed scripts can be accessed via the following link: `https://github.bath.ac.uk/it297/Urban-rural-clustering.git`.

4.4 Dimensionality Reduction

Humans are only capable of processing and understanding data usually in up to three dimensions. However, computers have no problems with processing data with dimensions in magnitude orders of hundreds, thousands or even millions.

Clustering algorithms have the ability to deal with multiple dimensions. Nonetheless, there are two main reasons why dimensionality reduction algorithms might be useful to be applied before running the clustering algorithms on the data. First, some variables in the dataset are usually derivatives of others, therefore, dimensionality reduction techniques can be applied to decorrelate the data and correct for such redundancy (e.g. PCA). The second reason is visualisation ability, as some advanced statistical dimensionality algorithms are very effective at separating the clusters for clear visual effects (e.g. t-SNE).

In this study, there are 32 different parameters used in combination, meaning that the data has a large number of dimensions. Technical information about PCA and t-SNE dimensionality reduction algorithms is discussed in this section with the full analysis following in Chapter 5.

PCA

Principal Component Analysis (PCA) is known as the oldest dimensionality reduction method for multivariate statistical analysis. Although invented in the early 20th century, it became widely used with the advent of electronic computers at the end of that century. PCA is a linear combination method of variables that derives Principal Components (PCs), which are dimensions representing the larges variation of the dataset (Jolliffe, 2002).

First important characteristic of PCA is its ability to capture the most important information from the data in the fewest dimensions possible (Jolliffe, 2002). Finding a linear combination of variables that corresponds to maximum variance takes place first (Pituch and Stevens, 2015). After, the process is repeated for the linear combination of variables that accounts for the next biggest variance for as many times as there are variables in the dataset. Hence, the first few principal components are the eigenvectors which contain the largest variances, therefore, capturing most of the variation present in the original data.

Second important feature of PCA is decorrelation (Pituch and Stevens, 2015). Since some of the variables in a dataset are often correlated (e.g. road length and road area). Often combining the results of those parameters leads to overfitting and produces unnecessary noise in the results. PCA, in contrast, eliminates co-linearity effects. Therefore, when transforming the data from original data into PCs, they automatically become decorrelated.

For illustration purposes, a 2D scatterplot of correlated variables is shown in Figure 4.1. The plot on the left shows the original data scatterplot. The plot in the middle (Figure 4.1) shows Principal Component (PC) 1, which has got the biggest spread of data and therefore accounts for the biggest variance in the dataset. The right-hand side plot in Figure 4.1 shows PC2 axis, the component which has got the smallest variance for a 2D dataset. The new axes are rotated, whilst perpendicular to each other and the distribution of data relative to the principal component axis is uncorrelated as a result. This is a simplified example used to demonstrate how a multivariate matrix with interrelated variables is transformed into a set of principal components which are uncorrelated.

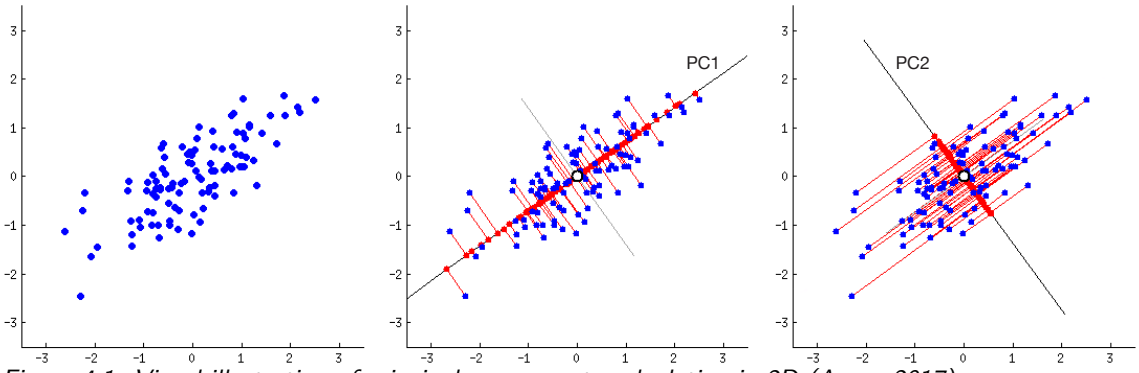

Figure 4.1: Visual illustration of principal components calculation in 2D (Anon, 2017)

For a better understanding of PCA, it is assumed that x-axis in Figure 4.1 represents the area and y-axis represents population size. The points on the graph are therefore population and area coordinates for several LSOA regions. Influence is considered as the projected coordinate of the observation on the PC axis. Therefore, the loading score on PC1 would be calculated as follows:

$$Area\ PC1\ score = (x\ read\ count\ 1 \times its\ influence\ on\ PC1) + (x\ read\ count\ 2 \times its\ influence\ on\ PC1) + \dots$$

$$Population\ PC1\ score = (y\ read\ count\ 1 \times its\ influence\ on\ PC1) + (y\ read\ count\ 2 \times its\ influence\ on\ PC1) + \dots$$

PCA's ability to reduce the number of dimensions and decorrelate all of the variables is crucial for extracting insights from the data mentioned in Table 4.1 and Table 4.2. It allows for correction in redundancy from correlated variables. Additionally, it also allows for inference to be made back to the input data, therefore, identifying which variables influenced clustering the most. However, PCA is only suitable for finding linear correlations between variables (Jolliffe,

2002), as it relies on orthogonal projections of the dataset that contain the highest variances possible. Therefore, its inability to find non-linear correlations is its biggest limitation.

t-SNE

t-Distributed Stochastic Embedding (t-SNE) is a highly advanced non-linear mathematical dimensionality reduction algorithm introduced by Maaten and Hinton in 2008. This algorithm is highly reputable for its capability to create easily interpretable visualisations of high-dimensional datasets (Maaten and Hinton, 2008). The goal of t-SNE is to map points in high-dimensional space onto a lower-dimensional space, usually a 2D plane, without losing much of the information.

t-SNE's advantage over PCA is it's non-linear ability to preserve local structure (Pathak, 2018), whereas PCA only aims to preserve global shape of data. t-SNE's perplexity parameter is an estimate of the number of close neighbours each data point has. It dictates how to balance between local and global aspects of the data and should be chosen according to the data size analysed (Maaten and Hinton, 2008). A recommended value for perplexity by Maaten and Hinton is between 5 and 50. For this study, the perplexity parameter has been specified as 10.

After t-SNE has been run, input features are no longer identifiable, making it difficult to interpret the results (Pathak, 2018), as it is not clear which variables were dominant when results were generated. Also, unlike PCA, t-SNE has limited ability of decorrelating the input data, which is a disadvantage if the input data has a certain degree of correlation. The time and space complexity increase with the number of dimensions, as it attempts to find similarity between pairs of points (Vadali, 2018). This makes the t-SNE algorithm computationally demanding to run on a standard computer, and overall, it requires a relatively high level of expertise to understand its complexity.

4.5 Cluster Analysis

Clustering is one of the most widely used and important methods used for extracting insights from multidimensional data. Identification of patterns within the input dataset of interest is the main goal of clustering algorithms. It is often referred to as "Unsupervised Machine Learning (ML)" in literature. "Unsupervised" because there is no prior knowledge about which variables belong to which cluster and "learning" because the machine uses the algorithm to "learn" how to cluster (Kassambara, 2017a).

In this study, the focus is on assessing readily available clustering algorithms and overall 3 partitional clustering methods and 2 hierarchical clustering methods have been taken for comparison. In this section, the basics of all the clustering algorithms assessed are discussed and their advantages/disadvantages outlined.

4.5.1 Partitional

Partitional clustering methods are used to classify observations in a dataset into multiple groups, based on their similarity. Partitional clustering algorithms require a pre-specified

number of clusters by the user in order to perform the clustering.

K-means

K-means (MacQueen et al., 1967) is the most widely used clustering algorithm for splitting the data into a set of k groups. It is a deterministic clustering approach (Likas, Vlassis and J. Verbeek, 2003), where the user is required to specify the number of clusters k. The algorithm randomly selects two points as centroids and assigns data points from the dataset to the nearest centroid. The mean value for each of the identified k clusters is then found and is set as the new centroid. Every observation is then checked again to see whether it is positioned closer to a different cluster. The process is iteratively repeated until there is little to no change in the cluster assignments (Kassambara, 2017a). Figure 4.2 is a simple illustration for the process described, where K-means clustering is applied on a two-dimensional Iris flower dataset (frequently used by data scientists for illustration purposes), where k = 2.

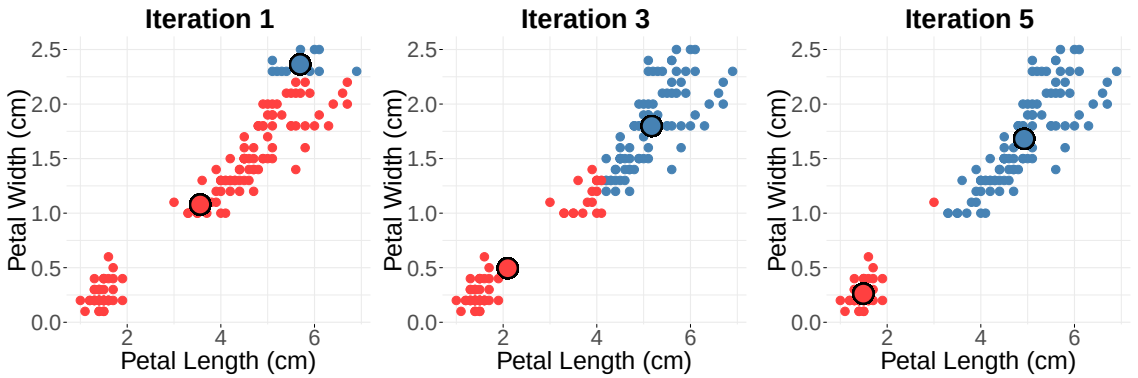

Figure 4.2: K-means clustering iterative process illustration

Once convergence is achieved, the result is considered final. However, this result may only be a local optimum, depending on where the starting points were and in this case, it wouldn't be the best possible cluster fit for the data (Trevino, 2016). Furthermore, the use of means suggests that the K-means clustering is highly sensitive to outliers (Kassambara, 2017a). It is generally known that K-means algorithm does not perform well for high-dimensional data (Johnson and Kim, 2012).

