Characterising urban processes using new forms of data and analysis

Thesis submitted in accordance with the requirements of the University of Liverpool for the degree of Doctor in Philosophy by

Krasen Petrov Samardzhiev

Department of Geography and Planning
University of Liverpool
United Kingdom

July 2023

Abstract

Cities are where the majority of the population live, the drivers of economic productivity and the places where actions to address global challenges are taken. Increasing urbanisation, changes in transportation, technological advancements, as well as the evolving behaviour of residents have brought about changes in urban form and function and therefore new challenges for researchers, administrators and planners. Alongside these developments, new sources of data are becoming increasingly available to track human interactions. Successfully using them can enable researchers to capture a more comprehensive picture of urban reality and provide insights into both old and new urban challenges.

This thesis combines new methodologies and new forms of data to analyse functional land usage within cities, morphological and functional integration, as well as the spatial distribution of people across different scales. The thesis consists of six chapters, the core of which are three independent pieces of research, each dealing with a specific aim. The other chapters are introduction, literature review and conclusion. A core aim of the research chapters is to combine data and methodologies in novel ways, in order to delineate areas of interest consistent with the phenomena under analysis. The specific focus on delineations is driven by: first, the fact that delineating the areas of analysis and the core units within them are one of the first steps researchers take in studying numerous phenomena; second, changes to the units can have effects on all subsequent analysis.

The first research chapter, delineates areas of similar activity using sound sensors, in order to define usage profiles within a city. The results show that non-acoustic, sound sensor data captures different patterns of human activity at high-temporal and spatial scales. In the second research chapter, tax and residence data are used to delineate economically integrated areas across (non-predefined) scales and examine the spatial distribution of jobs within them. There are three scales of activity emergent in the results - metropolitan-like, state-like and super-state-like - and decentralisation patterns are present in the first level. The final research chapter uses machine-derived building footprints to operationalise a minimalist definition of urban areas - areas of high building density surrounded by areas of low density. The resulting urban delineations differ locally in density, size and the types of urban features they contain, but are on average most similar to functional urban areas.

When combined, the results from all three chapters show the importance of parameter choices in analysis, a relationship between urban form and function and highlight the advantages and pitfalls of using new forms of data and data science methods for quantitative delineations. Specifically, the results from the three chapters show that

aspects of urban function are still reflected in urban form. They also show limited evidence in support of the integration of cities in the United States into interconnected large-scale clusters - megaregions and show more evidence, in line with the literature, of the decentralisation patterns within large urban areas. Furthermore, the results show the advantages and pitfalls of using novel data science methods and new forms of data - the importance of tailoring algorithms to specific geographic purposes, testing and reproducibility, as well as the viability and advantages of hierarchical based approaches to delineations.

Acknowledgements

Completing this thesis would not have been possible without the tremendous support of many people.

I would like to thank my supervisors Prof. Dani Arribas-Bel, Prof. Vitaliy Kurlin and Prof. Alex Singleton for all of their advice, support and patience. Prof Arribas-Bel has guided me throughout the PhD and has read, and been involved in, everything from this thesis and more. His feedback, suggestions and guidance have significantly improved my research and abilities. Prof Alex Singleton helped shape the direction of the thesis and has also read and given feedback on numerous drafts of my chapters and papers. Dr Vitaliy Kurlin helped me fill the gaps in both my mathematics and data science knowledge, so as to be able to carry out the technical aspects of this research.

I would also like to thank my family, friends and colleagues. The PhD journey was filled with doubts and hurdles but their support helped me stay on track. I would especially like to thank my parents Petar and Vladimira and my brother Deyan for all of their help. My friends and colleagues at the University of Liverpool, the Geographic Data Science Lab (GDSL) and the Data Science Theory and Application (DSTA) played a huge role in my life during the PhD, through peer review, writing workshops, discussing research ideas and conferences. I've made great friends and hope to keep in touch with everyone in the future.

I also extend my gratitude to my viva examiners, Dr. Les Dolega and Dr. Mingshu Wang for reading my work and providing helpful feedback which raised the level of this thesis. Finally, I would also like to thank my industry partner Carto for their support during the PhD, particularly Andy Eschbacher who showed much interest in my work while he was there.

Romulo, cum verbis quoque increpitans adiecisset "sic deinde, quicumque alius transiliet moenia mea,"

Liv. 1 7.2

Contents

List of Figures

List of Tables

1

Introduction

1.1 Background

Cities are increasingly seen as one of the most important structures through which to address the numerous social and environmental challenges facing people. One of the core reasons for this growing influence is that most of the human population currently lives in cities and there is an expectation that an even greater proportion will live in an urban area of one type or another (Batty, 2018, chapter 2). Cities are also the main drivers of economic activity in both developed and developing countries (Zhang, 2016) and challenges and solutions to transportation, water and waste infrastructure, as well as inequality, economics and other problems at the local and national level are focused on cities (Lobo et al., 2020). Furthermore, numerous proposals and and policies needed in order to respond to global problems such as climate change and sustainable development, are being articulated in terms of local urban actions (UN, 2020).

Bounded territorial units are crucial in carrying out the research, planning and implementation of solutions to all these challenges (Nelson, 2020). Bounded territorial units delineate the space over which phenomena of interest occur, and thus affect data gathering, analysis and theories (Wolf et al., 2020; Parr, 2007; Arcaute et al., 2015). This importance and wide-spread use of delineations has led to debates about their advantages, usefulness and limitations.

The impossibility of providing a single 'correct' boundary that encapsulates the full array of complex features which define human interactions, has long been recognised by geographers (Hartshorne, 1939). Therefore, there exist numerous proposals about the appropriate aggregation and delineation procedures depending on the specific phenomena analysed (Duranton, 2021). Some geographers even suggest that there is a

decreasing practical and interpretive value of bounded territorial units in light of the changes brought about from globalisation and new forms of societal organisation (Harrison, 2013). The arguments center on how more emphasis should instead be placed on the mobility patterns of ideas, capital, workers and commodities, analysed as a whole without the need of defining explicit boundaries and units within them.

Nevertheless, spatial proximity plays a role in organising both formal and emergent social, political, and cultural systems (Petrović et al., 2020; Nelson, 2020; Wolf et al., 2020) and defining the extent of the basic units and the area under analysis is the first step in quantitatively analysing spatial phenomena (de Bellefon et al., 2019; Nelson, 2020). Bounded units are also required for numerous practical purposes - defining administrative regions, voting districts as well as implementing policies (Coombes, 2014).

The present thesis focuses on this problem - the delineation of boundaries and creation of typologies for the analysis of urban phenomena, across scales ranging from the neighbourhood to the national level. Spatial scale is a multifaceted, core geographical concept. As it is used in this theses it refers to the size or extent of a process, phenomenon or investigation (Atkinson and Tate, 2000) and different urban scales such as urban centres, cities, larger urban areas, regions and megaregions are emphasised, whereas global or personal scales are discussed less. Similarly to changes in delineations, changes affecting scale can have large effects on analysis and results (Möck and Küpper, 2020).

To address these issues, researchers have developed numerous ways to define boundaries at different scales. A popular approach is to combine multiple smaller-scale fundamental units, in order to delineate boundaries for a larger-scale phenomenon. For example, small-scale census administrative areas can be used used in order to define the extent of urban areas (Khiali-Miab et al., 2019). Uniform cells or hexagons derived from gridding the territory under analysis, is another type of unit that can be used for the same purpose. These fundamental units can be combined in numerous ways - based on functional relationships such as flows of goods or people (Nelson and Rae, 2016); spatial contiguity (Rozenfeld et al., 2008), various characteristics of the units - population density, built density (Florczyk et al., 2019); or combinations of multiple approaches.

With the rising availability of new forms of data researchers are able to use other units and interactions such as tweets (Wei et al., 2020), mobile phone data (Secchi et al., 2015), location-based social networks (Calafiore et al., 2021). In addition to capturing new processes, these new forms of data can enable quantitative analysis at higher temporal and spatial resolutions. However, they come with a set of disadvantages related to data gathering, validity, generalisability and bias (Arribas-Bel and Tranos, 2018),. Additionally, processing and analysing the data requires the adoption

of new methods and considerations about computing power. This thesis makes use of new forms of data and attempts to address these challenges through the integration of geography and data science (Singleton and Arribas-Bel, 2021).

The main focus of the thesis is to delineate areas at various scales, in order to analyse three different urban phenomena - urban land use, large-scale functional integration and the spatial extent and organisation of cities. A secondary aim, which is achieved through the combination of results, is the analysis of the relationship between urban form and function and the internal spatial form of the delineated areas. To achieve this the analytical chapters of the thesis combine both traditional and new forms of data with machine learning methods, and attempt to modify the machine learning methods in accordance with the specific requirements of the phenomena analysed. The data sources used are captured decibel patterns from sound sensors, satelite-derived building footprints and geo-referenced work and home address tax records. The methodologies used differ across the chapters, however in general, they are all a form of unsupervised machine learning or clustering - the attempt to group data in a meaningful way.

The first analysed phenomenon is urban land use and aims to characterize the predominant usage of areas within cities into profiles. Examples of such profiles are industrial, residential or business. Identification of the different types of land use, changes across time, as well as comparisons across cities play an important role in understanding urban dynamics (Batty, 2018, chapter 6). Furthermore, delineating areas within cities with similar usage types is important for the analysis of the effects of zoning regulations, planning efforts (Toole et al., 2012) as well as other urban processes like sprawl (Zanganeh Shahraki et al., 2011). Land use pattern information also has numerous commercial applications - for example, it can be used for balancing usage on mobile phone networks (Cici et al., 2015).

The second phenomena is large scale-economic integration of urban areas into megaregions. A megaregion is a large-scale conceptual unit which represent a cluster of urban centres, integrated morphologically, culturally and economically (Glocker, 2018). A popular example of a megaregion in the United States is the 'Northeastern Megalopolis', which ranges from Boston in the North to Washington D.C. in the South and spans numerous official administrative boundaries, hundreds of kilometres of built environment and has tens of millions of residents. The importance of defining and delineating megaregional boundaries stems from proposals to focus economic, infrastructure and sustainability planning at the megaregional scale in order to increase economic competitiveness in global markets and to better tackle urban sustainability problems (Ross et al., 2016; Nelson, 2017; Lang et al., 2020; Amekudzi et al., 2012).

The third area of focus is the delineation of urban boundaries. The spatial extent of cities affect subsequent analysis results such as the calculation of unemployment,

population, economic performance and density statistics (Parr, 2007). Consistent and globally applicable city delineations and definitions capture the spatial extent of urban phenomena more accurately and enable the creation of general urban theories, better comparisons between cities and more accurate cost benefit analysis of policy outcomes (Lobo et al., 2020; Roberts et al., 2017). To address this problem researchers have developed numerous methods that define the spatial extent of cities based on economic interactions, morphological features or population densities (Duranton, 2021).

A secondary focus is the spatial form of the delineated areas and the patterns between form and function. Due to technological and economic developments, there have been patterns of employment decentralisation in North America and Europe (Dadashpoor and Malekzadeh, 2021). The exact form, consequences and whether this phenomena should be encouraged is an active area of research. The difficulty in consolidating results across studies comes in part, due to the fact that definition, delineation and scale choices have large effects on the analysis of urban form (Möck and Küpper, 2020; Barrington-Leigh and Millard-Ball, 2015). Lastly, the relationship between urban form - how cities are arranged spatially - and function - the human activity within them - is analysed. Traditionally urban form has followed function, however the relationship is also evolving due to advances in transport and communication technology (Batty, 2018, chapter 4).

1.2 Aims

The overall aim of the thesis is to delineate boundaries and create typologies in order to analyse urban phenomena. This is broken down into three more specific main aims, each dealing with one or more of the introduced urban processes - land use patterns, large-scale functional integration and the spatial extent of cities. Each aim encompasses an individual paper, presented as a chapter in the thesis. The three aims are:

- To identify areas with different urban land use patterns at a high spatial and temporal resolution using sound sensor readings.
- To identify potential megaregions, as well as to explore the distribution of people within them.
- To delineate urban areas morphologically and to explore the spatial structure of density variations within the resulting boundaries.

The firs aim complements recent developments in the land use analysis literature. The use of new forms of data such as mobile phone records or tweets has allowed researchers to explore land use patterns at high temporal and spatial resolutions (Cici

et al., 2015; Calafiore et al., 2021). However, such data is not readily publicly available since it is owned by private companies. Furthermore, some types of these new data such as footfall counters can only capture information about specific groups, due to technological challenges (Lugomer and Longley, 2018). In contrast sound sensor data from smart city projects, such as the Newcastle Urban Observatory, are publicly available and capture all activity within their radius.

In tackling the first aim, this thesis explores the viability of detecting urban land use from publicly available secondary datasets - sound sensors. Provided sound sensors capture activity information about the areas they are place in, they represent an additional source of data which could be used to infer land use patterns in place of and in addition to other new forms of data. A secondary outcome, is a comparison of methodologies that measure differences between pairs of sound sensor patterns. A new approach - topological data analysis - is tested against other methodologies used in the land use literature which focus on mobile phone and Twitter data. By carrying out this comparison, the thesis gives insights into potential usage of new methods for the analysis of new forms of data, and identifies areas of improvement for these new methods.

The second aim is achieved by analysing the emergence of large-scale, integrated structures in a tax records dataset in the United States. Previous work by Nelson and Rae (2016) used clustering (or community detection) techniques and commuter data to define megaregions and the approach taken in this thesis builds on that research. First, it uses a complementary dataset to define and verify emergent functional structures at several emergent scales. Functional here refers to the fact that the results are based on actual economic interactions directly captured by the data, rather than on morphological or built-environment information used as a proxy. Both the extent and scale of the results are inferred from patterns in the data and are not defined beforehand. Second, further analysis of the results is carried out following the exploratory spatial data analysis (ESDA) approach adopted by Arribas-Bel and Sanz-Gracia (2014). This is done in order to analyse the spatial structure of the resulting areas. The results delineate the spatial boundaries of potential emergent functional megaregions, and which cities and census tracts fall within the megaregional extent. This is an important problem since being part of a megaregion can result in significant benefits, provided megaregional planning agendas and policies are implemented, ie. Nelson (2017); Nelson and Rae (2016); Amekudzi et al. (2012). Furthermore, the multi-scale approach allows for the analysis of the spatial structure of employment patterns in the US - whether employment can be characterised as monocentric, polycentric or scattered and at what level.

The third aim is achieved by leveraging satellite-derived building footprints and machine learning to define urban areas. While numerous approaches have looked at

defining urban areas based on buildings (e.g. Arribas-Bel et al. (2021a); de Bellefon et al. (2019)) the approaches either rely on aggregating spatial units into grid cells or specifying explicit global density thresholds. Different choices of grid sizes and parameters affect both spatial extent and the resulting scale of the boundaries (Möck and Küpper, 2020; Statham et al., 2020, 2021; Balk et al., 2018). These issues are circumvented by using individual buildings and a machine learning approach tailored to the data. This leads to more stable urban definitions for the analysis and gathering of economic and policy data. The approach and data can also be applied to multiple countries, since building footprints themselves are a universal unit. Lastly, the analysis results in a nested hierarchy of potential delineations at numerous scales. The hierarchy can be used to analyse the potential integration of the final results into megaregions and the spatial form of the delineations at lower scales.

Through the specific approaches taken to address the three main aims, the thesis implicitly also shows the value of adopting new forms of data and methods to solve urban challenges. The three analytical chapters address the problems of delineation through the use of clustering methods derived from data science, as well as new sources of data. Chapter three uses new forms of data, as well as new methods and emphasizes the importance of good validation and comparison between methods. Chapter four is an example of how traditional datasets can be analysed with methods directly created for new forms of data. Chapter five one uses both 'new' datasets and 'new' methods. All three chapters aim to be examples of Geographic Data Science (Singleton and Arribas-Bel, 2021) - the explicit integration of quantitative geographical analysis with data science.

A similarity between the chapters is in the specific analysis approach taken to address the aims. All delineation methods used have no explicit scale requirements set - boundaries and scales are emergent from the data and the analysis. Furthermore, there are no spatial models or a predefined number of expected final delineations. Specifying apriori, any of these of these parameters affects the results and poor choices can lead to defining or defining away the phenomena under analysis (Möck and Küpper, 2020). Chapter three is limited to a inner-city scale only due to data availability, while the other two chapters explore scales ranging from the local to the national. By successfully applying these methods the papers show the viability of using unsupervised machine learning methods, which infer emergent structure in the data from local variations.

Another commonality between the chapters is that they produce not only final boundaries, but a hierarchy which shows how core units relate to each other. This enables the exploration of phenomena at different scales and also provides a detailed history of the creation of these final results and their structure. This feature of hierarchical methods is used in all chapters. Due to data limitations, chapter three only shows

how the same hierarchical multi-scale approach can be used with different data types such as sound sensors readings throughout time. The hierarchies are used extensively in the last two analysis chapters for the analysis of urban form. Furthermore, the different focuses of the two studies - home-workplace tax information and building polygons respectively, provide insights into both the morphological and built environment urban form spatial patterns and functional urban form spatial patterns.

1.3 Thesis structure

The structure of the rest of the thesis is as follows. The next chapter is a literature review, providing the relevant background for the analytical chapters. Afterwards, each of the three analytical chapters represents an individual paper, that address one of the three main thesis aims. The last chapter provides a summary and discussion of the results and concludes the thesis.

The next chapter of the thesis provides more background on the relevant urban concepts, developments and methods. The literature review broadly covers:

- Cities, urban form and function, changes and challenges
- The importance and effects of bounded territorial units
- New forms of data and related developments in geography
- Quantitative delineation approaches

Particular emphasis is placed on the four concepts presented in the introduction - spatial distribution, megaregions, urban spatial extent and land use. Additionally, there is a brief general overview of the methodological approaches used for data analysis.

The following three chapters each address a core aim of the thesis and represent independent pieces of research. They describe in detail the relevant literature, data, methodology and results.

Chapter three focuses on urban land use detection within cities. It describes the current approaches to land use and changes brought about by new forms of data, as well as how sound sensors can be used to capture land use patterns. The analysis is based only on the hourly recorded maximum level of noise, at the particular location where a sensor is placed. This is in contrast to other urban land use studies which use acoustic sound data - the actual record of the occurring sounds. The analysis in the chapter also includes use of topological data analysis (TDA) methods to address the challenges of processing new forms of data. Rather than directly using these novel methods, the paper carries out a comparison between several TDA methods as well as methods adapted from papers using other new forms of data, such as Cici et al.

(2015). The comparison is based on both internal statistical measures of quality, as well as external validation trough the use of POI and building data from OpenStreetMap. Additionally, the best set of results are compared to similarly derived activity profiles from tweets and mobile records.

Chapter four focuses on delineating megaregions using functional data. To this end LODES jobs data is used, which provides information on the residence and workplace of over 120 million Americans in the contiguous United States. This dataset is transformed into a network of nodes and edges where nodes represent census tracts and an edge between a pair of nodes - the number of people that live at one census tract (or node) and are employed at the other. The delineation of spatial units at multiple scales is based on the formation of communities within this network - groups of nodes which are more strongly connected with each other than with other nodes. The approach taken results in a nested hierarchy of delineations, with no predefined number of levels - all resulting scales are inferred from the data. The spatial structure of the delineations at each hierarchical level are explored and comparisons to other existing geographical units are carried out. Furthermore, the final results are compared to other defined megaregions in the literature - Nelson and Rae (2016) and Hagler (2009).

The last analytical chapter explores a morphological definition of urban areas in the United states, based on individual building footprints and no explicit density thresholds. These building footprints are derived from satellite imagery using computer vision algorithms and capture 130 million buildings in the contiguous United States. Similarly to chapter four, a hierarchy of nested units is created using a modified unsupervised machine learning algorithm. However, this hierarchy extents from pairs of individual buildings to megaregions and the whole country in scope. The final delineations are defined as the most 'persistent' that appear in the hierarchy. A comparison between the final results and seven other datasets is carried out to contextualise the results. Furthermore, the hierarchy is used to explore the patterns of density within each final delineations, as well as their potential morphological integration into megaregions.

The last chapter discuses the findings and draws overarching conclusions from the results of the individual papers. It provides a short summary of the main results from each paper and their limitations. It also combines the results with a focus on the relationship between urban form and function and compares the results from the two papers which use United States data. Furthermore, it discusses the appropriateness of using unsupervised, non-model, density-based approaches for delineation and the use of new forms of data and methods for addressing urban challenges.

2

Literature review

This chapter provides an overview of the importance, challenges and changes affecting urban areas with a focus on the issues relevant to chapters three, four and five. It is structured in six main sections. The first section is a general introduction to the urban processes and properties, relevant to the rest of the thesis - urban growth, urban density, city size and changes in urban form and function. It also discusses the relationship between urban areas and processes defined at different scales. The second section covers the effects of changes in boundary delineation on the above phenomena, as well as the general use and criticisms of bounded territorial units in quantitative geography. The third section describes the uses of new forms of data, the opportunities and challenges they present for urban analysis and approaches to using them. The fourth section covers the general approaches to delineating boundaries. The discussion is structured following the framework introduced by Duranton (2021), but adapted to cover urban areas, megaregions and land use delineations using new forms of data. The fifth section provides the background to the specific methodological approaches used in chapters three, four and five - clustering. The last section draws on all the discussed literature and broadly specifies the overall research gap that this thesis addresses. More specific formulations for each aim are provided within the relevant chapters.

2.1 Changes, challenges and processes related to cities

Across the social sciences there exist numerous definitions of what a city exactly is and what are its core features. These views range from cities as complex systems, or primary social and cultural, or economic and organisational structures to researchers who question the concept of the city as a unit of analysis, planning and research (Williams, 2012). The definitions reflect specific underlying purposes and organising principles of urban areas: as central places of service provision (Christaller, 1980); or facilitators

of human interactions (Mumford, 1937), where the core of the city are group activities at different scales, whereas the physical and market organisation is secondary; or other concepts. In turn these definitions of cities are tightly coupled to normative theories about what an ideal city is and how it should operate (Lynch, 1984). These theories speak to the organisation of form - the configuration of the built environment - and function - the actual usage of the built environment- within cities. For example, whether parts of the urban form should be specialised for particular functions such as residential living (Mumford, 1937) or instead, mixed-usage should be encouraged (Jacobs, 1961a). Others focus on what influence does form have on function and visa versa, such as the distribution of economic activity, land prices (Alonso, 1960). There also exist numerous questions about the appropriate way to manage urban growth and to identify the correct scale to tackle pressing challenges, such as transportation (Geddes, 1915). It should be noted that many of these debates about form and function go back thousands of years, due to the long-recognised importance of cities (Rykwert, 1988).

The diversity of views is also reflected in quantitative geography itself and more specifically in the many ways in which researchers operationalise the idea of a city. From minimalist definition of cities are places of high population density (O'Sullivan, 2011) to definitions incorporating aspects of the road network, and the distribution of jobs and people (Bertaud, 2018). This diversity is also extended to how urban form and function are quantified. There exist numerous measures of the built environment (Kropf, 2009), as well as quantitative definitions of particular form-related phenomena such as polycentricity (Derudder et al., 2021) or sprawl (Barrington-Leigh and Millard-Ball, 2015). Similarly, different aspects of function are emphasised for different applications - i.e. mobility (Moro et al., 2021).

The focus in this thesis is on urban definitions that can be quantitatively operationalised, especially using new forms of data and methods. This is done since, the analytical chapters either use new forms of data directly - chapters three and five - methods created for new forms of data - four and five - or combinations of both - chapter five. The urban phenomena described in the next sections focus specifically on changes and challenges to urban form and function, related to delineations and to the analytical chapters. The discussion does not cover the full array of challenges and functions of urban planning - inequality, segregation, crime, congestion, pollution, management and others (Glaeser, 2011). Nevertheless, delineation choices also affect these and other issues as is pointed later in section three of this chapter. For example, many of these issues are related to urban growth (Cohen, 2006), the analysis of which is directly related to urban delineations and form (Roberts et al., 2017; Barrington-Leigh and Millard-Ball, 2015).

Lastly, the discussion of scale is shaped by similar considerations. Scale is a core

concept in geography and can range from the personal to the global (Campbell, 2018). Generally, an urban phenomena under analysis is coupled with a scale of analysis and important characteristics of the phenomena can change as the scale changes (Wang et al., 2019). The discussion in subsequent sections is limited to different levels of urban scales - urban centres and some smaller-scale delineations, urban areas, functional urban areas, megaregions. This limits the large topic of planning and proposed planning scales to only a discussion related to these phenomena.

2.1.1 Urban definitions and delineations

Quantitatively defining urban areas is a difficult problem. Countries have individual political and administrative definitions for what a city is, however, using these for research and analysis is problematic for several reasons. First, individual countries have different official administrative designations. For example, in China a city officially has to have at least 100,000 people, whereas in Spain the minimum threshold is 5,000. Furthermore, a country's own definitions can sometimes change across time. Second, land size and land use within the units can be a mix of both rural and urban land and their spatial extent is optimised for the purposes of surveying (Wolf et al., 2020). Third, cities can extend beyond their borders into the surrounding area, however for political or economic reasons their official boundaries do not necessarily reflect this (de Bellefon et al., 2019). Therefore, one of the core problems of urban research that affects all analysis is how to provide a rigorous definition of urban areas and their spatial extent which is both theoretically and practically useful. What further complicates this challenge are calls for globally applicable definitions, which would enable comparisons across countries and time spans (Florczyk et al., 2019).

A core component of many urban definitions is the idea that cities are human settlements with a higher density of people or connections than the surrounding areas (O'Sullivan, 2011). This view captures an important aspect of cities, which many theories use to explain particular outcomes. Examples of this are the higher levels of innovation and economic productivity present in cities. As the density and number of urban dwellers increases so do the possible types of interactions between people, leading to an environment of innovation and idea generation (Ahuallachain, 2012). Furthermore, the increased population concentration creates the necessary consumption market and the enables specialisation which leads to the emergence of economies of scale (Glaeser, 2010). It should be noted that, cities cannot only be described in terms of density of people - other important aspects are an internal structure, interactions and perceptions (Lobo et al., 2020; Galdo et al., 2021). However, definitions which focus on density of different types results in more easily quantifiable and globally applicable definitions.

There are different operalisations of the idea that cities are areas of high density surrounded by lower density. These can take the form of defining the extent of cities as the contiguous built up environment, high population density, the reach of commuters or the extent of the local flows of services and goods (Parr, 2007). Other more complex definitions combine multiple aspects of the transport system, the population density and the economy - such as defining cities as the effective extent of the labour market (Bertaud, 2018). There are also hierarchical definitions which describe different units that encapsulate one another, such as urban centres within urban areas, (Florczyk et al., 2019). These different datasets and definitions can led to differences in core aspects such as the level of national urbanisation (Williams, 2012; Onda et al., 2019). In general, there is a consensus that no single definition captures all aspects of a city, and so definitions are context dependent (Lobo et al., 2020; Batty, 2018; Duranton, 2021; Parr, 2007; Williams, 2012).

2.1.2 Urban processes and challenges

One of the urban main challenges facing researchers, which affects numerous other phenomena, is analysing the process of urbanisation. Even though, there exists a consensus that the global urban population is growing - the exact degree of urbanisation and its rate of change are open questions (Balk et al., 2021; Duranton, 2021; Balk et al., 2018; Jochem et al., 2020). A better understanding of these two phenomena can shed more light on the effects of urbanisation on national and regional economies (Roberts et al., 2017; Bosker et al., 2018). Furthermore, the exact spatial patterns and consequences of increasing urbanisation are not well understood. This raises important questions about the future number of cities, their sizes and the distribution of population within them (Batty, 2018, chapter 2).

There is evidence that city sizes are distributed following a power law distribution, Zipf's Law more specifically (Duranton, 2021). This type of distribution of city sizes suggests that on average, the largest city is twice as large as the second largest city, three times as large as the third largest city, etc. The distribution also suggests that as the size of cities decreases the number of smaller cities raises exponentially. Furthermore, there is evidence that the distribution of city sizes has stayed constant over the last 200 years, with the largest 100 cities only taking a slightly larger proportion of the urban population recently (Batty, 2006). This is despite the fact that particular cities have had large population shifts themselves.

Another changing aspect of cities are urban form and function and the relationship between them. And furthermore how changes in form and function themselves, affect outcomes of interest such as employment rate, health and transport costs (Ewing and Hamidi, 2015). Form, most generally understood as the configuration of built envi-

ronment, is an important concept in urban research, planning and the understanding of cities (Kropf, 2009). In the past, most cities were compact and similar in shape, limited in their possible development by the transport and communication technology of the time (Batty, 2018, chap 6.). Currently, there is a wider diversity of urban form driven by the growth of cities beyond their official boundaries, the building of low density developments (sprawl) and vertical development such as the construction of sky scrapers (Batty, 2018, chap 6.).

Urban sprawl is a term that has many definitions and dimensions, similar to cities themselves. It is usually defined on a continuum with compact, high-density or monocentric development on one side and sprawl - low density, single use development - on the other (Ewing and Hamidi, 2015). The key management problem is whether low-density development is sub optimal, and wastes energy and resources, and the same quality of life can be achieved through better planning and development in denser urban areas. Some researchers describe sprawl as an outcome of market forces and resident preferences, whereas others argue that it a primarily a reflection of imperfections and externalised costs in land markets and regulations. The latter argue that government unfairly subsidise sprawl development at the expense of taxpayers, i.e. through housing subsidies and construction of highways, whereas the former say that its an expected reaction to growing urban population (Ewing and Hamidi, 2015). Research into sprawl is an active area of interest since there is a link between it and development goals such as segregation, health, economic productivity and sustainability (Lacy, 2016).

Similarly, there have been numerous changes in urban function - how the built environment is being used by people. Before the 19th century most cities acted as markets, comprehensive administrative and service centres for the surrounding farmlands, whereas currently, cities can emphasize and focus their economic development towards tourism, industrial production, education and other specialisations (Glaeser, 2011). Furthermore, there are is an ongoing repurposing of existing buildings and changes in usage patterns driven by changes in communication, transportation and computer technologies (Purkarthofer et al., 2021). Studies based on mobile phone data (Cici et al., 2015) or other smartphone data such as tweets (Frias-Martinez et al., 2013) or location-based social networks (Calafiore et al., 2021), have enabled the capture of more distinct types of specialised human activity than previously available. These shed light on the internal dynamics of urban areas at an unprecendent temporal and spatial scale.

These changes in form and function are reflected in the spatial distribution of services, employment and population within urban areas. For the majority of human history most cities were centred around a central business district, which contained all services and acted as a commanding centre. This has been the case since ancient times

- in Rome the forum acted as the central place, in Athens it was the agora. However, there is an established general trend of employment and population decentralization within urban areas in the 21st century (Dadashpoor and Malekzadeh, 2021) in Europe and North America. The exact patterns of this decentralisation has been described by different concepts - polycentricity, scatteration and edge cities. Polycentricity refers to urban areas in which employment is concentrated in several centres, operating in a complex network of interactions. There exist numerous ways of defining the centers and quantifying the relationship between them, which have an effect on the analysis of benefits and outcomes (Derudder et al., 2021). The "scatteration" view argues that employment is scattered throughout areas with no concentration in centres (Manduca, 2020). Edge cities describe a development pattern where new peripheral urban developments which fulfil certain size, retail, perception and office space criteria are appearing near established cities (Garreau 1991, p. 7.) Proponents of these developments and theories argue that they should be encouraged and that more localised systems better understand local needs and provide services better than a centralised and more distant authority. This would then lead to reduced strain on local governments and better productivity, service availability (Kwon and Seo, 2018) and other desirable outcomes such reduced traffic congestion(Wang and Debbage, 2021). Whereas criticisms focus on the inability of local actors to address other types of problems such as public health, sustainability development, global competitiveness which require more centralisation and coordinated efforts (Yang and Zhou, 2020; Meijers, 2008).

Related to this topic is the question of whether, as technology continues to develop and changes accelerate, the difference between form and function will increase. Numerous theories and studies rely on the fact that human interactions within cities are reflected in or influenced by the built environment (Batty, 2012). For example, the flows of consumption and production within the national economy are highlighted by the different types of transport networks and land use centres. Recent advances in communication technology and the shift to a knowledge economy can bring about a change in this relationship (Lobo et al., 2020). There is a chance that human interactions would become more complex and placeless, and not reflected in the physical form of cities. Therefore, digital and other traces of behaviours themselves become more important than urban form features for the study of urban processes.

One manifestation of these trends are the changes in the urban retail system. Retail centres, the places where retail activity is concentrated in urban areas, are currently facing long term structural pressures from online retail (Dolega and Celińska-Janowicz, 2015). In addition to being the focal points of physical consumption, the state of the retail system is linked to the wider economic state of the urban area (Dolega et al., 2021). As such there is a growing importance of using online sources of data e.g. - (Davies et al., 2018) - in order to capture consumer behaviours in full and to better understand

local economic factors. These experienced disruptions, caused by the disconnect of form and function - purchasing behaviour not reflected in physical shops and trips - vary spatially, as well as functionally at a national and local scale (Dolega et al., 2016). In some regions, larger and more attractive retail centres draw patrons from a more extensive areas and adapt in different ways, such as placing more emphasis on leisure offerings (Dolega and Lord, 2020). Therefore, there are still important aspects of the function of consumption captured by urban form and density.

The effects of transport technology on the relationship of urban form and function is similarly complex as evidenced by the ”death of distance” or metropolis paradox. The core of the paradox is that accessibility to central locations in cities has become more important even as average travel speeds have increased (Couclelis, 1996). In light of this, core locations, such as city centres and place more generally, still play an important role in urban economies and processes. This complexity is corroborated by the rising numbers of supercommuters - people who travel very long distances of 100km or more for work, albeit less frequently than normal commuters (Rae, 2015).

2.1.3 Analysis, planning and scale

The changes in urban form and function are closely related to questions of urban planning - what types of urban development to encourage and how. There has been a wealth of research on how different aspects of cities affect desirable outcomes such as income, health, transport and sustainability. However, there do not seem to be conclusive results and as such the ideal of what a city should look like changes with time (Batty, 2018, p.113). In addition to determining causal and related factors, there are questions about the best methods of intervention to achieve desirable outcomes. Typically, the focus has been on changing urban form and transport as the least intrusive way to improve the lives of citizens (Kropf, 2009). However, these types of interventions can become less effective in the future, due to the growing differences between form and function.

In addition to research into the describing urban processes, their desirability and outcomes, there are questions about the appropriate level of coordination or scale at which urban challenges should be managed and analysed. There are proposals to focus planning and research at different scales for different tasks - at the individual city level, at a larger city-region level which includes the surrounding towns and land, at a regional level which encompasses several cities and the areas between them, at a megaregional level which is made up of different parts of several regions or at a national level (Purkarthofer et al., 2021). The effects of globalisation, international transport and capital links also have an effect on urban form and function and further complicate planning and analysis (Harrison and Hoyler, 2015a). It should be noted

that units at these scales have little administrative and political power and mostly act as statistical units for analysis. Even if any policies are inspired by analysis at these scales, in practice they are implemented through already existing administrative units and scales (Nelson, 2017).

One popular example of a scale for research and analysis are functional urban areas (Möck and Küpper, 2020; Rappaport and Humann, 2021; Lobo et al., 2020). Functional urban areas are operationalised in different ways, but the overall goal is to capture local labour and consumption markets (Schiavina et al., 2019). Typically these are centred on a major urban centre, which expands outwards from its official urban boundaries, reflecting the fact that modern cities have expanded considerably (Rappaport and Humann, 2021). Examples of these are statistical areas in the US - metropolitan, micropolitan and combined statistical areas, in Europe - functional urban areas developed by Milego et al. (2019) and worldwide functional urban areas developed by Schiavina et al. (2019). An example of a metropolitan area is the Boston-Cambridge-Newton, which is centred at the city of Boston and covers all census tracts (a census unit in the US) which have a high proportion of commuters related to it. Employment decentralisation and urban economic performance is mainly studied at this scale as it represents local markets (Möck and Küpper, 2020; Dadashpoor and Malekzadeh, 2020; Bosker et al., 2018; Wang et al., 2019), however other types of research such as public health (Meijers, 2008) and inequality (Shen and Batty, 2019) are also regularly carried out at this scale.

There are also planners that seek to address challenges at a larger scale - the regional and megaregional. Regionalists argue that the analysis and governance of phenomena is best carried out at a large sub-national scale, reflecting changes in the current economic system that cover both urban and rural land (Harrison and Hoyler, 2015a). For example, regions such as California's Silicon Valley, England's South or Italy's North exhibit different patterns of development than those at their respective national level. Similarly to cities, defining and operationalising the spatial extent of regions is an active area of research, with criticisms aimed at the inability to define exact boundaries (Harrison, 2013).

There is also interest in delineating existing or future large urban agglomerations - which would contain millions of people and large percentages of the national economic activity (Hagler, 2009). An example of this area is the 'Northeastern Megalopolis' which spans the entire area from Boston in the North to Washington D.C. in the south. Numerous conceptualisations of these areas started being developed such as megaregions or megalopolitan areas - areas that cover parts of multiple regions (Lang et al., 2020). The interest was driven by the fact that these areas could act as focus points and increase the competitiveness of national economies in the global market (Glocker, 2018), as well as act as the right scale to tackle sustainability and other large-scale

cross-regional challenges (Ross et al., 2016). However, in order to reap the benefits, planning and collaboration between numerous partners is required (Wheeler, 2015). Critics of megaregions point out that attempts for collaboration at this scale has suffered from low stakeholder commitment and focus on single-issues, rather than a fully articulated vision (Glass, 2014). Nevertheless, there is evidence that the concentration of population and economic activity, as well as the functional integration between previously markedly more self-contained urban areas will continue and therefore, research at this scale remains relevant (Nelson, 2017).

On the other hand, there is also increasing interest in the analysis of smaller-scale intra-urban processes and the involvement of the local community in planning. On the analysis side, the interest is driven by new forms of data and an increasingly digitised world, which enable research at lower temporal and spatial scales. At the core of this development is the ability to capture the actual experienced behaviours of people, which means that fewer assumptions have to be made or modeled (Singleton and Arribas-Bel, 2021). Data such as movement trajectories or spatial social networks, link peoples behaviors directly to urban places and enables the exploration of short-term urban dynamics at precise locations. Furthermore, fusing these new forms of data to other more traditional datasets such as census information or surveys augments the results in other types of analysis such as geodemographics and land use studies.

These developments offer the chance of a new type of urban planning and smarter cities. As mentioned previously, urban planners have been able to exert limited influence over resident behaviours in order to improve the lives of citizens (Batty, 2018). The devices which collect new types of data can act as real-time sensors which can drive urban development at more granular and temporal scales . Analysis can range from information about the effectiveness of local government-provided services or gathering direct information about how people use public spaces. Furthermore, new forms of data also allow for studies into aspects of urban life which are less related to place. These can provide important information about how to influence desirable outcomes in the context of the changing relationship between urban function and form (Yang and Yamagata, 2020).