PAM

Partitioning Around Medoids (PAM) algorithm is also a deterministic clustering approach. It is designed to find k representative objects (medoids) from the observations in the dataset. After k medoids have been established, each observation is assigned to its nearest medoid and the clusters are constructed. Then, every medoid and non-medoid data point are swapped to minimise the sum of dissimilarities between the data points and their closest medoid (Kassambara, 2017a).

The main advantage of PAM over K-means described in section 4.5.1, is its robustness to noise and outliers, as it uses medoids instead of means as centroids of clusters (Murphy, 2014).

CLARA

Clustering Large Applications (CLARA) is an extension to PAM method, used to reduce the computing time and resources required for large datasets. It randomly splits the dataset into multiple subsets of fixed size and applies the PAM algorithm to produce the most appropriate set of medoids for the subset. Each data point is assigned to the closest medoid. Thereafter average dissimilarity between every object in the whole dataset and the medoid of its cluster is calculated. The process is repeated a multiple numbers of times until the average dissimilarity described above is minimised.

4.5.2 Hierarchical

An alternative to partitional clustering (Section 4.5.1) is hierarchical clustering. The result of hierarchical clustering is a dendrogram, which is a diagram representing a tree, based on similarities. An example dendrogram is shown in Figure 4.3, where the similarities between the cities are based on the distances between them. There's no requirement to pre-specify the number of clusters to be produced. However, once the quantity of clusters is specified the dendrogram is cut at a respective step and cluster results are extracted. In the example on Figure 4.3 if 2 clusters were required they would be identified as shown.

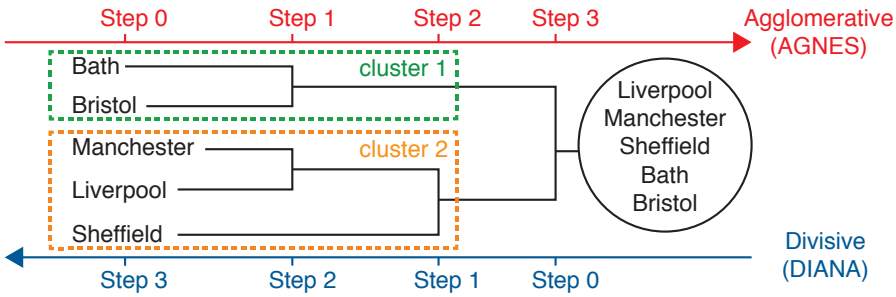

Figure 4.3: AGNES and DIANA hierarchical dendogram representation

The results from hierarchical clustering are usually easy to interpret and it is also relatively easy to spot any outliers. However, different levels of clustering make it difficult to identify an optimal number of clusters. Hierarchical clustering also requires significantly more computational power than the partitional algorithms (Tarabalka, Benediktsson and Chanussot, 2009). Another disadvantage of hierarchical algorithms is that they do not revisit already constructed clusters for the purpose of improvement (Chitra and Maheswari, 2017).

AGNES

Agglomerative Nesting Clustering (AGNES) is a form of hierarchical clustering that starts with a full set of items and combines the elements closest to each other into a cluster, at which point the cluster becomes an element. The process is repeated until all elements are merged into one big cluster containing all observations (Chitra and Maheswari, 2017). The red arrow in Figure 4.3 shows the steps taken in AGNES for building clusters .

DIANA

Divisive Analysis Clustering (DIANA) is another form of hierarchical clustering, where initially all data belongs to the same cluster, and the largest cluster is continuously split until until every element is separate (Kaufman and Rousseeuw, 2009). DIANA chooses two objects with maximum average dissimilarity (largest distance) and clusters the rest of the objects based on their similarity to the two most dissimilar chosen. The procedure is illustrated by the blue arrow in Figure 4.3.

4.5.3 Other clustering algorithms

There are other clustering algorithms which can be used for clustering of urban data, however, for the reasons described below they miss the scope of the current paper.

City Clustering Algorithm (CCA) is a fast and automated way to cluster specific values in a 2D dataset (Kriewald et al., 2019). It is used to construct cities based only on spatial geographical features (e.g. location of population, location of buildings). It is ,however, unable to account for any socio-economic features (Rybski et al., 2012) and has difficulty determining the radius of spatial search objectively (Chen et al., 2017).

There are a number of other clustering algorithms which are more computationally sophisticated, such as sparse manifold clustering and embedding (SMCE), Mumford-Shah segmentation, spectral-spatial diffusion learning (DLSS) and others (Murphy and Maggioni, 2018). But they are relatively newly developed and have their complexity, therefore are out of the scope of this paper, which is to assess readily available clustering techniques.

4.6 Goodness of fit

A rigorous validation procedure is required when comparing different combinations of dimensionality and clustering techniques proposed in Section 4.4 and Section 4.5. Additionally, an optimal number of clusters appropriate for the data analysed needs to be identified. Visually interpreting the plot of the clusters on the map is not deemed satisfactory, as the justification would be highly subjective.

Various validation measures have been proposed to determine the statistical properties of cluster analysis (Yeung, Haynor and Ruzzo, 2001; Datta and Datta, 2003; Kerr and Churchill, 2001). Internal validation metrics such as compactness and separation of the cluster partitions are the most important (Brock et al., 2008) when assessing clustering goodness of fit. Compactness evaluates the homogeneity of a cluster, by looking at the internal cluster variance. Separation assesses the extent of separation between clusters, generally, a distance measure between cluster centroids (Brock et al., 2008).

Compactness increases and separation decreases as the number of clusters is increased and therefore, the latest methods provide an individual score as a combination of both compactness and separation (Brock et al., 2008). Dunn Index (Dunn, 1974) and Silhouette Width (Rousseeuw, 1987) are the most well-known non-linear combinations of separation and

compactness. However, due to the noisy nature of the urban data, Silhouette Width measure is preferable to a more noise-sensitive Dunn Index measure (Handl, Knowles and Kell, 2005).

Silhouette width quantifies the degree of confidence in the allocation of an observations to a particular cluster. Silhouette Width coefficient takes values from -1 to 1. Values of observations that are well-clustered, with clear assignment of observations to cluster centres are closer to 1. In contrast, values being close to -1 mean that the algorithm is practically unable to find a significant cluster (Kaufman and Rousseeuw, 2009).

4.7 Comparing global clustering and local clustering quality

This section will describe the methodology of achieving objective number 3 specified in Chapter 3.

The established optimal clustering algorithm will be used for this objective. The procedure takes place in the following order:

1. PCA and t-SNE dimensionality reduction algorithms are applied to both global dataset (England) and a local dataset (BATHNES).
2. Two global datasets (PCA and t-SNE reduced) are subjected to the clustering algorithm identified as optimal in previous sections. Corresponding cluster map plots are produced.
3. Global map plots from PCA and t-SNE reduced datasets are assessed and benefits and disadvantages are outlined for both.
4. From the cluster map plots produced, the plots for the local area of BATHNES are extracted.
5. Two local datasets (PCA and t-SNE reduced) are subjected to the optimal clustering algorithm and corresponding cluster map plots are produced.
6. The cluster map plots from step 4 and 5 are compared and analysed based on their classification of areas, which depends on global versus local variance.

Chapter 5: Results and Analysis

As discussed in the earlier chapters, PCA and t-SNE dimensionality reduction methods have been applied to BATHNES dataset. Thereafter 5 different clustering algorithms (see Section 4.5) have been applied to the 2 datasets (PCA and t-SNE reduced). This way every configuration of dimensionality reduction and clustering algorithms can be assessed against each other.

5.1 Clustering on PCA dimensionality reduced data

5.1.1 Choosing the number of principal components

For PCA dimensionality reduced dataset, a scree plot is presented in Figure 5.1, which shows the proportion of the total variance that is represented by each PC (referred to as dimension on the plot). The cumulative scree plot in Figure 5.2 illustrates the proportion of the variance in the data that is explained by a certain number of PCs. For the chosen BATHNES dataset, six PCs explain approximately 75% of the variance in the data. This is deemed acceptable, as the remaining components do not add much value in explaining the variance of the dataset. Therefore, the first 6 PCs have been chosen for cluster analysis. Elimination of the remaining components also accounts for any redundancy created by interrelated variables.

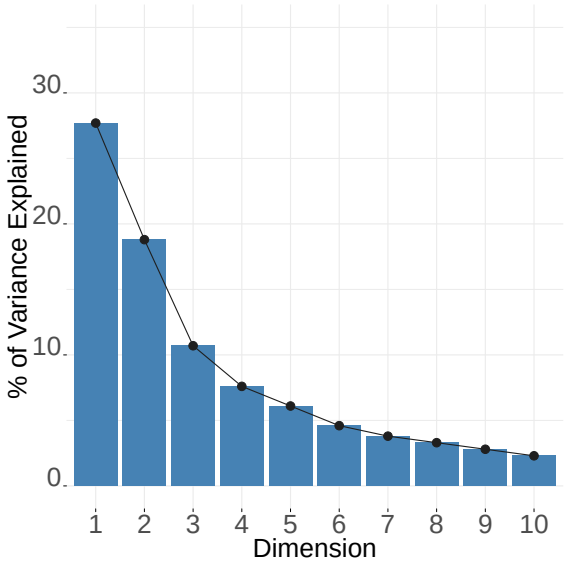

Figure 5.1: Scree plot on BATHNES data

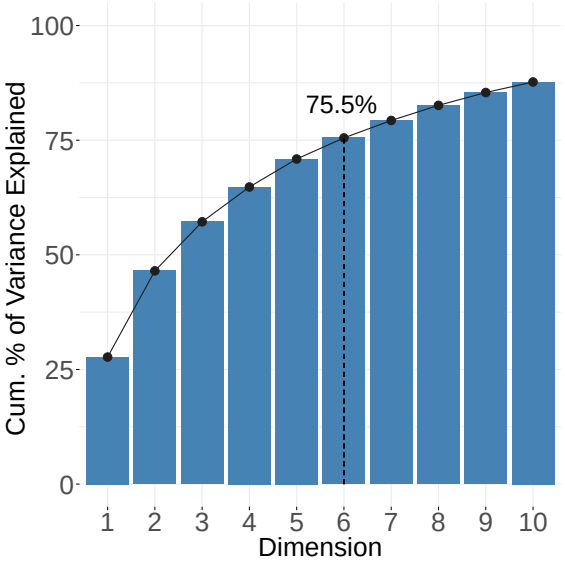

Figure 5.2: Cumulative scree plot on BATHNES data

5.1.2 Importance of variables on principal components

Before clustering is performed, an inference needs to be made into the importance of different variables when constructing PCs. Two factor maps have been produced in Figure 5.3, representing the contribution and cos2 factors. The factor plots only include the first ten PCs, because the others have very little significance in explaining the variance of the data.

Cos2 factor represents the quality of representation for the variables. The variables, in this case, are the loading scores calculated as described in Section 4.4. Contribution factor consists of contributions of the variables to the PCs. These factors are calculated in the following way:

$$cos2 = (coord)^2 \times 100$$

$$contribution = \frac{cos2}{\sum_{n=1}^{32} cos2}$$

Cos2 factor map (Kassambara, 2017b) in Figure 5.3 represents the quality of representation of the variables by a particular PC as a percentage. The summation of cos2 coefficients across all 32 PCs equals to 100%, since the combination of all PCs should explain 100% of a feature. In other words, the cos2 factor map explains how much of the variable is explained by each PC. From the cos2 factor map, it can be seen that variables representing densities as well as the sheer size of the areas are very well represented by PC1. Variables representing buildings and built-up areas are represented very well by PC2. The number of dwellings and household spaces have the strongest representation by PC3. Socio-economic data is most strongly represented by PC4. Using cos2 factors, the quality of representation of each variable by each PC can be identified. Overall cos2 factor map in Figure 5.3, shows that the majority of variance in variables is explained by the first few components, reinforcing what was mentioned in Section 5.1.1.

Contribution factor map (Kassambara, 2017b) in Figure 5.3 explains the percentage contribution a particular variable has on a given PC. The addition of contributions across all variables for a single PC equals to 100%. Henceforth, the contribution factor map explains how much of the PC is explained by each variable. It can be seen that for PC1, there is no dominant variable in terms of how much it contributes to this PC. Many variables contribute equally well. The same pattern is seen across the first four PCs. Starting with PC5, there tends to be one variable which is the strongest contributor. For example, male resident population is a significantly stronger contributor to PC7 than any other variable. This might suggest that there is a particularly high male resident population for a few LSOA regions analysed, which is not representative of the overall variance of the data. Discounting the PCs above the first six is therefore beneficial since they do not seem to represent the overall trend of the data.

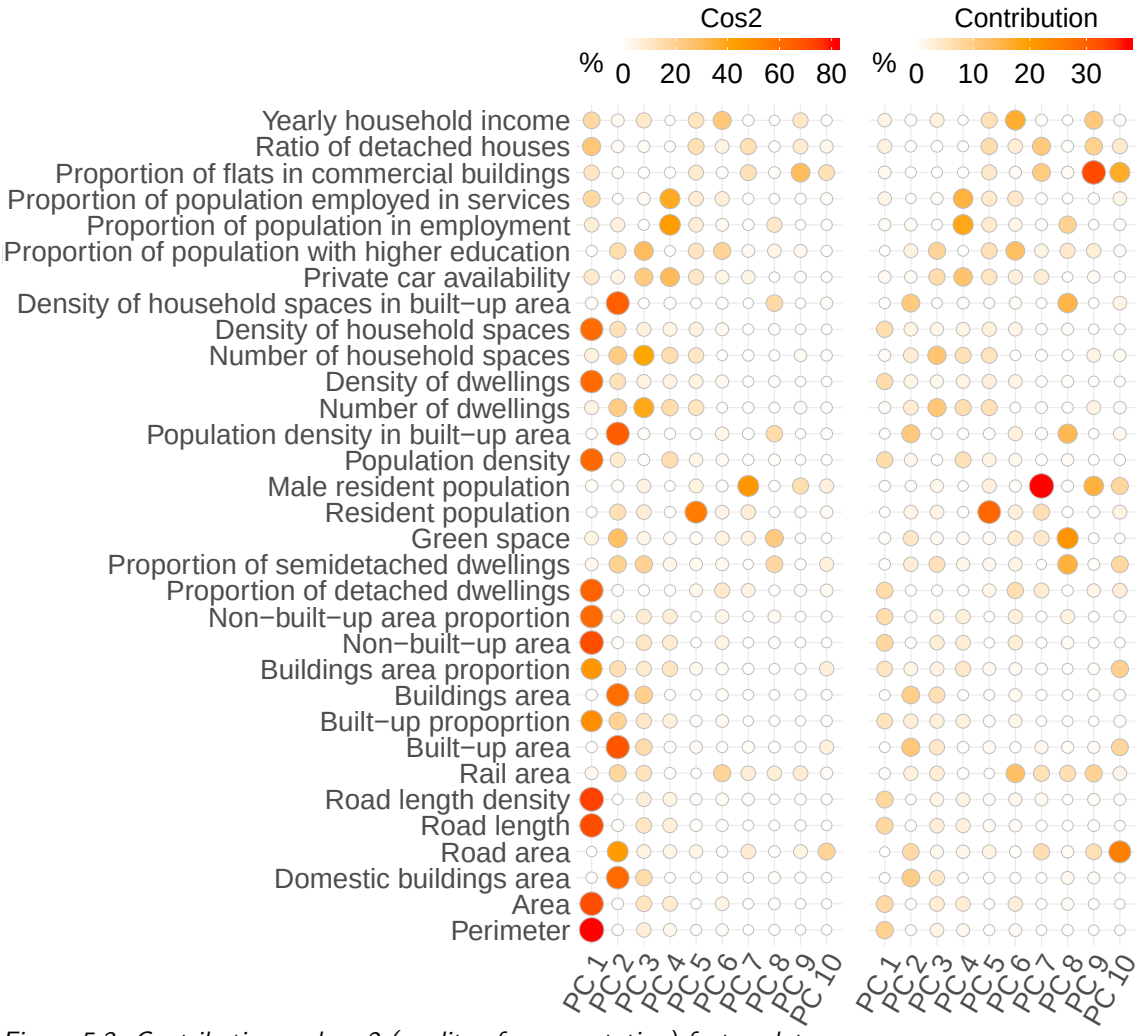

Figure 5.3: Contribution and cos2 (quality of representation) factor plots

5.1.3 Clustering analysis

The results of clustering on PCA dimensionality reduced dataset are shown in Figure 5.4. Visually, clusters produced from PCA reduced dataset are not presented very clearly. The clusters look rather convoluted, but this is a result of the plotting against the first two dimensions only (PC1 and PC2). The other four dimensions which have been taken into account are projected onto this 2D plot. It needs to be noted that the values presented on the axis represent the coordinates of each LSOA unit on the PCs1 and PCs2.

Silhouette plots in Figure 5.4 show that most of the data points have a clear assignation to a cluster since their values are well above zero. Nonetheless, for all clustering algorithms, there is at least one cluster in which data points have values below zero. This indicates that those points are not assigned properly, but it was the best fit the algorithm found. DIANA clustering performance is particularly strong among all the algorithms presented. DIANA average silhouette width of 0.31 is significantly higher than the K-means's 0.24. Furthermore, by analysing the DIANA silhouette plot visually, it can be noted that there are only two data points which have silhouette widths of below zero. Based on silhouette widths alone, DIANA clustering is comparatively the optimal clustering algorithm when PCA dataset is used for six clusters.