2.2 Boundaries in quantitative geography

Defining bounded territorial units is one of the crucial steps in quantitatively analysing the discussed urban phenomena. Generally, spatial boundaries are tightly coupled to the phenomena under analysis and the method used to operationalize it. Changes in the size and spatial extent of the core units - their scale and shape - affects subsequent statistics, analysis and results (Openshaw, 1979; Arribas-Bel et al., 2021a; de Bellefon et al., 2019; Balk et al., 2018; Parr, 2007).

This section focuses on how delineated units affect the urban phenomena previously discussed. The first subsection, shows the effects on basic properties such as city sizes, as well as more complex statistics like productivity and unemployment. The second subsection covers the advantages and disadvantage of using boundaries for the analysis and planning of urban phenomena in general. It highlights the criticisms of bounded territorial units, as well as their importance in various areas of research and administration.

2.2.1 Bounded territorial units and urban phenomena

Definitions of urban areas emphasise different aspects of urban life, depending on the goals of particular studies. At the most basic level, rates of national urbanisation, urban population counts and city size distributions are affected by these choices. Within the same country, different delineations can lead to differences of calculated urban population proportions of more than 10 percent (Roberts et al., 2017). Similarly, they can lead to excluding and including different areas into the urban delineation, or even to breaking up or combining urban areas (Duranton, 2021). These basic properties - size and population counts - have an effect on more complicated urban statistics - density, productivity, employment and others - which all effect subsequent analysis and theorizing (Parr, 2007). For example on the national scale, increased urban population has been associated with an increase in size of the service economy relative to the agricultural sector and generally better national economic performance. However, the applicability of this theory to all countries depends on the exact definition of urban area (Roberts et al., 2017). The analysis of economies and diseconomies of scale within cities and regions are also dependent on the exact operational definitions of a city and the area of study (Arcaute et al., 2016; Duranton, 2021).

The analysis of all the phenomena related to urban form and function, discussed previously, are also affected. There is a trend of decreasing urban density in the United States and United Kingdom, however the exact numbers depend on definitions and some researchers even point to evidence in the opposite direction (Dijkstra et al., 2021). Defining where sprawl begins and what is the structure of it are one of the core challenges in analysing it (Barrington-Leigh and Millard-Ball, 2015), made more difficult by the fact that population density in cities is hard to measure, since changes in the delineated extent of cities can cause large shifts in density calculations (Henderson et al., 2021). Analysis of urban function, at different scales can suffer from poor data quality, methodology or inappropriate definitions which result in the delineations of many small areas due to noise in the data (Furno et al., 2017). Research into employment and population decentralisation patterns is affected by both changes in the spatial extends of the areas of under analysis - e.g. the boundaries of a metropolitan area - and

centres within it (Möck and Küpper, 2020). Planning proposals at the megaregional level need to identify the exact beneficiaries and stakeholders within each megaregion, which are affected by the operational definition and delineated spatial extent (Nelson, 2017). The spatial extent of retail centres also affects subsequent functional analysis and the relationship to wider theories such as Central Place Theory (Ballantyne et al., 2022).

In the worst case, inappropriate delineations can erase or even reverse the impacts of the phenomena under analysis (Roberts et al., 2017; Batty, 2018; Möck and Küpper, 2020).

Similarly to changes in spatial extent, changes in scale affect subsequent analysis and results (Parr, 2007; Coombes, 2014; Möck and Küpper, 2020; Wang et al., 2019). A complicating factor is that scale can also be implicitly defined based on operational definition of spatial boundary delineations (Arcaute et al., 2015). For example, if cities are defined as contiguous areas of high population density, the threshold choice for high has an effect on the scale of the results. If the value is very low multiple cities are merged together resulting in the final delineation of large regions, whereas if the value is very high - individual cities will be split apart into neighbourhoods. Additionally, there are different strands of research which use the emergence of coherent spatial units at a particular scale as evidence of emergent underlying behaviour and interactions. Some research in megaregions, where the scale is not explicitly defined beforehand, are an example of this - i.e. Nelson and Rae (2016); Lang et al. (2020); Florida et al. (2008). There, strong economic ties at super regional scale, covering many cities and millions of people, is used as evidence of emergent economic behaviour which requires specialised planning and governance. However, in those and other cases scale and scale properties are again tightly coupled to specific definitions of phenomena and choices of parameters (Arcaute et al., 2015; Lang et al., 2020; Duranton, 2021).

2.2.2 Advantages and disadvantages of bounded territorial units

Geographers have long argued that no single boundary can fully capture all types of human interaction at a specified location (Hartshorne, 1939). Due to this and all of the effects delineation choices have, there are numerous criticisms against defining bounded regions of space for analysis and planning. Some focus on problems in specific operational definitions or quantitative analysis challenges. Other criticisms focus on the nature of bounded territorial units themselves. And some researchers go as far as saying that delineations are not useful at all.

Operational criticisms raise questions about the stability of results, the appropriateness of data sources, methodologies or combinations of these factors. These criticisms aim at improving delineations approaches and the validity of results obtained using

delineations. In general, they highlight the modifiable area unit problem - the fact that aggregating and disagregating data based on different spatial units affects the calculation of statistics (Openshaw, 1979). For example, if a city is ethnically segregated at the postcode level, aggregating the data at the neighbourhood level could yield the reverse results using the same dataset and methodology. More specific methodological criticisms focus on the choice of parameters for delineation methods such as density thresholds (Statham et al., 2021, 2020; de Bellefon et al., 2019; Duranton, 2021; Balk et al., 2018). Others highlight the use of limited datasets and methodologies that do not capture all aspects of the phenomena under analysis and argue for data fusion approaches (Ewing and Hamidi, 2015).

More conceptual criticisms focus on the structure of boundaries. One famous argument is that of Christopher (1965) which aims to show that non-overlaping boundaries are not appropriate units to capture functional reality in urban areas. For example, is not possible to draw boundaries around distinct neighbourhoods within cities, due to the numerous overlapping functions that exist between units such as schools, parks and retail spaces with common catchment areas (Alexander, 2017). Therefore, any analysis and especially planning proposals should incorporate these multi-faceted overlaps as core parts. In response to these criticisms, researchers have aimed to adjust existing or adopt new methodologies and datasets (Batty, 2018; Nelson, 2020).

Other geographers argue against the use of bounded territorial units for specific purposes. One area of research where this has been applied is urban population density analysis, where small changes in parameter thresholds for delineation methods lead to large changes in population density estimates (Duranton, 2021). In order to avoid using delineations, (Henderson et al., 2021) propose to measure population densities by using radii around expected populations locations on a grid without aggregating the grid cells into cities.

Finally, some researchers go further and suggest dropping boundaries all together. They suggest that disentangling the patterns of social and economic interaction into coherent units is becoming harder, but more importantly, is not necessary insightful (Nelson, 2020). Instead, more focus should be placed on the analysis of specific phenomena, analysed as a whole without the need of defining explicit boundaries or units within them (Bergmann and O'Sullivan, 2018; Gibadullina et al., 2021).

However, in spite of all these raised issues, delineated spatial units remain an important structure for analysing social, political and cultural systems (Petrović et al., 2020; Nelson, 2020; Wolf et al., 2020; Batty, 2018). They are a core part of research, administration, governance and planning and have always been practically important and widely used, even though how much they have been emphasised over time has varied (Harrison and Hoyler, 2015a; Paasi and Zimmerbauer, 2016; Purkarthofer et al., 2021). In fact, interest in urban research, dedicated towards delineating appropriate

spatial boundaries to analyse different phenomena, is growing (Duranton, 2021). This has resulted in the development of numerous new methods dedicated to addressing both technical and conceptual issues, as well as reflecting more facets of functional reality (Nelson, 2020; Duranton, 2021; Gibadullina et al., 2021; Wolf et al., 2020). This has in part been driven by new forms of data, which enable new aspects of urban life to be captured and analysed at novel spatial and temporal scales (Arribas-Bel et al., 2021a; Batty, 2018).

2.3 New forms of data

As human activity becomes increasingly digitised, people are leaving more digital trails which can be mined for insights. Initially, analysis of these datasets was focused on the balancing and monitoring of server logs and digital advertising, however these uses were outgrown as people realised more and more the value of these data (Singleton and Arribas-Bel, 2021). In quantitative urban research, this led to an explosion of 'new forms of data' used to analyse urban phenomena and processes. This section introduces examples of new forms of data and applications, their effects on delineation tasks and finally, covers their advantages and disadvantages. As with other sections, the focus is on examples of new forms of data relevant to the three analytical chapters, especially datasets and examples relevant to chapter three.

2.3.1 Examples of new forms of data

One of the most popular examples of new forms of data are mobile phone records and associated information (Cici et al., 2015). Mobile phone operators have long used the data in order to optimize load distribution and detect customer profiles. With increasing collaboration between telecom companies and universities, as well as public projects such as D4D and the Milan Telecom challenge (Italia, 2015), researchers have found many applications of mobile phone data to urban problems (Naboulsi et al., 2016). Time-series of calling records have been used to explore the frequencies and distribution of how people move within a city (Csáji et al., 2013), in the prediction of socio-economic indicators (Frias-Martinez et al., 2013) and studying urban dynamics at different temporal scales (Arribas-Bel and Tranos, 2018).

In addition to the mobile phone records themselves there is a wealth of data available from smartphone apps. Social media data, such as geo-tagged twitter data is another type of widely used 'new forms of data'. It has been used to infer land uses within urban areas (Lenormand et al., 2015; Frias-Martinez and Frias-Martinez, 2014), identifying natural disaster zones (Bruns and Liang, 2012), analysing tourist sentiment (Curlin et al., 2019), health analysis and public administration (Hu, 2019).

Other widely used app data has been different types of mobility traces from apps. These data represent full trajectories of human movement at specific times and locations or location-based services data which links social networks to physical locations. Movement trajectories have found applications in transport planning, retail analysis, inequality and activity spaces analyses (Toch et al., 2019). Whereas, location-based services data has been used in urban geography to analyse urban neighbourhood characteristics and create comparisons between cities (Calafiore et al., 2021).

Different types of sensors, which similarly capture various aspects of urban life, have also seen increasing adoption (Lau et al., 2018). These sensors capture information such as mobility counts, noise, pollution levels, temperature and others. Sensors which count the number of people in a vicinity have found extensive usage in retail analysis and neighbourhood analysis (Lugomer and Longley, 2018), whereas sound sensors are used to analyse noise pollution, sound pattern analysis and detection of anomalous events (Virtanen et al., 2018). Furthermore, different types of sensor data has been combined for transport analysis (Brambilla et al., 2019; Lau et al., 2018).

Another source of data comes from specialized card traces such as retail loyalty cards and transport cards such as Ouster cards. These datasets represent purchasing behaviours for people within different contexts - retail loyalty card data captures purchasing behaviours of individuals at specific stores and time, whereas transport card data captures public transit movements. In addition to the analysis of consumer behaviour for the retail sector (Rains and Longley, 2021), loyalty card data has been used to analyse various health outcomes (Davies et al., 2018). Similarly, in addition to being directly used for transport planning, transport card data have been used to analyse spatial and temporal mobility patterns relevant to urban planning more generally (Sulis et al., 2018).

2.3.2 Applications of new forms of data

New forms of data have enabled the operationalisation of new theories, as well as the integration of results across different timeframes. One such example is the effects of land use and urban vitality on urban health. Urban vitality broadly, is the notion that mixed land usage is related to better outcomes for urban residents and is further associated with higher levels of activity within neighbourhoods throughout the day (Jacobs, 1961b). De Nadai et al. (2016); Sulis et al. (2018) provide empirical support for this theory using travel card and mobile phone data respectively, as a measure of human activity. In addition to opening existing theories for analysis, these of data enable the development of new theories. For example, it allows the relationship between short-term resident behaviour and long-term urban outcomes such as changing land uses, to be explored for the first time (Arribas-Bel and Tranos, 2018).

Similarly, new forms of data have enabled the analysis of residents' movements and interactions within cities at increasing spatial and temporal scales, This has led to numerous analysis of short-term urban dynamics - the pulse of the city (Batty, 2018) - and how population density and land use changes during the day. Previously, movement data for all types of mobility studies related to transport or planning, were generally gathered through surveys and later GPS tracers (Shen and Stopher, 2014). The high cost of these information gathering methods meant that they were not employed often and at scale. New spatial datasets - transport cards, GPS traces and social networks - have enabled analysis at more precise locations, across different time spans and the capture of previously unexplored behaviours. Generally, research has been carried out in two ways: first, by extracting semantically meaningful behaviours from the raw trajectory data then analysing the resulting sequences; second, by linking the raw trajectory data to places and treating it as a feature of the places themselves (Shen and Stopher, 2014).

Examples of the first types of studies focus on common behaviours between groups and the factors that guide them. Example of this type of analysis are Schneider et al. (2013) which find that people within the same social group exhibit more similar movements patterns than people of other groups. By combining this data with socio-economic census data it is possible to look for factors which affect the inequality of access to high quality urban spaces. Wang et al. (2018) find that lower income groups spend more time during the day in low-income places and have less access to parks and business centres. Similarly, Shen and Batty (2019) find that higher managerial groups have access to spatially larger and more diverse places, compared to people in lower-earning occupations who have more segmented and local access. Recently, this type of analysis has even been extended to the point of interest level - parks, cinemas, retail outlets, etc. - with similar results (Moro et al., 2021) None of the results of these analysis are conclusive yet, not least because they all suffer from data availability and preprocessing issues, however they show the potential of new forms of data to capture the actual experienced urban life of different people.

The second way this type of data has been used is by treating it as a feature of places themselves. Examples of areas which have benefited from this development are geodemographic classifications and land use studies. Geodemographic classification groups geographical areas into clusters based on socioeconomic features, such that areas within a group are more similar to each other than to other areas. An underlying assumption behind these types of classifications is that people in the same group behave more alike than people in different groups. They are widely used both in Europe and the United States, in academia - to explore segregation, for example and in industry - for marketing and customer analysis (Singleton and Spielman, 2014). With the increasing availability of new forms of data, new facets of human behaviour such as

the exact purchasing patterns have been incorporated leading to a more accurate representation of the behavior of people within groups. Similarly, new forms of data have enabled the creation of granular temporal and spatial geodemograhic classifications, e.g. (Calafiore et al., 2021).

New forms of data also make it possible to improve delineation procedures. They have the potential to ameliorate established problems and biases, such as the modifiable area unit problem (MAUP), by providing data at very high spatial scales(Wolf et al., 2020). Due to this, granular new forms of data data have found a variety of purposes - i.e. creating urban delineations (Schiavina et al., 2019) and approximating populations at high spatial resolutions (Florczyk et al., 2019). Furthermore, developments in computer vision have enabled advanced image analysis of both satellite images such as extraction of building footprints (Huang et al., 2019) to analysis of crowd-sourced images to understand spatial leisure patterns (Chen et al., 2019).

2.3.3 Geographic data science

As they found adoption in the field of quantitative geography new forms of data are generally distinguished from other data sources by their origin, size and processing requirements (Arribas-Bel, 2014a). One of their most important characteristics is the secondary nature of the data - these datasets are not carefully collected and curated for the research or analysis purpose they are often used for. This is in contrast to other data like surveys which have a higher level of control that ensures better data quality. Thus, new forms of data are a 'byproduct' of some process and can have many issues related to representativeness and quality (Rains and Longley, 2021; Arribas-Bel, 2014a). However, they come with advantages such as more granular spatial and temporal coverage, which open up new avenues for research. Furthermore, with increased adoption the data becomes more representative and the results of the analysis improve (Batty, 2018; Cici et al., 2015; Arribas-Bel and Tranos, 2018).

Another differentiating aspect of these data are their handling. In many cases the secondary nature of the data leads to extended pre-processing which is not needed when using more traditional survey or census data (Arribas-Bel, 2014a). This is further exasperated by the fact that some types of data are guarded and have to be accessed through secure infrastructure which also prolongs the analysis process (Calafiore, 2021). Furthermore, the sheer amount of data and type of data can have an effect on the viable methodology and processing - due to computational requirements many methods simply cannot be applied to certain datasets (Duranton, 2021).

Geographic Data Science is one emerging practice aiming to address these issues by combining data science and geography (Singleton and Arribas-Bel, 2021). Data science approaches dealing with new forms of data have not necessarily been created

with geographic applications in mind and therefore this has resulted in limited tools available to social scientists. Similarly, the tools and GIS approaches available to geographers are not usually suitable for the volume and problems that new forms of data present. Geographic Data Science proposes to address this with the development of methods which place geographical concepts at the heart of new data science computational methods, thus encouraging quantitative geographers to engage more with data science tools and methods, in order to achieve richer analysis outcomes (Singleton and Arribas-Bel, 2021; Arribas-Bel et al., 2021b). An example of this approach is Chapter five of this thesis which modifies existing data science algorithms to achieve the specific aim of delineating urban areas with few explicit thresholds. Similarly, with the development of new methods and the growth of available computing power new, more computationally expensive methods can be applied to other foundational and widely used datasets such as census data, road networks or tax addresses. The analysis in Chapter Four is an example of this - it uses community detection approaches, developed to analyse phone calls, to regionalise the United States, based on tax records, without specifying a priori scale.

2.4 Quantitative approaches to delineating units

This section introduces the general quantitative approach to delineating urban areas at different scales, followed in chapters three, four and five. The literature review and methodology sections of each of these chapters provide more specifics, relevant to the respective analysed phenomena. The discussion follows an adapted framework used by Duranton (2021) to analyse approaches used to define urban areas specifically. This was done since, the Duranton (2021) framework is general enough and has overlapping stages with other frameworks used in the literature of land use delineation using new forms of data (Furno et al., 2017), as well as delineating megareigons (Glocker, 2018). However, in order to accommodate megaregions, urban areas and land use delineation approaches into a single narrative the third and fourth steps ”Normalising a definition of urban” and ”Aggregating units into urban areas” from Duranton (2021) are generalised and merged. Therefore, the discussion of delineation approaches is structured as follows:

- Deciding on the core units under analysis
- Determining the area under analysis and classifying units
- Aggregating units
- Verification of the results

Lastly, a final sub-section provides an overview of the specific methodology used in chapters three, four and five - clustering.

2.4.1 Deciding on the core units under analysis

The first step of delineation approaches deals with selecting the core units of analysis. This is a crucial step and limits the subsequent available classification and aggregation approaches.

Urban areas

Numerous units have been used in order to delineate urban areas - small-scale administrative units, roads networks, hexagon or grid cells and even individual buildings directly. These units are then associated with data, which could be morphological such as built-up density or street intersections; or functional - population counts, flows of goods and people; or even more specialised such as phone calls or app interaction data. The data can be associated with the units either as features of the units themselves or as relations between the units. An example of the former is population per census tract and of the latter - number of commuters between a pair of two census tracts.

Megaregions

The units and data used for defining megaregions are similar to those used to delineate cities - administrative units, grid cells, etc. The difference is typically in the expected scale of the results and initial size of the core units - quantitative megaregional definitions can directly use already delineated urban areas as initial core units to be aggregated. For example, Lang et al. (2020) directly uses metropolitan statistical areas as the core units. Similarly, flows of goods between cities can be used as the data associated with the units (Ross et al., 2009).

Land use delineations

The general procedure for delineating land use areas is similar to the delineation of cities, however the scale of the units is smaller, the scope of the analysis is more limited and temporal variation plays a more important role. Many of the new forms of data already discussed are used for this purpose - location-based social networks (Calafiore et al., 2021), mobile phone data (Cici et al., 2015), app data such as twitter (Frias-Martinez and Frias-Martinez, 2014) or flickr (Chen et al., 2019) and others. Examples of units are specific points of interest - cinemas, shops, etc (Re Calegari et al., 2015) or mobile operator tower catchment areas (Grauwin et al., 2015). And examples of data

for the former are visitation statistics and of the latter - number of processed mobile phone calls in the area.

In addition to being generally of smaller scale, the data is more limited in scope compared to megaregional or urban delineations. Typically, the data does not cover the whole territory of a country but is limited to one or more major cities. This limits the types of analysis which researchers can carry out. It is impossible to carry out comprehensive international analysis, or to study the whole typology of results within a country to derive statistical patterns such as those relating to city size (Zipf's law). Some international comparisons do exist, however they are limited to comparing the results between major cities in different countries (Grauwin et al., 2015; Furno et al., 2017). With the growing adoption of such datasets and their increased availability to researchers there are some examples addressing these issues such as Blondel et al. (2008).

2.4.2 Determining the area under analysis and classifying units

A further step some approaches take is to the classify the core units into groups. For example, Florczyk et al. (2019) use grid cells as base units and population, as well as built-up density as features with the goal of delineating several types of urban areas. The grid cells are classified into urban and non-urban based on minimum population density and percentage of built-up area thresholds. The appropriate classification thresholds for this and other methods are an active area of research (Duranton, 2021; Arcaute et al., 2015; Batty, 2018; Statham et al., 2021, 2020).

Duranton (2021) identifies several widely used approaches to selecting thresholds - global, relative, statistical, model-based and ambiguous for urban areas. This classification is applicable for megaregions and land use grouping approaches as well, since they describe ways of classifying or discarding core units and not specific thresholds or models.

When using a global threshold, all units are grouped, based on a single value regardless of their location, whereas relative thresholds take into account the local context such as the region or country. For example, Florczyk et al. (2019) use a global minimum of 1,500 people or atleast 50% built up area to classify a grid cell as urban. These thresholds are applied to all cells regardless of their location. Global thresholds such as these have the advantage of providing more consistency across the results, however there is difficulty in specifying an agreed upon value (Onda et al., 2019). Relative thresholds address this issue, but they make comparisons between subsequently delineated cities more difficult.

Model and statistical approaches can use the data itself to define different types of thresholds. One example of a statistical approach is the methodology used in Chapter

5, where the core units are building footprints and are classified based on the surrounding building density. An example of a model approach is the one carried out by Taubenböck et al. (2019) where grid cells are classified based on density, following a monocentric urban model. However statistical and model approaches are not always possible and can sometimes lead to inconsistent and hard to interpret results (Duranton, 2021; Coombes, 2014).

Because of these difficulties and trade offs, the last type of classification methods embrace ambiguity. This could mean researchers and planners splitting or merging groups based on non-quantified political and cultural context such as (Hagler, 2009; Nelson and Rae, 2016; Hamilton and Rae, 2018). Or incorporating non-expert opinions, for example local residential knowledge to create the classifications (Galdo et al., 2021).

2.4.3 Aggregation

The next step in the analysis aggregates all or subsets of units into delineated areas. The types of methodologies available for doing this depend on the underlying data and features. If the basic units capture non-relational information at a particular place, as such population within a grid cell, the aggregation method uses some notion of contiguity. Examples of this are Florczyk et al. (2019); Glocker (2018); Schiavina et al. (2019); Georg et al. (2018); Balk et al. (2021). There also exist approaches such as Arribas-Bel and Schmidt (2013), which delineate the areas based on differences in features and ignore geographic space. On the other hand, if the basic units contain relational data, e.g. flows of goods between two grid cells, there are two options. The aggregation method could ignore any spatial contiguity information and aggregate the units based only on patterns in the network of flows (Nelson and Rae, 2016). Or the approach could take into account both spatial and relational data (Ross et al., 2009).

Similarly to the unit classification there are important parameter decisions which have to be make at this stage. These concern how to define the limits to the aggregation of units and could also be categorised into - fixed, statistical and combined. Fixed define the aggregation criteria for the entirety of the data - these could be contiguity matrices or radii which determine that all units within them are to be aggregated - (Arribas-Bel et al., 2021a; Statham et al., 2021, 2020; Glocker, 2018; Florczyk et al., 2019; Schiavina et al., 2019). Aggregating small administrative units based on a fixed percentage of commuters, e.g. metropolitan areas, is another example of this approach. Statistical and model-based methods compare the distribution of the data itself against a model to determine whether the units should be aggregated. For example, Nelson and Rae (2016) uses such a method to aggregate census tracts based on commuter flows between them. If two census tracts have more commuter flows than is expected between

two tracts of the same size, assuming a uniform distribution of flows, then they are grouped together. de Bellefon et al. (2019) uses a similar approach with different data - grid cells with building density. The tradeoffs between the two types of methods are similar to those of classifying units, already discussed. Fixed thresholds and limits are more easily interpretable and have lower computational requirements, whereas statistical methods take into account patterns in the data itself, but are not always applicable and can be hard to interpret. Other approaches can combine fixed and some inferred aspects and even take into account expert or local opinion (Galdo et al., 2021).

Additionally, it should be noted that the scale of the final results of an analysis, using the same data, can be implicitly or explicitly chosen depending on methods or parameters used. For example, Florczyk et al. (2019); Schiavina et al. (2019); Glocker (2018) all use the same datasets - grids of population density and built-up levels to define core urban areas, functional urban areas and megaregions respectively. The difference between the three approaches is in the selection of parameters or methods to aggregate the core units. Similarly, Rappaport and Humann (2021); Lang et al. (2020) define metropolitan areas and megaregions respectively, using the same commuter flow census data. More technical analysis such as Balk et al. (2018); de Bellefon et al. (2019); Arcaute et al. (2015) show how varying parameter choices can also affect the scale of the resulting delineations implicitly.

Land use

Many land use aggregation approaches are similar, but a differentiating aspect is the prominent role temporal variations play in these types of delineations, which in turn affects the types of aggregation methods used. When delineating megaregions or urban areas the data typically covers one specific period or instance of time. Examples of this are population in a year, or commuter flows for a particular quarter gathered through a survey. In contrast, land use analysis with new forms of data incorporate much more temporally granular information such as hourly, daily or weekly activity, which is one of the core advantages of using new forms of data. However, this temporal variation makes the aggregation process more difficult since it requires more advanced methods to measure the similarity between units (Furno et al., 2015). Different approaches can treat temporal variations in the information as independent features or they can incorporate them into the analysis using more specialised time-series methods. For example, if the core units are grid cells and the data associated with them is hourly population over a year, approaches of the first type could aggregate all the temporal data into an average population per hour across the whole year (Soto and Frías-Martínez, 2011). Approaches of the second type would take into account temporal patterns such as the seasonality, noise and variation across the entire time span and would not use temporal aggregates, but compare hourly differences directly (Cici et al., 2015).

2.4.4 Verification

The last step in the analysis is the verification of the results.

Urban areas

Duranton (2021) identifies two common approaches across the literature to validate the quality of the results - comparisons with other existing delineations and comparisons against widely-studied properties of cities. The former type compares aggregation agreement between sets of delineations using similarity metrics such as Jacard index or Rand score. Another similar approach is to directly measure spatial statistics such as overlap percentages (Arribas-Bel et al., 2021a).

The second type of comparisons can take the form of comparing aggregate statistics - total delineated urban population or land, or they can be more granular and complex - the distribution of delineated city sizes against theoretical expectations such as Zipf's law. The verification of delineated urban units is an active area of research and large differences in comparisons are common (Duranton, 2021). This is due to the variety of data and methodological approaches used and the fact that there doesn't exist a single delineation which captures all aspects of urban activity (Duranton, 2021; Batty, 2018; Lobo et al., 2020; Hartshorne, 1939; Parr, 2007).

Another approach is to rely on expert evaluation of the results such as Florczyk et al. (2019); Hagler (2009); Galdo et al. (2021); Hamilton and Rae (2018), which makes it possible to evaluate the quality of delineations from multiple perspective, something the other tests cannot achieve. However, it takes considerable resources and time to carry out and furthermore, there is the potential for inconsistencies.

Megaregions

The validation process for megaregional delineations is similar overall - there is a reliance on comparisons and expert opinion. However, the validation is more limited due to three problems. First, there are no well-studied statistical properties of megaregions such as Zipf's law. Second, there are less readily available polygons of megaregions. Third, even if some comparisons are carried out - Lang et al. (2020); Glocker (2018); Nelson and Rae (2016) - they are not consistent and do not employ similarity metrics such as rand score, but are more ad hoc.

Land use delineation

Similar to the other types of delineations discussed, the final results of land use analysis are validated using different types of comparisons and aggregations. A popular approach is to compare the delineation results against official zoning plans and land

use types (Soto and Frías-Martínez, 2011). Another popular approach is validating the results through properties within the delineated areas, which are not directly used in the aggregation process. For example, Furno et al. (2015) use the number of registered businesses within delineated areas to show that areas, designated as 'bussiness areas' by their methodology, have more business activities on average than other delineated areas. However, direct comparisons between derived sets of results is more difficult due to data availability and processing issues. For example, many of the approaches use data from private companies which is not readily available (Grauwin et al., 2015; Soto and Frías-Martínez, 2011; Furno et al., 2015; Lugomer and Longley, 2018).

2.5 Clustering

All the delineation approaches presented in subsequent chapters are based on clustering - an unsupervised machine learning approach. Clustering is the attempt to group data in a way that meets with human intuition, however intuitive ideas of what makes a 'good' grouping are not formally defined and depend on the application context (Hennig, 2015). Clustering is a type of unsupervised machine learning, due to the fact that there is no predefined classification of interest in the dataset such as a target category. Instead, a common target is to group the data in such a way so that members of a group are more similar to each other, than to members of other groups.

The choice of clustering as the main methodology for delineating boundaries in the thesis is guided by two principles. First, the motivation behind clustering algorithms maps directly to the motivation for delineating coherent spatial units and typologies. In both cases the goal is to separate groups, where members within the group are more similar to each other than to members of other groups. Second, the two particular algorithms used make no assumptions about the number of final delineations needed or their scale. The number of final delineations and thresholds for grouping items together are inferred from the data. In none of the three cases where clustering is used, do the algorithms assume what size or how many of the final results there should be. Furthermore, both methods provide information about the internal structure of the delineations. In addition to placing items in groups, the final results of both algorithms produce a hierarchy that shows how each cluster is constructed.

There are two types of clustering algorithms used in this thesis: HDBSCAN in chapter three and five, and Louvain community detection in chapter four. HDBSCAN is a density clustering algorithm which requires the specification of only one parameter - the minimum number of elements within a group to consider it a clusters. Density-based clustering is a family of clustering algorithms which take into account not only the similarity between the data but the number of observations as well (Campello et al., 2020). A group of items is considered a cluster, only if it has 'sufficient' density for

some algorithm specific definition of 'sufficient'. When using such an algorithm not all items in a dataset would necessarily be assigned to clusters, in contrast to other approaches which partition the dataspace.

The second algorithm used is Lovain community detection. Community detection is the notion of clustering extended to graph data such as flows of goods between places. The only requirement for using this algorithm is specifying the notion of similarity between two places. In chapter four the dataset used is a network where the nodes are census tracts and the relationship between two tracts is the number of people who have either work or home addresses registered in the pair.

2.5.1 Advantages and disadvantages of the methods used

In practice, clustering requires the specification of parameters, similar to the different aggregation approaches. One of the first and most important decisions is defining the metric which measures the similarity between the data items being grouped. This choice is usually independent of the actual clustering algorithm used. Example metrics are euclidean distance, which is a generalisation of geographic, 'as the crow flies' distance to higher dimensions or 'great circle' distance which is an approximation of distance travelled along the earth's surface. The other parameter choices depend on the specific type of clustering algorithm being used.

Both of the algorithms used in the chapters are examples of statistical or model aggregation, as discussed in the previous section. As such they have certain disadvantages related to the interpretation and consistency of the results and choice of hyperparameters. First, as discussed, it is more difficult to interpret why delineations are calculated. The research chapters in the thesis aim to address this shortcoming by analysing intermediate results and contextualising them within the existing literature through comparisons with other delineations and phenomena. In all cases these intermediate results play a role in the analysis. Second, the choice of hyperparameters can have a large effect on the final results. The comparisons, alongside the analysis of the final delineations provide the justifications for specific parameter choices. Furthermore, ranges of hyperparameters as well as different clustering configurations are tested in the chapters.

Chapters three and five use the HDBSCAN clustering algorithm and euclidean distance as a distance metric. The advantage of HDBSCAN in particular is that it enables the discovery of clusters with different shapes, densities and does not require specifying the resulting number of clusters beforehand (McInnes and Healy, 2017). It was directly developed to address the chaining problem and density problems present in algorithms such as DBSCAN and single-linkage clustering and is widely used in different domains (Campello et al., 2020). Nevertheless, due to the local cluster extraction

the algorithm can struggle with certain data distributions (Malzer and Baum, 2020) and therefore careful validation of the results and testing out multiple hypterparameters is needed.

The similarity metric used in chapter four is modularity, which is defined as the difference between the observed and expected flows based on a random reallocation (Newman, 2006). The advantage of the combination of Louvain and modularity is that it is possible to feasibly process large amounts of data and it produces well-defined groups of nodes in experiments with synthetic and real data (Lancichinetti and Fortunato, 2009). However, disadvantages include the fact that all data is assigned to communities and that changes of paramaters can have large effects on the final results (Nelson and Rae, 2016). In chapter four this issue is addressed through an analysis of the intermediate results, in order to verify that the final clusters are made up of units which resemble other delineations in the literature, in addition to analysing the internal spatial structure of the results.

2.6 Research gap

As a whole, the research carried out in the thesis broadly aims to create delineations and typologies for the analysis of urban land use, economic integration, city size and shape, using new forms of data and addressing previous shortcomings in definitions. In the course of doing this the concluding chapter reflects on the relationship between form and function, as well as urban decentralisation and large-scale morphological and functional integration of urban areas.

More specifically, the third chapter addresses these two main gaps in the urban land use and urban sound literature - first, sound typologies are mainly focused on acoustic sound or those that are not such as Zambon et al. (2016) focus mostly on traffic; second, high resolution urban activity analysis primarily uses data from private companies. The fourth chapter, uses more recent large-scale data and focuses on a hierarchical delineation of the constituent units of megaregions, in order to confirm that they are made up of coherent units, in contrast to other research that directly produce final results such as Nelson and Rae (2016); Lang et al. (2020); Hagler (2009). The fifth chapter uses building polygons directly to delineate urban areas without having specified global density thresholds. This ameliorates the problems of grid aggregation and provides a real-time, globally applicable definition of urban areas, which is one of the core features required for the generalisability of results (Wolf et al., 2020). The literature review sections in each chapter contextualise these aims in more details.

In addition to this, the clustering procedures used in all research chapters are hierrarchical in nature and aim to show the advantages of these types of methods in analysing new forms of data. The hierarchical approach enables the exploration of

intermediate units of aggregation, as well as answering questions about why other potential groupings did not occur. These properties are used in all chapters to link the analysis to the wider literature.

Furthermore, the thesis shows the importance of contextualising results when using new forms of data and methods. This is done both in terms of comparing different methods and results. In all chapters external data is used to contextualise the results and to analyse what phenomena present in the wider literature are captured by the final clusters/delineations. In chapter three this is done through the use of Openstreetmap data and qualitative comparisons with urban sound and urban fabrics literature; in chapter four - through the analysis of polycentricity and the comparisons with Nelson and Rae (2016); Hagler (2009); lastly in chapter five - through the analysis of polycentricity as well as comparisons with seven other datasets - Florczyk et al. (2019); Hagler (2009); Leyk et al. (2020) and official boundaries.

3

Urban land use detection through sound

Abstract: This paper aims to show that sound sensors are an important source of secondary information for the analysis of urban dynamics. This is achieved through the use of a clustering method to identify areas with distinct functions based on the hourly sound patterns recorded by sensors throughout Newcastle and Gateshead. Three methodologies from the new field of TDA, as well as a baseline approach are adopted, in order to address the problem of measuring the differences between the sound patterns captured by each sensor. A comparison is carried out which shows that the baseline approach of comparing sound sensors patterns - normalizing the pattern and using a correlation measure- produces the best results. These clusters show that sound sensor patterns are strongly affected by the characteristics of the surrounding area and its usage. In general, sound sensors that are placed alongside the same streets are grouped together and the different clusters exhibit different sound patterns, comparable to the detected clusters from mobile phone and app data research. There are patterns that correspond to areas with significant nightlife, residential areas, leisure areas like parks and mixed usage areas.

3.1 Introduction

Understanding the ways in which people interact with the urban environment is a key challenge for urban planners and researchers. This information can be used to evaluate the effects of zoning regulations (Toole et al., 2012), for sustainability and transportation planning as well as in analysing urban processes like sprawl (Zanganeh Shahraki et al., 2011). Furthermore, there are numerous commercial applications available - predicting land and house prices (Duranton and Puga, 2015) or balancing usage on telecommunication networks (Cici, 2015). Within this context, urban land use detection aims to characterize the use of areas based on their functions such as business, residential, mixed and others. A better understanding of the activity of people within areas provides information about overall urban function and dynamics (Arribas-Bel and Tranos, 2018; Batty, 2018, chapter 6) and generalisable approaches to this problem enable comparisons between cities (Furno et al., 2017).

Traditionally, data on land use patterns has been collected by remote sensing, conducting surveys, or GPS tracking techniques. Each of these approaches comes with with its own advantages and disadvantages. Commissioned surveys or GPS tracking studies can have good experimental design - the collected data can be representative and statistically robust. On the other hand, the disadvantages are the high cost, participant misreporting their behaviour (in the case of surveys) and the collected data only providing a static image of the researched phenomenon (Shen and Stopher, 2014). The most popular land use detection approach has been remote sensing which relies on the analysis of satellite images. It has the advantage that the data it provides covers larger areas and the costs of obtaining and analysing it is comparatively low. However, such techniques do not directly capture human activity data and there is a small variety of urban land uses that can be identified. In fact a recent systematic review (Reba and Seto, 2020), found that over 85 % of remote sensing land use studies identify only one urban land use type, and highlighted the need to differentiate multiple urban classes.

Advances in technology and increased adoption of 'smart' devices have created a deluge of valuable 'new forms of data' for research (Arribas-Bel, 2014b). Examples of this type of data are mobile phone call records, app data , images, or sensor data - such as footfall. These data have already had an impact on urban land use detection. For example, Cici et al. (2015) show that it is possible to use volume of cell phone call records across time, in order to delineate the city of Milan into residential, office, nightlife and other areas with a distinct land use type. Frias-Martinez and Frias-Martinez (2014) use total volume of tweets per hour to the same effect. Furno et al. (2017) expand the scope and create a global comparison, which finds that there are common patterns among cities in different countries for transportation and business areas, however residential areas have more heterogeneous patterns. Such studies show

the advantages of new forms of data - through them it is possible to obtain a picture of urban activity at higher temporal and spatial resolutions, thus giving insights into short and medium term urban dynamics at various scales.