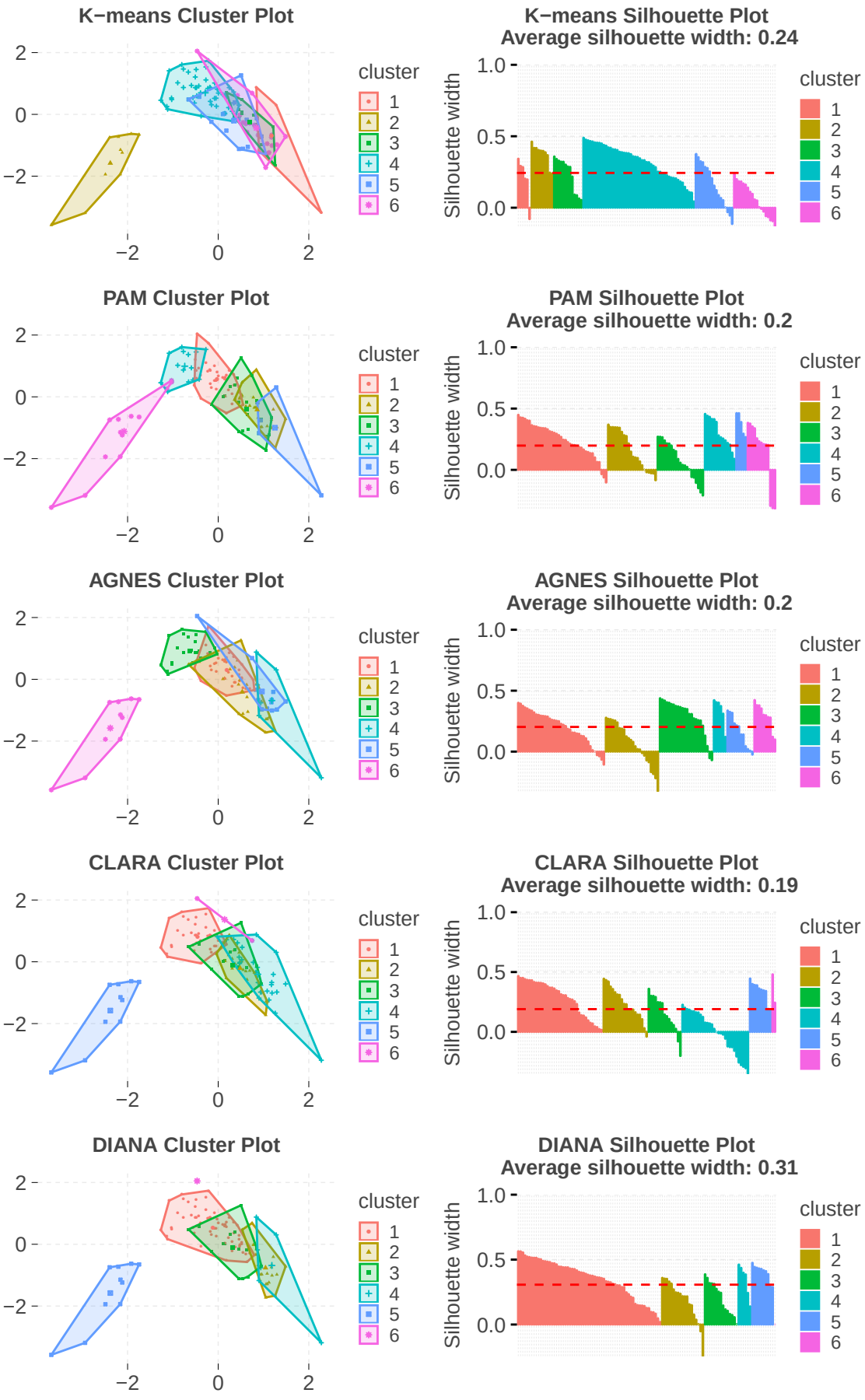

Figure 5.4: Clustering of BATHNES variables using PCA reduced data

5.2 Clustering on t-SNE dimensionality reduced data

Contrary to PCA, t-SNE is a newer and more advanced method and it uses all the dimensions of the data, of which there are 32 in this study. t-SNE attempts to take every dimension of the dataset into account and map this data onto a two-dimensional plane. Therefore, when clustering is performed on t-SNE transformed dataset, it takes the two-dimensional points from t-SNE output.

Clustering results together with silhouette coefficient plots for t-SNE reduced dataset are presented in Figure 5.5. From the cluster plots, it is clear that visually t-SNE clusters have much better separation and compactness than the cluster plots done on the PCA dimensionality reduced data presented in Figure 5.4. This is due to t-SNE's non-linear ability to embed data points neighbourhood in multiple dimensions and map it onto a two-dimensional plane.

The average silhouette coefficients are significantly higher than the ones presented for the PCA reduced dataset. A big contribution to those high silhouette values is silhouette assessment for a 2D clustered data points, unlike PCA's 6D space. The data points are generally much closer to each other in 2D space, which positively influences the silhouette coefficients. The performance of all clustering algorithms is very similar when comparing the average silhouette coefficients since they only vary between 0.56 - 0.58 across clustering algorithms. K-means algorithm is marginally a better fit than the other algorithms when judged using silhouettes metric.

However, because t-SNE is a stochastic method yielding slightly different results on every run, it is considered inadequate to judge a clustering algorithm's performance based on the t-SNE dimensionality reduced dataset. Especially because the range of silhouette coefficients is relatively small and the stochastic approach of t-SNE would most definitely lead to a different algorithm performing better on every run.

Unlike PCA, in t-SNE dimensionality reduction the input variables can no longer be identified after t-SNE has run (Pathak, 2018). Therefore, no inference can be made about the contributing variables from the output of t-SNE. Also, t-SNE's is less capable of decorrelating the variables, unlike PCA, therefore there might be a certain degree of redundancy in the final output.

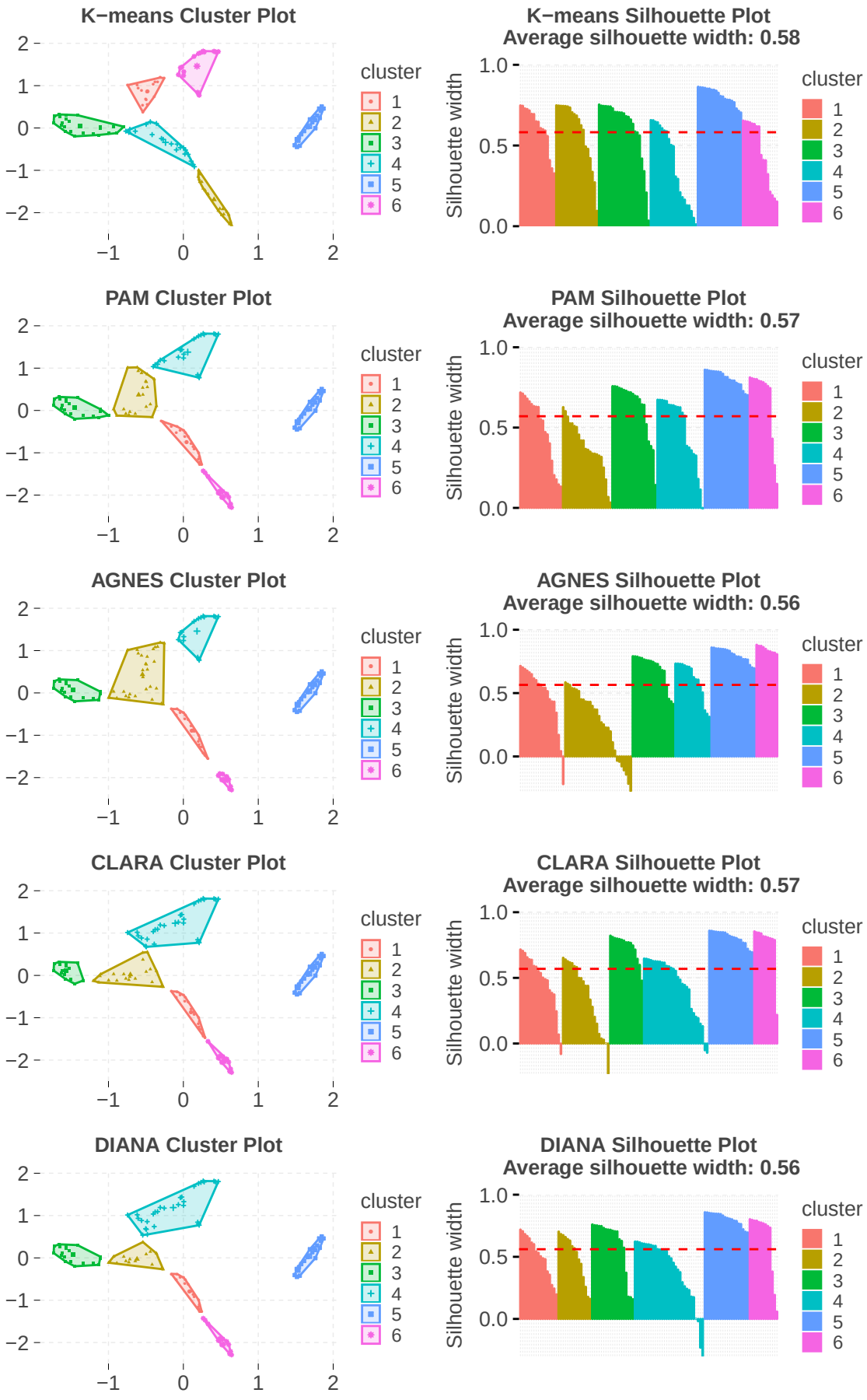

Figure 5.5: Clustering of BATHNES variables using t-SNE reduced data

5.3 Goodness of fit assessment

5.3.1 Validation based on silhouette width

In Section 5.1 and Section 5.2 clustering has been set to six clusters, for the initial run, and DIANA clustering algorithm proved to be comparatively better than others. However, to identify the optimal amount of clusters and an optimal clustering algorithm, silhouette coefficients need to be compared for different combinations of numbers of clusters and clustering algorithms. Figure 5.6 clearly illustrates the result of such analysis.

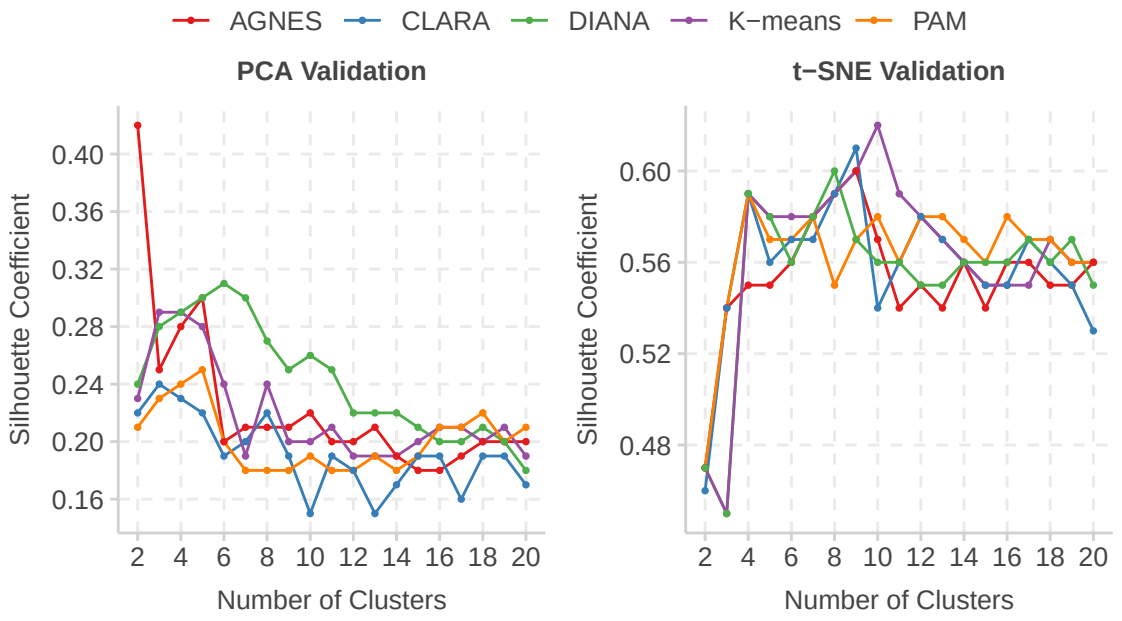

Figure 5.6: Silhouette coefficient comparison across various combinations of number of clusters and clustering methods

From Figure 5.6 it is immediately evident that silhouette coefficients for t-SNE dimensionality reduced data are significantly higher than the silhouette coefficients for PCA dimensionality reduced data. t-SNE's ability to embed neighbourhood from multiple dimensions and transform it into a two-dimensional space is certainly contributing to those high silhouette coefficients. Nonetheless, the 2D dataset in this case, as opposed to 6D for PCA, is what contributes to those high silhouette coefficients the most. This observation holds true because the points in a 2D dataset are usually much more closely spaced, than in multiple dimensions.

Furthermore, because t-SNE is a non-linear algorithm which performs different transformations on different regions stochastically (Wattenberg, Viégas and Johnson, 2016), the t-SNE dataset cannot be used for an objective measure of performance of clustering algorithms. Henceforth, only the PCA validation is used in order to validate the optimal number of clusters and the optimal algorithm.

By examining the PCA validation plot in Figure 5.6, it is clear that AGNES algorithm has the highest silhouette coefficient of 0.42 when the number of clusters equals to two. However, it is well-known there are more than two types of urban areas and for this study, the scope

is to classify multiple types of urban areas. Hence, the next best performing algorithm is DIANA, which has a silhouette coefficient of 0.31 for 6 clusters. This suggests that the optimal combination for clustering is DIANA clustering method with 6 clusters, for which the silhouette plot and the cluster plot can be found in Figure 5.4.

5.3.2 Qualitative check of the map cluster plots

Qualitatively assessing the performance of the clustering on the map cluster plots is very important. Patterns and errors can be spotted very easily using the qualitative approach.

Although it is difficult to adjudicate whether the clustering pattern is reasonable in rural areas, there is a pattern by which clustering can be assessed visually. Attention should be dedicated to the similarities between cities, towns and large villages. Figure 5.7 shows the plots for both dimensionality reduction techniques (PCA and t-SNE) in combination with five clustering methods under assessment. The cities and towns have been given large text labels, with villages given medium text labels and Bath regions given small text labels. Visually analysing cluster map plots for BATHNES dataset, in particular, is advantageous, because of familiarity with the area.

In all the plots, the light blue colour coded LSOA regions are the most rural of the kind. It needs to be noted that even though two LSOA regions belong to the same cluster on two different maps, the colour coding might differ.

Overall the correct pattern is seen on all the maps, as there are multiple cluster types in and near Bath, which is also the biggest city in BATHNES. This is undoubtedly correct since the city of Bath contains much more varied settlements than other towns and villages. The other three of the biggest towns in the county (Keynsham, Midsomer Norton and Radstock) are also surrounded by LSOA's of various cluster types, although fewer than for Bath. Relatively large villages like Paulton and Timsbury are usually surrounded by a single cluster type. Small villages form part of the rural, light blue cluster in the majority of maps on Figure 5.7. All of those patterns suggest that the overall clustering methodology is correct, although the more specific delineation of boundaries is hard to judge qualitatively.

In all configurations of dimensionality reduction techniques and clustering methods, it can be noted that LSOA's north-east of Weston and the LSOA's surrounding Clutton belong to the same cluster with one exception being PAM method on the PCA reduced dataset. LSOA's surrounding Pensford have been clustered together with areas north of Batheaston in all the clustering algorithms except K-means, PAM and DIANA for t-SNE dataset. This can be valuable information when comparing two LSOA areas to each other. However, establishing which one of the maps represents the optimal clustering method is a highly subjective task. This is because the influencing values for clustering are a combination of the variables specified in Table 4.1 and Table 4.2. The clusters can't be directly back-correlated to a meaningful value. But this is also the primary reason we have used algorithms for this task because manually combining the variables is deemed impossible.

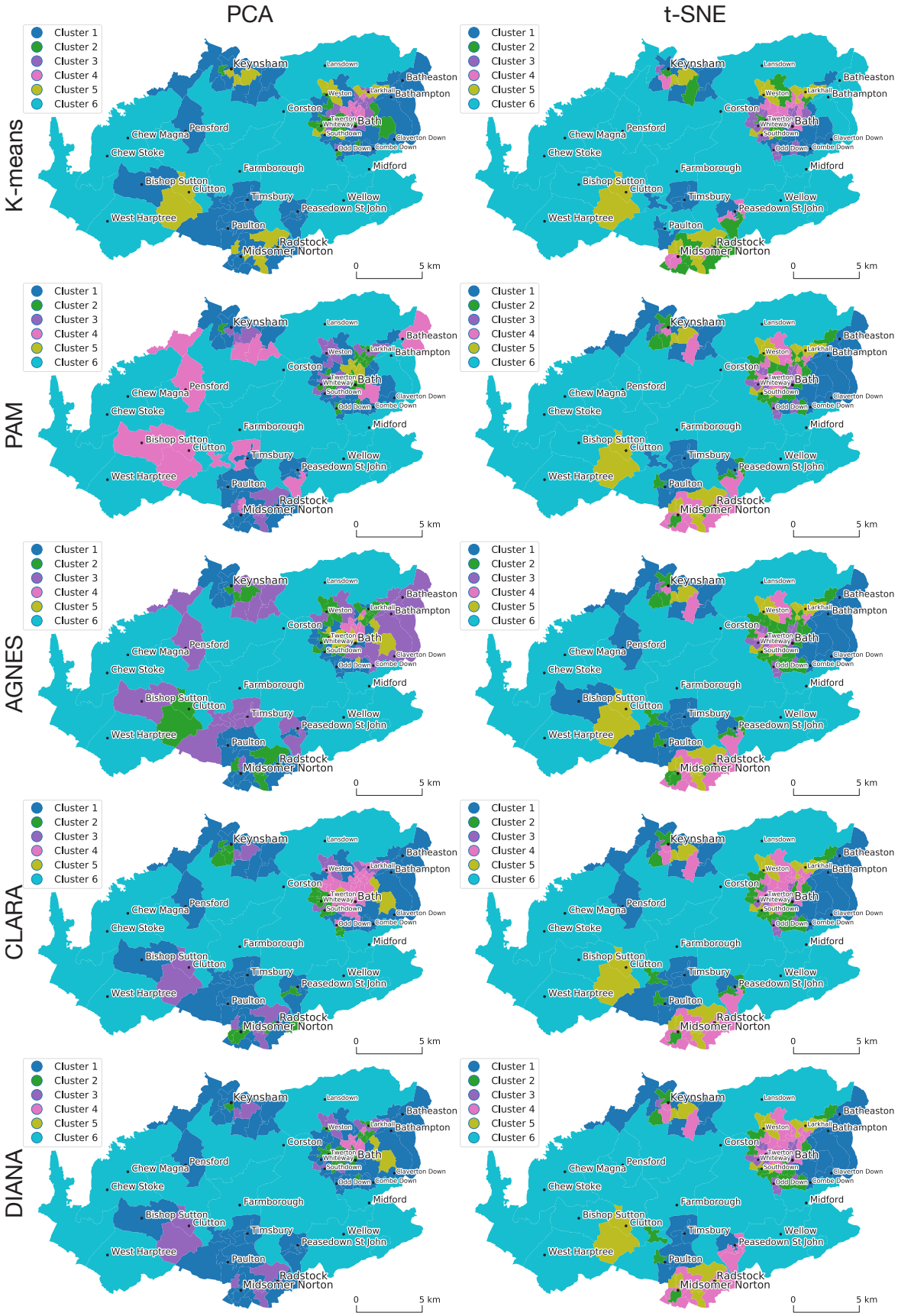

Figure 5.7: 5 clustering methods plots using PCA and t-SNE dimensionality reduction techniques on BATHNES dataset

5.4 Comparison of global clustering versus local clustering on local scale

5.4.1 DIANA clustering on PCA global dataest (England)

Figure 5.8 shows the cluster map plot on the global (England) PCA dimensionality reduced dataset. DIANA clustering has been applied in this case, as it is considered a comparatively better algorithm from the previous analysis.