The main contributions of this paper are twofold. First, to the best of the authors' knowledge, it is the first exploratory study into the use of sound sensors to detect different functional land use areas within a city. The specific dataset used for the analysis comes from the Newcastle urban observatory (James et al., 2014). It captures, over a 30 day period, the hourly sound levels from 40 sensors placed alongside roads in Gateshead and Newcastle in the United Kingdom. The sensor readings are not acoustic sound recordings, widely used in the urban sound analysis literature, but represent decibels levels at a certain location and time.

Urban sound sensor data offers several advantages when used for land use detection. First, sound readings are affected by both human activity (Groos and Ritter, 2009) and built environment (Zuo et al., 2014), thus urban areas delineated based on noise patterns capture aspects of both. Second, the analysis captures hourly, or even shorter term, variations in land use dynamics due to the high temporal resolution of the data. Third, sensor readings are affected by the activity of all people in the catchment area of the sensor. Other 'new data' sources, such as tweets or footfall counters detect only the activity of users of a specific app or a certain type of mobile phone. Fourth, aggregated sound sensor data, which does not record actual residents, can be made more readily available to researchers in contrast to other sources. For example, mobile phone records and app data could be limited by the willingness of private companies to share the data and by privacy protection regulations such as the General Data Protection Regulation (GDPR). Finally, sound sensors have the potential capture noise at a higher resolution and more consistent areas than other sources, i.e. mobile phone records, where the catchment areas are affected by tower placements and are of varying sizes.

The second contribution is that the paper carries out a comparison between methods from the emergent field of topological data analysis and the 'new forms of data' land use literature. Similarly to other forms of data, the sound data used for the analysis has problems of effective and accurate processing (Arribas-Bel, 2014b). Additionally, the sensors have issues related to measurement, processing, battery and other types of errors (Smith and Turner, 2019). In an effort to address these, this paper adopts topological data analysis (TDA) methods. TDA techniques have found applications in analysing signal or time series data, such as temporal sensor readings, in various domains such as medicine (Emrani et al., 2014), finance (Gidea and Katz, 2018) and epidemiology (Piangerelli et al., 2018). A main feature of TDA methods is their robustness (Cohen-Steiner et al., 2007), which refers to their handling of missing or spurious data during the analysis process. The successful applications, along with the methods' robustness makes TDA techniques a good choice for the analysis of the sound sensor

data, since they can account for the problems of missing or corrupted readings. The comparison is carried out in order to validate TDA effectiveness and to create the best performing set of clusters as measured by internal, external and spatial metrics.

The rest of the paper is structured as follows. The next section places this research into the context of urban sound analysis, urban land use detection using new forms of data and the topological data analysis literatures. The data and methodology section describes the datasets used in detail and the approach taken to the land use problem. Specifically, the methodological approach is to group together the different sensors based only on the similarities between their sound patterns. A way to measure these similarities, is chosen by a comparison between a TDA method and a quality baseline, adopted from the literature. The TDA method itself is chosen based on a quantitative comparison between candidate TDA methods, carried out in Apendix A. In the results and discussion sections the best performing set of results are analysed based on their sound patterns and the surrounding points of interest. Lastly, the resulting areas are situated both in the urban sound analysis and the land use detection literature.

3.2 Literature review

3.2.1 Urban sound analysis

Advances in urban sensor systems have led to more aspects of city dynamics being captured (Zanella et al., 2014). Projects such as Newcastle urban observatory[1] and Array of Things [2] record levels of CO2, noise, temperature and traffic among others with the aim of providing data for more efficient decision making and urban management.

Urban sound analysis aimed at understanding the types, duration and impacts of various sounds found in cities, has benefited from these advances. One of the primary topics of research in this area is noise pollution, which aims to analyse the impacts of noise on human activity and health and to propose ways to mitigate its harmful effects (Zuo et al., 2014). Other topics include event monitoring and acoustic scene classification (Virtanen et al., 2018). Event monitoring deals with creating smart systems that monitor for and respond to specific sound events - gunshots, fights breaking out or home invasions. Whereas acoustic scene classification is concerned with classifying urban sounds into categories based on their source - park, cafe, office, restaurants and others. The difference between this paper and these types of research is that first, we do not have predefined categories and second, actual acoustic sound records are not used for the analysis, rather the data is the recorded max decibels volumes at a specific time and place.

[1] https://urbanobservatory.ac.uk/
[2] https://arrayofthings.github.io/

This paper is closely related to other noise pollution studies from the urban sound analysis literature. Their aim is to classify the types of roads within urban areas, based on traffic noise patterns for the purpose of noise pollution monitoring. Zambon et al. (2016, 2017) group roads into two types based on their hourly sound patterns - one cluster which has two peaks around the beginning and end of working hours and another which has a similar pattern, but with a higher activity during the nighttime hours. Furthermore, Orga et al. (2017) similarly analyse urban sound sensors and find two clusters, identical to the ones described in Zambon et al. (2016, 2017). They also discover that one contains more noise samples related to pedestrians, whereas the other contains more noise samples related to transport. This hints at the potential of the sound sensors to capture underlying characteristics of different areas. Lastly, Brambilla et al. (2019) again find two clusters with similar patterns to the above, based on a metric which measures adverse noise exposure. The consistent finding among all of these studies is that urban sound near roads is dominated by two patterns with similar peaks during working hours, but one cluster shows more activity during the nightime.

The focus of this paper is on a more granular grouping of the sound sensors than the above work and one which captures functional information about the surrounding areas. First, the goal is not to classify streets, in order to find optimal placements of sensors for the purpose of noise pollution monitoring. But rather, the focus is on analysing the link between the functional usage of the surrounding area and the captured sound patterns. To this end, different methodological approaches are adopted - Topological data analysis (TDA) methods to measure the differences between sensor patterns and HDBSCAN clustering to achieve a finer grained, higher quality grouping of the clusters. Second, the positions of the sound sensors used in this paper study were not necessarily chosen for the monitoring of traffic noise. Placements were effected by resident demand and by air pollution considerations (James et al., 2014).

In spite of these differences, the use of this type of sound sensor data for land use detection is motivated by several developments in the urban sound literature. That human activity has a large effect on sound sensor patterns has already been suggested by Groos and Ritter (2009). In addition, studies from traffic analysis suggest that the surrounding built environment affects the sensors readings. Hupeng et al. (2019) show that different street characteristics affect sound propagation. Zuo et al. (2014) finds 'ubiquitous traffic noise exposure across Toronto and that noise variability was explained mostly by spatial characteristics'. These results suggest, that at the very least, areas with varied functional usage and vastly different built characteristics, such as parks and nighttime entertainment areas, are identifiable. Furthermore, there is scope for future improvements in the field that would enable the better capture of human activity directly. Recent experiments in sensor technology (Lau et al., 2018) show that modified sensors are capable of capturing only sounds resulting from human

activity.

3.2.2 Urban land use detection

Popular approaches to urban activity and land use detection include the use of survey data, commissioning GPS trackers studies, and remote sensing. Remote sensing in a land use context refers the analysis of satellite images through GIS techniques in order to infer different land use types. The focus is not necessarily restricted to urban areas and the extent of such studies can cover the entire landmass of a country deriving other non-urban land use types such as forest, agriculture, watersheds. Remote sensing is currently the most widely adopted method of land use detection. There is no single methodology and the analysis can focus on different aspects captured by the satellite images such as night lights, transport links or buildings. Advantages of this method are it can be globally applied and that there is temporal data available, making it possible to track changes in land use over time. Two related disadvantages of remote sensing are that there is no actual functional data captured and that the majority of studies discern only one urban land use class (Reba and Seto, 2020).

Activity surveys are a popular tool and have been used to address a variety of urban land use research questions. They require dispatching surveyors in order to collect data either in person, though the mail or online apps. With the advancement of technology GPS trackers are being adopted in addition or to entirely replace surveys . These trackers give longitude–latitude at short time intervals and do not allow for participant misreporting to affect the data. One of the main uses of both activity surveys and GPS trackers has been for transport planning to better understand travel behaviour, route choice and traffic safety (Shen and Stopher, 2014). Other uses include the characterization of neighbourhoods or city areas based on citizen activities. For example, Sung et al. (2013) use activity surveys in the Seoul to evaluate whether diversity of functions and activity within a neighbourhood has effects on its vitality, while Marquet and Miralles-Guasch (2015) evaluate the importance of walkable environments to neighborhood wellbeing in Barcelona. Data derived from these surveys has the advantages of good experimental design - it can be representative of the studied population and any inferences made from the data can be statistically robust. The main disadvantages of surveys are the high cost and the complex planning required to carry them out successfully, as well as participants misreporting data. However, as a new method, the GPS survey also has some shortcomings, such as unstable signal acquisition in certain areas and difficulties in GPS data processing (Shen and Stopher, 2014). Additionally, it is expensive to track large numbers of individuals and usually the scope is limited to the low hundreds of participants (Frias-Martinez and Frias-Martinez, 2014). Additionally, due to the high cost and complexity such surveys are not carried out frequently.

As new forms of data have gained popularity, land use researchers have started adopting them in order to address the disadvantages mentioned above. These new forms of data are different in type and come from distinct sources such as street level imagery (Zhang et al., 2019a), app data such as twitter or foursquare (Frias-Martinez and Frias-Martinez, 2014), or mobile phone records (Cici et al., 2015). The advantages of these data types over more traditional data are that they provides a real-time picture of urban dynamics. With increased adoption the data becomes more representative and the results of the analysis improve. It also address several of the issues of using surveys to gather data, such as respondents misreporting activities. The disadvantages are that this type of data might be subject to different privacy laws in different countries, making it difficult to analyze or even collect. It can also be hard to get access to this data, since it is closely guarded by companies due to its commercial value. Lastly, preprocessing and analysing the data requires care, since the choice of methodology and preprocessing can effect the results significantly.

There are two popular approaches to land use detection using these data: one where the focus is on functional relationships and another where its on intensity of activity. Relational 'new forms of data' include among others geolocated twitter mentions, location-based social network (LBSN) data, ride sharing or hauling data, or mobile phone calls between places. These have found applications at various scales and for different problems. Calafiore et al. (2021) use location based social network to characterize and compare the neighborhoods of different cities. A relevant paper to this one is the study performed by Lenormand et al. (2015) in five Spanish cities. It delineates four areas of different land use: residential, business, logistics/industry, nightlife. These approaches are possible anywhere where there is enough data available and represent an important addition to traditional data sources, since they directly capture functional relationships at high spatial and temporal resolutions.

The other type of data, focuses on intensity of activity at predefined places - cells in a grid, streets, mobile operator tower catchment areas and others. This is in contrast to studies that try to understand land use function from the position of the individual, i.e. surveys or GPS. The most popular examples of this type of data are number of tweets per hour or number of phone calls per minute recorded at a geographic location. Sound sensor data is also an example of this data type.

Geo-tagged twitter data is one of the most widely used 'new forms of data'. It has been used to infer urban land uses by clustering different areas in the city based on profiles of tweet activity. With this methodology Frias-Martinez et al. (2012) and Frias-Martinez and Frias-Martinez (2014) identify areas with specific tweet activity signatures that corresponded to different types of land use. The analysis is carried out in different places - Manhattan, London and Madrid , and both studies are able to detect clusters that correspond to four urban land use types: business, leisure/weekend,

nightlife and residential, while Frias-Martinez and Frias-Martinez (2014) detects an industrial pattern of activity unique to London. Using a similar methodology, Zhan et al. (2014) inferred four types of land use areas in New York City: residential, retail, open space/recreation and transportation/utility. Other studies have been carried out using other geolocated data similar to tweets, i.e. Chen et al. (2017) using the social media "Tencent".

Mobile phone records is another 'new form of data' that has been used in order to capture land use patterns in different cities (Pei et al., 2014; Toole et al., 2012; Furno et al., 2015; Miao et al., 2018; Zhang et al., 2019b). The studies differ in scope with some focusing on single cities, while others compare patterns identified in various cities to each other. Similarly to results from twitter activity analysis, there are at least four types of land use areas identified - residential, business, leisure and entertainment. These show up across different methodologies and data types in the studies. It should be noted that, there are papers that as a result of the specific clustering methodology used end up with a significantly larger number of clusters, i.e. (Furno et al., 2015).

Each of these land use types is characterized by a temporal activity pattern - measured by the number of tweets per hour in the first case and the number of phone calls per hour in the second . For example, business districts have an activity signature characterized by high level of activity during working hours and low activity the rest of the time (Cici et al., 2015). This activity pattern is reversed for residential areas and there is altogether another usage pattern for entertainment or mixed usage areas. Several papers have found areas with similar business patterns in different cities, but different patterns for residential areas (Grauwin et al., 2015; Furno et al., 2017). Most studies focus on the pattern of activity and completely ignore magnitude, with some exceptions that give marginal improvements (Pei et al., 2014). This suggests that differences in the areas are captured by the variations in the activity patterns, rather than in population effects.

This paper aims to use sound sensor data for land use detection. Sound sensor data is closely related to both the mobile phone activity data and the twitter activity data types described above. Both measure the intensity of actions of different types, related to human activity, at fixed places throughout an urban area. In addition to having all of the general advantages of new forms of data already discussed, generalised noise data as used in this paper, is more available to the public in contrast to other sources. Sound sensor readings are not limited by the willingness of private companies to share the data or by privacy protection regulations such as the General Data Protection Regulation (GDPR). Another advantage sound senors have over other data sources, is that they capture human activity at a higher resolution, than say mobile phone records. Furthermore, sound sensors capture the sounds generated by all people in the respective sensor's catchment area. This is in contrast to most mobile phone or

twitter data, that capture only the activity of people that use a specific mobile operator or app. A disadvantage this data shares with other 'new forms of data', is that it requires additional effort to process and analyse, in order to get meaning full insights. (Arribas-Bel, 2014b). Additionally, the specific sensors used to gather the sound data suffer processing errors, battery issues and other problems (Smith and Turner, 2019).

3.2.3 Topological Data Analysis

In order to address the issues with processing and analysing the data and this paper turns to methods of Topological data analysis. To effectively analyse the differences between sound sensors, a way to compare the captured sound patterns is necessary. In general, the sound patterns can be considered as time series or longitudinal data - readings by the sensors, of the same phenomenon, at different points in time and TDA techniques have been used to analyse signal or time series data in various domains such as medicine, finance and engineering.

To measure the difference between two time series they are either analysed directly or converted to a time-lagged representation (Perea et al., 2015) and persistent homology is applied (Chazal and Michel, 2017). Persistent homology provides an object, a 'persistent diagram' that describes the multi-dimensional coarse shape of the time series. The zeroth dimension corresponds to connected components or clusters, the first - cycles present in the time series data. The second and above dimensions represent higher dimensional generalizations of cycles. This information is reflective of various underlying periodic patterns in the time series, as well as critical points - peaks and valleys. Persistent diagrams of different time series can be compared to each other through 'bottleneck distance', which is a measure robust to noise (Cohen-Steiner et al., 2007). This approach has been shown to be capable of classifying volatile time series data (Umeda, 2017). Costa and Škraba (2014) uses it to compare the spread of influenza like diseases in seasons of influenza in Italy and Portugal. Emrani et al. (2014) uses it to classify breathing patterns according to the presence of wheezes, while Piangerelli et al. (2018) successfully detect when a patient will have an epileptic seizure. Gidea and Katz (2018) uses a similar method to predict transitional market events, like financial crashes. There are also direct applications that show improvements in clustering results when using TDA techniques (Perea et al., 2015). These results show the potential of TDA techniques to capture differences between time series.

3.3 Data

3.3.1 Sound data

The data used in the analysis comes from the Newcastle Urban Observatory Environment system of sensors (James et al., 2014). It collects data from numerous sensors across across Newcastle and Gateshead in the UK. The sensors record continuously quantities such as temperature, air quality and noise levels. For this analysis only the maximum decibel levels at a particular time inteval are used. The sensors are generally placed near central locations - city centre, business parks and others, however residents can apply for the placement of sensors in their neighbourhoods, parks or other areas through the related project - 'SenseMyStreet'[3]. Some of the major projects the sensors are already being used for is designing a visualization engine for smart cities (Holliman et al., 2017) and monitoring of urban runoff (Jonczyk et al., 2016).

The downloaded raw data used for this paper covers all the available sound sensors over the two month period between the 1st of January 2019 and the 1st of March 2019. Figure 3.1 shows the geographical position of the sensors in Newcastle and Gateshead. It can be seen from the figure that, most sensors come in groups along a specific street and naturally form a clustering. There are readings from 62 sensors in the raw dataset, which record the maximum captured sound level at similar time intervals. Figure 3.2 shows the processed sound patterns recorded by four sensors placed in different areas - this is the only type of data that will be used for analysis. The preprocessing of the sound patterns was necessary in order to address potential issues with the data stemming from problems with the sensors - battery problems, data record problems and measurement errors (Smith and Turner, 2019).

First, all readings were aggregated hourly. The data was then limited to a period of one month - from 6th of January 2019 to 6th of February 2019. The period was chosen since it was the timeframe that had available data for most sensors. Next, all sensors with a higher proportion of missing to available data were dropped. For all remaining sensors, the missing values were mean interpolated based on day of the week and the hour. Lastly, all of the sensor data is normalised, in order to focus the subsequent analysis on the patterns of sound rather than magnitude. The processed dataset, consists of 40 sensors with hourly noise readings covering the entire 1 month period.

[3] https://sensemystreet.uk/

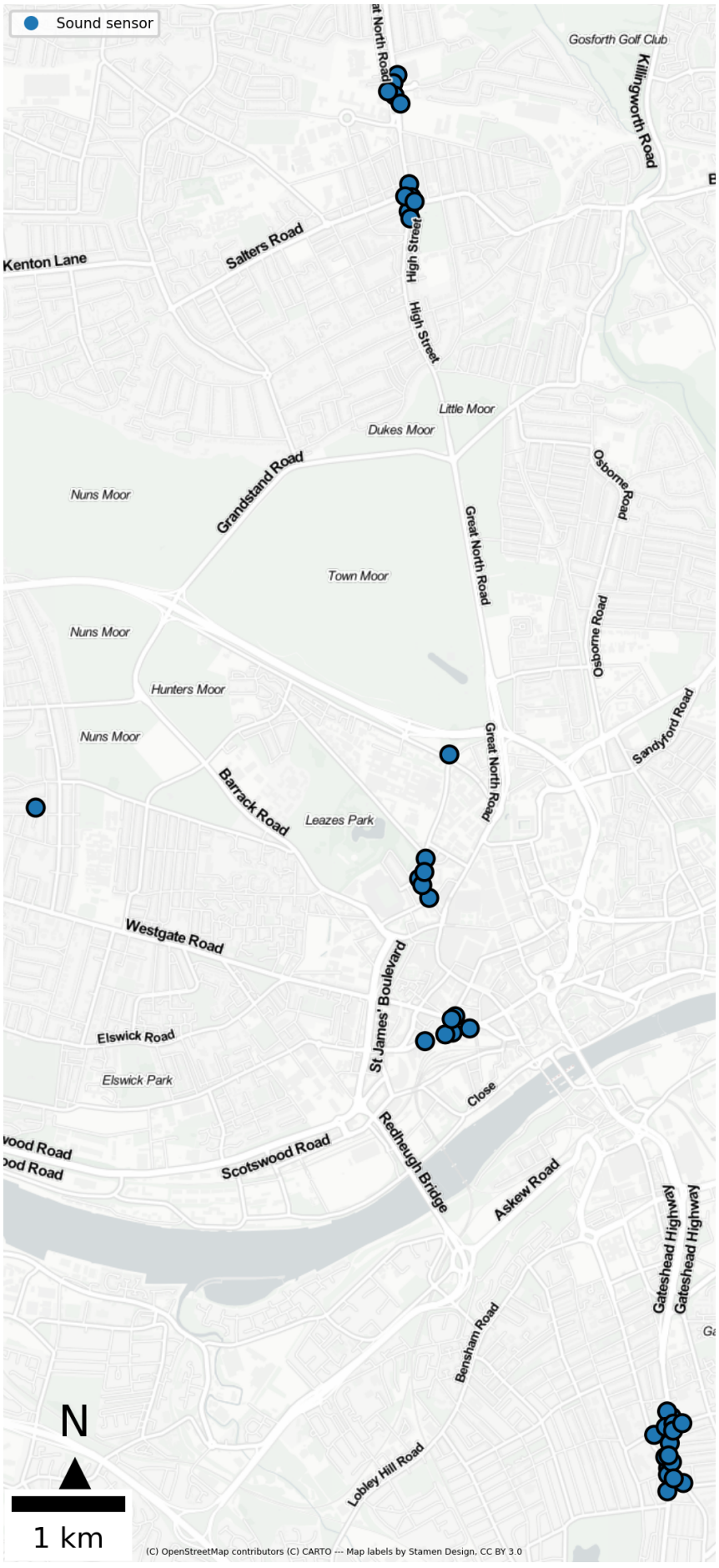


Figure 3.1: Sound sensor positions in Newcastle and Gateshead

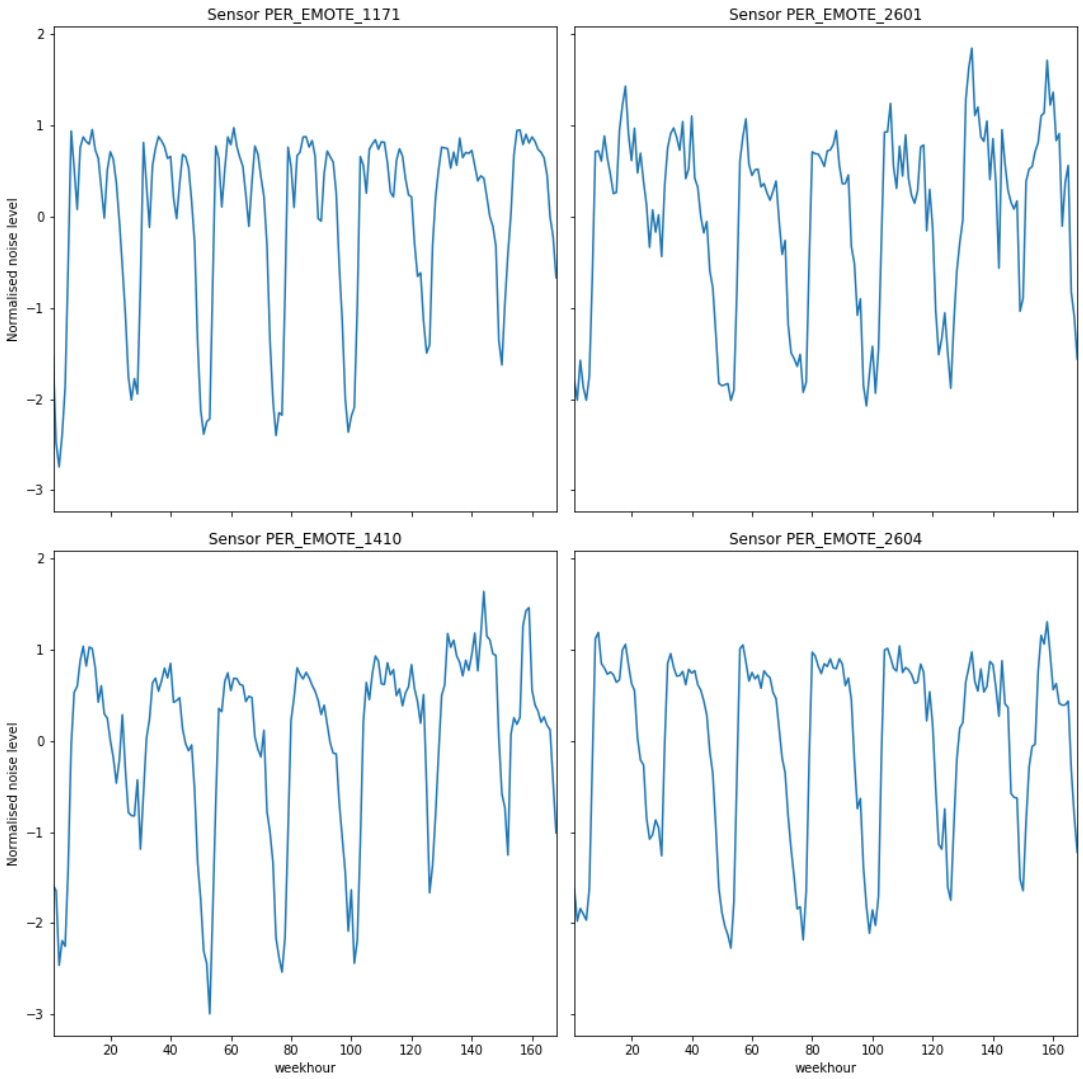


Figure 3.2: Sample sound patterns from four sensors

3.3.2 Openstreetmap data

Data from Openstreetmap (OpenStreetMap contributors, 2017) is used in the analysis of the clusters and sound sensor placements in order to better characterize the areas the sensors are in. Specifically, since each of the sensors is located on a street, characteristics of the street and surrounding available Openstreetmap amenities, as well as buildings in a 100m radius are gathered.

There are three types of roads the sensors are placed in - primary, secondary and tertiary. The available points of interest in Openstretmap come in six broad categories. In order to better understand these six categories, they are presented along with example points of interest that would fall within them:

1. Sustenance - bars and restaurants

2. Education - schools and universities

3. Transportation - bus and taxi stops
4. Financial - banks and change bureaus
5. Healthcare - hospitals and pharmacies
6. Entertainment - cinemas, nightclubs and theaters

3.4 Methodology

The methodological approach can be broadly broken down into two parts. The first part is concerned with the building of the optimal grouping of sound sensors into clusters that represent areas of similar functional usage. Two sets of clusters are derived in order to evaluate the effectiveness of the new TDA approaches. The first set of clusters uses a TDA method to differentiate sound patterns, while the second an approach developed by Furno et al. (2015) for the analysis of mobile phone signals for land use detection. These sets are then compared using external, internal and spatial coherence measures. In the second part the best performing grouping is analysed in detail.

3.4.1 Clustering

The main aim of the paper is test the viability of sound sensor data as a secondary source of information for urban dynamics. Specifically the goal is to see whether the captured sound patterns differ, depending on how people use the surrounding areas. First, there is no prior knowledge how many functional areas exist in the data and second, it is expected that if the sensors do in fact capture activity data they will show patterns similar to the core types of previous land use studies. A clustering approach was chosen for these two reasons, and in accordance with other papers mentioned in the previous section, e.g - (Pei et al., 2014; Cici et al., 2015; Furno et al., 2015; Miao et al., 2018; Zhang et al., 2019b).

The methodological starting point in those studies is calling records, collected at mobile tower stations or geo-tagged tweets in different parts of the city. This gives a representation of activity patterns through time and space. These time series are then pre-processed and prepared to be analysed using clustering methods. This is done with the goal of finding out how many distinct activity patterns there are and where the geographical places where these patterns occur are. As a last step, external data such as nearby points of interest is used to validate the results.

This paper follows the same broad approach when analysing the sound sensor data. The difference is that the choice of single-linkage hierarchical clustering and a choice of cutoff metric will be replaced by clustering using HDBSCAN (Campello et al., 2015). HDBSCAN offers improvement over hierarchical clustering in three ways.

First, it takes into account the presence of outlier sensors - that is sensors that display anomalous or very different behaviour. This is appropriate for the sound sensors data due to problems such as measurement or processing errors, battery issues and others (Smith and Turner, 2019). Second, it is able to detect clusters with a different number of members (varying density). And lastly, it gives a flat clustering. Although the approach is slightly different, the two methods are directly related and HDBSCAN was developed as an extension of single-linkage hierarchical clustering, to address some of its shortcomings (McInnes and Healy, 2017).

In order to make the comparisons between the TDA and the baseline methods more exact, the same clustering method - HDBSCAN - with the same parameters will be used. HDBSCAN requires only one choice of parameter and that is the minimum number of sensors in cluster, in order to consider the group a valid cluster. The choice for every method is two, due to the spatial dispersion and total number of the sensors.

Measuring differences in sound patters

In order to apply the HDBSCAN clustering algorithm, a measurement of the difference between sound patterns needs to be defined. This paper computes differences between sound patterns in two ways - first using a TDA algorithm and second with an already established approach. These differences are then used to create clusters of sound sensors by applying HDBSCAN and the results are evaluated as described in the next section.

As mentioned before, the captured sound patterns by the senors are a type of time series data - readings of the same phenomena across time. Clustering of time series is an active area of research and is done for exploratory data analysis, to get summaries of large datasets or as a intermediate step in other methods (Aghabozorgi et al., 2015). As this is an active area of research there exist numerous ways to measure the differences between sound sensors. Furno et al. (2015) carry out a comparison between different methods with mobile phone data for the purpose of functional land use detection. The best performing method from that study is used as a baseline in this paper. It consists of two steps. The first step is to extract a median, weekly activity for each sensor. Afterwards, each of these typical weeks are normalised and the pairwise distance between all sensors are computed based on their correlation. These pairwise distances are then used by the HDBSCAN algorithm to compute the clusters.

3.4.2 Cluster evaluations

Clustering is the attempt to group data in a way that meets with human intuition (McInnes and Healy, 2017). In this paper we use numerous metrics that try and capture aspects of this intuition - internal clustering measure, external data and spatial

autocorrelation.

Internal validation metrics measure desirable properties of the resulting clusters. This paper uses one of the most widely metrics - silhouette score. It measures how separated the clusters are - i.e. whether a sensor could reasonably fall within another cluster as defined by the clustering method. It ranges in value from one to minus one, where scores near zero indicate overlapping clusters. In general, negative values can be interpreted as having sensors assigned to the wrong cluster, as there exists a more appropriate assignment choice. Positive values close to one mean a perfect separation of clusters.

Internal evaluation metrics are helpful in providing information about clustering results, however this information is not always practically relevant to the task (Campello et al., 2015). To account for this, each set of clusters are further validated externally and characterized with points of interest and building data data from Openstreetmap. If the resulting clusters represent different functional area profiles, then this should be reflected in the distribution of the amenities and buildings of the surrounding area. This is a common assumption and validation approach among many twitter and mobile phone data land use studies - (Cici et al., 2015; Furno et al., 2015; Miao et al., 2018; Zhang et al., 2019b). To measure this difference for each set of clusters, the individual clusters are treated as observations and the distribution of different types of amenities, as described in section 2, are their characteristics. The differences between them is measured by Euclidean distance. The larger the values the more functionally separated the clusters are. This distance is based on external, not used in the formation of the clusters data. Smaller values indicate similar clusters, meaning that different clusters do not capture different functional information. For each set of clusters both the mean and minimum distance is reported - large mean distances indicate that there is atleast one cluster different than the others within the same set, whereas a large minimum value means that the most similar clusters are dissimilar.

Another external validation metric used is spatial autocorellation. Since most of the sensors are placed along the same streets, they capture similar or the same noise patterns and a good cluster assignment should group at least some of the sensors in a spatially coherent manner. To measure how much this is the case for each set of results its clusters are compared against a naive spatial clustering. This naive spatial clustering simply groups sensors on the same street into the same cluster. Adjusted mutual information score (AMI) is used to measure how much each set of clusters resembles this naive spatial clustering. The measure's values range from zero to one, with a score close to one meaning that there is strong spatial autocorrelation in the analysed set of clusters.

It should be noted again that the only data used for the clustering is the sound patterns captured by the sensors, (a sample is shown in in Figure 3.2). Neither spatial

information, nor points of interest or building data will be used. If the analysis results in clusters that respect the geography and street characteristics, then this is evidence in support of the idea that sound readings can be used to capture distinct land usage patterns.

3.4.3 Sound patterns and TDA

As used in this paper, Topological data analysis methods allow for another way of measuring signals. The TDA method used in this paper is based on the ideas of persistent homology, diagrams and bottleneck distance (Chazal and Michel, 2017). The specific method used for the comparison is called 'Lowerstar' and it focuses on the peaks and valleys in the time signal. In applying the methodology to our data we follow the successful application of this method by Knyazeva et al. (2016). We directly compute the persistence diagram for each sound sensor based on its entire sound pattern. Each persistence diagram describes the shape of the time series in terms of critical points - local minimums and maximums. The diagram itself consists of components, equal to the number of local minimums and each component has two values - a birth and death times. The birth times correspond to local mins and the death times correspond to local maxes. These diagrams are used to compute the differences between each pair of sound sensors based on 'bottleneck distance', which is a measure of similarity between persistence diagrams that is robust to noise and sensitive to small changes (Cohen-Steiner et al., 2007). These pairwise distances are then used by the HDBSCAN algorithm to compute the clusters.

In addition to numerous successful applications, this TDA method was used since it was the best performing TDA method in the comparison carried out in Appendix A. This cluster evaluation comparison followed the same format as the one that is carried out between the TDA and baseline approach, described in the previous section.

3.5 Results

3.5.1 Comparison results

Table 3.1: Clustering results

	Number of clusters	Number of Outliers	Silhouette score	AMI	Mean difference in POI distribution	Min difference in POI distribution
Baseline	5	12	0.558	0.365	0.425	0.107
TDA Lowerstar	3	17	0.304	0.128	0.104	0.075

Table 3.1 shows the results of the comparison procedures. The application of the Furno et al. (2015) sound pattern differentiating procedure results in 5 clusters that capture

28 sensors and 12 outliers. The 'Lowerstar' TDA method on the other hand, splits 23 out of the 40 sensors into three clusters and 17 outliers.

Overall, the methodology adopted from Furno et al. (2015)('Baseline' in the table) performs better than the 'Lowerstar' TDA approach across all measures. The mutual information score (AMI) measures the similarity between the resulting clusters and a spatial grouping of the sensors. The higher mutual information score for the baseline method shows that it captures the spatial dimension of the sensors better than the TDA approach. Similarly, the higher silhouette score, means that clusters derived from this methodology are more separate than the clusters resulting from the TDA comparisons. Lastly, the mean and minimum external validation metrics show that the clusters differ more in the distribution of points of interest for the baseline methodology. These higher values suggest that the clusters capture areas with more distinct distributions of points of interest, representing different functional usages. Since the baseline clusters perform better against the metrics, the rest of the analysis focuses on them.

3.5.2 Clustering results

The baseline methodology produces five clusters that contain 28 out of the 40 sensors and 12 outliers. An advantage of the HDBSCAN algorithm is that it enables the discovery of distinct points based on the outliers. Table 3.2 presents the average distance of the outliers to every other datapoint. As it can be seen two sensors have a noise pattern that is on average almost twice as different as the rest. It is very likely that these sensors represent a different type of behaviour to the other clusters, but cannot themselves form clusters since they are lone sensors. The fact that they are outliers and not clusters is a limitation of the data. These sensors are analysed along with the other clustering results under the names 'Claremont Road' and 'Baldwin Avenue'. The rest of the outliers are discussed as a group.

Table 3.2: Average distance to other points for each detected outlier

Sensor	Average difference to other sensors
PER_EMOTE_1004	0.078
PER_EMOTE_1005	0.058
PER_EMOTE_1171	0.127
PER_EMOTE_1204	0.055
PER_EMOTE_1205	0.061
PER_EMOTE_1208	0.062
PER_EMOTE_2601	0.122
PER_EMOTE_2605	0.078
PER_EMOTE_2606	0.076
PER_EMOTE_2763	0.057
PER_EMOTE_2766	0.048
PER_EMOTE_2902	0.054

Figure 3.3 shows sensors colored by cluster membership, with the clusters named after the street or place the majority of sensors are in. In general, sensors that are on the same street are in the same cluster (with the exceptions of outliers). For example, the 'High Street' cluster covers the most area and all the sensors on its street. It is interesting to note that most outliers lie on intersections, which suggests that that is where the noise pattern changes on a street. There is also a cluster named 'Corner sensors' which contains 2 sensors on the corners of the 'Durham Road' cluster.

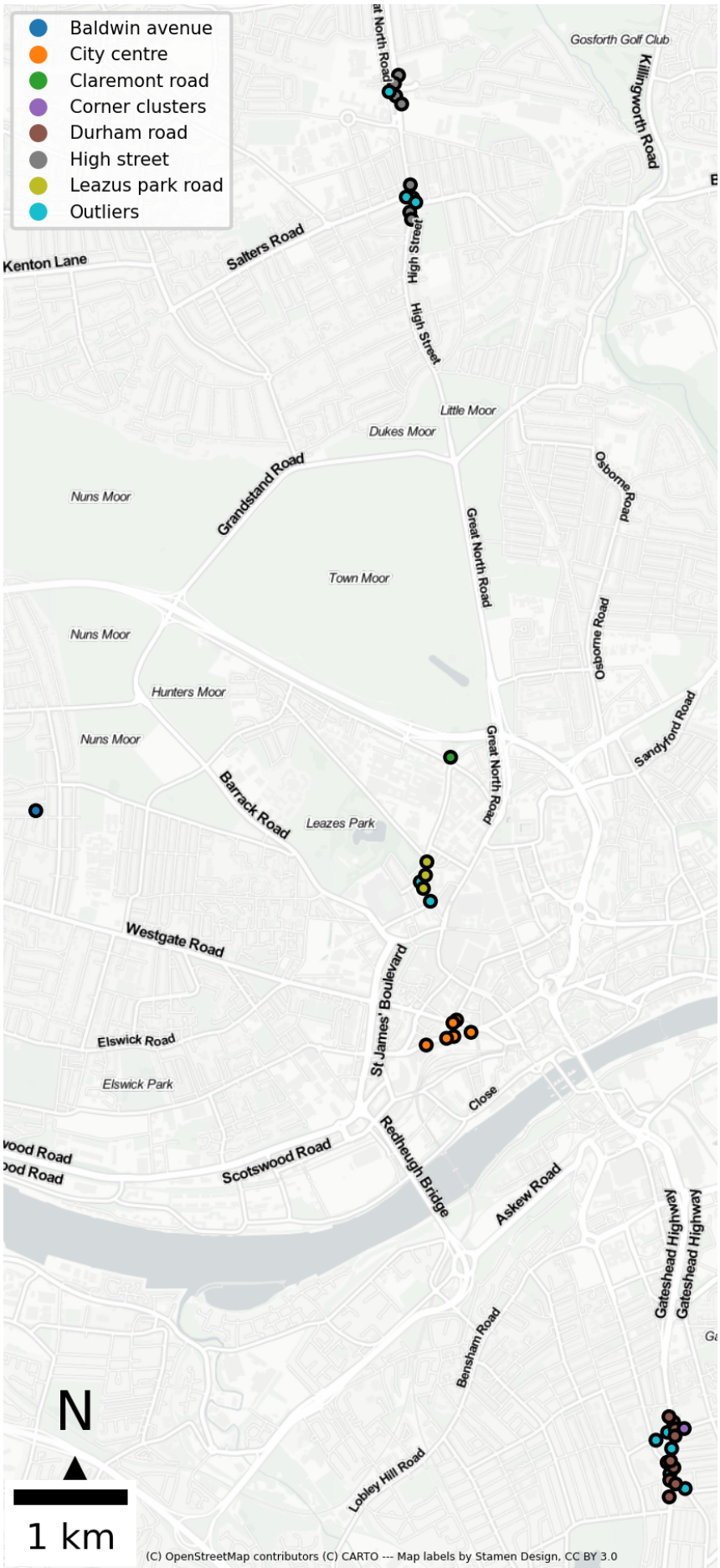


Figure 3.3: Baseline clusters

Figures 3.4 and 3.5 show the normalized median weekly noise pattern for each cluster. Most of the clusters have a similar dominant pattern that follows typical working hours - with two peaks, followed by a drop in activity. The differences in the patterns come from considering the relative times of the peaks as well as the differences between the weekend and weekdays. For example, the most distinct cluster stands out because of the large spikes in noise during weekend nighttime hours.