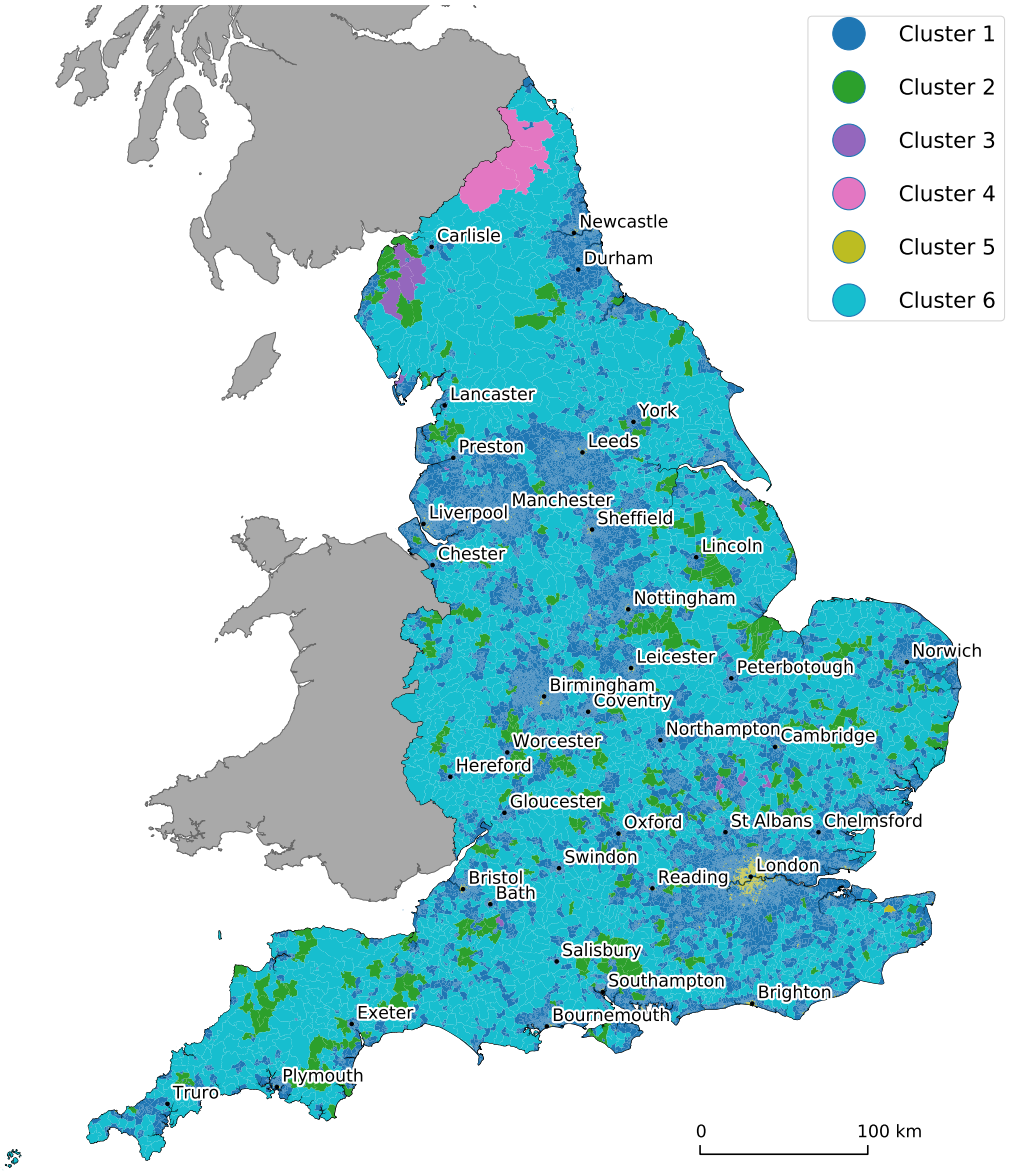

Figure 5.8: Global England DIANA clustering on PCA dataset

An overall pattern of distinguishing between rural and urban areas can be spotted in Figure 5.8 by differentiating the light blue and dark blue colours. The dark yellow colours seem to represent city centres. As expected PCA dimensionality reduction preserves the global shape, but not much definition can be seen locally on a smaller scale. Furthermore, cluster 3 and 4 seem to suggest that those are outliers, which if eliminated from the sample might produce a better representation plot.

5.4.2 DIANA clustering on t-SNE global dataest (England)

Figure 5.9 presents a cluster map plot on the global (England) t-SNE dimensionality reduced dataset. DIANA clustering has been applied in this case, as it is considered an optimal algorithm from the previous analysis.

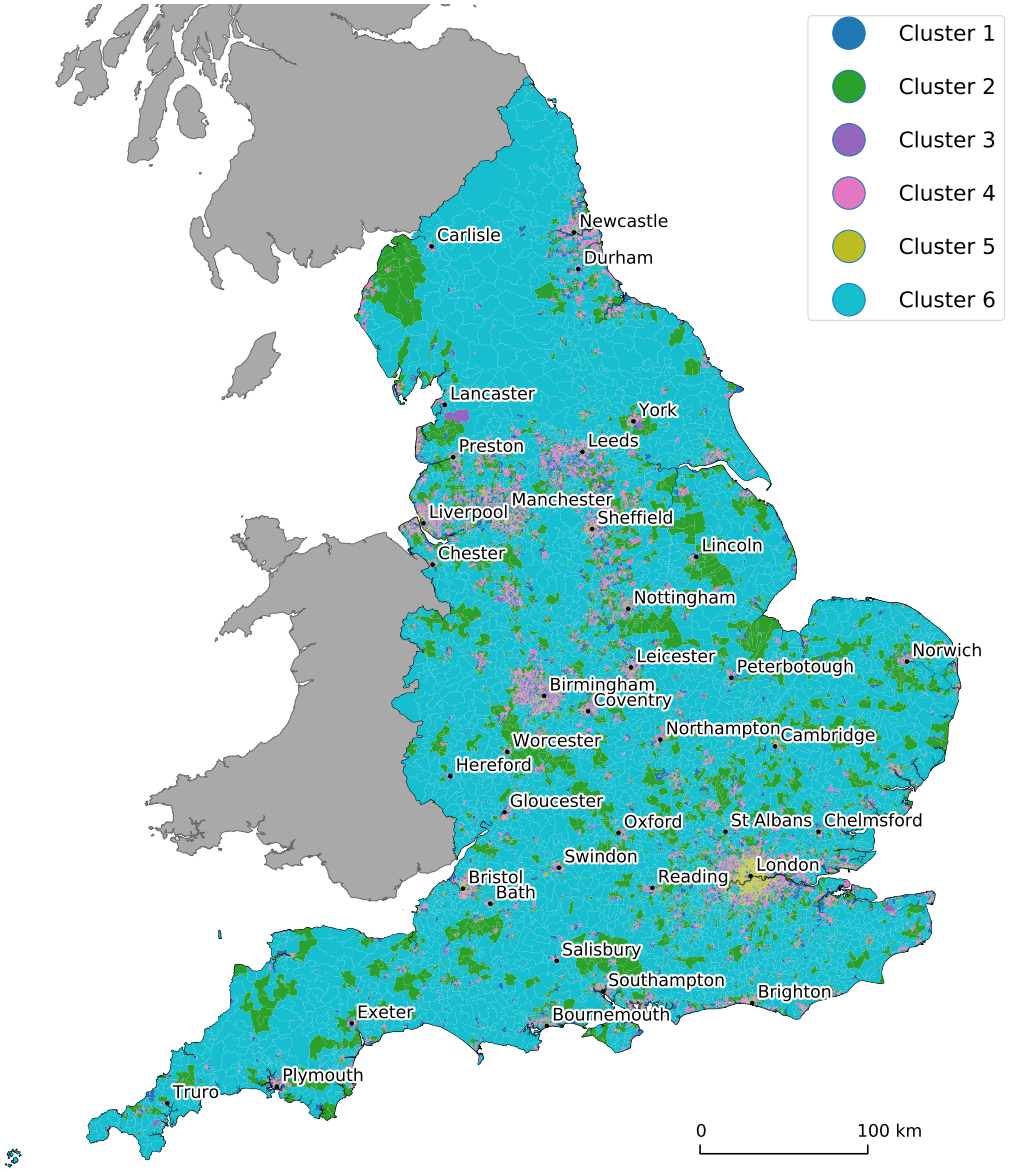

Figure 5.9: Global England DIANA clustering on t-SNE dataset

In Figure 5.9 it is clear that the rural areas are represented by cluster number 6. Unlike the cluster plot on PCA data (Figure 5.8), there seems to be more differentiation between the urban areas. Instead of one cluster (cluster 1 in Figure 5.8) representing the majority of urban areas, these urban areas are also subdivided into several groups. Therefore, as mentioned in Section 4.4, this plot proves t-SNE's great ability in preserving the local structure of the data as well as capturing the global shape.

5.4.3 Comparison of global clustering and local clustering

The global maps have been presented in Section 5.4.1 and Section 5.4.2 for PCA and t-SNE datasets respectively. In this section BATHNES cluster map is extracted from the global (England) cluster maps in Figure 5.8 and Figure 5.9 and compared to the local cluster maps from Figure 5.7. This has only been done for DIANA clustering algorithm, as it is considered optimal. Figure 5.10 shows the cluster map plots representing the comparison described.

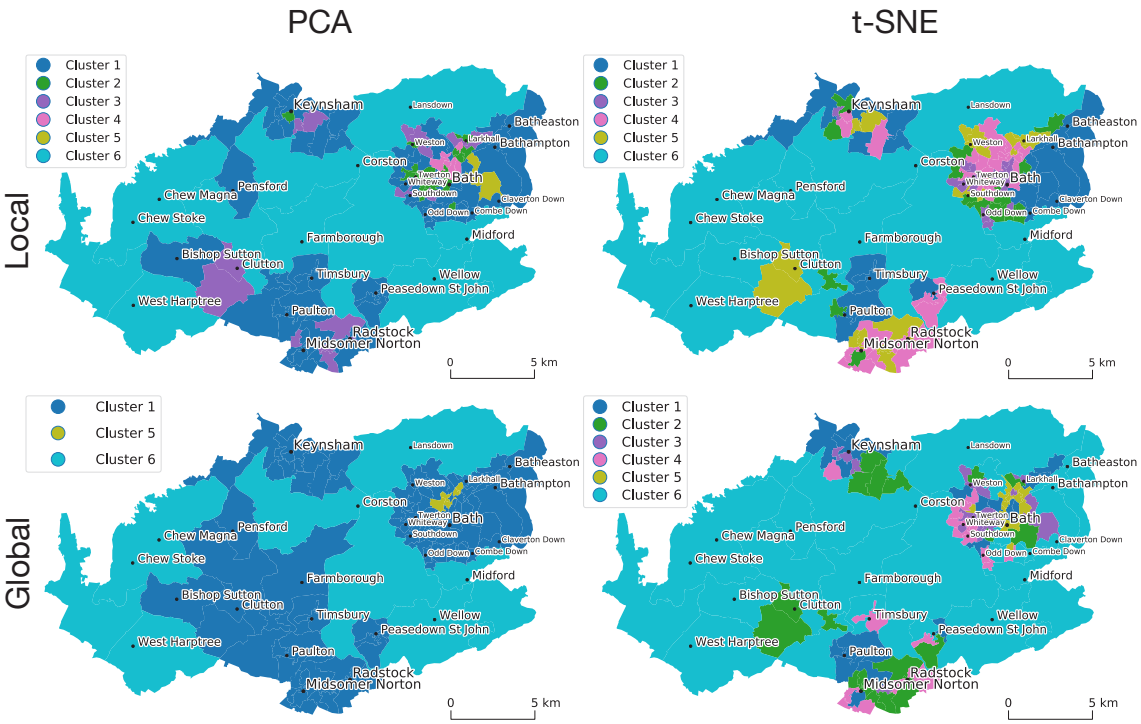

Figure 5.10: Global versus local clustering comparison

It is evident on the global PCA cluster map in Figure 5.10 that in this case clustering only identifies rural areas (cluster 1), somewhat urban areas (cluster 6) and dense urban areas (cluster 6). LSOA regions under cluster number 5 are considered as dense urban areas, as they fall in the same cluster as central parts of London on Figure 5.8. However, when clustering is done on the local scale, various types of urban areas are identified. This difference between the global and local cluster plot on the local scale can be attributed to PCA's inability to preserve the local structure of the data. In other words, it is not able to assess the similarities on a fine scale. This is because PCA uses the total variance of the dataset to construct PCs, and therefore, the larger the global variance, the less likely it will be able to preserve the local structure.