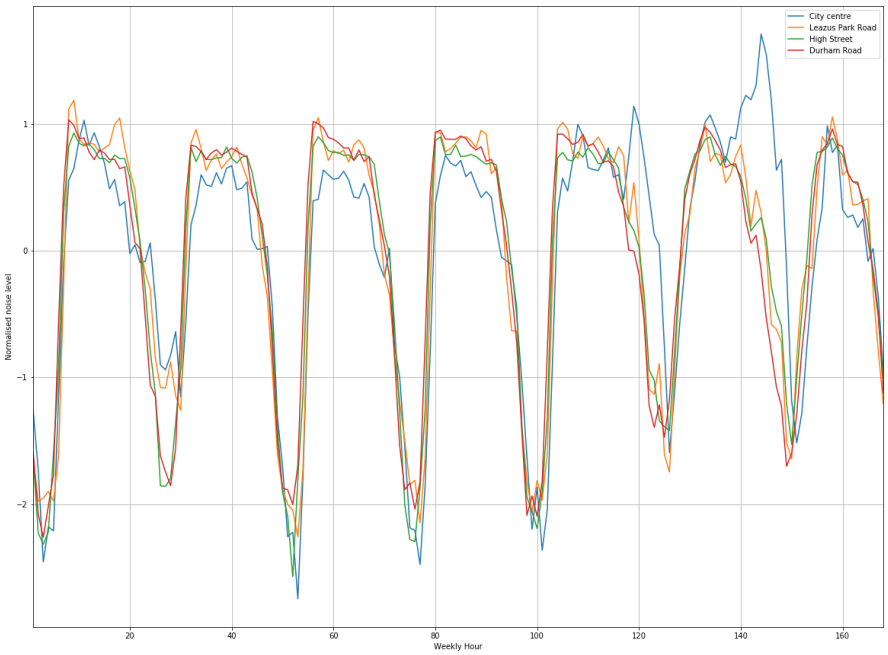


Figure 3.4: Average week for the first 4 clusters

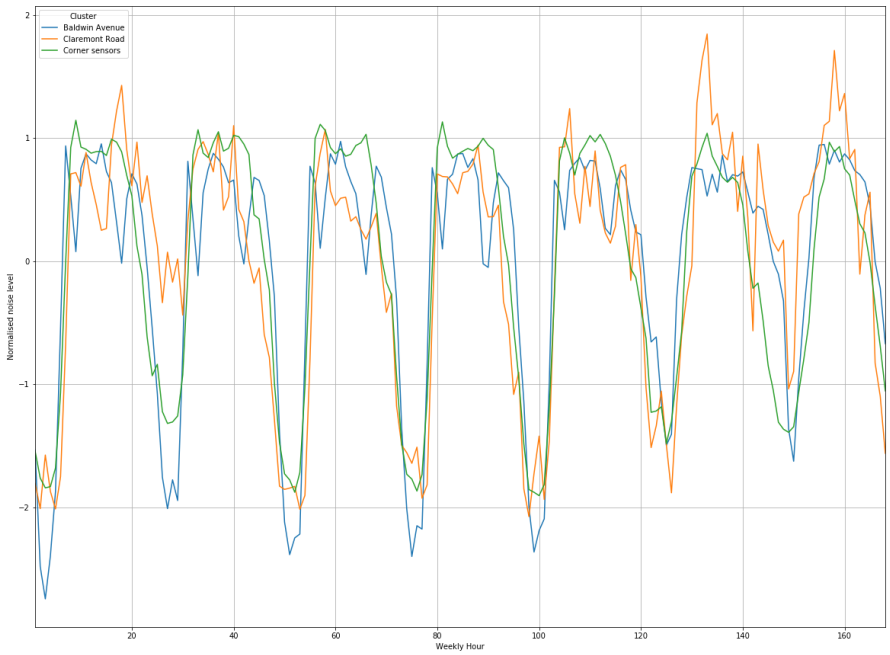


Figure 3.5: Average week for the last 3 clusters

Tables 3.3 and 3.4 shows the distribution of points of interest and buildings, available from Openstreetmap, across the clusters. The number of residential buildings varies greatly among the clusters, even though they cover geographical areas of different sizes. For example, 'Baldwin Avenue' a lone sensor has 44 residential buildings in a 100 metre radius around it, whereas the 'High Street' cluster covers a geographical area, at least 5 times larger, but only has 16. There are also noticeable differences in the POI data. For example, the city centre cluster is the only one with any entertainment POIs. Another thing to note is that the sustenance and transport points of interest dominate most clusters with varying levels.

Table 3.3: POI results

	total	sustenance	education	transportation	financial	healthcare	entertainment	sensors
City centre	102	0.470	0.009	0.460	0.019	0.000	0.039	6
Leazus Park Road	25	0.360	0.040	0.560	0.000	0.040	0.000	3
High Street	32	0.343	0.000	0.531	0.093	0.031	0.000	8
Durham Road	13	0.615	0.000	0.230	0.000	0.153	0.000	9
Baldwin avenue	1	0.000	0.000	0.000	0.000	1.000	0.000	1
Claremont road	7	0.000	0.142	0.714	0.000	0.142	0.000	1
Corner sensors	2	0.000	0.000	0.500	0.000	0.500	0.000	2

Table 3.4: Building data

	Residential buildings	Total buildings	Sensors
City centre	2	15	6
Leazus Park Road	14	16	3
High street	16	26	8
Durham Road	99	100	9
Baldwin avenue	44	44	1
Claremont road	0	0	1
Corner sensors	108	109	2

The two largest clusters are 'Durham road' and 'High street'. In terms of the road network 'Durham road' is a traffic heavy secondary road, while 'High street' is a primary road. This is reflected in the larger number of transportation POI around the sensors on 'High street', which feature a parking and a metro link. 'High Street' also has relatively less sustenance POIs than 'Durham road' and is the only cluster with a significant number of financial POI around it. 'Durham road' on the other hand has a high percentage of healthcare buildings such as doctors and pharmacies around it. In terms of residential buildings 'Durham road' has significantly more. These differences suggest distinct primary usage of each area, however the noise patterns of the two clusters are similar as seen in Figure 3.4. In order to better see the differences the average week and weekend day are shown in Figure 3.6. Both clusters follow the same general dominant patterns. On weekdays, there are peaks in the morning around 8:00 A.M. and noise remains consistent until 19:00 and then starts to drop off. During the weekends there is a peak at 12 and a leveling off, which indicates that sensors are placed near areas that have high activity during working hours. 'High street' overtakes 'Durham Road' at around 18:00 during the weekdays and 20:00 during the weekends. It is possible that this is a result of the high level of traffic POIs. The two cluster patterns identified here are similar to the ones described in a number of studies in Italy (Zambon et al., 2016, 2017) - similar working hours pattern, but the differences between them come in the nighttime activity. Other than this, there are no clear differences in the patterns of these two clusters. This suggests that general noise patterns dominate busy traffic roads and make it hard to distinguish between the types of areas the sensors are in.

One of the most present clusters in studies using twitter and mobile phone data are the worktime or business areas identified in most studies - i.e. Pei et al. (2014); Cici et al. (2015); Furno et al. (2015); Miao et al. (2018); Zhang et al. (2019b). These are characterized by high activity during the working hours and low activity during the weekends. In our dataset, one area has a lot of businesses and offices - 'High Street', however it is very similar to another one 'Durham road', which is a primary road with mainly houses, parks and shops. This speaks to some limitations of using

sound sensors in that for sensors on or near heavy traffic roads, the type one traffic noise pattern dominates and differences are hard to detect. Nevertheless, the algorithm managed to detect the differences, which shows promise for future studies.

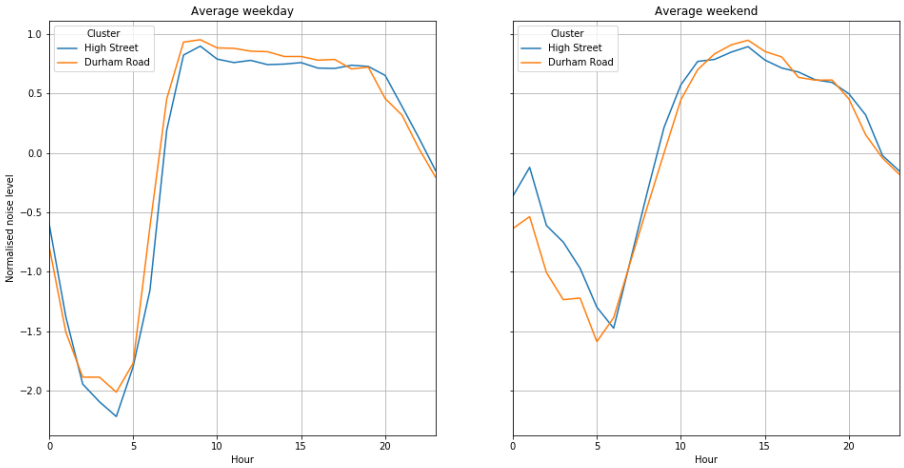


Figure 3.6: Average weekday and weekend comparison between the clusters on Durham road and High street

The next cluster - 'City centre' - consists of six sensors placed on three streets near the Newcastle city centre. The sensors are located near busy primary roads, hotels, transport links and the only nightlife present in the data. The cluster has the most distinct pattern with high levels of noise during weekend nighttime hours. It is a representative of the second type of roads found in the Dynamics studies (Zambon et al., 2016). A cluster with a pronounced night activity and relatively similar daytime pattern to the worktime clusters. The separation of sensors into city centre and others is expected since there are results that suggest that in England city centres have more urban vitality than other areas within the city, meaning more activity at all times (Sulis et al., 2018). Furthermore, a city centre or nightlife cluster is present in other studies that use mobile phone records or geo-tagged twitter activity profiles - i.e. Soto and Frías-Martínez (2011); Frias-Martinez and Frias-Martinez (2014).

The next clustering group is 'Leazus Park Road' and consists of 3 sensors. Leazus Park Road is a tertiary road, which has some houses, a number of shops and restaurants around it, as well as a football stadium nearby, but outside of the range of the sound sensors. Figure 3.4 shows the median weekly noise patterns for the sensors in this cluster. It can be seen that this cluster has a mixture of the patterns of the previously discussed clusters. There is a general worktime activity, but there are also increases during night time, as well as spikes of activity on the weekends starting at 20:00. There is also a number of residential houses around the sensors. This suggests that the area around the cluster has more of a 'mixed usage' - there are houses as well as businesses

in the area. This cluster type is present in many twitter or mobile activity studies (Miao et al., 2018; Furno et al., 2017; Toole et al., 2012; Soto and Frías-Martínez, 2011; Frias-Martinez et al., 2012; Frias-Martinez and Frias-Martinez, 2014). These cluster types have an activity pattern that lies somewhere between residential and business or work areas - constant usage during the weekdays and the weekend days. This is in contrast to our case, where 'Leazus Park Road' has high worktime activity, which increases during night time, as well as spikes of activity on the weekends starting at 20:00.

The next cluster - 'Corner sensors' - represents sensors placed on streets that intersect 'Durham road'. It is difficult to analyse this cluster precisely since it has the same building and POI characteristics as the 'Durham Road' cluster, due to its proximity to it. Its pattern is presented in Figure 3.5 and is again similar to the 'Durham road' one. The same is true for the other outliers - usually they lie on the corners as well or are too close to the other sensors.

The last two clusters 'Baldwin Avenue' and 'Claremont Road', only have one sensor each. Figure 3.5 shows the pattern of noise in the typical week for 'Baldwin Avenue'. It has three relative peaks during the workdays - the first occurring before working hours start around 7:00 and the last smaller peak after working hours finish. Another thing to notice is the drop of activity from 7 to 11 and from 16 to 20:00. On the weekend there is a more activity during the working hours and the peaks start after the workday peeks. The distribution of the POIs is different to the rest of the clusters, and the only nearby amenity is a pharmacy. The pattern, the lack of points of interest and the high number of residential buildings suggest that this is a residential area. This is in contrast to the residential clusters defined from tweets and mobile phone records, where the peak activity of residential areas is only after working hours. Examples of these clusters can be found in most studies, i.e. Miao et al. (2018); Cici et al. (2015).

Figure 3.5 also shows the weekly pattern for the sensor on 'Claremont road'. It can be seen that there is relatively low usage during the weekdays compared to the weekend. It is different from the city centre cluster by the fact that the noise levels peak at 1:00 and drop when approaching nighttime. The POI distribution for this sensor is completely dominated by infrastructure points of interest, however most of those are bike kiosks and parking. This sensor is identified as a representative of a the weekend or leisure type of area. These groups, again present in new forms of data land use studies - Furno et al. (2015); Frias-Martinez et al. (2012) - have sometimes been defined to encompass museums as well as parks. However, in our dataset it refers only to parks.

3.6 Discussion

The results from the cluster analysis show that there is a strong spatial correlation in sound patterns - most sensors placed on the same street are assigned to the same cluster. However, sensors on different streets with similar sound patterns are also grouped together. This is the case for the 'City centre', 'High street' and 'Leazus Park Road' clusters. In other cases, sound sensors that lie on the same street are split - some are grouped in a cluster, while others are identified as outliers. This happens for the 'Durham road' cluster, for example - out of the eleven sensors on the street, two are identified as outliers. Both of these results suggest that the delineations capture more information, than simply the underlying geography of the sensors. The second result, specifically indicates that sound sensors have the potential to identify areas of interests at a higher than street level scale.

Another interesting spatial pattern is the existence of a 'Corner Sensors' cluster, as well as the fact that many sensors identified as outliers lie near street corners. This result suggests that the sound patterns on some streets significantly change at the intersection with other streets and that identified cluster boundaries lie on street intersections. Since the sound levels on streets are effected by both the built environment (Hupeng et al., 2019) and the types of amenities on them (Yildirim et al., 2019) it is possible that street intersections present points of change - i.e. when a residential street such as 'Shipcote terrace' cuts into the 'Durham road' clusters, which lie on a primary road. However, this is not always the case, for example, 'Durham road' contains at least three sensors that lie near intersections with other streets.

The results also show that the sound patterns of the resulting clusters are all effected by the two cluster types found in traffic noise analysis (Zambon et al., 2016, 2017). The first pattern, with two peaks mimicking working hours, dominates the 'High Street' and 'Durham road' clusters - both of which are traffic-heavy roads. The second pattern is present in the 'City centre' and 'Leasuzes park road' (to a lesser degree) clusters - two daytime peaks with increased nightime activity relative to other areas. However, there are other types of patterns that emerge from the analysis when the focus is placed on relative peaks and differences between weekend and weekday noise activity. This results in areas that have similar patterns to clusters derived from 'new forms of data' such as mobile phone and twitter activity profiles. The distribution of points of interest and buildings in the area surrounding each cluster was used to characterize what different functional urban areas they represent.

There were six activity profiles detected, excluding outliers. Clusters such as 'City centre' and 'Claremont road' represent city centre areas and parks that have direct analogs in the literature with identical activity profiles. A 'city centre' or a mixed, high-usage area is present in most studies that use mobile phone (Soto and Frías-Martínez,

2011) and twitter records(Frias-Martinez and Frias-Martinez, 2014). In England, city centres can be focal points of retail, dining and social activities, meaning more activity at all times, which also contributes to the unique sound profile of this cluster type (Dolega and Lord, 2020). Similarly, profiles representing parks like 'Claremont road' are present in the literature (Frias-Martinez et al., 2012). Other clusters such as 'Leasus park road' and 'Baldwin Avenue', identified as comprehensive and residential clusters, show some differences in their activity profiles compared to their analogues from the twitter and mobile phone data studies - e.g Soto and Frías-Martínez (2011); Furno et al. (2015). Although the 'High Street' and 'Durham road' clusters are split, both their activity patterns are heavily dominated by working hour traffic noise, or type one traffic noise patterns. 'High street represents a office or work cluster with few residential places, whereas 'Durham Road' a traffic heavy road with more shops and houses. This speaks to some limitations of using sound sensors for differentiating functional areas based on sound sensors on or near heavy traffic roads, which is consistent with the literature since noise pollution from road traffic dominates noise patterns and is a major concern of urban planners (Morillas et al., 2018). However, it should be noted that this is not the case for sensors near the city centres.

The resulting activity profiles and the two types of dominant sound patterns are also related to urban sound typologies. Sound typologies are actively used in noise pollution and urban soundscape research, however the studies typically focus on acoustic aspects of sounds not just maximum decibel levels (Torija et al., 2013). The activity profiles results in this paper are fewer and less granular than in urban soundscape literature, which is reflective of the limited information used to form the clusters - maximum decibels per hour. However, they are more numerous than the road noise typologies that typically show fewer noise profiles - (Zambon et al., 2016, 2017) and and more similar to most activity data studies that use mobile phone or twitter data. This is due to both the position of the sensors and the methodology used. This positioning of the results in the literature gives the indication that incorporating more aspects of sound data is likely to result in a more granular typology of profiles. Alternatively, the same methodology used in this paper, alongside more acoustic data can be used to help classification of supervised acoustic sound sensors tasks in urban 'soundscape' research (Virtanen et al., 2018).

Furthermore the results as a whole, highlight the unique value sound sensors offer. First, they show the secondary benefits the sound sensors bring - capturing human activity information at a high resolution - in addition to their usage for monitoring noise pollution and anomalous events. Second, the granularity of the data is at the sub-street level and in fact can be controlled using the effective range of the sensor. Third, the data gathers aggregate information (which is not accoustic) and therefore preserves the privacy of the public whose actions are recorded. The results suggest,

that sound sensors could be used as a substitute for human activity data, for example, as a proxy measure of 'urban vitality', similar to how mobile data is used by De Nadai et al. (2016). Furthermore, the effectiveness of this approach will continue to increase given increased adoption (more sensors) and developments in sound sensors (better sensors) as well as machine learning methods (better methods) (Virtanen et al., 2018).

Lastly, another contribution of this paper was the comparison between the baseline methodology, used in other land detection research, and the TDA techniques from the new field of topological data analysis. As mentioned by Arribas-Bel and Tranos (2018); Singleton and Arribas-Bel (2021) it is important to create and adopt methods for the analysis of new forms of data that account for their specific properties. The comparison between TDA approaches carried out in Appendix A, and the comparison between the baseline and the best-performing TDA approach carried out during the analysis, also show the importance of external clustering metrics and the benchmarking of new approaches. The final results suggests that at least in the area measuring differences between time series, TDA techniques show promise but should be further refined in order to produce the best results.

3.7 Conclusion

New forms of data offer a dynamic view of urban life at great detail. Studies using 'new forms of data' - mobile phone data with different methodologies (Soto and Frías-Martínez, 2011; Furno et al., 2017, 2015), app data such as twitter or foursquare data (Frias-Martinez and Frias-Martinez, 2014; Calafiore et al., 2021) - find there are distinct activity profiles of urban areas. This paper shows the value of sound sensors in doing the same. This is achieved through the use of a clustering method to identify areas with distinct functions based on the hourly sound patterns recorded by sensors throughout Newcastle and Gateshead. The results represent six distinct activity profiles - mixed, residential, two types of business, nightlife, leisure are captured by sound sensors. These definitions were ascribed with in accordance with POI data from Openstreetmap. Some of these areas have a direct analog in other new forms of data land use studies, while others are dominated by traffic noise patterns found in the urban sound analysis literature.

In general, this work serves as another example of the potential of 'new forms of data' to enrich urban research. Given more sensors around different street types, the current methodology could be used in order to classify different street types and compare them against their official designations. The results also suggests that sound characteristics could be used in supervised approaches to improve area classification tasks, similar to Hermosilla et al. (2014). Furthermore, the clustering comparison can guide future developments in TDA techniques. For example, it suggests that more

focus should be put into developing more robust methods related to the Lowerstar technique, since it came the closests to the baseline technique.

One limitation of the study is the number of available sensors. The spread of sound sensors is currently not enough for a full coverage approach, similar to the ones where data from phone towers is used. Although the data covers several streets with different functions, the number of sensors is small relative to the studies done in other areas such as tweets and mobile phone records. This impacts the generalisability of the results to larger areas, or ,in fact, to the whole of the Newcastle area. The small number of sensors, further limited the analysis of some interesting results such as the analysis of the 'Corner' and intersection outlier sensors. However, the results still show the promise of combining sound data and our approach to capture different functional areas within a city. The sound patterns of the detected clusters correspond to the activity patterns of clusters derived from tweets or phone towers. It should be noted that the analysis did not explicitly look for any specific cluster patterns previously identified - residential, leisure, city centre or mixed. Rather these separate profiles were the result of the unsupervised cluster analysis.

4

Dynamics and emergence of megaregional structure in US employment data

Abstract: . There are numerous definitions of megaregions, but the common thread among them is that these new geographical units are made up of clusters of either merging or closely interacting distinct urban centres and their surrounding areas and that they represent an emerging economic and policy scale. This paper quantitatively builds a hierarchy of functional spatial units to analyse emergence of megaregional dynamics and structure in United States employment data. It does so by using LODES, a large dataset of 'origin-destination flows' aggregated at the census tract level, and a community detection methodology to build up a hierarchical tree of spatial units at different scales. The validity of these new units is tested against several characteristics - spatial coherence, their population distribution and similarity to established geographical units from the metropolitan to the megaregional level. The results show that there is evidence of the emergence of 6 out of 11 widely popular megaregions. With the exception of the Cascadia megaregion, all detected megaregions in the employment data are in a single state. The rest of the megaregions commonly discussed in the literature break down across state, but not metropolitan boundaries. This suggests that there is limited evidence for the large-scale cross-border economic interactions required to define megaregions.

4.1 Introduction

There are numerous proposals to focus planning, promote cooperation and to create new political structures at a 'megaregional' level in the United States (Friedmann, 2019; Ross et al., 2009; Lang et al., 2020; Nelson, 2017; Hagler, 2009; Ahuallachain, 2012; Ross et al., 2016). A megaregion is a large-scale unit that describes a cluster of urban centres with high levels of economic output, population, as well as infrastructure and cultural integration (Glocker, 2018). Examples of proposed megaregions are the 'Northeastern Megalopolis' encompassing an area from Boston in the North to Washington D.C. in the South, or the 'Piedmont Atlantic' going from Raleigh, North Carolina in the east to Birmingham, Alabama in the west. Each of these and other megaregions, span many official administrative boundaries, hundreds of kilometres of built environment and have millions of citizens.

The proposed advantages of focusing on this megaregional scale include an ability to better harness the benefits of economies of scale and to tackle problems with rapid urbanization. The megaregions are almost always their country's most important economic, political, and cultural centres, and represent the key points linking regional and national economies to global networks (Nelson, 2017). Their constituent parts can share transport infrastructure for people and goods, enabling robust housing markets and the development of offices, science and technology parks and are able to support multiple and varied economies of scale (Florida et al., 2008). Furthermore, the megaregion can act as the appropriate scale for cities to align their policies to more effectively reach common goals on resource depletion, environmental pollution and ecological damage, which are increasingly becoming crossborder problems (Ross et al., 2016).

To achieve these benefits large scale planing and coordination efforts are required across many existing administrative boundaries. At a smaller scale, there have been problems with regional planning for sustainability, although there have also been some successes in terms of environmental planning, growth management and transportation planning (Wheeler, 2015). Planning at the megaregional level is even more challenging due to the exponential growth in the number of interested parties. For example, Glass (2014) talks about a Midwest megaregion consisting of "hundreds of competing governance spaces, all with different legacies, authorities and sociospatial constituencies".

To address this issue researchers have looked to emergent or projected economic, social or infrastructure interactions at a megaregional scale to provide the justification behind a megaregion's spatial extent, and the motivation to build the required collaboration efforts between its constituent members (Sorensen and Labbé, 2020). This has resulted in various operational definitions emerging. The commonality between them is that they define megaregions as built up from established units - counties, metropoli-

tan areas, etc - based on similarities in criteria such as population density, nightime light intensity, overlapping built environment, commuter flows, business interactions (Glocker, 2018).

This paper uses the LODES dataset of economic interactions (Graham et al., 2014) to algorithmically build a hierarchy of nested delineations, in order to analyse the emergence of megaregions in the US. LODES is a comprehensive dataset of residence and workplace tax information derived from the records of 140 million people in the contiguous 48 states. The hierarchy of delineations is created by appliying a community detection approach, the Louvain algorithm (Blondel et al., 2008) to this data, resulting in delineations with non-arbitrary boundaries, which are nevertheless crossed by numerous flows. Each level of delineations within the resulting hierarchy represents a different scale of economic interaction, and gives a different meaning to its units. In order to interpret these, the communities at each level are compared to established geographical units - such as metropolitan areas, states or combined statistical areas and megaregions. Furthermore, the paper explores the spatial patterns of employment in the delineated areas, as it changes at the different levels.

The advantage of this approach is in the flexibility it affords. It does not require setting up explicit thresholds, specifying urban models or using other additional data, in contrast to other approaches (Kockelman et al., 2019). The units at each levels of the hierarchy are built up from census tracts, based only on the relative strength of the connections between them, compared to other existing connections. Similarly, the levels of the hierarchy are not pre-defined and new levels emerge based on the relative connections at lower levels. Therefore, this approach can detect emergent urban phenomena at various scales, including the megaregional, however no scale requirements are explicitly specified.

The rest of the paper is structured as follows. The next section provides a literature review of the concept of a megaregion and its operalisations. The 'Methodology' section describes the data and community detection approach, the comparisons carried out, as well as the way in which the spatial concentration of economic interactions is explored. Finally, the results section describe the resulting hierarchy of delineations, while the conclusion & discussion section place them within the context of the literature.

4.2 Literature review

4.2.1 Megaregions

The conceptualizations of large scale multi-city regions in America started in 20th century, however these were mostly ignored by practicing urban planners and policy

makers, until they gained more traction in the early 21st century with the release of the millennial census (Lang et al., 2020). Geddes (1915) presented one of the first analysis of merging ubran areas. He used the term 'conurbation' to describe the cluster of Newark and Jersey City and New York's Manhattan and Brooklyn, integrated via supply chains over 100 years ago. The French geographer Jean Gottmann first described a long-extended metropolis in the entire Northeastern US. He argued that the large morphological growth of the cities between Boston and Washington led to their suburbs clashing into one another, and resulted in a unique cluster of metropolitan areas which extended beyond traditional and political borders (Gottmann, 1957). There were also calls to begin promoting transportation planning, such as intercity passenger rail service in the Northeast Megalopolis, focused on this new scale of geography (Lang et al., 2020). At the beginning of the 21 centrury, with the coming of the new census it was discovered that many US metropolitan areas as measured by the census had grown to such a large scale that they were running into one another. Combined statistical areas (CBA) were introduced as a way of better understanding these areas and the underlying trends that drove their creation.

There are 388 metropolitan statistical areas, 541 micropolitan statistical areas that make up the 929 core-based statistical areas (CBSAs) and 169 Combined metropolitan areas (CBAs) in the US census (Ross et al., 2016). CBSAs represent urban areas with a metropolitan centre core and their surrounding commuter flows, subject to population restrictions, while CBAs represent merging core-based statistical areas. These new units, along with the underlying phenomenons, led to more interest from researchers as well as urban planners, in the concept of megaregions and finding other megalopolises in the US. One of the most influential sets of results came from the Regional Plan Association, where megaregions are defined as 'interrelated population and employment centers or MSAs that share common transportation networks, cultures, and environmental features' (Hagler, 2009). The identified megaregions are shown in Figure 4.1.

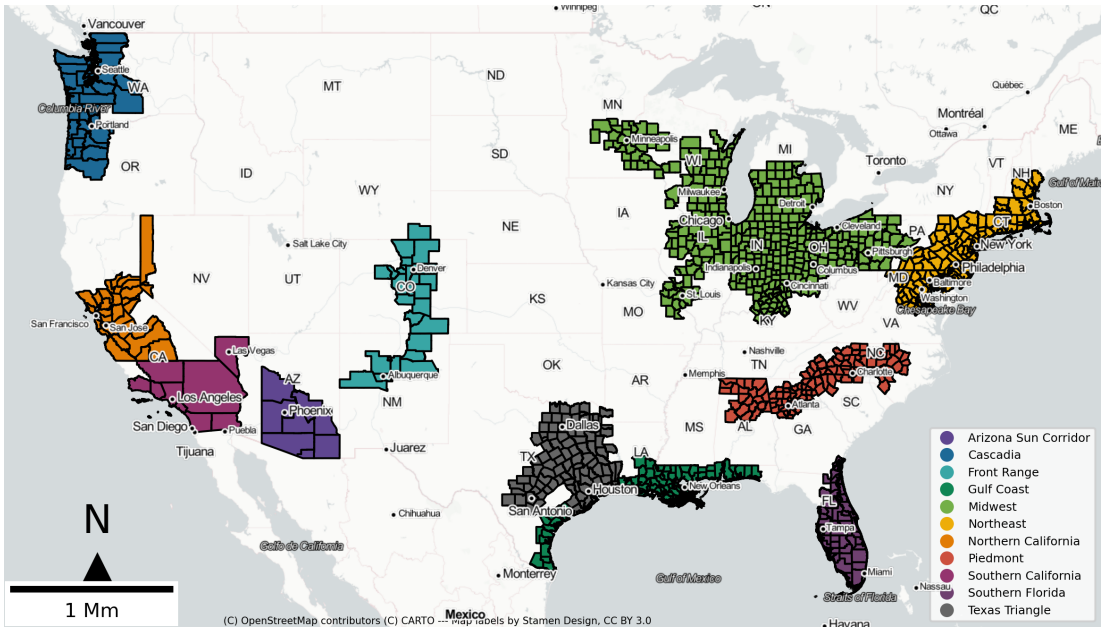


Figure 4.1: America2050 megaregions

In the following years many quantitative definitions, delineations, as well as conceptualizations and critiques of super-urban areas emerged, building on the definitions made by the America2050 project. However, a common theme across definitions is that they represent large scale areas which encompass several urban centres along with their surroundings, subject to qualifying properties such as population, build density, infrastructure and cultural integration (Glocker, 2018). These definitions result in conceptual marco-scale structures which capture the idea that large cities are merging together, spanning hundreds of kilometres of built environment and millions of citizens.

Some of the driving factors behind these changes are urban population growth, frictionless capital flows, advances in technology, and changes in commuting behaviours. Urban population is projected to continue to grow, however the spatial distribution of this increase is not uniform (Batty, 2018). Nelson (2017) predicts that ten megaregions specifically, 'will account for about 80% of the US's total growth in population' and 84% of the regional GDP growth. This concentrated growth has led to expansions of city boundaries, and in some cases has meant that urban areas have started merging into each other morphologically (Harrison and Hoyler, 2015b). Furthermore, it has contributed to more complex commuting patterns (Rae, 2015).

Similarly, increases in digitisation and advances in information technology allows for services and knowledge to be accessed without the need for mobility, increasing the scale of potential economic transactions between cities (Georg et al., 2018). A megaregion is defined as a polycentric cluster of urban areas which emerges within this network, characterised by strong economic, cultural and infrastructural ties between its constituent urban centres (Hall and Pain, 2006; Feng et al., 2018; Glocker, 2018). Polycentric areas are those in which economic activity and employment is con-

centrated in several centres, operating in a complex network of interactions (Möck and Küpper, 2020). This is in contrast to monocentric urban areas where most activity is concentrated on a single commanding place, e.g. a Central Business District (CBD) , which dominates local labor markets.

The importance of the megaregions comes from proposals that better coordination within this complex structure can lead to better planning to tackle economic, transportation, global competitiveness and sustainability challenges (Harrison and Hoyler, 2015b). This can take the form of more effectively sharing transport infrastructure, science and technology parks and supporting the development of economies of scale (Sassen, 2010). These benefits would be, more pronounced in the case of the smaller cities in the megaregions, due to the fact that its easier for residents and businesses to access services that would require much larger population densities and workforce specializations - so called 'borrowed-size' effects (Marull et al., 2013). In fact, Ahuallachain (2012) argues that cities already function within sets of related megaregions, and policies aiming to improve manufacturing and production should focus at the megaregional level. The megaregion can also act as the scale at which cities align their policies to tackle sustainability and environmental issues since they are increasingly becoming trans-border problems (Ross et al., 2016). There is also some precedent for action at this scale – for example with regard to environmental protection in the Great Lakes area and the Regional Greenhouse Gas Initiative (RGGI) consortium establishing a market in greenhouse gas emissions in the northeastern US and adjoining Canadian provinces (Sorensen and Labbé, 2020).

4.2.2 Delineating Megaregions

Since megaregions describe a new geographic scale of social and economic interaction that span over several administrative boundaries, it is important to know who the potential partners and members could be and to engage with them at the right scale. However,researchers disagree on many megaregional boundaries. For example, Ross et al. (2009) proposes one combined California megaregion, while the America 2050 project by the Regional Plan Association (Hagler, 2009), identifies two separate Northern and Southern California mega regions. Ross also proposes a Central Plains mega region that does not exist in the America 2050 typology, while the latter shows Front Range and Gulf Coast mega regions not found by Ross. The two teams have sharply different boundaries for Piedmont and Midwestern mega regions. Lang et al. (2020) present a version of the America 2050 map with most of the mega regions divided into sub-regions, for example 'Twin Cities', 'Chicago', 'Michigan Corridor', 'Steel Corridor' and 'Ohio Valley' portions of the Great Lakes Megaregion. Given these problems, evidence of emergent interactions at the megaregional scale, is still an important fac-

tor for the megareginoal research agenda (Harrison and Hoyler, 2015a) and provides important motivation to build collaboration efforts between potential partners.

To address these concerns researchers have developed different delineation methodologies. The commonality across these is that they assume that megaregions are not simply a large scale city, but a agglomeration of cities and their lower density hinterlands, spatially and functionally linked through environmental, economic, and infrastructure interactions (Ross et al., 2009). It should be noted that there is overlap in spatial extent of various definitions (Glocker, 2018), suggesting that they capture different aspects of the same underlying phenomenon.

One group of definitions focuses on the characteristics of the areas within the potential megaregion such as morphology, population size, travel time and others, e.g - (Georg et al., 2018; Hagler, 2009; Glocker, 2018; Habitat, 2013; Ross et al., 2009). The motivation behind this approach is that if the labour markets and business interactions of distinct urban centres become so integrated that they overlap, the build environment between them would also overlap. Thus contiguous development at a large scale results from an area functioning as a megaregion. However, many of the morphological studies have the disadvantage that they completely discard interactions. Different land use systems and regulations at the local level lead to very different ways in which the built-up areas of cities relate to functional reality, while increasingly long and complex commuting and economic patterns further complicate that relationship (Wu et al., 2019).

These issues are addressed by explicitly using functional approaches to delineation. They define the megaregions as made up of some type of functional units depending on the choice of relationship. In most cases the focus is not on direct relationships between all the constituent cities, since the distances at the megareigonal scale are too big, but on continuous integration and strong chain effects. Examples of such relationships are commuter flows or business interactions (Lang et al., 2020; Nelson and Rae, 2016; Batten, 1995; van Oort et al., 2010; Geddes, 1915). The methodologies require treating the origin-destination flows as a (usually non-spatial) network where nodes are some geographical units such as counties, and the links between them represent functional relationships such as commuter flows. This approach can result in spatial outliers and depending on the geographical units used it can add a lot of excess landmass such as farmland or hinterland to the megaregional definitions.

The third approach aims to reconcile both and be a compromise between functional integration and morphological features. For example, (Ross et al., 2009) combine both morphological dependencies and functional relationships (transportation links) to better capture the extent of megaregions. More recently, advances in data analysis techniques enabled the use of secondary data (Arribas-Bel, 2014a), which have led to even more complex combined approaches, for example, based on satellite data and

taxi cab rides (Wei et al., 2020). A disadvantage of this combined approach is the data availability limitations, since large amounts of diverse data are needed, as well as methodological problems of how to combine it (Glocker, 2018). The data validity in the case of the new forms of data can also be a problem.

Most of the approaches in these three groups require setting up explicit thresholds or parameters and expand already predefined large-scale units. For example, most of them use the already pre-defined metropolitan and micropolitan areas from the census and add the surrounding areas, subject to them having a stronger connection than a pre-defined threshold. A disadvantage of this approach is that different thresholds lead to different results, and the methods are biased towards already defined areas (Kockelman et al., 2019). One way to mitigate some of these problems is to adopt 'bottom-up' analysis techniques from network analysis such as community detection. Examples of applications of this methodology in the US are Nelson and Rae (2016), in China - Wu et al. (2019), in Central Europe - Khiali-Miab et al. (2019). The advantages of this family of approaches is that they they build up the relevant units quantitatively form the data, and require less structure imposed beforehand. For example, the polycentric model that megaregions could follow does not need to be defined prior to the analysis. An example of this method carried out in the US is Nelson and Rae (2016). It uses commuter flows along with a community detection methodology to delineate the US into self-contained areas. The results of the analysis are show in Figure 4.2. There are 57 detected areas, centred around large cities. Lang et al. (2020) argues that the detected communities more closely resemble his megalopolitan area concept, rater than the megaregional concept as, for example, defined by Hagler (2009). And, in fact it can be seen that the scale of the detected communities resembles extended CSAs and that multiple areas are needed to cover a single megaregion defined in America 2050.

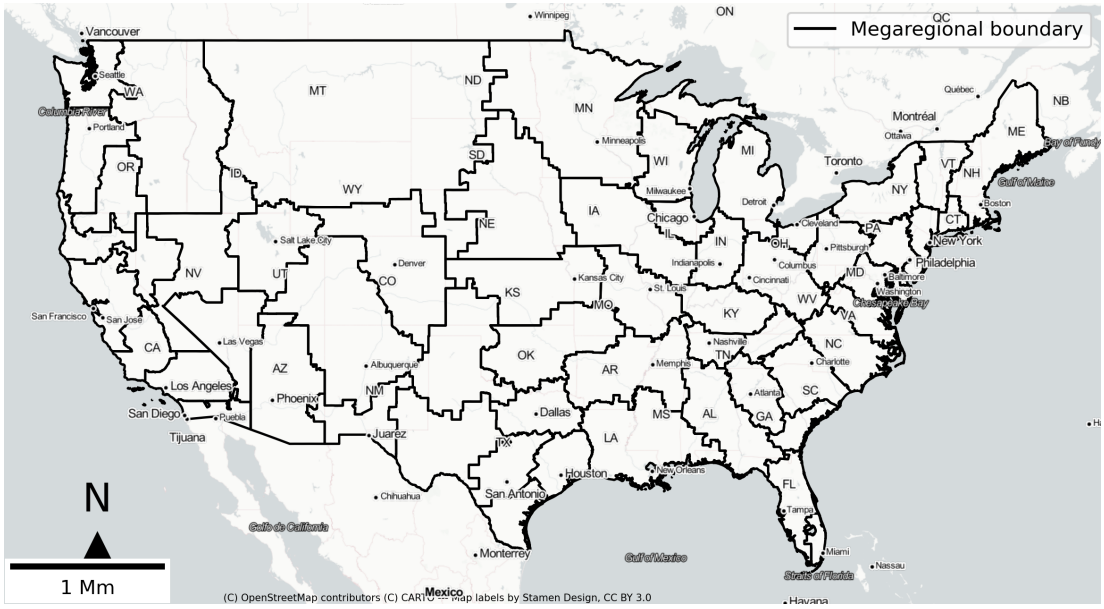


Figure 4.2: Nelson and Rae (2016) megaregions

This paper builds on the results of Nelson and Rae (2016) by addressing some of the issues raised by the authors and other researchers (Glocker, 2018) and by adding a hierarchical aspect to the analysis. First, the LODES dataset used in this study is not exactly commuter data and captures more business interactions than the commuter data used by Nelson and Rae (2016). It is larger in terms of volume than the commuting data used by Nelson and Rae (2016) and has a more complex structure - 140 million people vs 100 million and 27 million links vs 4 million. Additionally, the data is newer - it is collected in 2017 as opposed to 2010 and is not affected as much by the recession in the late 2000s. Second, the exact community detection method used in this paper, the Louvain algorithm (Blondel et al., 2008), produces a hierarchy of spatial units at different levels for the entire US, not just a flat delineation. It is an algorithmic combination of on the one hand the bottom up approach - by delineating small scale regions based only on commuter flows, and on the other the top down approach - by combining these regions based on interactions. This way it can produce insights into large-scale phenomena present in the data. The hierarchy can also be compared with other spatial units to show that at each level the delineated community actually capture meaningful spatial entities, which reflect phenomena from the literature.

4.3 Methodology

4.3.1 Data

Available interaction data

The main type of data used is a variant of origin-destination data. These datasets represent a functional relationship between two points - the origin and destination. Common types of these relationships are commuter flows, interactions between businesses or even calls between people. In addition to an origin-destination pair, these data also record a measure of the strength or 'weight' between the origin-destination pairs - for example, number of commuters. Generally, this data is analysed first by converting it to a (usually non-spatial) network where nodes represent origins and destinations and the flows are edges between them. This way tools from network analysis can be applied to the data to gain insights.

There are two high-quality large-scale datasets available for the United States, which directly capture economic interactions spatially - American Community Survey (ACS) and Longitudinal Employer-Household Dynamics Origin-Destination Employment Statistics (LODES). The differences between the two is in the collection procedure, scope, definitions and coverage (Graham et al., 2014). ACS data comes from a yearly nationwide survey of 2.5 million addresses and specifically asks for the address where a person worked last week. LODES data is collected from various administrative data and surveys and is available at the census block level. It has 'origin-destination flows' which might not correspond to an actual physical trip, but to registered home and work addresses, inferred from sources such as tax forms. This distinction is especially important for industries like catering and construction where the actual work location changes often. It is also important for businesses that have multiple offices, since only the main one is considered as the work location. The LODES data also has the disadvantage that the scheme is opt in, therefore states volunteer their data and not all data is available for all states at the same time.

This paper uses the LODES data since it can capture more complex interactions at a larger scale. This is because recorded links do not necessarily correspond to a physical trip, but they represent an employment or economic interaction. This way the problem of whether a person works from home or on location, or at a different location that the registered or main one, becomes less important since the focus is on the connection itself. A person living in D.C. and working from home for a firm in New York will be captured by the data and will be an important contribution towards the existence of the Norteast megaregion. This is reflected in the fact that the data has more links than the ACS data and potentially more complex patterns.

Preprocessing

The data used for the megaregions community detection is downloaded from Census Bureau (Graham et al., 2014). It covers the 48 mainland US states and DC, excluding Hawaii and Alaska. The data records 2017 residence and workplace information for almost 140 million (139053815) people with approximately 120 million pairs of home-work locations. Tables B.1 and B.2 in Appendix B show a breakdown of this data by state and by the interactions between pairs of tracts that are within the same state and between different states. The average percentage of interstate flows is around 13% with 27 out of the 50 states having at least 10%. This is further evidence that the LODES data captures more economic patterns than just commuting data, and that its well suited when analyzing large scale economic units such as megaregions.

Before analysing the data only two preprocessing steps are taken. It should be noted that no data was discarded with these preprocessing steps, neither because of large distances, nor based on small commuter sizes. First, the data is aggregated to the census track level. This was deemed necessary since at the block level over 70% of links only recorded a single person. Furthermore, census tracts are a more robust starting unit of analysis, since in general they are defined to be stable across time and have an average population of 4000 people. This is in contrast to the block and group level, where there is much more variability in physical size and population. Additionally, it significantly reduced the required computational time and resources.

The second and final preprocessing step is to transform the resulting aggregated datasets into an undirected weighted graph. Each census track is considered a node and each 'origin-destination flow' between tracks - an 'economic' link. The fact that the graph is undirected means that there is no difference which tract is the home location and which is the work location. The direction of the flows is not important for the detection of megaregions, what matters the most is the volume and pattern of interactions. To this end incoming and outgoing links between pairs of tracts (or counties) are added together to form a single undirected link. The weight of each link is therefore equal to the number of people that use the 'origin' node as a home place and the 'destination' node as a workplace, plus the number of people that have the 'origin' tract registered as a workplace and the 'destination' as home. This step also addresses the problem mentioned by Glocker (2018) that switching the home and work locations can change the results of the algorithm.

Processed data

Figure 4.3 shows a map of a 1,000,000 random sample the processed data for the lower 48 states. There are some general patterns that emerge just with visual inspection. First, some of the previously mentioned megaregions are visible. For example there is

an outline of the Norheastern Metropolis corridor extending from Boston in the north to Washington D.C. in the south. The outline of the Texas triangle - the area between Houston, San Antonio and Dalas - is also visible. Second, the further west you go from the Mississippi river the less economic links become visible, until the West Coast is reached. This reflects the population density of the United States. Third, it is hard to specifically delineate the exact boundaries of each region just by looking at the map. For example, the entirety of California looks completely connected as one big megaregion.

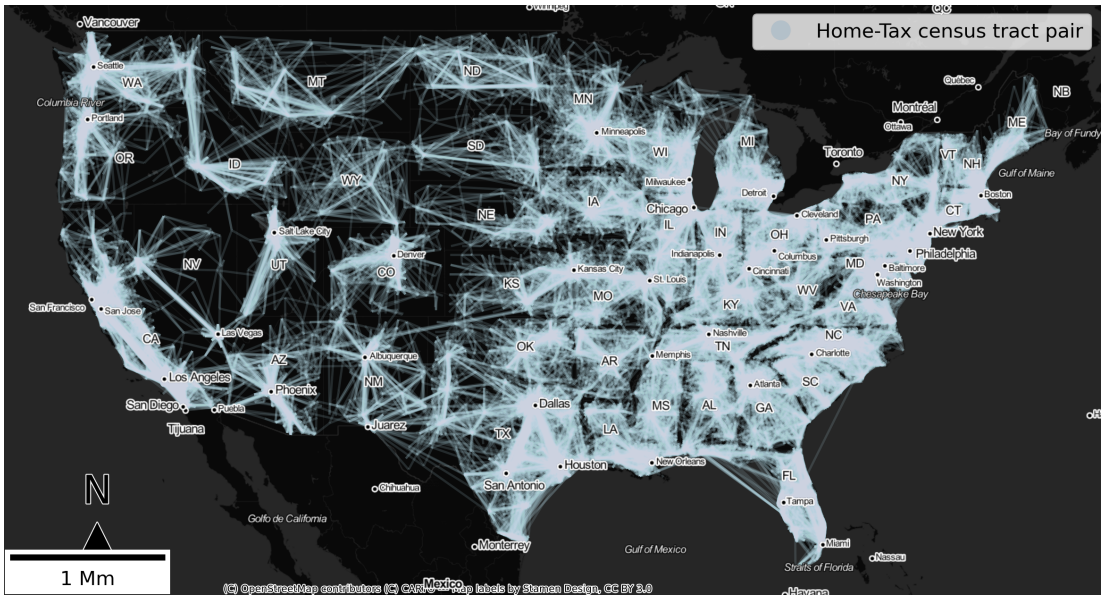


Figure 4.3: LODES origin-destination flows

4.3.2 Network analysis

Treating a set of relationships such as the 'origin-destination flows' between spatial units, calls between people, or social media interactions as an abstract graph makes it possible to draw on the rich theory and tools of network analysis. The main idea is that nodes can be grouped in such a way, where nodes within a group interact with each other more than with nodes outside of the group. In network analysis community detection is the problem of splitting the nodes of the graph in several 'similar' communities based on the connections, in order to gain insights into the networks dynamics.

Community detection has been used in several works in order to delineate regions or areas. Examples of applications of this methodology in the US are Nelson and Rae (2016), in China - Wu et al. (2019), in Central Europe - (Khiali-Miab et al., 2019) and in Scotland - (Hamilton and Rae, 2018). Glocker (2018) cites some disadvantages of focusing solely on functional delineations such as our approach. One is related to the quality and reliability of the available data. Often travel surveys and commuter surveys are not 100 percent accurate. The data does not cover all types of work - project work,

or home locations. The boundaries are effected by whether the network of commutes is constructed using places of work or home locations. Lastly, transportation links play an outsized role in the analysis since they directly influence commuter flows and employment patterns.

These criticisms are addressed by the specific choice of methodology and data. The LODES dataset covers the records of more than 140 million people in private sector work, therefore problems of representability should have limited impact. Furthermore, the focus is on megaregions where the density of the population is the highest. The problem with transportation links is addressed by the methodology. It is possible for two cities, say New York and Philadelphia to be in the same community even if potentially they dont have any links between them, provided there are intermediary nodes with sufficient links.

Furthermore, this paper uses a different community detection algorithm and modifies its output to produce a hierarchy of geographical units at various levels. This hierarchy makes it possible to order the results by integration based on the level they appear in. It also makes it easier to validate the results of the analysis, by comparing the resulting delineations at various scales to other existing geographical units. Lastly, the connections between tracts at different levels of the hierarchy have different interpretations.

4.3.3 Louvain algorithm

The algorithm used in this paper is Louvain (Blondel et al., 2008). In several community detection comparison papers it is found to produce state of the art results (Rahiminejad et al., 2019; Lancichinetti and Fortunato, 2009) . It finds structure in the data based on a notion of similarity between census tracts called modularity (Newman, 2006). Modularity is a measure of how much stronger the observed links within a community are than it would be expected if the network was connected randomly. Formally, it is defined as the difference between two fractions. The first is the fraction of flows between a pair of nodes within the potential community to the total number of flows in the graph. The second one is the expected fraction of flows between a pair of nodes within the potential community to the total number of flows in the graph, if the flows between census tracts were random and did not follow any specific pattern. It ranges from -1 and 1 and a score of closer to one indicates a graph that has strongly expressed communities, otherwise the network structure is close to random.

The Louvain algorithm works in two steps. In the first, each of the nodes is paired up in the same community with immediate neighbours based on how much the assignment increases modularity. This continues until no further assignments increase the total modularity. The second step merges into a single node all of the nodes in each

community. Then the first step is repeated on this new 'induced graph' and after it the second step. The whole process is repeated until there are no more modularity gains. This process is illustrated in Figure 4.4.

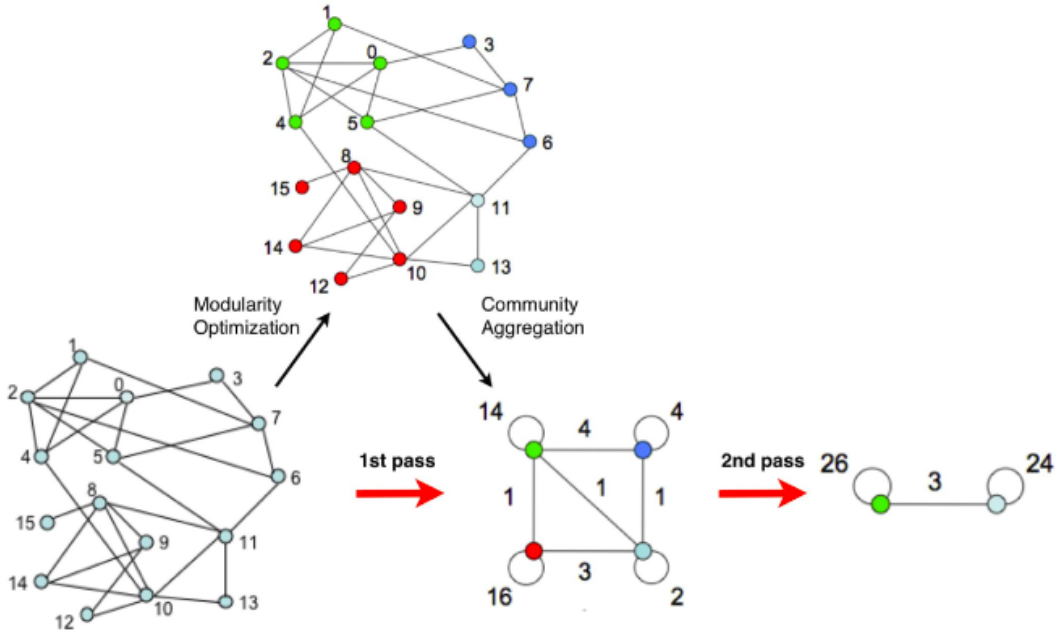


Figure 4.4: Summary of the steps in the Louvain algorithm (Blondel et al., 2008)

In the case of the processed LODES data this means that as a first step the method delineates, what Nelson (2020) calls 'conchronations' - spatial units made up sets of census tracts, that have a non-arbitrary boundary based on employment interactions. The boundary is well-defined but is not strict, since the different conchronations still interact with each other. Next, it repeats the same process, however for this step it uses the conchronations as a base geographical unit, rather than the census tracts. This results in a hierarchy that shows the interaction between sets of census tracts at various scales.

4.3.4 Delineation analysis

The hierarchy is analysed at the various levels to see the extent to which the delineations correspond to other geographical units and to explore whether there is megaregional integration and when it starts. To do this the detected areas at the different scales are compared to the government defined statistical areas such as CBSAs and CBAs, state boundaries, as well as megaregions identified in Nelson and Rae (2016); Hagler (2009). The CBSAs and CBAs were chosen since the are widely used and studied in practice. There are also two megaregional comparisons carried out. First, with the megaregions defined in the America2050 project (Hagler, 2009), which are chosen due to their wide usage and availability. There is academic (Stich and Webb,

2019; Ross et al., 2009), federal [1] state and local level (Oden and Sciara, 2020; Bellisario et al., 2016) research into them. Furthermore, most case studies that exist focus on the regions specifically defined in the America2050 project. For example, Ross et al. (2009) considers the whole of California as one big megaregion, however the Bay Area economic council splits the state into northern and southern megaregions (Bellisario et al., 2016) similar to the America 2050 project. Lastly, the shapefiles are readily available online. The second comparison is with the results from Nelson and Rae (2016), since they are the closest to this study in terms of methodology and data.

Two types of comparisons are carried out between the resulting set of delineations from the hierarchy and the other areas - one based on rand score and one based on spatial intersections. The latter treats the two sets of delineations which are being compared as polygons and measures percentages of spatial overlap. The former, is a standard measure of how similar two sets of assignments are and does not take spatial information into account. The rand score measures the similarity in census tract grouping, between the delineations and the other areas. It has a range from -1 to +1, with higher values indicating more agreement between assignments.

Lastly, for each level of the hierarchy we analyse the spatial distribution of employment within the detected community at every level. This is done through the local Moran's statistic (Anselin, 1996), which constructs local spatial statistics to measure "significant spatial autocorrelation for each location." The method allows us to distinguish hotspot areas within each delineated community with high employment surrounded by high employment values(HH), low value surrounded by low values (LL), areas with low values surrounded by high values (LH) and high values surrounded by low values(HL).

Similarly to Arribas-Bel and Sanz-Gracia (2014), we focus on the distribution of employment based on employment centres : "An employment center is a contiguous set of spatial units within an urban region, conditional on each spatial unit exhibiting a spatial concentration of high employment density that is significant at the $p < 0.10$ level". Where spatial units that classify as high employment are those of high employment surrounded by areas of high employment, and areas of high employment surrounded by areas of low employment. And contiguity is set by the queen criteria - two units are considered spatially contiguous if they have a point in common. The spatial units in each case are the units one level below in the detected hierarchy. For example, the spatial units in the first level will be census tracts. The employment level for each unit is derived by aggregating the number of employees within it, based on the units' spatial extent. This classification of the area within communities allows us to analyse how spatially centralised or decentralised employment is at different scales.

The results from this spatial analysis allows for further verification of results and

[1] https://www.fhwa.dot.gov/planning/megaregions/what_are/

comparisons with the literature. A distribution of the economic activity similar to those identified in the literature provides a further signal that the delineations represent aspects of underlying economic processes and are not artefacts of the data or methodology. Furthermore, if some of the delineated units created resemble other units, such as metropolitan areas, the comparison also highlights the effects of boundary delineation choices on final results, which is an active area of research (Möck and Küpper, 2020). Lastly, if the delineated areas represent megaregions, the analysis of the spatial distribution at multiple scales provides some insights into the polycentric structure of megaregions.

4.4 Results

The algorithm converged after three iterations, producing a three-level hierarchy. There are no spatial outliers, even though no spatial data was used in the detection procedure. The communities detected in Figure 4.7 are built up from the communities at the second level shown in Figure 4.6, which in turn are made up of the communities in Figure 4.5, which represent collections of census tracts. Level one has a modularity score of 0.673, level two - 0.905, while the last level has 0.949. It should be noted that there exist numerous flows between pairs delineation which are not shown on the map. For example in the third level, there are flows between pairs of delineated communities shown in the map which represent millions of people.

4.4.1 Level one communities

The first level of community detection, shown in Figure 4.5, corresponds to areas built up from census tracts based on the flows of the data. The census tracts are grouped together in self contained communities that have more connections within them than is expected based on the definition of modularity.

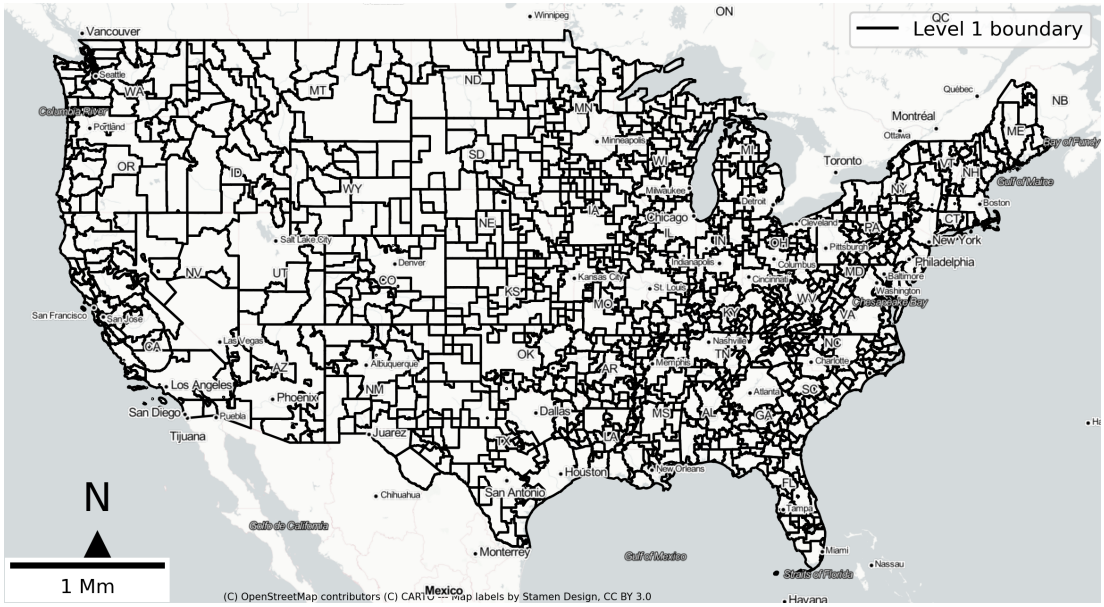


Figure 4.5: Level 1 communities based on LODES aggregated by census tract

The detected communities roughly correspond to a mixture of metropolitan/micropolitan and combined statistical areas in terms of size. There are 1349 detected communities and 929 official core statistical areas defined in the census respectively. For the most part the algorithm merges census tracts into counties and merges the counties together, however there are exceptions. Census tracts are grouped together into their official county boundaries in 75% of cases. In the rest of the cases, official county boundaries are split between detected communities. For example, only 10% contain numerous communities, where the second largest community covers at least 30% of the of the county.

The adjusted rand score between the level one communities and CBSAs is 0.86. This value, close to +1, suggests that the majority of census tracts in our delineations are assigned in a similar way to official metropolitan and micropolitan areas. If we consider only detected areas that geographically intersect with defined CBSAs, 70 % of core statistical areas have at least 80% of their geographical area falling within a single community, further suggesting a close resemblance to CBSAs. Some of the differences are related to places like Salt Lake City, whose community resembles more its combined statistical area. On the other hand some densely populated CSAS like New York-Newark-New Jersey are separated. Other differences can be explained by the fact that there are no population restrictions placed on the data and that every tract has to be part of a community. This means that low density places in states like Montana and Texas, which do not have enough population to qualify for CSAs also form communities.

Table 4.1 shows a summary of the areas with more than one centre, detected after applying the LISA algorithm to each community at this level. There are 341 commu-

nities which have more than one detected centre. They differ in the amount of workers they account for, however together they account for over 120 million or 0.87 % of the flows. Which suggests that employment within the delineated communities at this level is decentralised. The rest of the areas are split into two groups - a group with no centres and one with a single centre. There are 297 communities with no detected cluster, which have relatively sparse population with a mean of 10,000 population, where 75% have a population less than 14,000. The last group consists of 710 communities with a single cluster, again these are also sparse in terms of population. The mean of the group is twenty thousand with 75% of the data having less than 24,000 people.

Table 4.1: Summary statistics for level one communities with more than one detected centre

	Census tracts	**Total Employment**	**Centres**	**Census tracts in centres**	**Proportion of employment within centres**
mean	177.96	353812.32	5.41	16.98	0.33
std	396.42	820655.22	7.97	31.91	0.16
min	3.00	2470.00	2.00	2.00	0.03
25%	28.25	44244.00	2.00	4.00	0.22
50%	58.50	103560.00	3.00	7.00	0.31
75%	151.75	289184.50	5.00	15.00	0.41
max	4218.00	8319571.00	79.00	302.00	0.92

4.4.2 Level two communities

Figure 4.6 shows the second level of the hierarchy, which represents mergers of the Level 1 communities. The resulting areas are in general much larger than core based statistical areas and in fact, most are made up of more than one combined statistical area. At this level only 2% of officially defined counties encompass two different communities, which shows that the communities at this level are almost always a combination of official counties. There are 72 detected communities as opposed to the 48 states, however there are clear examples of detected states such as Wyoming or Colorado. The rand score between the states and the level two communities is 0.72 also suggesting a close similarity. In terms of spatial overlap, there are 29 states that have atleast 80% of their area covered by a single delineated community. The rest of the detected communities correspond to a grouping of many states in a cluster (Maryland and Virginia), while the others are places like Florida where the communities are centred around major cities.

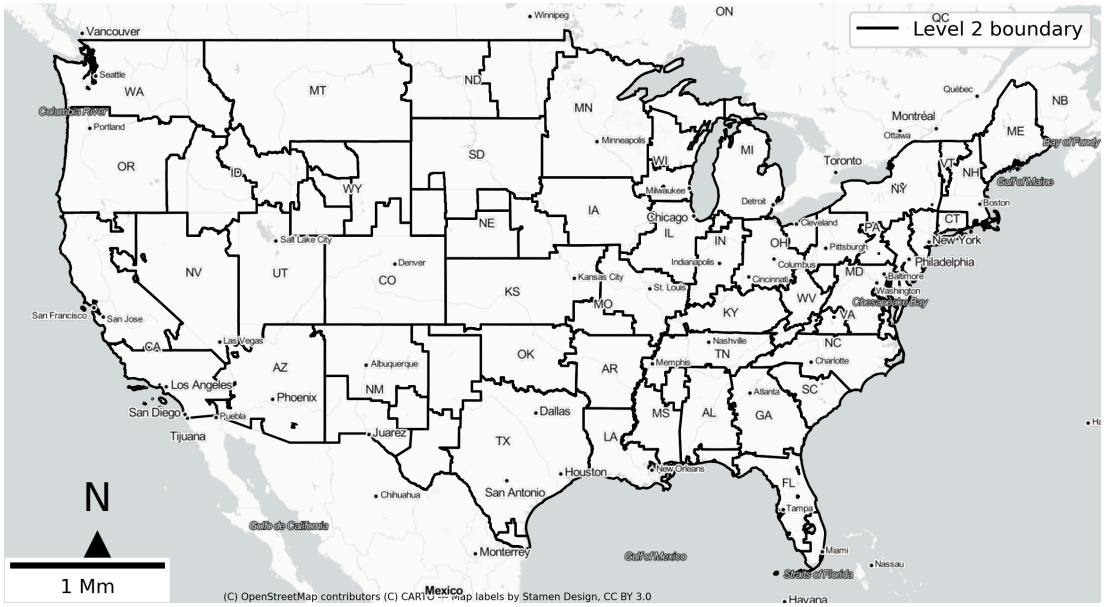


Figure 4.6: Level 2 communities based on LODES aggregated by census tract

Table 4.2 shows the results of running the LISA analysis on the second level communities. Here a centre is set of contiguous level 1 communities with high employment. There are 49 communities with a single centre, which account for 81 % of the flows, 14 with no centres and 5% of the captured flows and 9 with multiple centres which account for 14% of flows. In contrast to the distribution of the firs level, most people reside in places with a single centre, suggesting that as the scale grows employment is concentrated to central areas within the delineation.

Table 4.2: Summary statistics for level two community centres

	Number of level one communities	**Total Employment**	**Centres**	**Census tracts in centres**	**Proportion of employment within centres**
mean	18.71	1930493.49	0.94	1.56	0.37
std	14.41	2319694.73	0.60	1.24	0.30
min	2.00	15443.00	0.00	0.00	0.00
25%	6.75	329664.25	1.00	1.00	0.09
50%	15.00	1214232.00	1.00	1.00	0.31
75%	28.25	2641038.25	1.00	2.00	0.66
max	56.00	10402941.00	3.00	5.00	0.95

4.4.3 Level three communities

Figure 4.7, shows the last level of the community detection with 34 detected communities. The communities group multiple states together and mostly respect state boundaries. Exceptions are Idaho, Illinois, West Virginia, New York and California, which have multiple communities crossing their borders. Population density also plays a role in the delineation. The highest number of communities appears in the north east, while the largest clusters in terms of area appear in the less populous mountain range. There is a different pattern in the most populous states - in California there are two

detected communities, and Texas and Florida form one community each. The detected regions have within them as few as 500 thousand employees to well over 11 million, with most being between 1 and 6 million.

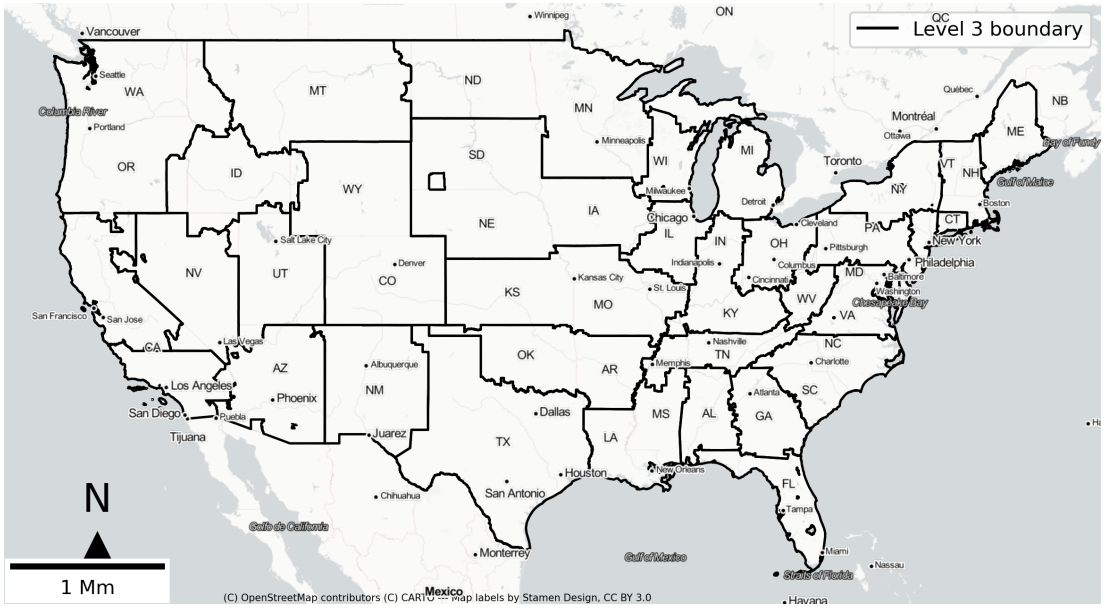


Figure 4.7: US megaregions based on LODES aggregated by census tract

LISA analysis is not carried out at this level, since it will not yield useful results due to the fact that most communities are created by merging zero, two, three, or at most five spatial neighbours. This leads to a low number of spatial connections and non-informative results.

4.4.4 America2050 comparison

The communities from the last hierarchical level are compared to the established megaregions from the America2050 project, shown in Figure 4.1. The rand score between the delineations and the megaregions is only 0.2, in contrast the score between the delineations and the state boundaries is .35. This result suggests that the delineations resemble states boundaries more than megaregional boundaries. However, there is a lot of regional variation in the results.

Table 4.3 shows what percentage of tracts from each America2050 megaregion fall outisde the largest intersecting community overall. A low percentage means that most of the census tracts of an America 2050 megaregion fall within a single community and that there exist strong connections, as measured by modularity, between a megaregions' constituent census tracts. Thus, a low percentage suggests evidence of emergent economic interactions at a megaregional scale. On the other hand, high percentages show a split of the megaregion between several communities and stronger local preference - there are groups of census tracts that are not strongly connected to

the megaregion, again as measured by modularity. The higher the value the more of these groups there are.

Over 90% of all of the tracts from the Arizona Sun Corridor, Cascadia, Northern California, Southern California, Southern Florida and Texas Triangle fall within a single community at the last super-state level of the hierarchy. Of these six, Arizona Sun Corridor, Northern California, Southern California, and Texas Triangle appear at the second hierarchical level, suggesting that they are more integrated than Cascadia and Southern Florida. Furthermore, all of them, apart from Cascadia are megaregions within a single state. The rest - Front Range, Gulf Coast, Midwest, Northeast and Piedmont - are split between a number of communities at all levels.

Table 4.3: Percentage of tracts outside of largest intersecting community in each megaregion

Megaregion	Percentage of tracts outside of largest intersecting community
Arizona Sun Corridor	0.00 %
Cascadia	0.00 %
Front Range	0.21 %
Gulf Coast	0.40 %
Midwest	0.78 %
Northeast	0.58 %
Northern California	0.04 %
Piedmont	0.45 %
Southern California	0.09 %
Southern Florida	0.01 %
Texas Triangle	0.00 %

Additionally, we compare our delineations to Nelson and Rae (2016). There is a .62 rand score similarity between our delineations and Nelson and Rae (2016), shown in Figure 4.2. However, the rand score rises to .77 when compared with our second level communities, which shows that our final communities are of a larger scale. Some differences are the emergence of communities that almost fully correspond to the Northern California, Southern California, Arizona Sun Coridor and Texas triangle megaregions, which are not present in the Nelson and Rae (2016) delineations. Other minor differences come from sparse density areas, where there are multiple possible assignments due to the low number of flows.

4.5 Conclusion & Discussion

4.5.1 Large-scale economic interactions in the US

Overall, the analysis carried out in this paper show that large-scale economic connections break down across state boundaries, whereas cross-state boundary interac-

tion is more common at lower-scales. In the emergent communities, at the first level 75% of census tracts were combined into their official counties, whereas at the second level that percentage was 98%. Afterwards, the counties were not necessarily combined in their respective core-based statistical areas. This suggests that as the scale increases official state boundaries played a more prominent role. Nevertheless, there are multiple examples of pairs of states which have strong economic interactions and were merged at the second or third level in the hierarchy. However, there is evidence of functional integration at the megaregional level in only six out of the eleven proposed megaregions by Hagler (2009). Arizona Sun Corridor, Texas Triangle, Southern Florida, Southern California and Northern California are the single state megaregions that have corresponding detected communities. Of these, Arizona Sun Corridor, Texas Triangle, Southern California and Northern California appear earlier in the hierarchy suggesting that they are better integrated than Southern Florida. Cascadia was the only detected megaregion that crosses state boundaries. Even in the case of Cascadia, there are stronger interstate preferances than megaregional, since communities that correspond to the states appear first.

This does not mean that there are few interactions between the constituent parts of the other proposed America2050 regions - Piedmont, Midwest, Northeast and Gulf Coast. For example, the Midwest (or Great Lakes) megaregion is split between 9 communities that roughly correspond to states. In over one fifth of census tracts that make up the megaregion definition, 10% of the links point to different communities. Similarly, the America2050 Northeast megaregion is split between 5 communities and a quarter of the census tracts that define it, have more than 10% of their links pointing to outside communities. In these two cases, the links represent millions of economic interactions across state boundaries. However, relative to the number of interactions within the individual communities, there are not enough connections to warrant merging them. There are even stronger connections in the Piedmont and Gulf Coast megaregions, however they are still not strong enough to cross all state boundaries. In the case of the Piedmont, there are enough connections between its constituent parts in the Carolinas, but not between them and the counties in Georgia, or between the parts in Georgia and Alabama. Similarly the Gulf Coast megaregion has strong connections between its parts in Louisiana and Mississippi, but not Alabama or Texas.

A megaregion that breaks the state boundary dependence is Cascadia, the megaregion in the north-west that has Seattle and Portland as its major cities. Cascadia is also different from the other megaregions in that it a bioregion - an area with common plants, animals and environment, unique at a global scale - and there was an effort to unite Cascadia for tourism in 1996 [2]. However, this was not successful partly because each state has its own marketing plans and budgets. Later, there has been a series of re-

[2] http://www.america2050.org/pdf/ecolopoliscascadia.pdf

search papers looking at the megaregion starting with an extension of the America2050 project (Hagler, 2009). Recent attempts found evidence of an emergent industry at the megaregional scale centred around electronics production, information technology and communitication services (Ahuallachain, 2012). The results from the hierarchical analysis, give further evidence that there is emerging employment behaviour that corresponds to the scale and definition of the Cascadia megaregion. However, there are stronger interstate preferences than megaregional, since communities that correspond to the states appear first. The fact that the megaregion only appears at the third level, suggests that there are enough connection across the entire border between Oregon and Washington, but not enough between the America 2050 counties or census tracts that are supposed to define it.

Another finding is the communities that make up the two California megaregions. Both appear in the second level of the community detection hierarchy. Even in the last level, Nevada represents its own community and there are no parts of it in the other megaregions. This is in contrast to most studies that group Las Vegas with Southern California and Reno with the Northern California megaregion, e.g. (Hagler, 2009; Harrison and Hoyler, 2015b; Glocker, 2018). Again this result suggests that state boundaries play a very important role. The division between Northern and Southern Californian megaregions in the detected communities is consistent with the decision taken in America2050 to split California in two. This is in contrast to other research, for example Nelson (2017) that treats the whole state as one big megaregion. Another finding is the difference in the hierarchical structure of the two megaregions even through they are in the same state. The Southern California megaregion is highly polycentric and mostly already present in the first level of the dataset. While the Northern megaregion consists of more independent units that form up the megaregion at a later scale.

The analysis also showed that our level two communities are similar to the megaregions defined by Nelson and Rae (2016). Lang et al. (2020) argues that the Nelson and Rae (2016) are more similar to 'megalopolitan areas', rather than megaregions. However, our final delineations are larger than Nelson and Rae (2016) and by extension 'megalopolitan' areas. Communities that encompass Arizona Sun Corridor, Texas Triangle, Southern California and Northern California already appear at the second hierarchical level of our results, whereas the final third level communities represent an even larger scale, with two other communities encompassing Hagler (2009) megaregions emerging.

4.5.2 Spatial distribution at different scales

The results from the spatial analysis of employment suggest that at a scale, similar to the metropolitan one, most employment in the US is decentralised. A general pat-

tern of intra-metropolitan decentralisation is well-established in United States polycentricity research (Dadashpoor and Malekzadeh, 2021; Manduca, 2020). This correspondence between our results and the literature is a further signal that the hierarchy of delineations captures aspects of economic reality at different scales, which were not pre-defined, but inferred from the data. More specifically, our analysis suggests a strong pattern of employment scatteration or dispersion. The 'Proportion of employment within centres' column in Table 4.1 shows the percentage of census tract employment that the centers capture from the total number of workers in the rest of the delineation. The mean percentage of workers within centres is 33 %. In 75% of the level one delineations the proportion is below 41 % and communities with a higher proportion capture just 22 million of the all workers. This means that for most places, which account for 100 million workers, there is some degree of employment scatteration, not just concentration within centres. This finding contrasts with previous research using the same dataset, which found decentralisation of employment but heavy concentration in few centres (Manduca, 2020). Part of this difference can be attributed to the differences in scale, which can have a large impact on polycentricity results - the phenomenon can be defined or defined away depending on the initial delineation of the territory under analysis (Möck and Küpper, 2020). For example, in our analysis New Jersey and New York are areas, whereas in most polycentricity research that uses metropolitan statistical areas they are parts of the same area.

The decentralisation pattern generally disappears when the scale is set to the second level of the hierarchy, where communities are large state-like delineations. This phenomenon - decentralisation and a lower scale and more centralisation as the scale of the units increases - is present in studies of employment concentrations in both Europe and China (Hall and Pain, 2006; Liu et al., 2018). The exception to this general pattern is the Phoenix Sun Corridor megaregion, for which the analysis showed there is no identifiable centre at this larger scale. This result, combined with the fact that the megaregion appears as early as the second level in the hierarchy, makes the Phoenix Sun Corridor the most integrated and decentralised potential megaregion in the United States.