Clustering performed on the t-SNE dimensionality reduced dataset is shown on the right part of Figure 5.10. Although the clusters differ from the global to local maps, a certain consistency in the local structure of the data is observed, as many LSOA regions fall in the same category. Evidently, t-SNE algorithm has better capabilities at preserving the local similarities between observations than PCA, therefore, when clustering is performed it can identify those similarities. This performance is attributed to t-SNE's perplexity parameter, which tells it how to balance out the concentration on global versus local structure.

Chapter 6: Discussion

The research presented introduces a new understanding of some of the readily available clustering algorithms and dimensionality reduction techniques when applied to urban data.

To reiterate, PCA was employed for two main reasons. Firstly to reduce the number of dimensions required for clustering and secondly to reduce the colinearity between the variables in the data. It was found that six principal components explain about 75% of the variance for the BATHNES dataset. From the cos2 factor map in Figure 5.3 it could further be deduced which variables were best represented by those six principal components chosen. Therefore, one can use this process alone to decide which variables represent most of the variance in the dataset. The advantages of PCA include its ability to linearly decorrelate the variables, reduce the number of dimensions and the option of referring back to the input variables. However, when the global versus local clustering on the PCA data were compared it was found that PCA is highly dependent on the global variance, therefore, not able to preserve the local structure. Also, from the cluster map plot for England (Figure 5.8) it can be deduced that clustering on PCA dataset makes it susceptible to outliers. This paper proposes using PCA dimensionality reduction in instances where inference back to the variables is extremely important and the interest lies solely in the global structure of the data.

t-SNE dimensionality reduction algorithm proved to be the optimal dimensionality reduction technique to be used when the local structure is just as important as the global shape of the data. England cluster map plot of the t-SNE dataset (Figure 5.9) clearly illustrates this advantage over PCA. However, there is no direct way to infer to the input variables, therefore, one can only rely on the cluster map output in an attempt of identifying what variables might have influenced clustering the most. t-SNE algorithm should be employed when global and local structure of the data is equally important, but no inference to the input data is required.

In terms of computational resources required, t-SNE is much more power intensive. For comparison, in dimensionality reduction of the England dataset, t-SNE needed to be run on a high-performance computer, whereas PCA was run on a standard personal computer. Nonetheless, clustering algorithms on the global England dataset were run on a high-performance computer, as these algorithms are very resource intensive. Therefore, computational resources available is an important factor to consider when performing dimensionality reduction and clustering.

After assessing various clustering algorithms, it was found that DIANA clustering algorithm with six clusters performed comparatively better than the other clustering algorithms. The validation analysis was done only on the PCA dimensionality reduced data since it is stable and outputs the same silhouette measure on every attempt. The conclusion that DIANA clustering is optimal is however limited, due to the use of PCA dimensionality reduced dataset only and the use of a single validation metric.

Chapter 7: Conclusion

Continuous high rates of urbanisation are leading to irreversible damage to the world's ecosystems. Nonetheless, literature suggests that urban areas can be some of the most efficient places to be living in. Therefore, it was established that an accurate definition of urban/rural boundaries is essential to future sustainable planning and development policies.

This research has provided valuable insights into the arrangement boundaries between rural, urban and other type areas. Previous researchers have taken a simplified approach of selecting the most evident characteristics of urban/rural areas, which included a limited number of either land use or population-related features. All of those metrics have their own limitations. This study, in contrast, has used a more broad definition, which used a great range of economic, social and geospatial factors. The inclusion of such a great variability of factors makes this approach a multi-disciplinary one, which is convenient for planning and policy design. Furthermore, the data used in this study was at the highest resolution possible, as it contained the smallest census units available.

Two dimensionality reduction techniques have been applied to the dataset, after which five different clustering methods have been applied to the transformed datasets. In the scope of this paper, the optimal number of clusters was determined to be six and the optimal clustering algorithm has been identified as DIANA.

The validation was done using silhouette widths and on the PCA dataset only, since t-SNE transformed dataset had silhouette values within a very small range. Furthermore t-SNE is a stochastic method, so would yield marginally different silhouette results every time clustering is run on the t-SNE dimensionality reduced dataset. Therefore, t-SNE could not be used for validation of clustering algorithms.

Global versus local clustering was also compared, where it was found that t-SNE has a much better ability to preserve both local and global structure and makes clustering less susceptible to outliers. However, this advantage comes at a great computational resources cost and inability to infer back to the input variables.

Overall, the aims and objectives of this study were met - in that, a thorough comparison of clustering algorithms was performed and an optimal one has been established. Global versus local clustering was compared and the results discussed. Furthermore, a very crucial difference in performance between the two dimensionality reduction techniques used was discovered.

There is a need for further research into other validation metrics to be used to validate clustering algorithms. Other dimensionality reduction techniques should also be explored in the future. In this study global dataset was only used for global versus local variance comparison, due to limited computational resources available for the algorithm to run the England dataset. With computational resources becoming increasingly powerful in the future, it is a possibility to repeat this analysis on a larger dataset (e.g. England).

Chapter 8: Bibliography

Anon, 2017. Principal Component Animation [Online]. Available from: https://gist.github.com/anonymous/7d888663c6ec679ea65428715b99bfdd [Accessed 20 March 2019].

Baxter, L.K. and Sacks, J.D., 2014. Clustering cities with similar fine particulate matter exposure characteristics based on residential infiltration and in-vehicle commuting factors. *Science of the Total Environment*, 470-471, pp.631–638.

Brock, G., Pihur, V., Datta, S. and Datta, S., 2008. clValid: An R package for cluster validation. *Journal of Statistical Software* [Online], 25(4), pp.1–22. Available from: http://www.jstatsoft.org/v25/i04/ [Accessed 2 March 2019].

Calcott, A. and Bull, J., 2007. *Ecological footprint of British city residents*. Surrey: World Wide Fund for Nature.

Cao, Z., Wang, S., Forestier, G., Puissant, A. and Eick, C.F., 2013. Analyzing the composition of cities using spatial clustering. *Proceedings of the 2nd ACM SIGKDD International Workshop on Urban Computing*. ACM, p.14.

Census Data, 2011. [Online]. Available from: https://www.ons.gov.uk/census/2011census/2011censusdata [Accessed 20 December 2018].

Chen, Y., Wang, J., Long, Y., Zhang, X., Liu, X. and Li, X., 2017. *Defining Urban Boundaries by Characteristic Scales* [Online]. Available from: https://arxiv.org/abs/1710.01869 [Accessed 3 March 2019].

Chi, G., Liu, Y., Wu, Z. and Wu, H., 2015. Ghost Cities Analysis Based on Positioning Data in China.

Chitra, K. and Maheswari, D., 2017. A Comparative Study of Various Clustering Algorithms in Data Mining. *International Journal of Computer Science and Mobile Computing*, 6(8), pp.109–115.

Coombes, M., 2016. Travel to work area analysis in Great Britain:2016. *Office for National Statistics*, pp.1–27.

Datta, S. and Datta, S., 2003. Comparisons and validation of statistical clustering techniques for microarray gene expression data. *Bioinformatics*, 19(4), pp.459–466.

Dunn, J.C., 1974. Well-Separated Clusters and Optimal Fuzzy Partitions. *Journal of Cybernetics*, 4(1), pp.95–104.

El Garouani, A., Mulla, D.J., El Garouani, S. and Knight, J., 2017. Analysis of urban growth and sprawl from remote sensing data: Case of Fez, Morocco. *International Journal of Sustainable Built Environment*, 6(1), pp.160–169.

Elson, M.J., Walker, S., Edge, J. and Macdonald, R., 1993. *The Effectiveness of Green Belts*. London: HMSO Books.

Fang, L. and Wang, Y., 2018. Multi-disciplinary determination of the rural/urban boundary: A case study in Xi'an, China. *Sustainability (Switzerland)*, 10(8), pp.1–13.

Foley, J.A., DeFries, R., Asner, G.P., Barford, C., Bonan, G., Carpenter, S.R., Chapin, F.S., Coe, M.T., Daily, G.C., Gibbs, H.K. et al., 2005. Global consequences of land use. *Science*, 309(5734), pp.570–574.

Gao, H., Chen, J., Wang, B., Tan, S.C., Lee, C.M., Yao, X., Yan, H. and Shi, J., 2011. A study of air pollution of city clusters. *Atmospheric Environment*, 45(18), pp.3069–3077.

Handl, J., Knowles, J. and Kell, D.B., 2005. Computational cluster validation in post-genomic data analysis. *Bioinformatics*, 21(15), pp.3201–3212.

Hasse, J.E. and Lathrop, R.G., 2003. Land resource impact indicators of urban sprawl. *Applied Geography*, 23(2-3), pp.159–175.

Huang, J., Lu, X.X. and Sellers, J.M., 2007. A global comparative analysis of urban form: Applying spatial metrics and remote sensing. *Landscape and urban planning*, 82(4), pp.184–197.

IBM, 2014. IBM SPSS Statistics V23.0 documentation [Online]. Available from: `https://www.ibm.com/support/knowledgecenter/ru/SSLVMB{_}23.0.0/spss/product{_}landing.html` [Accessed 10 March 2019].

Jensen, J.R. and Cowen, D.C., 1999. Remote sensing of urban/suburban infrastructure and socio-economic attributes. *Photogrammetric engineering and remote sensing*, 65, pp.611–622.

Johnson, C. and Kim, M., 2012. Clustering of Cities by Craigslist Posts. *Stanford University - CS 229*, pp.1–5.