4.5.3 Implications

Megaregions describe a new geographic scale of social and economic interaction that span over several administrative boundaries. However, they do not map cleanly to any jurisdictional element and the benefits of planning and administration at the megaregional scale requires collaborative efforts at various levels. Several papers, such as Ross et al. (2016); Nelson (2017); Lang et al. (2020); Friedmann (2019), suggest that formal administrative structure and cooperation between different governing bodies

at the megareional scale will be necessary to drive megaregional creation and optimization. This can take the form of better transportation links that lead to greater megaregional coherence (Yu and Fan, 2018), coordinated economic policies aimed at leveraging megaregional economies of scale (Marull et al., 2013) or large-scale sustainability planning (Ross et al., 2016).

Such efforts are, however, not widely adopted. For example, Stich and Webb (2019) find that planning in Lousianna counties, could be aimed to leverage the traffic from the Gulf Megaregion, but no such initiatives exist. Similarly, research by the Bay Area Council Economic Institute show that there are increasing flows of people, goods and services within the megaregion, but there is a lack of cross-county collaboration (Bellisario et al., 2016). Oden and Sciara (2020) find that at the national scale, only in the case of the Arizona Sun corridor, Northern and Southern California megaregions, the megaregional scale and language are used in transport planning. Evidence of already emergent megaregional interactions is one way to provide motivation to build the required collaboration efforts between potential partners.

The results of the analysis show that there are two types of the proposed US megaregions.

The first are the megaregions that are delineated in our analysis - Arizona Sun Corridor, Texas Triangle, Southern Florida, Southern California and Northern California. The results suggest that there exist strong economic links between these regions, reflected in employment patterns. These results lend support to researchers, planners and administrators's suggestions to develop specific planning and administrative efforts for these megaregions, in order to make the most out of the already emergent economic interactions (Nelson, 2017; Lang et al., 2020; Purkarthofer et al., 2021). This includes 'Economic Development Structures' which cross county lines as proposed by Bellisario et al. (2016) for the Southern California megaregion. Additionally, infrastructure projects, such as better transportation links, can further increase megaregional coherence (Yu and Fan, 2018).

The second set of megaregions are those proposed in the literature but not present in our set of results - Piedmont, Midwest, Northeast, Front Range and Gulf Coast. The inability to delineate areas encompassing these regions suggests that the existing economic interactions are more locally concentrated. Our analysis shows that if large scale economic interaction are to be encouraged for these megaregion, the focus should be on developing structures, both political and economic, that target neighbouring counties from different states that fall within the same megaregion. This is the case, since megaregional integration breaks down across state boundaries.

Additionally, the results from the different scales could be used directly or combined with other data sources for planning and research purposes. First, the results can be combined with other data sources, such as the social interactions used by (Calafiore

et al., 2021). Second, analysis that uses metropolitan areas, or focuses on local labour markets, could use the first level of delineations from our results instead. Any differences in the resulting analysed phenomena will highlight the effects that the 25% fixed cutoff used in delineating metropolitan areas have on the current results in the literature. The polycentricity analysis carried out in this paper is one example of this, and studies that focus on health (Meijers, 2008), sprawl (Barrington-Leigh and Millard-Ball, 2015) or scaling laws (Lobo et al., 2020) in urban areas could use our results in a similar manner.

4.5.4 Future work

One limitation present in this and previous community detection approaches is the fact that methodology relies only on one type of data. This could be business transactions or commuter flows, but even if these flows implicitly capture some information about historical and cultural ties important aspects are missing. Hamilton and Rae (2018) show this in the context of Scotland, where historically separate areas are merged to increase modularity, however for a variety of reasons, some of which political, this is not desired. New forms of data (Arribas-Bel and Tranos, 2018) can help in addressing this by capturing more varied interactions between people such as leisure and social visits which reflect cultural cohesion.

There are also several limitations of the methodology. First, community detection gives a separation of the graph, rather than a clustering - i.e. all communities have to be assigned to a deliniation, even if this is not desirable. Second, sometimes as the scale of the communities grows, problems arise from the definitions of modularity - the value of edges between nodes increases as the scale increases. This means that there are census tracts with a low number of associated employees that can be attached to any community without it much effecting the overall modularity score. Also as the scale increases the relative value of, for example 1,000 commuters between places increases. This issue is somewhat mitigated in this paper, since the focus is on the megaregions which represent high density areas and there is validation carried out at each level of the hierarchy to ensure that all the intermediary results are valid. Nevetheless, advances in methodology which specifically take into account this application can improve the results and adress these limitations (Singleton and Arribas-Bel, 2021).

This study could be further extended by considering the evolution of megaregions over time. This is possible since there is historical LODES data available. The changing megaregion boundaries could give even more information about the dynamics of large scale economic units. Further the Lodes data can be broken down into income levels to see what differences income makes to the emergence of megaregions in the US.

5

Delineating urban areas through satellite-derived building footprints

Abstract: This paper takes advantage of new forms of data and delineates urban boundaries in the United States, based only on the footprints of individual buildings and a machine learning approach. For this purpose it uses 129 591 852 building footprints, generated by applying computer vision algorithms on satellite imagery. The delineation approach separates areas of high building density surrounded by low density, where the exact values for low and high are locally inferred from the data. Overall our results show that its possible to use building footprints, derived from satelite imagery to define the extent of urban areas in place of population data The resulting delineations are 4009 and capture urban areas with population ranging from 10000 to 17 million. In total they account for 82% of the population in the Contiguous United States and 78% of all the building footprints. The number of building footprints in the clusters is highly correlated with the actual census population with statistically significant results - spearman .95 and pearson .95. On average the resulting boundaries are most similar to census defined urban centres and functional urban areas and are much larger than official city boundaries. The delineated areas have varying internal densities, which change in different places – LA has numerous dense cores, whereas the San Jose/San Francisco delineation has a relatively consistent density. Furthermore, in addition to internal variations we analyse the building density in the US at different scales

5.1 Introduction

The majority of people currently live in cities and it is projected that by 2050, there will be an additional 2.4 billion to do so (Batty, 2018, chapter 2). Given the current and projected growth, urban areas have become a principal driver of most social, institutional and technological innovations and are the focus of the solutions to numerous pressing challenges facing society (Lobo et al., 2020). The growing levels of urbanisation have led to new emergent urban phenomena and many challenges related to transport, sustainability and urban management. Defining bounded territorial units for quantitative analysis, at different scales, is one of the core steps in carrying out the necessary research to address and manage these changes. At the most basic level, different approaches to delineations of boundaries lead to a different number of cities, city sizes and urban population (Roberts et al., 2017). Furthermore, the spatial extent of urban areas affects the calculations of numerous properties - employment, labour productivity, population density - which in turn influence subsequent analysis at various scales (Arcaute et al., 2015; Batty, 2018; de Bellefon et al., 2019; Lobo et al., 2020; Parr, 2007). Since relying on official urban designations is problematic, researchers have attempted to provide consistent quantitative urban boundary definitions in numerous ways (Duranton, 2021), aided by the rising availability of new forms of data and methods (Wolf et al., 2020; Arribas-Bel et al., 2021a).

In general, these delineation approaches rely on combining fundamental units, based on similarity of attributes or relationships. The attributes can either represent functional - commuter flows or number of workers at a place - or form data such as built environment characteristics. Popular choices of units are small scale administrative areas - census tracts, counties (Nelson, 2020) - or cells and hexagons derived from gridding the territory under analysis (Florczyk et al., 2019). These can be grouped together to form urban areas in three ways: first, based on relationships such as commuter flows; second, on spatial contiguity and characteristics of the units - population density or night lights; or third, on combinations of the previous two approaches. With the rising availability of new forms of data (Arribas-Bel and Tranos, 2018), researchers are able to look at other units and interactions such as tweets (Wei et al., 2020), mobile phone data (Secchi et al., 2015), location-based social networks (Calafiore et al., 2021).

All of these approaches typically require that a number of parameters or models be specified beforehand. Some approaches may require the assumption of a standard urban form, e.g. a monocentric model (de Bellefon et al., 2019; Taubenböck et al., 2019). Other approaches, for example delineations which rely on population grids, require density thresholds or grid sizes to be explicitly set by the researcher. In general, lower minimum density requirements increase the scale of delineated areas, while

higher density values decrease it (Balk et al., 2018; Arcaute et al., 2015; de Bellefon et al., 2019). Different parameter choices can lead to separating a city apart or merging together metropolitan areas. Furthermore, the choices of parameters affect what types of urban forms, such as suburbs, ex-urbs or towns are included in the delineations. Overall, the different parameter choices have an affect on the final size and scale of delineated areas, which in turn affect the subsequent analysis of phenomena (Möck and Küpper, 2020; Balk et al., 2018). Currently, there is no agreement on an optimal scale, specific optimal threshold values or whether a single value should be applicable in all contexts (Statham et al., 2020, 2021; Duranton, 2021; Möck and Küpper, 2020).

The main aim of this paper is to operationalize a minimalist definition of what an urban area is - a geographical area with a concentration of individuals and activities higher, relative to the surrounding area (O'Sullivan, 2011) - and derive the spatial extent of cities across the whole United States. We aim to apply as few restrictions and assumptions to this operational definition as possible by using a modified HDBSCAN approach (Garcia-Pulido and Samardzhiev, 2022). Specifically, we do not explicitly parametarise the number of areas, scale, density thresholds, CBDs, models or use intermediary aggregations such as grids. Instead we delineate areas through density clustering of individual buildings polygons, based on variable density thresholds, inferred locally from the data. Furthermore, we do not explicitly specify the final scale of the units, the distinction of what is urban and what is not and instead derive these quantities from the data. By doing this there are no explicit restrictions on what constitutes part of an urban area - our results be a mix of units of different urban scales and encompass multiple ex-urban and peri-urban components locally. The only pre-set parameter is the minimum number of buildings within a delineated area, for which we test several options.

To delineate the final areas, we place all the buildings in the contiguous United States into a nested hierarchy of parcels of land, based on the location of individual buildings and their local built environment density. The final results are derived by extracting the most consistent delineations in this hierarchy and are compared against six other research and administrative urban units of different scales. Furthermore, we use the hierarchy to achieve two additional aims - first, to analyse the density variations within individual boundaries and second, to analyse the relationship between the delineated units.

The main advantage of our approach is its flexibility. Urban areas are delineated using relative values of 'high' and 'low' which vary locally as inferred from the surrounding building footprints at a particular place. There is no single density, radius or model requirements imposed on the identified boundaries. The final results can be anything from highly dense and large scale polycentric urban areas such as metropolitan areas, to sparser and smaller scale towns and they can be of a different scale in

different parts of the country.

Furthermore, a second set of advantages comes from the use of the 129 591 852 individual building footprints, generated by applying computer vision algorithms on satellite imagery (Heris et al., 2020). First, there is no need to aggregate administrative units or grid cells. This means that the analysis minimises problems related to the modifiable area unit problem (Openshaw, 1979) or issues related to the choice of contiguity matrix selection (Statham et al., 2020, 2021). Second, there is no need to rely on surveys or census to obtain population data, since the approach relies only on the footprints. Third, the methodology can potentially be applied globally, since buildings are homogeneous units across different countries and the footprints can be extracted from satellite images. Fourth, the approach has the potential to track temporal changes in real-time, provided there are up-to-date satellite images available.

The rest of the paper is structured as follows: Section 2 provides background on related approaches to urban area delineation, for different problems and scales. Section 3 and 4 describe in detail the dataset and methodology respectively. Finally, Section 5 and 6 describe and discuss the resulting urban areas, comparisons to other delineations and the results from the other analyses carried out.

5.2 Literature review

5.2.1 Urban delineations and scale

Different urban challenges are tackled using delineations of different scales and sizes. Questions about the effects of urbanisation and urbanisation policies on national economy are tackled using delineations, which represent cities or dense urban centres (Roberts et al., 2017; Florczyk et al., 2019). In these cases boundaries affect statistics such as employment and labour productivity, which in turn influence theories about the effects of urbanisation. For example, several analyses from Argentina, using delineations based on local administrative units, challenge a long-held theory that as urbanisation increases the proportion of services in the national economy also grows (Roberts et al., 2017). In contrast, analysis using urban delineations based on night-light intensity, which ignore local definitions of 'urban', show a smaller percent of urbanisation and therefore a proportion of the service economy in line with theoretical expectations (Roberts et al., 2017).

Other types of analysis, which focus on for example, the spatial distribution of employment patterns, use larger scale delineation units (Parr, 2007). These functional urban areas, such as metropolitan areas, consist of cities along with nearby towns, and in some cases other cities, with the goal of capturing the spatial extent of local labour and activity pools. In addition to directly studying spatial labour patterns (Arribas-Bel

and Sanz-Gracia, 2014), functional urban areas are also used to study the effects of spatial urbanisation patterns on sustainability, culture, public health and social integration outcomes (Meijers, 2008). In all of these cases, research into the above phenomena are affected by the exact delineation methodology of both of the area under study and the number and boundaries of proposed centres within it (Möck and Küpper, 2020).

There are numerous other proposals that the appropriate scale for analysis and planning of economic and sustainability policies is larger still and should focus on megaregions (Hall and Pain, 2006). Megaregions or related large-scale urban concepts represent urban agglomerations - numerous cities, their surrounding towns and areas - that capture huge areas of built up environment, population and economic activity. In the US, it is estimated that over 80% of the population and economic growth until 2045 will happen in megaregions (Nelson, 2017). Due to the large population and economic output, it is further stipulated that megaregions can support varied economies of scale and can drive the competitiveness of national economies in global markets (Glocker, 2018). Additionally, megaregions can act as the appropriate scale for sustainable development, by tackling problems brought about by increasing urbanisation, which cross current administrative boundaries (Ross et al., 2016). However, in order to reap these benefits large scale planning and collaboration between numerous partners is required (Wheeler, 2015). Different megaregional definitions and delineations lead to different numbers of megareigons, spatial boundaries, sizes and scales (Hagler, 2009; Lang and LeFurgy, 2003; Ross et al., 2009; Glocker, 2018).

There is no agreed upon best delineation method, data or appropriate scale to use for analysis of these and other aspects of urbanisation and cities (Statham et al., 2020, 2021; Duranton, 2021; Lobo et al., 2020; Batty, 2018; Möck and Küpper, 2020; Glocker, 2018). Additionally, the same datasets and methodologies can be used to delineate units at different scales depending on explicit parameterasations. Different delineation approaches come with advantages and drawbacks, which affect subsequent analysis and calculations.

5.2.2 Delineation approaches

Local administrative and official city boundaries have proven to be sub-optimal for quantitative analysis. Cities can extend beyond their borders into the surrounding area, however for political or economic reasons their official boundaries do not necessarily reflect their growth (de Bellefon et al., 2019). Furthermore, using administrative boundaries makes comparisons across countries and even within the same country, across time, difficult due to inconsistent definitions.

To address this, one popular group of delineations builds the urban areas based on spatial contiguity and the characteristics of smaller scale administrative units. Such ap-

proaches focus on attributes such as population size, built environment, travel time and others. One example of this in the US is the Census Urban Centres (CUC) boundaries. After each census, the Census Bureau delineates urban areas that represent densely developed territory, encompassing residential, commercial, and other nonresidential urban land uses based on local administrative and political units - census blocks and block groups - and qualifying criteria. There are two types of urban areas delineated: urbanized areas (UAs) that contain 50,000 or more people and urban clusters (UCs) that contain at least 2,500 people, but fewer than 50,000 people. The technical methodology is described in detail in the Federal Register of August 24, 2011.

Furthermore, local administrative units and associated characteristics can be used to delineate much larger megaregions. One of the most influential sets of such delineations came from the America 2050 study where megaregions are defined as 'interrelated population and employment centers or MSAs that share common transportation networks, cultures, and environmental features' (Hagler, 2009). The actual boundaries were constructed by merging census counties based on spatial contiguity, projected and existing population thresholds, as well as expert opinion.

Another approach researchers have taken relies on functional relationships such as commuting patterns, taxi rides or flows of goods. The assumption is that if the interactions between two places capture information about an economic system's performance and the daily life of individuals, it can be used to assess whether they form part of the same urban area (Duranton, 2015). The most popular and widely-used example in the US is metropolitan statistical areas which aim to capture local urban labour markets, that expand beyond central urban cores. Metropolitan statistical areas are constructed using urbanised areas, described above, as anchors and nearby counties which have a high percentage of commuters flowing into or out of them. The commuter threshold for the 2010 census is set as 25 percent.

Additionally, flow data can be used to delineate larger scale megaregional or supermetropolitan structures such as megaregions and megalopolitan areas. The delineations are again based on the strength of relationships between the underlying units. This can measured based on the density of connections using community detection approaches (Nelson and Rae, 2016) or by varying the commuter threshold calculations (Lang et al., 2020).

The disadvantages of these approaches come from the aggregation units and the methods used. First, land size and land use within local administrative units can be a mix of both rural and urban land and their spatial extent is optimised for the purposes of surveying, not delineations (Wolf et al., 2020). Second, concerns about urbanisation phenomena and their impacts have led to calls for globally applicable and consistent definitions (e.g., Florczyk et al. (2019)). Furthermore, when using explicit thresholds different methods and values lead to different results and scales. Additionally, inter-

action data such as commuter flows is not readily available globally or at all scales (Glocker, 2018).

Grid delineation approaches aim to address some of these issues. One popular example of delineations using population and built environment data, extrapolated to grid cells, is the DEGURBA classification (Florczyk et al., 2019). This dataset represents urban areas derived by aggregating one square kilometre GHSL population and built environment cells, where each group of cells has a population of at least fifty thousand. Furthermore, each cell that makes up part of the urban centre has to have a population of at least 1500 or it should have a built density of more than 50 %. The delineated areas and accompanying datasets are widely used in the analysis of urban phenomena at both a global and local scale (Florczyk et al., 2019). The advantages of this approach is that it is spatially consistent due to the grid and it provides a globally applicable definition of dense urban areas. There are other complimentary approaches which also use population grids with different methodologies. Examples of such are the city clustering algorithm (CCA) (Rozenfeld et al., 2008) or the approach taken by Statham et al. (2020, 2021).

Grids can also be used to delineate larger scale areas such as megaregions or urban functional areas, which aim to capture local labour markets similar to metropolitan statistical areas. A popular example is the combinations of multiple GHSL urban centres into functional areas (GHS FUA), based on spatial contiguity, population thresholds and a logit generalised linear model (Schiavina et al., 2019). Another example is Dingel et al. (2019), where the authors use a grid and contiguous nighlight intensity in neighbouring cells to define metropolitan areas. At a larger scale, (Glocker, 2018) combine the previously defined GHS FUA areas using a mean shift algorithm, population thresholds and a distance threshold of 300km, as well as infrastructure calculations to define megaregions. Similarly, Florida et all. (2008) uses grids and nigh light intensity thresholds to define megaregions.

One of the main disadvantages of grid approaches are problems of aggregation and contiguity choices. Different grid sizes can lead to different aggregation statistics which affects subsequent calculations (Openshaw, 1979; Wolf et al., 2020; Arribas-Bel et al., 2021a). Other disadvantages come from the usage of specific thresholds and contiguity criteria. Both affect delineations and there is no universally agreed global threshold or contiguity aggregation method (Statham et al., 2020, 2021; Duranton, 2021; Batty, 2018).

There are also other grid methods which focus on interaction data such as taxi flows, human flows and mobile phone networks (Deng et al., 2019; Grauwin et al., 2015; Wei et al., 2020). Similarly, there are approaches which aim to use both relationship data and morphological features. For example, (Ross et al., 2009) combine both morphological dependencies and functional relationships (transportation links)

to delineate large scale urban areas. However, these approaches are rarer due to the data availability limitations, since large amounts of diverse data are needed, as well as methodological and theoretical problems of how to combine it (Glocker, 2018).

Other studies only focus explicitly on morphological features such as the patterns of intersections of the road network without using grids or administrative units (Arcaute et al., 2015, 2016). The disadvantages of these approaches are similar - threshold parameters have to be defined or aggregations have to be used, which in turn affect the final results and global data availability is a problem. However, with the development of satellite technology and computer vision algorithms, new forms of data have started to better capture population distributions and activity (Roy Chowdhury et al., 2018).

This paper directly uses individual building footprints derived from satellite images to delineate urban areas. There is prior research that show that individual buildings can be used to delineate urban areas and analyse their spatial form. Such examples are the studies carried out by Arribas-Bel et al. (2021a) in Spain, de Bellefon et al. (2019) in France, Adolphson (2009) in Sweden, Krehl (2015) in Germany and Usui (2019) in Japan. The difference between these studies and this paper are threefold.

First, the buildings used in this paper are derived from satellite images, whereas the buildings footprints used in those studies come from government or private company records. As such, our approach has lower data requirements and is more globally applicable. It should be noted as well that, satellite derived individual building footprints have started being used in other areas of urban analysis. For example, Huang et al. (2019) uses the same building footprints used in this study, enhanced by OpenStreerMap data, to infer population at high spatial resolutions more accurately than by using other data - nighlight intensity, land cover or impervious surface layer data.

Second, there are methodological differences in how we delineate urban areas. In the cases of Arribas-Bel et al. (2021a), Chaudhry and Mackaness (2008) and Usui (2019) there is use of a density threshold implicitly or explicitly defined through the choice of one or more parameters. For example, Arribas-Bel et al. (2021a) use a minimum number of buildings within a specified radius as a density threshold. In contrast, our approach is based on a locally variable density threshold. Furthermore, where de Bellefon et al. (2019); Adolphson (2009); Krehl (2015) use a a grid, we directly use individual building locations and therefore avoid the need for spatial aggregation or de-aggregation, such as population data to a grid cell or data grid cells to urban areas.

Third, due to the locally adaptable density threshold, our approach does not have a fixed scale and it is possible to delineate areas of mixed scales. For example, the North East of the United States is very densely populated and built up. As such, different density thresholds set a priori can have large affects on the final delineations. By changing preset density requirements, it is possible to attach peri-urban and ex-urban areas to cities and even merge different dense urban cores together into larger scale

units such as megaregions (de Bellefon et al., 2019; Arribas-Bel et al., 2021a; Arcaute et al., 2016; Statham et al., 2021). In contrast, our approach allows for the detection of urban footprints at the official city, dense urban core, functional area, megaregional or other scales. However, there is no scale explicitly or implicitly defined and the final results can have a mix of scales, reflecting the underlying building density in the data.

5.3 Data & Methods

5.3.1 Data

The data used in this study comes from the Microsoft Cognitive Toolkit (CNTK) and consists of 129 591 852 computer vision generated building footprints extracted from Bing imagery, covering all 50 US states. Each footprint is a polygon that represents the geographical position of the building. The building footprints used in this study cover the entire contiguous United States with 48 states (D.C included). The geometries were projected using U.S. Albers equal-area conic projection (EPSG: 5071) to obtain their areas in metric units and their centroids were used for pairwise distance calculations. Figure 5.1 shows the available data for Washington D.C. The dataset is a highly accurate representation of the built environemt in the US - a sample of five million footprints were measured against Openstreetmap geometries which resulted in 99.3% precision and 93.5% recall accuracy metrics (Heris et al., 2020). Similar datasets were released publicly for Australia, Canada, Uganda and Tanzania showing the global applicability of the computer vision approach.

Figure 5.1: Building footprints data for Washington D.C.

In spite of the high accuracy the data comes with several limitations. First, the dataset contains only the pure geometry of extracted building footprints, meaning that information such as building height and building type is not included. Second, Bing imagery is a composite of multiple sources, so the date of extracted building footprints varies. Third, there are many errors in densely built city centres where entire blocks are detected as a single building. Figure 5.1 is a demonstration of this - footprint area increases as the building density increases. Sometimes this is representative of the actual morphology of cities, however in many cases multiple buildings are represented as a single footprint. One of the worst cases in this regard is Manhattan, where entire blocks are represented by one footprint. Similar problems arise for other city centres. Despite the aforementioned limitations, this dataset provides comprehensive open-source building footprints available for the entire U.S.

Additionally, we use six datasets to compare and evaluate our delineated areas. The first one is the one square kilometre GHSL population grid (Florczyk et al., 2019). This grid is derived from census and built environment information in different countries and gives population estimates in cells for the whole world. The dataset is used to

estimate the population for each set of delineated areas - both the reference ones and the ones obtained from this paper. This is done to provide consistency across the comparisons.

The second dataset is the GHSL Urban Centre boundaries (Florczyk et al., 2019). The third - the Census Urban Centres (CUC) boundaries. The fourth - the megaregion delineations from the America 2050 project (Hagler, 2009). The choice of this particular set of megaregions was motivated by the widely available research into the America 2050 megaregions, from different researchers (Oden and Sciara, 2020), the Federal Highway Administration megaregions project[1] and state authorities (Bellisario et al., 2016). The fifth dataset is the GHS Functional Urban Areas (Petrović et al., 2020). Lastly, the sixth dataset consists of the official local incorporated place boundaries in the United States(towns and cities) and census metropolitan area delineations.

5.3.2 Clustering approach

This subsection describes how the building footprints will be analysed. Its sections cover two areas - first, a focus on the specific methodological approach to delineations and second, on the analysis of the results. The clustering sections focus on density-based clustering, the details of HDBSCAN and the changes made to account for the computational complexity and footprint size issues. The delineation sections describe the comparisons between datasets and additional cluster analysis.

Density-based clustering & HDBSCAN

In order to define the urban boundaries this paper turns to machine learning clustering algorithms. Specifically, the algorithm used in this paper is an extension of HDBSCAN (Campbell, 2018), which is a method that produces state-of-the-art results for density based clustering (Campello et al., 2020). Density-based clustering algorithms are a family of clustering algorithms that focus on the counts of data items, in addition to the distances between them and use this information to group together some or all of the data items. One of the most popular density clustering methods is DBSCAN, which is a precursor to HDBSCAN.

DBSCAN is a widely used algorithm in numerous fields (Schubert et al., 2017), however parameter choices can have a large impact on its performance. Examples of papers that have used DBSCAN or DBSCAN-like algorithms for delineations of urban areas are the city clustering algorithm (CCA) (Rozenfeld et al., 2008), Statham et al. (2020, 2021) and Arribas-Bel et al. (2021a). DBSCAN requires two parameter choices - a distance radius and a minimum neighbours value. The combination of these

[1] https://www.fhwa.dot.gov/planning/megaregions/what_are/

parameters ensures that only similar data items in areas of high density are considered clusters. The leftover data is labeled as noise.

There are two limitations of applying DBSCAN directly to our dataset. The DBSCAN parameters explicitly impose a minimum density threshold on all delineated areas and implicitly define a scale of the final units. By decreasing the minimum density threshold otherwise different cities are merged together in the results, while by increasing it parts of the same city can be separated (Balk et al., 2018; Arcaute et al., 2015; de Bellefon et al., 2019; Duranton, 2021). Another issue is that these parameters apply to the whole dataset and do not take into account local context. These two problems are especially evident in the North East of the US, which is very densely built, and changes in parameters can lead to delineated areas of different scales.

HDBSCAN aims to limit these issues by computing all possible DBSCAN clusters across all possible distance thresholds for a fixed minimum neighbours value (McInnes and Healy, 2017). This way only one parameter a minimum neighbours value, instead of two, has to be estimated. We chose the algorithm since it does not need the number of urban areas or their scale specified beforehand and makes little assumptions about the shape of the potential urban form areas. The algorithm separates data points in areas of high density and groups them together based on proximity from data points in areas of low-density. Additionally, it does not partition the whole datataset and it can mark points in low-density areas as being outside any cluster (noise). In this paper the data points represent individual buildings and the proximity is defined as geographical distance. The application of HDBSCAN on this data, leads to an operational definition of an urban area as a place with high levels of building density surrounded by areas of low density. Both 'low' and 'high' are relative values inferred for specific locations from the dataset itself.

HDBSCAN works in three steps - first it orders all points/buildings into a hierarchy, then it converts the hierarchy into a an approximation of the probability density function of the data - a condensed tree, and finally, it uses the condensed tree to extract the final clusters. The order and connections of points into the hierarchy is based on the distance threshold value at which they have more than the specified minimum number of neighbours. Figure 5.2 shows the top half of the hierarchy for the buildings in the D.C. area, shown in Figure 5.1. Vertical lines are either individual buildings or groups of buildings that fall under the same horizontal line. The horizontal lines represent possible DBSCAN clusters. Horizontal lines appear at the distance threshold when two or more points become mutually reachable and both have more than the specified minimum number of neighbours, or in other words when they become part of the same DBSCAN cluster.

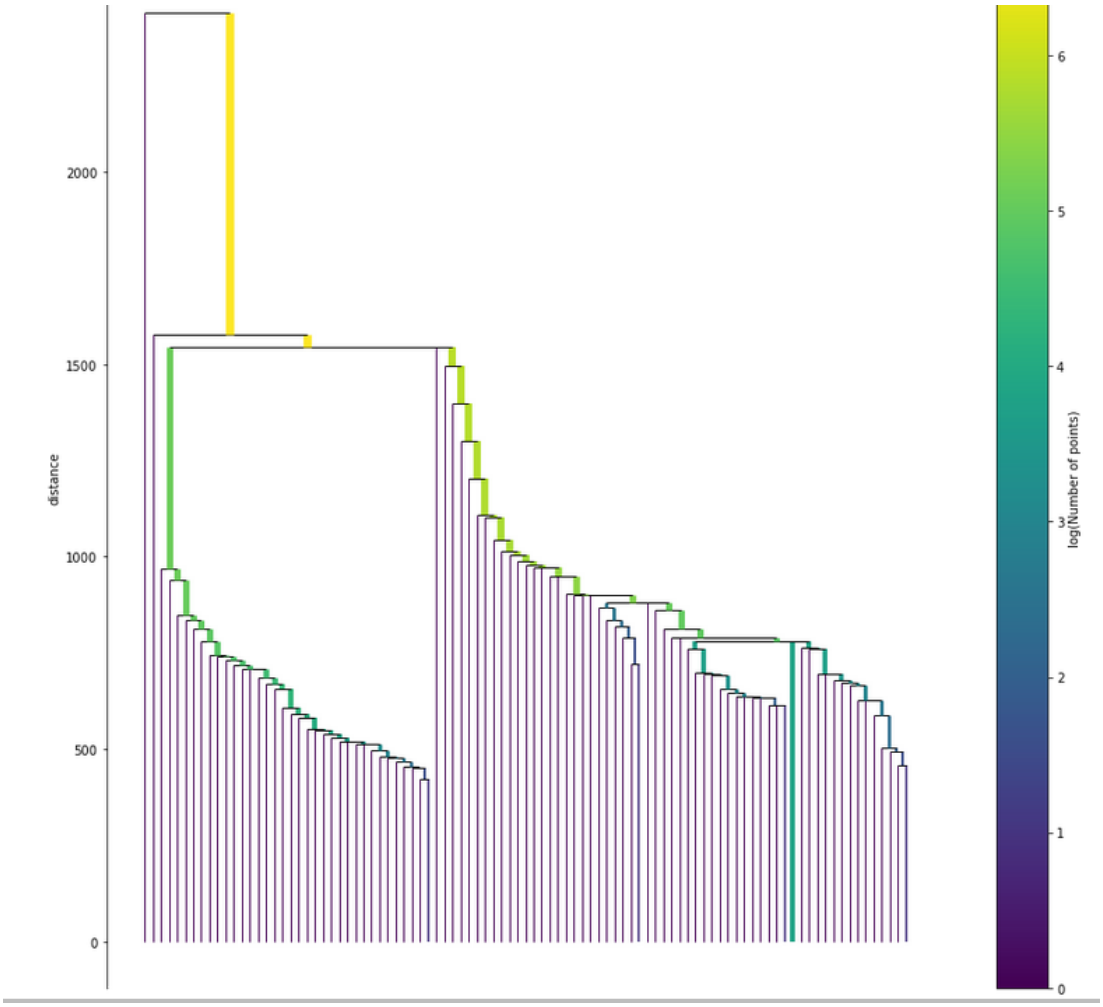


Figure 5.2: The top half of building hierarchy constructed for Washington D.C.

The second step converts the hierarchy into a condensed tree, by inverting it. This turns the hierarchy into a an approximation to the probability density function of a random variable, defined only over the existing data (McInnes and Healy, 2017). When applied to the dataset used in the paper, the hierarchy approximates the probability of the appearance buildings at specific locations in the continental United States. The specific probability of a building appearing at a specific location is the inverse of the distance it joins a DBSCAN cluster in the hierarchy. The condensed tree for all buildings in D.C. is shown in Figure 5.3.

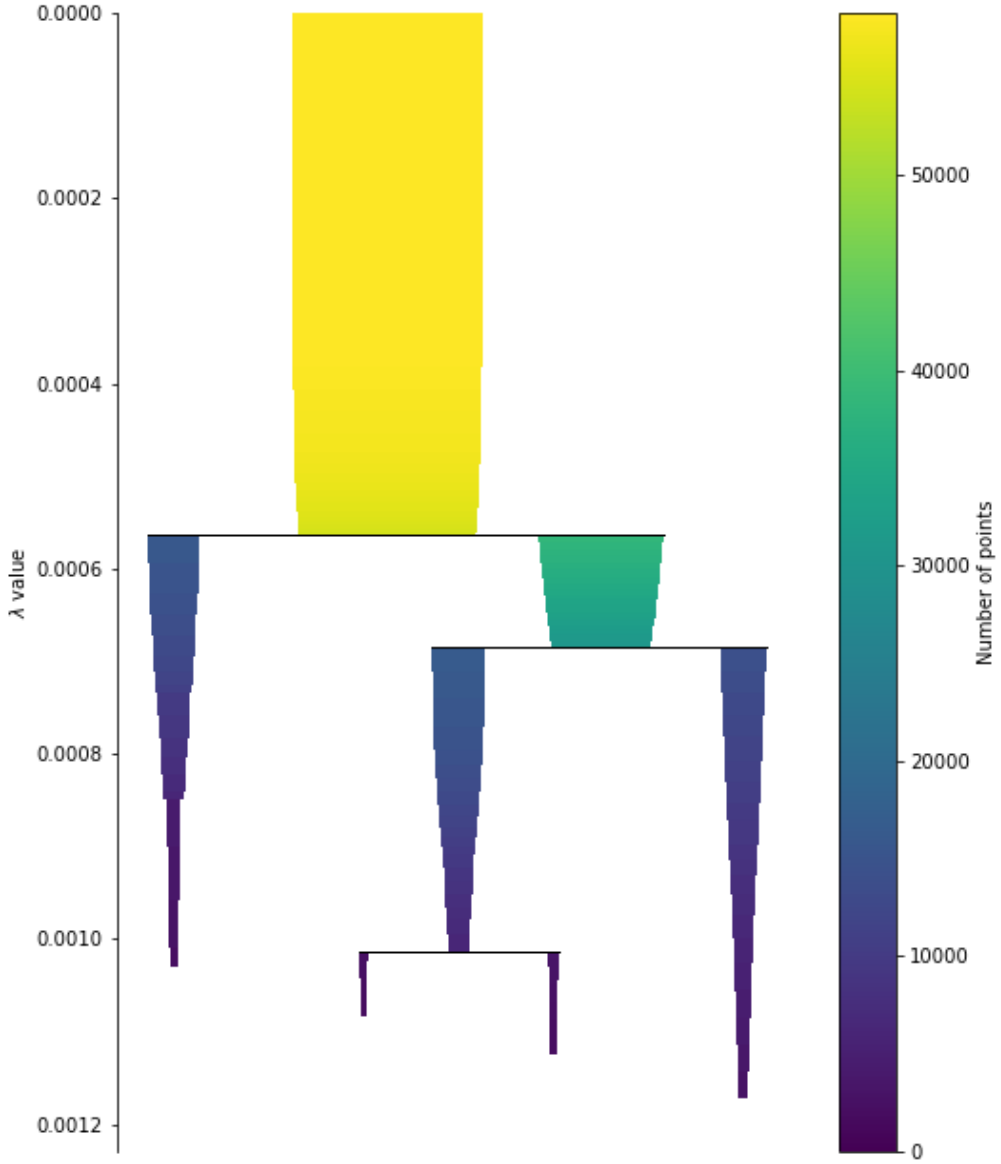


Figure 5.3: The density tree constructed for Washington D.C.

The third step extracts the final clusters based on the condensed tree. A cluster forms when more than the specified minimum number of buildings are mutually reachable. It disappears at the value (or horizontal line) when it is reachable by another cluster. The final clusters are those that persist the most throughout the hierarchy, where persistence is defined as the difference between the inverse distance threshold where

the cluster forms and where it is merged with another cluster. There also exists the restriction that if a cluster is picked as final, clusters it merges into cannot be final. In this way the final clusters take into account local changes in density, making it possible to merge groups of buildings together only if they have similar densities and are close together. Therefore, the algorithm aims to separate densely built-up areas in the United States, from sparsely built-up areas.

HDBSCAN extension

Directly applying existing HDBSCAN algorithms to our dataset is problematic due to the large number of footprints and the size of the footprints in dense areas. To address this, we extend the HDBSCAN algorithm, by incorporating the area of the buildings into the hierarchy and cluster extraction calculations. The area is included to account for the problems computer vision methods have in detecting individual buildings in densely built cities (Jochem et al., 2020). Specifically, the area is included by defining the median area of all buildings in the dataset. In the case of the Microsoft data this is 177.5968 square metres. Then the area of every building is divided by this number and the result is rounded up. The resulting numbers are incorporated in the first and second stages of HDBSCAN calculations by giving each building a 'size' or 'multiplicity'.

Figures C.1 and C.2 in the Appendix, shows the differences this makes to the results in Manhattan when running the HDBSCAN algorithm with the same parameters. In the first case, where the area is not included in the calculations, New York is split into parts by Manhattan. This is because blocks in Mannhathan are considered single buildings and, based on the surrounding area and the number of detected buildings in them, Manhattan is classified as a place of low density. When the area of the buildings is taken into account this problem disappears,as shown in the Figure C.2. With our adjustment a street block in Manhathan, which is erroneously classified as a single building in the dataset, counts for 22 buildings if its area is 22 times larger than the median building area.

Due to the large amount of data and the need to take into account building area, we used a fast modified version of HDBSCAN in order to derive the cluster hierarchy for the whole dataset (Garcia-Pulido and Samardzhiev, 2022). The modified algorithm uses a fast, scalable DBSCAN implementation to separate the dataset into clusters where HDBSCAN can be applied independently and afterwards integrates the partition results together. More detail, experiments and proof of correctness are available in (Garcia-Pulido and Samardzhiev, 2022). The resulting hierarchy is then used as input to the tree extractions and clustering procedures implemented in the HDBSCAN package developed by McInnes and Healy (2017). Finally, in order to delineate the urban boundaries this paper uses the alpha shape complex to define a concave polygon that encompasses all the buildings within the same cluster (Edelsbrunner et al., 1983).

HDBSCAN parameter choices

HDBSCAN requires only one parameter to be specified - the minimum number of buildings, within an area so its considered an urban area (or cluster). The choice of parameter is guided by three considerations - interpretation, performance and accuracy. To account for interpretation we follow the approach set out by Arribas-Bel et al. (2021a) - a minimum urban population is set and the parameter value is chosen based on it. Furthermore, the value of the parameter has to be large enough not the be affected by the local geography - low values lead to results that cut in half cities separated by rivers, for example. Lastly, smaller parameter values are computationally easier to process, while larger ones take more time or are currently impossible with the size of the data and current implementations of the algorithm.

Arribas-Bel et al. (2021a) sets the minimum neighbours value based on the assumption that a building accounts for 2.2 people and that urban areas have a minimum total population of five thousand. This paper follows the same methodology and sets the minimum size parameter based on minimum population considerations. The parameter value is approximated using the number of persons per household, which according to the 2010 census is 2.63. We run the delineation procedure five times with values of 2000, 2900, 3800, 7600, 19000 which aim to approximate urban areas with minimum population sizes of 5,000, 7,500, 10,000, 25,000 and 50,000. A final set of results for the rest of the analysis is chosen based on comparisons between these runs.

5.3.3 Delineations analysis

This paper uses both the hierarchy and the final clusters in the analysis of the results. The final clusters are identified with a principle city, based on their geographic intersections with official state boundaries. When more than one administrative boundary intersects a delineation, the most populous city is chosen. The population and area of each delineated urban centre are defined using the US conic projections and the GHSL population grid. It should be noted that the population grid data is only used to count the people within an area and plays no role in the delineations.

Afterwards a series of comparisons are carried out with other urban boundary definitions of different scales - GHSL, official administrative units, Census defined Urban Centres and GHS Functional areas. First, the analysis is limited to the largest 15 delineations from our results. They are compared to the largest 15 delineations from the other sets in terms of population and size and their internal density variations are analysed. Next, a more robust comparison is carried out using a random sample of 10 million buildings. Each building is assigned to a delineated official, GHSL, FUA or CUC boundary based on its geographic location and a point in polygon test. Then the similarity between each of these assignments and the HDBSCAN delineations are

computed using the adjusted rand index. The rand index is a comparison metric that ranges from -1 to 1 and describes how similar groupings of data are. A higher rand index value indicates that the buildings in the assignment are similarly grouped to ours and that the two sets of delineations overlap. Since different delineations have different minimum population thresholds we create a series of comparisons where only areas above a certain threshold play a role in the rand index calculations.

In addition, the delineations are further analysed using polygon intersection tests. Three sets of spatial intersections are calculated between the HDBSCAN and the most similar set of delineations based on the above results. The first set consists of HDBSCAN areas fully encompassed in corresponding delineations of the other set. The second set is the opposite - HDBSCAN areas which are fully within the other set. And the last set contains polygons which overlap to varying degrees.

Lastly, the hierarchy is used to analyse the relationships between the delineated areas, as well as their relationship to metropolitan statistical areas and megaregions. The delineated areas are assigned to their corresponding metropolitan boundaries to analyse the distribution of delineated urban land. Finally, the hierarchy is again used in order to rank the megaregions based on relative morphological integration and to explore how close the delineations are to achieving a megaregional scale.

5.4 Results

5.4.1 Consistency across parameter range

Table 5.1 shows the similarity between delineations with different minimum samples using the adjusted rand score. When comparing two sets of delineations we assign clusters with size smaller than the minimum size of the larger set to -1 (or noise). For example, when comparing the 2000 and 2900 delineations all clusters in the 2000 delineation that have less than 2900 members are assigned to -1. This is done since the minimum samples parameter controls the minimum size of delineated areas and the lower this value is, the more clusters can get delineated in areas where there are no clusters for higher values.

Table 5.1: Adjusted rand index between pairs of delineations with different minimum sample parameter

	2000	**2900**	**3800**	**7600**	**19000**
2000	1.00	0.78	0.72	0.65	0.60
2900	0.78	1.00	0.80	0.69	0.63
3800	0.72	0.80	1.00	0.73	0.64
7600	0.65	0.69	0.73	1.00	0.71
19000	0.60	0.63	0.64	0.71	1.00

Table 5.1 shows that there is a large degree of agreement between the results. In fact, an increase of an order of magnitude in the minimum samples parameter, 2000 versus 19000, results in delineations with a high similarity score of 0.6. Most of the differences come from areas around the largest cities such as Denver, New York and Orlando. With New York specifically, the differences are due to a split along the Hudson Bay, obtained using a minimum selection of 2000 buildings corresponding to 5000 residents. The New York example is shown in the Appendix in Figure C.3. Furthermore, the population in the largest fifteen areas based on 2000 minimum samples is around 86 million, whereas for 19000 is 93 million. This shows that an increase in the only required parameter of 1,000 % results in a population change of 10%, which shows the robustness of our method.

For the rest of the analysis we pick 3800 which roughly corresponds to a minimal population size of 10,000. This set of delineations is the middle of all values and is the most similar to all other delineations on average. Furthermore, the value of 3800 is large enough to take into account natural barriers such as water bodies, and at the same time is computationally feasible. [2]

5.4.2 Number and size of delineated areas

The algorithm, with a minimum samples parameter of 3800, delineates 4039 urban areas of various scales and sizes in the contiguous United States. The final delineations are shown in Figure 5.4. Specific urban areas are shown in Appendix C. Table 5.2 shows descriptive statistics for the areas.

The population in the detected areas, as inferred from the GHSL population layer, ranges from zero to more than seventeen million. This is in part due to problems with missing values in the GHSL population data and due to groups of large buildings such as warehouses or other places like camping sites being delineated as urban areas. We propose several ways to address the latter issue below. The number of buildings ranges from 3800 up to nine million. Their total area is 23895.98 square kilometres from a total area of 32276.29 square kilometers of buildings for the whole US or 74 percent. The population captured by the HDBSCAN boundaries is 263,852,178 or 82 % of the contiguous US population captured in the GHSL grid. These close fractions suggest that the number of buildings in the delineations closely capture population data. In fact, there is a 0.91 Spearman and 0.95 Pearson correlation, between the building medians(building area divided by the median building area) and the population, both statistically significant with p-values of zero.

[2]For reference the set of results with a minimum samples parameter of 19,000 took more than a day to process end to end.

Table 5.2: Descriptive statistics for the delineated areas

	Area	Building medians	Density	Cluster size	GHSL Population
mean	295.81	44204.83	170.28	21581.11	65814.96
std	644.42	257823.24	195.23	119552.13	509398.56
min	3.00	3800.00	1.21	77.00	0.00
25%	60.45	6055.00	45.37	3266.00	5955.64
50%	140.95	10404.00	96.74	5443.00	10744.91
75%	298.23	20827.00	219.50	10936.00	23798.59
max	19137.55	8915402.00	2885.53	4042986.00	17001425.48

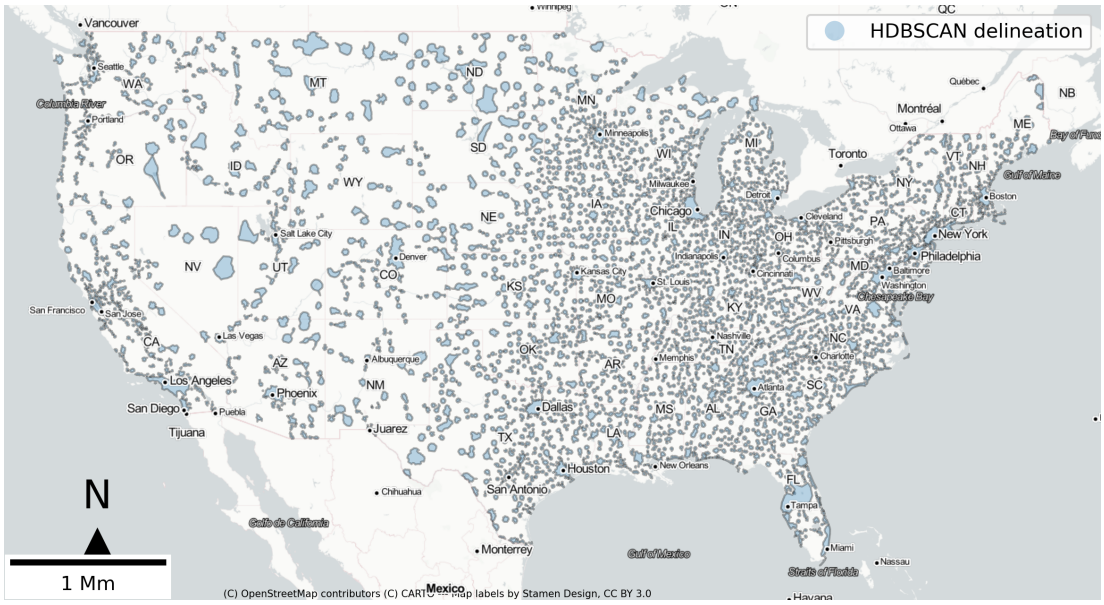


Figure 5.4: Urban HDBSCAN delineations

One way to account for the issue with the zero population is to drop any areas from the results with less than the specific minimum building size, in this case 3800. Weighing the building footprints was done to account for the problems in very dense city centres due to the data shortcomings discussed previously. By weighing the building footprints outside of urban centres, the algorithm can erroneously detect campsites and much smaller urban areas than the desired minimum population of ten thousand. There are 1307 delineated areas that have less than 3800 total footprints and 2702 that have more. There is no need to know the population of the areas in advance, this procedure only relies on the number of building footprints in each area.

The density column in Table 5.2 shows one additional problem. The density column represents the total building area divided by the area of the final alphashape delineation. Low values in the column show that there exist large delineations in terms of area, in sparsely populated places such as Nevada and Montana. This suggests that very sparsely populated small towns in remote areas get delineated as a single entity, which is a result of the algorithm not having a set minimum density threshold. De-

pending on the application, such areas can be further processed or dropped altogether if the analysis requires it. Again, this procedure relies only on information already available and does not require population data. It should also be noted that both this and the previous changes mostly affect areas population below 30,000.

For the rest of the analysis in this paper, the whole unprocessed set of delineations is used. Even though the problematic areas represent a large number of the final delineations, they are a much smaller percentage of the buildings or population which plays a role in the subsequent analysis. Nevertheless, we aim to highlight the potential drawbacks of applying our novel methodology to other researchers. Specific applications of our methodology, for example for calculation of local economic productivity, can use the above two methods to discard the problematic areas. Alternatively, our delineations can be combined with datasets such as the GHSL population grid and problematic areas can be dropped. It should also be noted that increasing the minimum neighbours parameter, decreases the number of problematic areas - for example, the 19,000 set of results suffers less from these problems. Additionally, when calculating statistics such as population density the polygons can be dropped altogether and the groups of buildings can be used directly.

5.4.3 Most populous cities

We focus the first part of the analysis on the 15 largest areas, in order to provide a more detailed view into the delineation process. This is done in two ways. First, we compare them to the 15 most populous cities in the other four sets of urban boundaries - Census Urban Centers (CUC), GHSL, GHS FUA and the official ones. Second, we analyse the density variations within each of the areas to explore the way in which the buildings were combined into clusters. Afterwards, we extend both types of analysis - the comparisons and within density variations - to the whole set of delineations.

Table 5.3 shows the area and population for each of the largest fifteen cities by population. The population is in millions and the area is in square kilometres. Our delineated areas show the most resemblances to the Census Urban Areas and the GHS Functional Areas, both in terms of population and area. It can be seen that 11 out of the 15 cities in our method appear in the top 15 Census Urban Areas. A noticeable difference is the city of Tampa, which in the HDBSCAN delineations is much larger since it forms a large cluster with the the city of Orlando. Similarly, 13 of our delineations are the same as the top 15 Functional urban areas, however the ordering is different.

There are also some similarities to the GHS Urban Centres and official boundaries. 11 out of the fifteen most populous areas are in the set of 15 most populous official boundaries. However, it can be seen that in all cases both the population and the area of our delineated clusters are much larger than the official city boundaries. As before,

Table 5.3: Descriptive statistics of the most populous 15 cities from each delineation

HDBSCAN			TigerLine Urban Centres			GHSL Functional Urban Areas			GHSL Urban Centres			City Boundaries		
Name	**Population**	**Area**	**Name**	**Population**	**Area**	**Name**	**Population**	**Area**	**Name**	**Population**	**Area**	**Name**	**Population**	**Area**
New York	17	8154.5	New York	18.98	9468.8	New York	19.52	17489	New York	15.95	5384	New York	8.16	1212.62
Los Angeles	16.61	10722.7	Los Angeles	12.34	4558.83	Los Angeles	15.66	10407	Los Angeles	14.28	5633	Los Angeles	3.82	1302.05
Chicago	8.66	6877.21	Chicago	8.87	6431.43	Chicago	8.8	13185	Chicago	6.78	3830	Chicago	2.58	606.42
Washington D.C.	7.34	7009.75	Dallas	6.01	4701.97	Dallas	7.08	19826	Miami	5.41	3040	Houston	2.43	1724.87
Tampa	6.61	19137.5	Philadelphia	5.73	5259.33	Houston	6.44	15114	Dallas	5.17	3699	Phoenix	1.58	1343.99
Dallas	5.94	5303.56	Miami	5.72	3400	Philadelphia	6.11	11525	Houston	4.87	3418	Philadelphia	1.5	369.61
Miami	5.72	4368.81	Houston	5.7	4387.78	Miami	5.81	6494	San Jose	4.6	1717	San Antonio	1.46	1208.72
Houston	5.31	4294.67	Atlanta	5.4	6945.19	Washington D.C.	5.64	7446	Phoenix	3.61	2304	San Diego	1.37	963.35
Philadelphia	5.1	4163.61	Washington D.C	5.04	3493.81	Atlanta	5.59	12348	Washington D.C.	3.37	1550	Dallas	1.24	996.67
Atlanta	4.79	7146.11	Boston	4.49	5054.24	San Jose	5.12	4038	Detroit	3.29	2545	San Jose	0.95	467.56
San Jose	4.72	2714.95	Phoenix	4.21	2981.52	Phoenix	4.57	9707	Philadelphia	3.13	1520	Austin	0.91	847.94
Phoenix	4.33	4858.01	Detroit	3.87	3554.79	Detroit	4.07	8184	Seattle	2.68	1885	Jacksonville	0.86	2265.3
Denver	3.55	8033.59	San Francisco	3.35	1380.99	Seattle	3.87	8328	Denver	2.23	1362	Charlotte	0.83	798.21
Detroit	3.33	2785.24	Seattle	3.31	2789.12	Minneapolis [Saint Paul]	3.27	9833	Boston	2.06	936	Indianapolis city (balance)	0.82	953.18
Boston	3.31	2973.01	San Diego	3.1	1971.57	Denver	2.93	6917	Las Vegas	2.03	878	Fort Worth	0.81	914.58

13 of our delineations are the same as the top 15 GHS Urban Centres, but the our delineated areas and populations are again larger and closer to the GHSL functional urban areas.

One of the core advantages of our methodological approach is that specific global density requirements do not have to be set, but are inferred locally from the data. Table 5.4 shows a summary of density statistics within each of the 15 largest urban footprints. The population column is again in millions, while the centres column refers to dense cores within the deliniations that contains at least 3800 buildings. It should be noted that this exact number is a consequence of the choice of parameter. The 'density' columns refers to the number of buildings in a one kilometre radius around each building in the centre. A higher value indicates higher building density, whereas smaller values - lower. For most areas the density standard deviation is around 500 buildings. The 'rank difference' column shows the difference in the ordering of HDBSCAN results, based on the number of dense cores compared to the ordering based on total population.

Table 5.4: Density statistics within the 15 largest delineations

City	**Population**	**Centres**	**Rank difference**	**Density std**	**Min. density**	**Max. density**
New York	17.00	94	-2	787	2538	5621
Los Angeles	16.61	151	1	758	2155	5407
Chicago	8.66	87	-2	911	2032	5430
Washington D.C.	7.34	63	-5	519	1539	3788
Tampa	6.61	114	3	718	1249	4170
Dallas	5.94	82	0	506	2110	4458
Miami	5.72	70	-1	537	1717	4719
Houston	5.31	90	4	599	2283	4784
Philadelphia	5.10	52	-3	882	2190	5918
Atlanta	4.79	54	-1	497	1697	3741
San Jose	4.72	38	-3	223	4178	5097
Phoenix	4.33	79	5	809	1671	4474
Denver	3.55	60	3	1092	997	4603
Detroit	3.33	41	1	488	2661	4517
Boston	3.31	31	0	758	2105	4666

It should be noted that in all cases the dense areas within a delineation exist in

a complex relationship to each other. The lower density centres are not simply extensions to a central dominating one, but have various interactions in the density and distance hierarchy. For example, in the case of New York, the 94 centres merge into 17 areas, which subsequently join together. It is not that case that density simply declines concentrically from Manhattan, and all centres are joined to it.

This table further shows the advantages of using a locally adaptive value of high and low density. Setting a global density threshold of equivalent to 900 buildings per km, for example, is required to delineate the Denver area in its current form, however this value if applied to the whole dataset, it will result in a delineation of the North-eastern megalopolis, which encompasses the area from Boston to Washington D.C. as shown in Table 5.5.

There are several other patterns that can be discerned from the table. In the case of San Francisco/San Jose footprint there is a very small density standard deviation. This suggests a consistent density throughout the area with no dominant centre. This is in contrast to cases such as Phoenix where the standard deviation is larger, which suggests that a lot of local centres are included in the delineation. Furthermore, it can be seen that the Los Angeles footprint has the largest number of centres, at nearly double that of the New York footprint which is slightly larger in population. This speaks to the difference in the density patterns and urban form of the delineations.

5.4.4 Comparison with other delineations

The city comparisons suggest that the most similar delineations to ours are the Census Urban Centre and the GHS Functional Area delineations. To verify this finding we compare the HDBSCAN delineations with the other approaches using the rand index. Since the reference delineations have different minimum thresholds, i.e. an urban area in the CUC has a minimum population of 5000 whereas in GHS - 50,000, we compute rand scores at different population thresholds. To do this, first we compute the population deciles for each delineation type - GHS, CUC, GHS FUA and official boundaries. Then all delineations below the threshold are assigned to the noise cluster, so as to not play a role in the similarity (rand score) calculations. The results are shown in Figure 5.5. The Y-axis shows the similarity (rand score) between the delineations and is fixed, while the X-axis shows the different population thresholds and varies depending on the type of delineation.

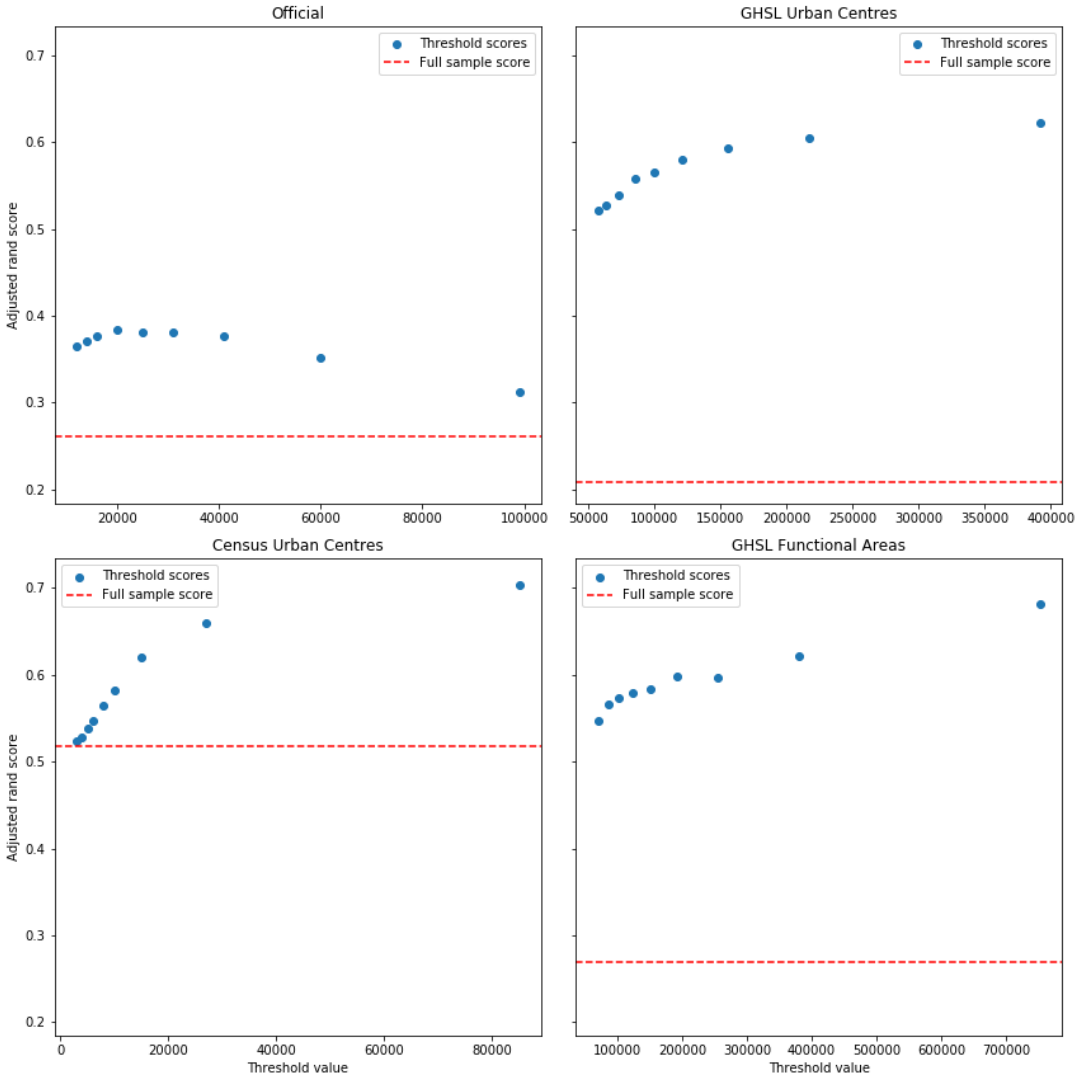


Figure 5.5: Rand Index comparisons between the different delineations at different population thresholds

The results suggest that, similarly to the analysis of the 15 most populous cities, our delineations overlap the most with the CUC and FUA ones. From the specific examples shown in the previous section it seem that our delineations are in the middle in terms of area - larger than GHSL and smaller than CUC and FUA - and as the population in the delineations increases so does the similarity to our delineations. The comparisons with Census Urban Centres show the highest overall rand scores, suggesting the most overlap.

Additionally, as the minimum population size increases the scores get higher. This suggests that our larger delineations are more similar to other types of large urban areas, whereas less populous delineations are more different. Interestingly, this pattern is reversed for the official boundaries - our larger delineations are much larger than the official administrative boundaries of large US cities. The most similar boundaries to the official ones are of small cities on average, with a population between 20,000

and 40,000, since after 40,000 the similarity starts to drop, while it increases until that point.

5.4.5 Spatial analysis across datasets

To further explore the set of delineations, three types of purely spatial comparisons are calculated between the HDBSCAN and CUC delineations. The first type is the number of HDBSCAN polygons that fully fall within corresponding CUC polygons. The second set is the opposite - the CUC polygons which are within the HDBSCAN ones. Third, the remainder of the polygons from the two types of delineations are analysed using information about their intersections.

There are 313 HDBSCAN areas that fall fully within a corresponding CUC area. This set of HDBSCAN delineations are mainly located within denser large areas, such as the New York CUC area, evidenced by the fact that the CUC areas which contain them have an average population size of 3,719,585. This result reflects a property of our method that does not just merge cities based only on geographic proximity. These 313 areas are separate in the final results, since there exist large deviations in local density.

There are 1810 areas in our delineations which fully encompass CUC delineations. These are for the most part smaller areas with the exception of Miami, Florida and Denver, Colorado. In fact, 75% of the CUC delineations have less than 11,318 thousand population. This suggests that our method overestimate the spatial extent of small urban settlements. The result can be attributed to our method strictly delineating areas with high density apart from areas of low density. Due to the very low built environment density these areas are in, the methodology can merge several of them, such as small towns, together. Furthermore, poor data quality can play a role as well, and a large proportion of these delineations are the problematic areas discussed previously.

The last set of spatial intersections is the polygons that have intersection relationships, and which are not part of the previous two sets. These were calculated as the percent overlap between the clusters as measured by the intersection of the two polygons to the ratio of the union. The percentage of overlap goes from a low of 8% to a high of 90%, with more than fifty percent of the points having atleast 27 % overlap. The mean population for these areas is around 381,200 thousand people, meaning that they are larger in terms of population than the previous within set. In addition, there are several larger boundaries that almost fully overlap such as Chicago, Atlanta, Houston and Dallas all with ovelaps above 75%.

5.4.6 Density variations within delineated areas

To generalise the analysis of density variations to the whole dataset and contextualise the results, we limit the analysis to areas that cross metropolitan statistical areas (MSA) boundaries. This is done since there is a known overall trend of employment decentralisation within US metropolitan areas (Dadashpoor and Malekzadeh, 2021) and this phenomena provides context against which we can compare our results.

Figure 5.6 shows the number of MSAs which contain delineated areas, where the largest delineated area is at most a specified threshold of the total delineated land mass within the metropolitan boundaries. The figure shows that for the majority of cases there is one dominant urban footprint delineated for each MSA. Furthermore, the MSAs without a dominant delineated area are smaller in terms of population. For example, when the threshold is specified as 50% , only 95, or less than a third, of MSAs have multiple large delineated areas within them. In other words, there are 281 MSA where a single delineated area covers more than 50% of the total urban footprint land mass within an MSA. The majority of the 95 areas have an estimated GHS population under 231,170. Only 11 of the 95 areas from the first set have population more than a million as opposed to 44 of the other type.

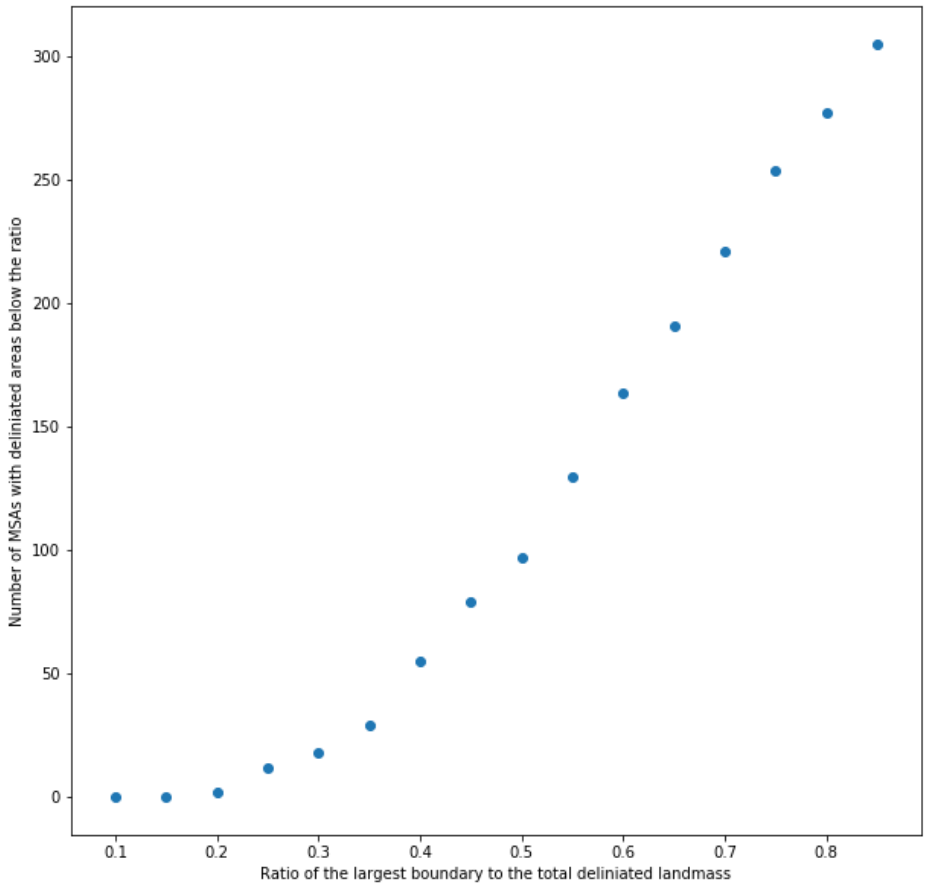


Figure 5.6: Number of Metropolitan areas, with a ratio of the largest delineated area below a threshold

The clustering hierarchy makes it possible to explore the number of dense areas inside each urban footprint and their relative importance, as well as how they evolve to form the whole footprint. The delineated areas themselves cross metropolitan boundaries and can have a multi-modal density. For example, the metropolitan areas of Washington-Arlington-Alexandria, DC-VA-MD-WV and Baltimore-Columbia-Towson, MD are delineated as a single cluster. Out of the 2259 urban footprints that fall within metropolitan areas 258 cross two metro boundaries, 22 cross three, 3 cross four and 1, the Tampa urban footprint crosses 8. And out of the 2259 urban footprints themselves 388 are identified as having multiple dense cores in the clustering hierarchy. A dense core is an area with at least 3800 buildings, which is denser than the surrounding area, where 'denser' again is a locally dependant value. The value of 3800 is due to the choice of parameter for the HDBSCAN algorithm. However, the majority of the population in all delineated areas - 77% - are in delineations with multiple dense

cores, whereas only 13% live in delineations with a single dense core. The rest of the population is located in delineations that do not intersect metropolitan statistical areas, the majority of which contain only one dense core themselves. It should be noted that this pattern holds across the range of different values of the minimum samples parameter.

5.4.7 Megaregions

Functional areas, metropolitan areas and urban census areas are some of the building blocks used to delineate larger scale units such as megaregions. In fact, the GHS, FUA and CUC already almost overlap from Boston in the north to Washington in the South, which corresponds one possible definition of the Northeastern megaregion. None of HDBSCAN results, using any of the tested minimum samples parameters are of a megaregional scale. This suggests that the local variability in the density of buildings is too large for megaregions to be delineated. In other words there are sharp declines in building density past the borders which resemble the functional area scale, since our delineations are most similar to them. However, by using the megaregions definitions from Hagler (2009), described in the data section, and the nested clustering hierarchy we can rank the potential morphological integration of all buildings in the dataset into megaregions.

Table 5.5 shows the megaregions and the corresponding density at which they would cover all the counties identified in the America 2050 (Hagler, 2009) definition. The entire set of building footprints was processed in order to calculate the densities, not just the building footprints in the already delineated 4009 areas. The minimum density (minimum buildings per km) value was calculated using the hierarchy and growing all delineations by individual buildings, or merging pairs of delineations together, until a delineation covered or exceeded the extent of a specific (Hagler, 2009) megaregion.

Table 5.5: Number of delineated areas within megaregions and density statistics

Megaregion	**HDBSCAN delineations**	**Minimum buildings per km**
Northeastern Megalopolis	300	970.72
Atlantic Piedmont	182	736.91
Cascadia	94	578.93
Midwest	789	567.14
Florida	57	514.68
Texas Triangle	193	428.63
Arizona Sun Corridor	33	263.89
North California	125	225.52
Frontrange	36	197.15
South California	64	170.13
Gulf Coast	100	155.93

Several interesting findings emerge from the exploration of the hierarchy which suggest that is impossible to delineate some megaregions based only on the density of built up area. The megaregions that can exist as morphological entities based on our approach and data are Northeastern, Atlantic Piedmont, Arizona, Texas Triangle, Cascadia, Florida and Gulf Coast. Out of those, the most morphologically integrated region is the Northeastern Megalopolis, while the most disintegrated is the Gulf Coast one. In order for an urban footprint to exist that covers the extent of the other megaregions it necessarily encompasses much larger areas. For example, the table shows that for an urban footprint the size of the Midwest megaregion to emerge, parameter choices have to correspond to approximately 567 buildings per km. However, at that point the Midwest megaregion becomes integrated with the Northeastern Megalopolis. This is also the case for the Front range megaregion - in order for it to exist the density has to be set so low as to cover the whole of the eastern United States as one urban footprint. Similarly, for Southern California to exist and encompass Las Vegas the whole of the Midwestern United states has to become one footprint. Lastly, in order for the Northern California megaregion to exist it has to encompass the entirety of California.

5.5 Discussion

The results show that this approach is able to capture meaningfully urban settlements in the United States of America, which in the absence of pre-specified thresholds, models or spatial aggregations, are much larger than official city boundaries. On average they are most similar to Census Urban Centres and GHS Functional Urban Areas, delineations designed to capture functional urban extent. However, the size and scale of the delineations varies in different parts of the country. Furthermore, even the density in different parts of an urban area varies. For example, there is so much consistently dense built up area between Washington and Baltimore that there is no sharp drop in local density which separates them, suggesting therefore that the cities have started growing into each other. The dense areas within the Washington D.C/Baltimore delineation, have different local densities of their own, whereas dense areas within the San Francisco/San Jose delineation have consistent density.

An analysis at the metropolitan level found that majority of metropolitan areas contain one dominant delineated area, however the majority of areas themselves contain multiple cores - areas denser than their surroundings. The results also show that there are large urban clusters forming at a scale larger than metropolitan areas and in some cases combined statistical areas. Nevertheless, although some cities are merging morphologically, in all cases the scale of the delineated areas is much lower than proposed megaregions. In fact, further analysis showed that certain megaregions cannot be delineated morphological based on our building data, since they are necessarily must

include much larger areas.

In general, the delineated areas are much larger than the official boundaries, similarly to the results of other delineation approaches (Duranton, 2021). The scale of the resulting areas ranges from large clusters of cities - i.e. one area covers both Washington D.C., Alexandria, Fairfax and Baltimore - to small towns with populations of less than ten thousand like Butler, New York.

Several studies have emphasized the difficulty of automatically detecting built-density changes when using intermediary aggregated units (Duranton, 2021; de Bellefon et al., 2019). In addition, other studies emphasize the difficulty of selecting a population or built-up density threshold manually. Balk et al. (2018) find that how much of the official urban population in the US is captured by the GHSL delineations is affected by threshold choices and that global thresholds have difficulty capturing peri-urban and ex-urban areas and mid-sized cities as a whole. de Bellefon et al. (2019); Statham et al. (2021); Arcaute et al. (2016) show similar results with different datasets and methods - explicit or implicit density thresholds affect what urban forms are included in the final results as well as their size. Table 5.1 showed that an increase of 1,000 % in the only parameter which needs to be specified resulted in a difference of 10% in the number estimated urban population, which demonstrates the robustness of our method.

Furthermore, the comparisons carried out in this paper show that our delineations capture gradual local building density changes at different scales. The final delineations are much larger on average, based on the comparison in Table 4.3, than the official boundaries and even larger than dense urban cores delineated with explicit thresholds. For example, in our results the cities of Baltimore and Washington D.C fall within a single delineation. These results further echo findings that the peripheries of some neighbouring urban areas have grown both in size and density so much that they are clashing against each other forming much larger urban units (Hagler, 2009; Lang et al., 2020).

The comparison results show that on average, the emergent scale from our analysis is most similar to that of functional urban areas and census urban areas - delineations designed to capture local economic activity and labour markets. Direct building comparisons showed that our delineations are highly similar to Census Urban Areas , with a rand score ranging from .55 for the whole dataset to 0.7 for units with more than 100,000 population. There is also a high similarity with GHS functional urban areas - with rand scores between 0.6 and 0.69 for places with more than 100,000 people.

The similarly between our delineations and other sets of results designed to capture local economic activity, as well as the high correlation between population and buildings, suggests that building density variations captures some information about spatial patterns of local labour markets. Employment patterns have a complex relationship with city size, population, built environment and distance to central businesses districts

(CBDs) (Craig et al., 2016). Nevertheless, detailed built environment data can provide information about employment monocentricity, polycentricity and scatteration patterns (Taubenböck et al., 2017; Krehl, 2015). Further comparisons showed that the majority of metropolitan areas in the contiguous US contain one dominant delineated area with many of these areas crossing the census defined metropolitan boundaries. However, the majority of population is found in delineations built up from multiple dense cores, which suggests that some form of population decentralisation dominates US urban life. Our results also suggest that the majority of the population is found in areas built up of multiple dense cores in a complex relationship to each other, even when the area under analysis is delineated with no reference to other existing units or scales or explicit density thresholds, which can affect polycentricity analysis (Möck and Küpper, 2020). This finding agrees with extensive body of literature that directly measures decentralisation of US employment through other delineations, data and methods (Manduca, 2020; Dadashpoor and Malekzadeh, 2021).

In spite of the similarity with GHSL Functional Urban Areas and Census Urban Areas, some of the results are more similar to other different types of delineations as shown in the table of the top fifteen more populous areas - Table 5.3. For example, pairs of major cities such as Baltimore and Washington D.C., San Francisco and San Jose, Rayleigh and Durham are merged together and their combined population resembles combined statistical areas rather than individual urban areas. In contrast New York and Long Beach, Santa Cruz and Watson Ville, Vallejo and Fairfield are split into separate areas smaller than metropolitan areas in population. Not all of these mergers or splits are the same in the GHS Functional Areas or the Census Urban Centres. There are also many cases where small towns are separated from the major urban aggregation, as shown in the spatial overlap analysis. This shows that our results vary spatially and show the advantage of out method in taking local context into account.

When looking at built density within individual delineations, shown in Table 5.4, there is a complex gradation of built density changes. This means that our results incorporate diverse urban forms of varying densities, in contrast to methods that have specified a priori density parameters which can exclude them. This is not possible to capture with preset global thresholds without changing the scale significantly for at least some delineations, e.g. the New York one. Furthermore, the density variations within specific delineations, shown in table 5.4 are reflective of labour patterns in specific places. The Phoenix delineations has the highest number of dense cores relative to its population, while the LA area is the one with the most total cores by far. This is reflective of the both economic labour and sprawl patterns in those specific areas (Dadashpoor and Malekzadeh, 2020; Ewing and Hamidi, 2015; Barrington-Leigh and Millard-Ball, 2015; Arribas-Bel and Sanz-Gracia, 2014; Angel and Blei, 2016). However, further research is required in classifying all of the delineations. For example,

some interesting cases are the San Francisco/San Jose delineated area which was has a consistent density pattern, Washington and Baltimore which have less dense cores than delineations with similar populations and Houston being one of the most polycentric areas.