Jolliffe, I.T., 2002. *Principal Component Analysis*. 2nd ed. New York: Springer.

Kasanko, M., Barredo, J.I., Lavalle, C., McCormick, N., Demicheli, L., Sagris, V. and Brezger, A., 2006. Are european cities becoming dispersed? A comparative analysis of 15 european urban areas. *Landscape and Urban Planning*, 77(1-2), pp.111–130.

Kassambara, A., 2017a. *Practical Guide To Cluster Analysis in R: Unsupervised Machine Learning*. 1st ed. STHDA.

Kassambara, A., 2017b. *Practical Guide To Principal Component Methods in R: PCA, M (CA), FAMD, MFA, HCPC, factoextra*. 1st ed. STHDA.

Kaufman, L. and Rousseeuw, P.J., 2009. *Finding Groups in Data - An introduction to Cluster Analysis*, vol. 344. John Wiley & Sons.

Kendig, H., 1976. Cluster analysis to classify residential areas: A Los Angeles application. *Journal of the American Planning Association*, 42(3), pp.286–294.

Kerr, M.K. and Churchill, G., 2001. Bootstrapping cluster analysis: Assessing the reliability of conclusions from microarray experiments. *Proceedings of the National Academy of Sciences*, 98(16), pp.8961–8965.

Kontgis, C., Schneider, A., Fox, J., Saksena, S., Spencer, J.H. and Castrence, M., 2014.

Monitoring peri-urbanization in the greater Ho Chi Minh City metropolitan area. *Applied Geography*, 53, pp.377–388.

Kriewald, S., Fluschnik, T., Reusser, D. and Rybski, D., 2019. *Orthodromic Spatial Clustering* [Online]. R package version 1.0.4. Available from: `https://CRAN.R-project.org/package=osc` [Accessed 15 March 2019].

Likas, A., Vlassis, N. and J. Verbeek, J., 2003. The global k-means clustering algorithm. *Pattern Recognition*, 36(2), pp.451–461.

Liu, Z., He, C., Zhou, Y. and Wu, J., 2014. How much of the world's land has been urbanized, really? A hierarchical framework for avoiding confusion. *Landscape Ecology*, 29(5), pp.763–771.

Long, Y., 2016. Redefining Chinese city system with emerging new data. *Applied Geography*, 75, pp.36–48.

Maaten, L.v.d. and Hinton, G., 2008. Visualizing data using t-SNE. *Journal of machine learning research*, 9(Nov), pp.2579–2605.

MacQueen, J. et al., 1967. Some methods for classification and analysis of multivariate observations. *In proceedings of the fifth berkeley symposium on mathematical statistics and probability*. pp.281–297.

Madlener, R. and Sunak, Y., 2011. Impacts of urbanization on urban structures and energy demand: What can we learn for urban energy planning and urbanization management? *Sustainable Cities and Society*, 1(1), pp.45–53.

Masucci, A.P., Arcaute, E., Hatna, E., Stanilov, K. and Batty, M., 2015. On the problem of boundaries and scaling for urban street networks. *Journal of the Royal Society Interface*, 12(111).

Moreno, E.L., 2017. *Concepts, definitions and data sources for the study of urbanization: the 2030 Agenda for Sustainable Development* [Online]. New York: United Nations. Available from: `http://www.un.org/en/development/desa/population/events/pdf/expert/27/papers/II/paper-Moreno-final.pdf` [Accessed 12 March 2019].

Murphy, C., 2014. Mod-03 Lec-27 K-Medoids and DBSCAN [Online]. Available from: `https://nptel.ac.in/courses/106106046/27` [Accessed 5 February 2019].

Murphy, J.M. and Maggioni, M., 2018. Unsupervised Clustering and Active Learning of Hyperspectral Images With Nonlinear Diffusion. *IEEE Transactions on Geoscience and Remote Sensing*, pp.1–17.

National Research Council, 2009. *Driving and the built environment: The effects of compact development on motorized travel, energy use, and CO2 emissions*. Washingotn, D.C.: Transport Research Board.

Ordnance Survey, 2017a. OS MasterMap Integrated Transport Network Layer [GML geospatial data] [Online]. Available from: `https://www.ordnancesurvey.co.uk/business-and-government/help-and-support/products/itn-layer.html`

[Accessed 20 December 2018].

Ordnance Survey, 2017b. OS MasterMap Topography Layer [GML geospatial data] [Online]. Available from: `https://www.ordnancesurvey.co.uk/business-and-government/products/mastermap-products.html` [Accessed 20 December 2018].

Osorio, B., 2017. *Characterizing the relationship between energy and urban form using data, scaling and combined metrics*. Ph.D. thesis. University of Bath, Bath.

Pacione, M., 2009. *Urban geography a global perspective*. 3rd ed. New York: Routledge.

Pateman, T., 2011. Rural and urban areas: comparing lives using rural/urban classifications. *Regional Trends*, 43(1), pp.11–86.

Pathak, M., 2018. *Introduction to t-SNE* [Online]. Available from: `https://www.datacamp.com/community/tutorials/introduction-t-sne` [Accessed 3 March 2019].

Pituch, K.A. and Stevens, J.P., 2015. *Applied multivariate statistics for the social sciences: Analyses with SAS and IBM's SPSS*. 6th ed. London: Routledge.

Python Development Core Team, 2015. *Python: A dynamic, open source programming language.* [Online]. Python Software Foundation. Available from: `https://www.python.org/` [Accessed 15 March 2019].

R Development Core Team, 2008. *R: A language and environment for statistical computing* [Online]. Vienna, Austria: R Foundation for Statistical Computing. Available from: `http://www.R-project.org` [Accessed 15 March 2019].

Rousseeuw, P.J., 1987. Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. *Journal of Computational and Applied Mathematics*, 20, pp.53–65.

Rybski, D., Rozenfeld, H.D., Andrade, J.S., Batty, M., Stanley, H.E., Gabaix, X., Makse, H.A. and Zhou, B., 2012. *City Clustering and Applications* [Online]. Available from: `http://diego.rybski.de/files/RybskiD{_}2012-02-23.pdf` [Accessed 10 March 2019].

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

Schneider, A. and Woodcock, C.E., 2008. Compact, dispersed, fragmented, extensive? A comparison of urban growth in twenty-five global cities using remotely sensed data, pattern metrics and census information. *Urban Studies*, 45(3), pp.659–692.

Schwarz, N., 2010. Urban form revisited-Selecting indicators for characterising European cities. *Landscape and Urban Planning*, 96(1), pp.29–47.

Seto, K., Fragkias, M., Guneralp, B. and Reilly, M., 2011. A Meta-Analysis of Global Urban Land Expansion. *PloS one*, 6(8), pp.1–9.

Shepherd, J.M., 2005. A Review of Current Investigations of Urban-Induced Rainfall and Recommendations for the Future. *Earth Interactions*, 9(12), pp.1–27.

Tarabalka, Y., Benediktsson, J.A. and Chanussot, J., 2009. Spectral-spatial classification of hyperspectral imagery based on partitional clustering techniques. *IEEE Transactions on Geoscience and Remote Sensing*, 47(8), pp.2973–2987.

Tayyebi, A., Pijanowski, B.C. and Tayyebi, A.H., 2011. An urban growth boundary model using neural networks, GIS and radial parameterization: An application to Tehran, Iran. *Landscape and Urban Planning*, 100(1-2), pp.35–44.

Tratalos, J., Fuller, R.A., Warren, P.H., Davies, R.G. and Gaston, K.J., 2007. Urban form, biodiversity potential and ecosystem services. *Landscape and Urban Planning*, 83(4), pp.308–317.

Trevino, A., 2016. Introduction to K-means Clustering [Online]. Available from: `https://www.datascience.com/blog/k-means-clustering` [Accessed 15 December 2018].

Tsai, Y.h., 2005. Quantifying Urban Form : Compactness versus Sprawl. *Urban Studies*, 42(1), pp.141–161.

Tsompanoglou, S. and Photis, Y.N., 2013. Measuring urban concentration: a spatial cluster typology based on public and private sector service patterns. *World Review of Science, Technology and Sustainable Development*, 10(4), p.185.

Turner, B., Hegedüs, J. and Tosics, I.., 1992. *The Reform of Housing in Eastern Europe and the Soviet Union*. London: Routledge Press.

Uchiyama, Y. and Mori, K., 2017. Methods for specifying spatial boundaries of cities in the world: The impacts of delineation methods on city sustainability indices. *Science of the Total Environment*, 592, pp.345–356.

Vadali, S., 2018. Dimensionality Reduction with PCA and t-SNE in R [Online]. Available from: `https://medium.com/@TheDataGyan/dimensionality-reduction-with-pca-and-t-sne-in-r-2715683819` [Accessed 8 March 2019].

Wattenberg, M., Viégas, F. and Johnson, I., 2016. How to Use t-SNE Effectively. *Distill* [Online]. Available from: `https://distill.pub/2016/misread-tsne/`.

Wei, F., 2010. Cluster analysis of entrepreneurial developing level of Chinese cities. *The 2nd international conference on information science and engineering*. IEEE, pp.178–182.

Yeung, K.Y., Haynor, D.R. and Ruzzo, W.L., 2001. Validating clustering for gene expression data. *Bioinformatics*, 17(4), pp.309–318.

Zhang, W., Ping, Z.W., Zhang, H.Y. and Zhang, Y.F., 2013. Application of gray clustering method in the ecological classification of the cities in China. *Proceedings of 2013 IEEE International Conference on Grey systems and Intelligent Services (GSIS)*. IEEE, pp.325–329.

Zhikharevich, B., Rusetskay, O. and Mladenović, N., 2015. Clustering cities based on their development dynamics and variable neighborhood search. *Electronic Notes in Discrete Mathematics*, 47, pp.213–220.

Zhou, Y., Smith, S.J., Elvidge, C.D., Zhao, K., Thomson, A. and Imhoff, M., 2014. A cluster-based method to map urban area from DMSP/OLS nightlights. *Remote Sensing of Environment*, 147, pp.173–185.