The final part of the analysis focused on the relationship between the delineated units. The non-existence of delineated areas at the megaregional scale indicates that there is too much variability in the local built environment density for these large urban units to appear. This result holds for the full range of minimum size parameters tested. Some cities such as Baltimore and D.C. merge together and form single delineations which span thousands of square kilometres, however, the required large scale high-density areas to cover proposed megaregions, such as those in Hagler (2009) do not emerge in our results. The analysis shown in 5.5 suggests that based on our approach and data, in order of relative morphological integration, the megaregions which can individually emerge are: Northeastern metropolis, Atlantic Piedmont, Cascadia, Florida, Texas Triangle and Arizona Sun Corridor.

In addition, the densities analysis presented highlight the difficulty of selecting a density threshold explicitly, similarly to other research (de Bellefon et al., 2019; Duranton, 2021; Statham et al., 2020, 2021; Balk et al., 2018). This is especially evident in the North East of the US. If the delineation procedure relied on explicitly defining minimum density, relatively small increases of around 200 buildings per square kilometre can result in large scale difference such as the emergence of megaregions or even one area which covers the whole of the North East and Midwest. Overall, this sensitivity to small parameter changes reflects the high-density of the area and its interconnectedness, however the fact that there are no emergent megaregions shows that more local density variations dominate.

Interestingly, our results also suggest that certain America2050 (Hagler, 2009) megaregions cannot be delineated at all. These are the Midwest, South California, Front range and Gulf Coast. In order for any of these to exist, based on our method and data, it has to merge with other areas in the United States, much larger in scale than megaregions. In total our results, echo other functional analysis delineations such as Nelson and Rae (2016), that show that there are large-scale emergent patterns, but at a scale, smaller that the one proposed in (Hagler, 2009). Furthermore, the results in general agree with research (Glocker, 2018; Lang and LeFurgy, 2003; Nelson and Rae, 2016; Ross et al., 2009) that propose smaller scale megaregions than the America2050 project.

5.6 Conclusion

The results from the paper have several implications for researchers and planners. First, the inability to delineate the Midwest, South California, Front range and Gulf Coast megaregions, suggest that the building density is lower within these proposed megaregions that between parts of them and parts of other megaregions. For example, there is significantly higher building density between parts of the Midwest and Northeast, than within the Midwest. This result is important for researchers and planners that use other higher resolution morphological data such as population grids or census tracts to delineate these megaregions - e.g Hagler (2009); Glocker (2018); Ross et al. (2009); Georg et al. (2018), since it highlights potential MAUP problems (Openshaw, 1979) and that at the highest resolution level there are signification density variations within proposed megaregional boundaries. This inability to delineate the megaregions could also be due to a breakdown in the relationship between form and function, in which case, the results highlight the need for functional data such as commuter flows, to be incorporated in megaregional delineation and analysis.

Second, the results highlight the fact that consistent building density is larger than official city boundaries and that multiple delineations cross statistical areas boundaries. Therefore, applications which emphasize morphological aspects of urban phenomena could make direct use of our boundaries in order to contextualise their results and explore the influence of statistical boundary choices and fixed density thresholds - i.e de Bellefon et al. (2019); Arribas-Bel et al. (2021b); Duranton (2021). The polycentricity analysis carried out in the paper is an example of how this type of work could be carried out.

Third, the results highlight the usefulness of buildings and provide a good case for investment in building-level datasets for the purposes of urban planning and research. The results from this paper speak to their usefulness in analysing the extent of cities, megaregions and polycentricity. Other researchers have already used them to create high resolution population density estimates (Huang et al., 2019). Developments in this area - population estimation using building footprints - are particularly important for developing countries where high-quality population data is unavailable (Wardrop et al., 2018). Additionally, if building heights or building functional data are available they can be incorporated in the delineations similarly to the polygon area, as (Arribas-Bel et al., 2021b) and Huang et al. (2019) do respectively .

Fourth, the usage of buildings as the units of the delineations enables the global applicability of the methodology with several considerations. First, the particular advantages and disadvantages of the specific buildings dataset have to be analysed. For example in the case of the dataset used in this paper, the buildings in city centres were of worse quality than the other buildings, whereas this will not be the case for offi-

cial cadaster data (de Bellefon et al., 2019). Second, different parameter choices of the algorithm need to be explored. In this paper, the parameter choices past a certain threshold - 3800 - proved to be more stable than the results below it , which were affected by natural barriers. However, in other geographies this result could be different.

Lastly, the identified limitations of the methodology have to be addressed. One of the limitations discussed in the paper is the problems with merging small towns in sparse areas together, due to the low surrounding built density. Both the spatial and the population analysis showed that problem delineations are in very sparse areas and have a population of less than 30,000 people. In this analysis the drawbacks of the method were highlighted for the benefit of other researchers and problematic areas were not dropped. However, for some applications for example scaling law analysis (Batty, 2006) or anything to do with population density (Henderson et al., 2021), the highlighted areas could be influential to the study. We proposed two ways to tackle these, relying only on building polygons, however some analysis might benefit from data fusion approaches that combine the delineations with other data such as population or amenities. Another promising approach is to use a more sophisticated cluster extraction scheme. Different methodologies have started being proposed such as ways to speed up the calculations for a range of parameter values at once (Campello et al., 2020).

6

Conclusion

This chapter provides a summary of the results from the analytic chapters and discusses the outcomes, challenges and potential improvements of the thesis. The next section broadly summarises the main results and limitations of each analytic chapter. The two subsequent discussion sections combine the results from the thesis with a focus on the delineation problem and the interplay between spatial boundaries, urban challenges and new forms of data. The last section concludes the thesis with a short overall summary.

6.1 Summary of results

The third chapter addresses the first aim of the thesis - *To identify areas with different urban land use patterns at a high spatial and temporal resolution using sound sensor readings.*

There are two general findings and one limitations of chapter three. First, sound sensor data, specifically maximum decibel recordings across time, capture activity patterns of the area surrounding the sensor. Second, the novel topological data analysis techniques performed worse than other methods in the comparison. The main limitation of the chapter was the number of sound sensors analysed and their placement, which limited the generalisability of the results.

The fourth chapter addresses the second aim of the thesis - *To identify potential megaregions, as well as to explore the distribution of people within them.* There are three main findings in chapter four and two limitations. First, three emergent scales of interaction were identified in the origin-destination employment data (LODES) - one corresponding to metropolitan/combined statistical areas, one corresponding to states and one super-state scale. Second, there is evidence of emergent economic integration at the megaregional level for six out of the eleven megaregions proposed by Hagler

(2009). Third, the results from the spatial analysis of employment suggest that at the first hierarchical level, where the scale is defined by the extent of the detected communities and the units are the census tracts, the majority of employment in the US is decentralised.

The first limitation is related to the data used in the study, while the second to the methodology. First, the community detection approach used in this chapter relies only on one type of functional relationship - employment. The second set of limitations relates to the specific use of modularity as a measure of emergent community strength and the limited set of comparisons carried out.

The fifth chapter addresses the third main aim of the thesis - *To delineate urban areas morphologically and to explore the spatial structure of density variations within the resulting boundaries.*

The main result of the analysis is that it is possible to capture accurate urban footprints using only individual building polygons derived from satellite images. Furthermore, on average, there is a decrease in building density at a scale most similar to functional urban units. A secondary result, is that the majority of delineated areas have one dense core, however over 70% of population lives in delineations with multiple dense cores which exist in a complex relationship to each other. Lastly, none of the footprints were at the megaregional scale and interestingly, some megaregions cannot be delineated based on building density at all.

Similarly, to chapter three the main limitations of the analysis are related to the data - only two-dimensional building footprint data is used; and methods - issues with delineating sparsely populated areas.

6.2 Urban delineations, form and function

The main focus of the thesis and a common theme across all three analysis chapters, is the classification of data and delineation of boundaries, in order to facilitate the analysis of different urban phenomena. Chapters three and four deal with functional urban behaviour - captured sound patterns produced by urban activity and home and tax addresses respectively, whereas chapter five deals with form - building footprints. Integrating the different results into a coherent whole and drawing overall conclusions is challenging due to the specific operalisation approaches, scales of the results and types of data used. However, there are two ways in which the results from the chapters can be combined.

First, the results provide a more detailed picture of large-scale economic integration, the growth of cities, population and employment decentralisation and the relationship between form and function. Chapters four and five show that overall, there is only limited evidence of emergent megaregional structure in the US. Furthermore,

in both chapters urban areas extent beyond official boundaries and have decentralised urban spatial forms. Results from chapters three and five show evidence in favour of a complex relationship between urban function and form - sound sensor patterns change depending on their location on a street and building densities drop at the boundaries of functional urban areas, on average. However, there exists multiple differences between the two sets of results. These results are analysed in more detail in the next section.

Second, the results from the chapters provide insights into the disadvantages and advantages of a specific approach towards delineation at different scales - one based on density of connections, with few model restrictions. Furthermore, the analytical chapters focus on building on the limitations of previous operationalisations of specific phenomena and the discovery of unexpected behaviour when constraints are relaxed. Chapter five operationalises a minimalist definition of urban areas based on density without using intermediary aggregations or fixed density thresholds. In chapter four, data sources and methods different to previous research - Nelson and Rae (2016) and Arribas-Bel and Sanz-Gracia (2014) - are used in order to create comparisons for megaregional delineations and labour spatial patterns respectively. Similarly, chapter three uses novel data and methods - sound sensors and topological data analysis - to detect profiles of urban activity.

Nevertheless, the interpretation of these results comes with caveats. Both chapters four and five deal with data that spans the whole of the contiguous United States, however on average, chapter four detects much larger emergent units than chapter five and furthermore, the two chapters differ in their conclusions regarding specific megaregions in the US. Part of the reason for this is the large difference in the core units of analysis - chapter four uses census tracts, whereas chapter five - individual building footprints. This is a data limitation issue since the flow data is not available at the building level due to survey costs, privacy concerns and data accuracy problems. The differences are also in part due to the focus of the analysis - functional versus morphological. Integrating these results is still possible if all of these considerations are taken into account and is further explored in the next section. In contrast, integrating the conclusions of chapter three with chapters four and five is more difficult, even if the same methodology is applied in the US instead of the United Kingdom. This is because neither of these two chapters takes into account time, which plays a central role in chapter three.

6.2.1 Urban form and function

There are three main conclusions which can be drawn from the combined results, which relate to the urban form and function debate. First, sound patterns are affected by urban form. Second, there is a general pattern of population decentralisation in

the US, reflected in both the morphological and functional data. Third, for the United States, chapters four and five show only limited evidence for the emergence of functional integration at the megaregional level, and none for morphological integration.

Overall, the results from the thesis suggest that urban function is reflected in urban form to various degrees, however the relationship is complex. Chapter three provides evidence that some street sound sensor patterns, which are similar to other human activity data, show strong spatial autocorrelation and change in activity patterns at street intersections. This is an important result due to the wide usage of the street network to infer functional behaviours, such as delineations of cities or population density estimations (Arcaute et al., 2016). The fact that sensors along the same street are grouped together, is inline with urban sound use literature that shows that the built environment has an effect on urban sound (Zuo et al., 2014). Furthermore the analysis of the points of interest present alongside different clusters, explicitly showed that the urban infrastructure - roads, transport stops, buildings, etc - differs based on the different identified activity profiles. However, it should be noted again that the generalisability of these results are limited due to data coverage.

The results from chapters four and five are also related to the function-form debate. The delineations of chapter five are on average most similar to functional urban areas, without incorporating any functional data. The first level of delineated units in chapter four resemble MSAs. Both of these functional and morphological delineations are larger than official city boundaries, which shows that the growth of cities is reflected in both form and functional data. This is consistent with the general delineation literature (Duranton, 2021) and is one of the core reasons for the importance and growing interest in urban delineation approaches.

Furthermore, the results from chapter four and five also show that the concentration of labour pools are at least somewhat reflected in building density fluctuations in the United States. First, both chapters show general agreement about the spatial distribution of people within the delineated areas themselves. The results from chapters three and four show that decentralised urban forms dominate in the US, and this is reflected in both functional and morphological data. In chapter five, the majority of urban land in the majority of metropolitan areas was contained within a single footprint. However, the largest delineations, which account for more than 70% of the estimated population, had multiple dense cores. Similarly, the majority of employment activity in chapter four was decentralised at the first emergent scale, which is similar to the existing metropolitan level. These results are consistent with the wider literature (Dadashpoor and Malekzadeh, 2021; Arribas-Bel and Sanz-Gracia, 2014), even though polycentricity analysis was not the main focus of the chapters and both of them had significant methodological and data differences with other studies. Chapter four and five did not specify a scale of analysis a priori, but instead inferred it from the data. In numerous

studies, changes in scale lead to changes in characterisations of polycentricity - it can increase, decrease or disappear (Möck and Küpper, 2020; Hall and Pain, 2006; Wang et al., 2019). Therefore, the results from both chapters are an additional empirical example in the literature in favour of urban decentralisation.

Additionally, in both chapters there is evidence that there is little morphological or functional integration across state boundaries at the megaregional scale. This is an important result for many researchers which take the emergence of economic or other interactions at the megaregional scale, as evidence that focused planning at this scale is required (Sorensen and Labbé, 2020). Furthermore, this result is also reflective of the diverse administrative and political structures that exist within proposed megaregions (Glass, 2014). The only delineated megaregion that crosses state boundaries is the Cascadia megaregion, which is delineated in chapter four only. Furthermore, there is agreement between the two sets of results, that two megaregions - Texas Triangle and Arizona Sun Corridor - are relatively well-integrated compared to the rest. These agreements show that the relatively limited cross border interactions present in the home-tax functional data are similarly reflected in the density of the building polygons data.

There are also differences between the two sets of results in chapters four and five, reflective of the complex relationship between form and function, as well as the differences in data and methodology. First, the analysis of functional data shows stronger signs of scatteration for the majority of people. Second, the number of centres identified in the fifth chapter is by and large greater than the number of centres identified in the fourth chapter. This difference is reflected in other research which finds that morphological decentralisation is larger than functional - (Burger and Meijers, 2012; Taubenböck et al., 2017). In part, these differences are due to the particular methodology, core units and modifiable area units problems. Another factor is that the morphological data does not differentiate between residential and office buildings. In spite of these underlying differences, in some cases such as the Bay area, both approaches are in agreement that there exist strong scatteration patterns.

These similarities and differences are also reflected when a comparison is made at the megaregional level. The Texas Triangle and Arizona Sun Corridor are the least morphologically integrated of the possible megaregions in chapter five, whereas they are as functionally integrated as the other single state megaregions. In chapter five, the hierarchical analysis shows that of the proposed megaregions the 'Northeastern Metropolis', which spans nine states, is the most morphologically integrated one and the second megaregion in the ranking is the 'Atlantic Piedmont' which spans five states. The analysis in chapter four, showed these megaregions as less functionally integrated than any of the single-state megaregions or Cascadia.

6.3 Delineation approaches and new forms of data

6.3.1 Hierarchical approaches

Although the methodological approaches used differ, their overarching aim is to define final spatial boundaries based on a nested hierarchy. The position of constituent units within this nested hierarchy is based on similarity and is central to the grouping of units. None of the methods used in any of the three papers rely on global parameters - there is no value specified beforehand, which determines whether two units should be grouped together. The boundaries derived from buildings, sound sensors and employment/residence relationships are based on local variations, inferred from the data and there are no spatial models or a final number of delineations specified. The results of the three chapters show the viability of creating delineations without these constraints. Furthermore, they show the advantages of calculating the final delineations and relationships between them based on hierarchies.

The hierarchy provides a readily available 'history' of how the final delineations were computed and why certain core units were excluded. The relationship between the units in the hierarchy played an important role in all chapters. In the third chapter, the hierarchy was used to analyse the sound sensors placed outside clusters. In the fourth chapter the delineated units at the first level of the hierarchy, which are similar to metropolitan statistical areas, were used to analyse the spatial patterns of commuters. Similarly, in the fifth chapter, the hierarchy was used to analyse the variations of building density within the individual delineations, as well as integration of the delineations into megaregions. A limitation of the third chapter was that there was not enough data to create a national hierarchy. However, the land use delineations are derived in exactly the same as in the fifth chapter and with more available data, the same types of analysis can be applied.

Taken all together, the results also show the advantages these types of delineation methods, and the importance of advances in this area. In chapter three, additional areas, representing different activity patterns, were found compared to analysis which use Twitter and mobile phone data, such as Furno et al. (2017). The first level of delineations derived in Chapter four, shows a different pattern of spatial distribution of employment to that found by Arribas-Bel and Sanz-Gracia (2014). This is due in part to the fact that spatial employment analysis and polycentricity research in particular are affected by both the delineation of the areas of study (i.e. metropolitan statistical areas) and the delineation of the centres within them (Möck and Küpper, 2020). Chapter five's delineations of urban areas show differences in population counts and density to even the most similar other delineations. All of these differences affect all comparisons between cities and all downstream analysis such as policy evaluation or

economic productivity research.

6.3.2 Bounded territorial units

Furthermore, the calculated hierarchies provide a way to address some criticisms of bounded delineations in general, such as Christopher (1965); Alexander (2017), that emphasize that no spatial process is completely isolated within its boundary. This is most evident in the fourth chapter, where there are multiple links crossing the delineated boundaries and connecting them into a larger network. In fact Nelson (2020) argues that the resulting units explicitly 'retain the both practical and conceptual utility of distinctly bounded entities without needing to assert that these territorial borders are inimical to patterns of flow and connection'.

The delineations of chapters three and five can be adjusted to incorporate overlaps in a similar manner. In the fourth chapter the notion of similarity is based on the density of buildings in geographical space and their geographical distance (given a map projection). In the third chapter similarity is measured by the density of sensors in an ambient 'sound sensor data space'. The position of each sensor within this space is based on the recorded sound pattern and similarity to other sound patterns, which is defined by a correlation distance metric. In both chapters a hierarchical graph of connections is built based on the density and distance between units (buildings or sound sensors). The final delineations are based on positions of units within this hierarchy and a minimum (local) density. In addition, the HDBSCAN approach leaves some units as noise, which are in fact reachable within the hierarchy by the final delineations. These can be potentially assigned to one or more final delineations, thus creating zones of overlap. In chapter five examples of this could be less densely populated places between two major urban areas and in chapter three - sound sensors on the intersections of streets that have activity patterns patterns that fall between two neighbouring areas. This approach can be further extended to define areas of overlapping density, centred on core units, again based on the hierarchy at every step of the deliniation procedure.

6.3.3 Validation and reproducibility

Taken together, the results from the analytical chapters also highlight the importance of validation and testing when creating delineations using new forms of data or approaches. Improvements in the availability and reproducibility of delineations can improve research results and outcomes by making it easier to connect the outcomes of an analysis to theory and other results from the literature (Wolf et al., 2020). In this thesis, the majority of analysis time was spend on the validation and interpretation of the results. Furthermore, as the analytic methods used become more complex, the importance of validation, testing and comparisons grew. This was especially relevant

in chapter three, where only OpenStreetMap data was available for external functional validation of the results. Similarly, in chapter four although there are numerous delineations of megaregions available in the literature, there are few publicly available shapefiles of them. The introduction of more comparisons, which can provide richer contextualisation, would strengthen the conclusions of the analysis in both cases. The wider availability of urban delineations is what made the relatively more numerous comparisons in chapter five feasible. The usage of data products and better tools, such as coding libraries can help deal with these issues and save researchers computational costs and time spent doing data preprocessing (Singleton and Arribas-Bel, 2021; Arribas-Bel et al., 2021b; Calafiore, 2021).

6.3.4 Advantages and disadvantages of using new forms of data for delineations

In general, all delineation approaches followed the Geographic Data Science (GDS) framework - integrating geographical knowledge, computational power, data and methods (Singleton and Arribas-Bel, 2021). In all three chapters numerous ideas and methodological approaches could not be directly applied due to computational requirements. Therefore, a significant proportion of analysis time was spent on dealing with computational complexity. This is one of the challenges of using new forms of data identified by Arribas-Bel and Tranos (2018).

Chapter three adopted a novel dataset and novel methodologies. Through this approach, the analysis resulted in the detection of areas with activity profiles which similar methodologies and research did not differentiate. Furthermore, the comparison carried out demonstrated that new methods are not necessarily the best performing when it comes to the analysis new forms of data - the results showed that TDA methods perform worse than other more widely adopted approaches. However, the sound sensors dataset used was limited to two months of data across several neighbourhoods, which limited the generalisability of the results.

Chapter four aimed to address the generalisbility issue through the use of the LODES data. Out of all data used in this thesis, LODES is the most widely used and tested dataset by the wider research community. The limitations of the TDA methods influenced the choice of community detection algorithm as well. The Louvain algorithm has found successful application in various fields and was explicitly designed for new forms of data - to analyse large amounts of mobile phone calling records (Blondel et al., 2008; Rahiminejad et al., 2019). Through this combination of data and methodology it was possible to identify emergent structures in US employment patterns at various scales. However, the methodology suffered from several drawback identified in the limitations section of the chapter. Issues stemmed from the direct application

of methods not customised to the particular task of employment pattern analysis - the similarity metric used by the Louvain algorithm itself was developed for abstract graph comparisons (Newman, 2006),

The fifth chapter aimed to address both sets of limitations in chapters three and four - the data set used has good coverage and the methodology is explicitly tailored to the dataset and aim - to operationalise a definition of urban areas with few imposed restrictions. The analysis is based on a novel dataset with national coverage in contrast to the limited dataset in chapter three. However, this data did not come without issues - its large size and inaccuracies in dense urban areas. To address these a machine learning algorithm had to be adapted in accordance with the aim of the project. Existing open source implementations of the clustering algorithm used - HDBSCAN - could not be directly applied to the dataset due to its size, and the inability to incorporate building footprints' areas. Furthermore, specific properties of geographic distance and map projections had to be exploited in order to make the required calculations computationally feasible. Specifically, geohasing and planar projections helped in calculating the nearest neighbours of the points and their first connections in the hierarchy, which significantly reduced the computational time and resources required, thus making the analysis possible to be carried out on a moderately powerful desktop computer (Garcia-Pulido and Samardzhiev, 2022).

6.4 Empirical, theoretical, technical contributions

The combined results from the thesis have several empirical, theoretical, technical implications. First, each chapter provides ready delineations, which can be directly used for the analysis of different urban phenomena at various scales. The second set of empirical contributions concerns the results of the analysis. Chapters four and five provide further functional and morphological support respectively, for the decentralisation of employment patterns in the United State (Dadashpoor and Malekzadeh, 2021), for the growth of cities beyond their urban boundaries (Duranton, 2021), as well as for the large scale integration of the Arizona Sun corridor and Texas Triangle megaregions. Furthermore, the results from all chapters show the increasingly complicated relationship between urban form and function - certain aspects such as decentralisation and large scale integration and small scale functional activity are reflected in both, however differences do exist.

The theoretical contributions are primarily concerned with the usage of hierarchical methods and new forms of data. First, maximum decibels from sound sensors aggregate into profiles similar to those in other activity data - mobile phones calling records, app data, etc. As such this data could be used in place of other information and captures patterns of human activities at high spatial and temporal resolutions. Second,

the thesis emphasized the usage of hierarchical clustering methods since they enable the validation of, not just the final results, but also intermediary units. Furthermore, the hierarchical methods provide less constrainted ways of operationalising different theories - this was demonstrated in chapter five, where a definition of urban areas as places of high density (O'Sullivan, 2011), surrounded by low density was used without explicit density thresholds. Furthermore, it is possible to incorporate overlaping ideas of places using hierarchical methods as discussed in the sections above. Lastly, the research carried out in the thesis provides examples of how Geographic Data Science (Singleton and Arribas-Bel, 2021) can be an effective approach to tackling urban problems.

The main technical contributions are the processing of large scale polygon data and the comparisons carried out between the TDA methods in chapter three. Garcia-Pulido and Samardzhiev (2022) provides a fast and scaleable DBSCAN and HDBSCAN clustering algorithms, alongisde extensions, which can be used by researchers when they need to analyse spatial data with hundreds of millions of observations. In addition, the comparisons of chapter three focused on three widely used TDA methods for finding the differences in time series analysis. Topological data analysis is an emerging paradigm and there are numerous methods developed in the literature, however the aprroaches tested performed worse than a simpler baseline method and this points to potential improvements TDA reaserchers can gain by switching to this methodology.

6.5 Implications and future work

These contributions have several pathways of impact in research and planning. First, the results encourage the use of hierarchical methods for the creation of different delineations and typologies for urban processes, using different types of data - relationships, time series and tabular and can scale to handle millions of observations. Second, the delineations produced could be used directly to identify the effects of fixed global thresholds on the analysis of phenomena, as was done in chapters four and five for polycentricity and in chapter three when the results were compared to the noise pollution typology literature. Examples of such applications are the effects of urban form on health (Meijers, 2008), the identification of patterns of sprawl (Barrington-Leigh and Millard-Ball, 2015) or the validity of the scaling laws hypothesis (Lobo et al., 2020). Furthermore, the polycentricity analysis can readily be extended using the provided delineations by incorporating more complex decentralisation metrics (Derudder et al., 2021). Third, all of the methodologies and datasets can be used to track changes across time since all of the data used in the analysis is temporal with different release periods. Fourth, the adopted approaches in chapters five can be used globally, since buildings are universal units in all countries and the chapter explicitly outlines some of

the challenges researchers may face in doing so. Similarly, chapters three and four can be adopted, provided there is available data.

The results have several implications for planning as well. In general, urban planners and policy makers use bounded territorial units, even though how much they are emphasised varies (Harrison, 2013). Therefore, the delineations created here can be used by planners directly or can be adapted if more porous boundaries and units are needed (Paasi and Zimmerbauer, 2016), using the hierarchy and the methods discussed above. Example uses can be as evidence in favour of the adoption of new administrative units at different scales (Purkarthofer et al., 2021), new cross border initiates (Bellisario et al., 2016) or to increase investments in new forms of data such as building polygons and sound sensors. The results show that these data sources are provide useful supplementary information, that enable the analysis of previously inaccessible aspects of urban form and function. These types of data could be especially good for developing countries, due to their relative cheap cost of adoption (Wardrop et al., 2018). Similarly, investments in data fusion approaches can improve data quality and increase analysis options.

Furthermore, all delineations can be used to track real versus planned development of urban phenomena. For example, the sound sensors can be directly used to track the differences and similarities between actual activity and the urban land use plans and regulations (Toole et al., 2012). Similarly, the methodology and data in chapter five can be used for the analysis of sprawl (Barrington-Leigh and Millard-Ball, 2015). The emergent megaregional boundaries and the proposed methodology from chapters four and five can be used across time to validate the emergence of new large-scale economically-integrated areas, and to track the predictions and proposals of researchers and planners such as (Hagler, 2009; Nelson and Rae, 2016; Nelson and Lang, 2018; Ross et al., 2016; Nelson, 2017, 2020; Ross et al., 2009) as well as the changes in the relationship between form and function and its implications for planning and research (Batty, 2018).

In order to facilitate this usage and improve results, these areas of future work could be emphasized - enhancing and combining different data sources, as well as developing methods and theories to integrate all results. First, data fusion and data quality improvements can extend the analysis both in scale and function. For example, more sound sensors in more areas could result in the discovery of new activity profiles and it would enable the exploration of land use changes across different scales, similar to the analysis carried out in chapters four and five. Additionally, the validity and cluster characterisation can be enhanced through the use of other functional data which enriches the OpenStreetMap POI and building data. For chapter five, different types of building information could be incorporated into the delineation procedures, such as the building heights used by de Bellefon et al. (2019) or Arribas-Bel et al. (2021b).

Furthermore, the resulting final delineations could be additionally combined with other widely available data sources such as population, vegetation or watergrids that could directly help resolve some of the limitations discussed in chapter five. Different types of relational data such as the one used by Calafiore et al. (2021) or Rowe et al. (2023) can be used to augment the economic interactions used in chapter four.

In general integrating all of these new types of data, as well as results into coherent theories and operalisations of these theories are crucial to increasing the impact of the research. This thesis demonstrates that one promising avenue of achieving this is through hierarchical methods. The advantages of this family of methods is that they can handle different data types and can provide intermediate results which shed light on the constituent units of the final clusters, which can themselves be used for analysis and quality control. Chapter five is an example of implementing this in practice.

6.6 Concluding remarks

It is expected that cities will continue to grown and change in the future, giving rise to new challenges for researchers, urban planners and policy makers (Lobo et al., 2020). Alongside these changes, more data, reflective of the diverse human interactions taking place within them, will be become increasingly available (Arribas-Bel and Tranos, 2018). Successfully using this data to address new and existing challenges, alongside the growing importance of cities, opens up opportunities for planners and policy makers to support to an unprecedented extent improvements in the lives of citizens (Batty, 2018; Singleton and Arribas-Bel, 2021). The results from the thesis show how advances in delineation approaches using new forms of data and approaches such as GDS can help achieve this. First, by capturing different aspects of urban reality, which previously were obfuscated by data or methodological limitations and second, by analysing phenomena across different scales and datasets. Combining these results and more traditional analysis, alongside historical, cultural and political context (Hamilton and Rae, 2018) as well expert and local resident opinion (Hagler, 2009; Galdo et al., 2021) into coherent theories is the next step in achieving timely and varied urban interventions and ensuring sustainable urban living.

A

Comparison between clustering methods

TDA methods

A comprehensive mathematical introduction to TDA can be found in Wasserman (2011). All three TDA methods used in this comparison are based on the ideas of persistent homology, diagrams and bottleneck distance. As previously mentioned, persistent homology provides an object, a 'persistent diagram' that describes the multi-dimensional coarse shape of the time series - the zeroth dimension corresponds to connected components or clusters, the first to cycles present in the time series data, while the second and above dimensions represent higher dimensional generalizations of cycles (Chazal and Michel, 2017). This information is reflective of various underlying periodic patterns in the time series, as well as the time series' critical points - its peaks and valleys. Persistent diagrams of different time series can be compared to each other through 'bottleneck distance', which is a measure robust to noise (Cohen-Steiner et al., 2007).

The first TDA method used in the comparison is called 'Lowerstar'. It focuses on the peaks and valleys in the time signal as the most important part. This method is the equivalent to basing the comparisons between different sound sensors on differences at which their critical points - local minimums and maximums - occur. It is described in more detail in the methodology section of this paper.

The second method is based on Perea et al. (2015). It is a particular implementation of a class of TDA transformations for time signals that have seen successful applications in fields such as medicine and engineering. The main focus is on the regularly repeated patterns in the sound signals, such as the daily and weekly periodicities. In that the goal of the transformation is similar to the Fourier decomposition used in Cici et al. (2015). To this end persistent homology is applied to a time lagged (delay embeddings) representation of the sound sensor pattern in order to obtain its persistent diagram. The transformed data is constructed by using 'delay embeddings' of the sig-

nal. That is the representation on which persistent homology is applied. This method is referred to as 'TDA Sw1pers'.

The last method is based on Pereira and de Mello (2015) and is similar to the one above, however several topological features are extracted from the persistent diagram manually - number of detected cycles, persistence of the most prominent cycle and others. These extracted features are then used to directly cluster the points. The methodology is followed as described in the paper, with the only difference being in the choice of clustering methodology used at the end. This was done in order to make the comparison between the different methods as close as possible. This method is referred to as 'TDA Pointcloud features'.

Comparison results

Table A.1: Clustering results

	Number of clusters	Silhouette	Adjusted Mutual Information	Mean POI difference	Min POI difference
TDA Lowerstar	3	0.304	0.128	0.104	0.075
TDA Sw1pers	3	0.271	0.140	0.098	0.057
TDA Pointcloud features	2	0.748	0.004	0.067	0.067

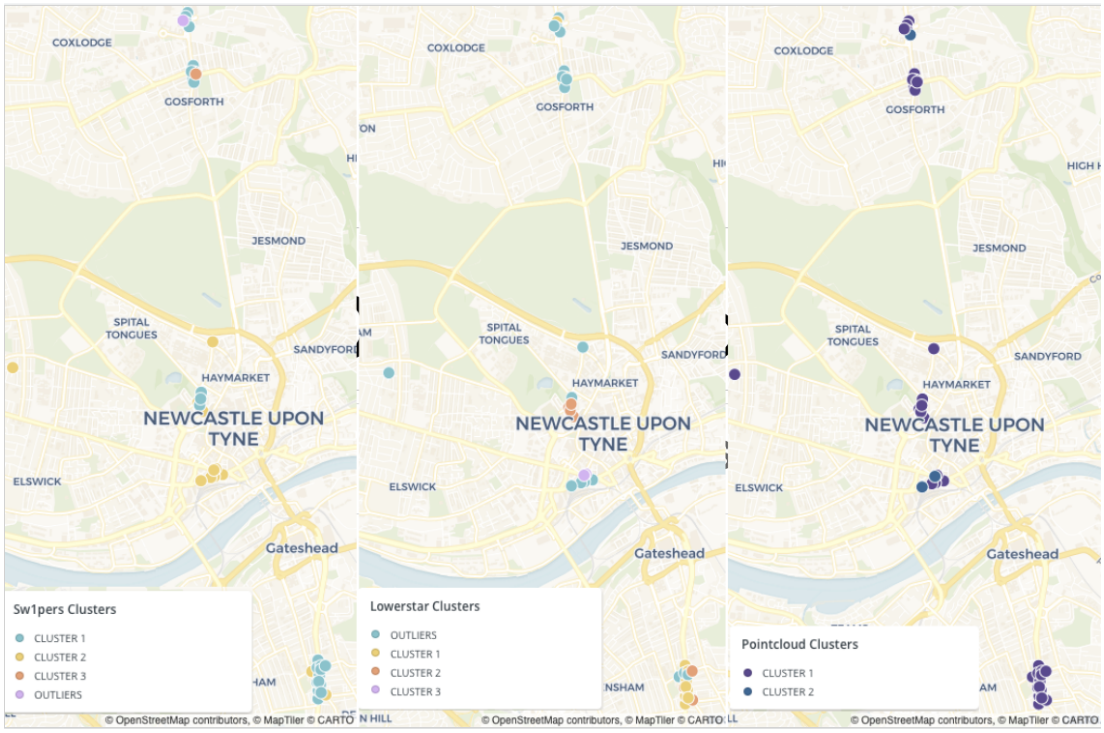


Figure A.1: Clusters obtained using TDA methods

Table 3.1 shows a summary of the number of clusters, silhouette score and mutual information comparison with the naive spatial clustering for each of the methods. First, it can be seen that the number of clusters varies from 2 to 5 for the different methods.

The outliers are not included, only the clusters with more than 2 members. Looking at just the silhouette scores the two best results are the baseline model and the TDA Pointcloud features one. However when the the mutual information comparison with the naive spatial clustering is considered the TDA Pointcloud features is the worst clustering. It has a score of almost zero, whereas all the other results exhibit some spatial patterns.

Figure A.1 shows the clustering results obtained by applying the TDA methods. Using the Pointcloud method results in 37 out of 40 sensors in the same cluster, with 3 spatially random ones in another. This result is not informative, since almost all of the data lie in a single cluster. Another problem is that there are no outliers detected, whereas all the other clustering methods find at least some. Furthermore, the positive mutual information scores of the other clustering methods suggest some degree of spatial correlation. Their results are more inline with evidence that shows sound patterns and propagation are spatially sensitive (Zuo et al., 2014). Therefore, this method doesn not produce the best result.

The next set of results is the one obtained with the TDA Sw1pers algorithm. It has positive values for both the mutual information comparison and the silhouette scores. It ranks second in the mutual information comparison and third in silhouette score. It splits the data into 3 clusters, with 2 outliers.

The last TDA method is the TDA Lowerstar. It has the second highest silhouette score and second mutual information score. It splits 23 out of the 40 sensors into three clusters and 17 outliers with strong spatial correlation.

Table A.2: Lowerstar features points of interest

	total	sustenance	education	transportation	financial	healthcare	entertainment	sensors
Cluster 1	118.0	0.542373	0.000000	0.406780	0.016949	0.016949	0.016949	7.0
Cluster 2	114.0	0.429825	0.026316	0.482456	0.008772	0.043860	0.008772	11.0
Cluster 3	75.0	0.466667	0.013333	0.453333	0.026667	0.000000	0.040000	2.0

Tables A.2, A.3, A.1 shows the distribution of the different types of types of interest in the clsuters. The categories represent points of interest from Openstrepmap data in a 100m radius around each pont in the cluster. The types of places that go into each category are described in the introduction to the data used in this paper in section 3.4. All sets of clusters show a similar distribution of points of interest - a high level of transportation and sustenance amenities. There are no discernible differences in the clusters of the point cloud clustering, echoing the conclusion from the previous metrics that this clustering does not provide much information. The Sw1pers clustering gives one cluster, cluster 3 that, stands out with some entertainment and financial services and a high level of sustenance in cluster 1. There is a similar pattern in the Lowerstar clustering, however it is more pronounced.

Given this comparison the best performing TDA method is the 'Lowerstar'. The

'TDA Pointcloud features' clusters do not provide much information, since most sensors get assigned to the same clusters. 'Lowerstar' has the higher silhouete , and a comparable mutual information score when compared to 'Sw1pers'. It also has the highest values of Mean POI difference and Min POI difference, suggesting that the amenities surrounding its sensors are the most distinct ones. Because of these results is the TDA method used in the paper.

Table A.3: Sw1pers points of interest

	total	sustenance	education	transportation	financial	healthcare	entertainment	sensors
Cluster 1	14.0	0.571429	0.000000	0.428571	0.000000	0.000000	0.000000	2.0
Cluster 2	150.0	0.460000	0.020000	0.446667	0.020000	0.046667	0.006667	25.0
Cluster 3	121.0	0.479339	0.008264	0.413223	0.041322	0.024793	0.033058	12.0

Table A.4: Point cloud features points of interest

	total	sustenance	education	transportation	financial	healthcare	entertainment	sensors
Cluster 1	194.0	0.443299	0.015464	0.469072	0.025773	0.036082	0.010309	37.0
Cluster 2	52.0	0.442308	0.019231	0.423077	0.019231	0.038462	0.057692	3.0

B

Overview of LODES data

Tables B.1 and B.2 provide an overview of the LODES Longitudinal Employer-Household Dynamics Origin-Destination Employment Statistics (LODES) data for 2016. The tables show the total number and percentage of interstate and intrastate flows.

Table B.1: Intra and Inter state flows

Name	Total	Intrastate flows	Interstate flows
District of Columbia	657538.0	0.280227	0.719773
Delaware	506453.0	0.686417	0.313583
Rhode Island	559570.0	0.712567	0.287433
New Hampshire	765545.0	0.721107	0.278893
West Virginia	791563.0	0.750396	0.249604
Maryland	2942352.0	0.752729	0.247271
New Jersey	4617436.0	0.767449	0.232551
Kansas	1493999.0	0.817227	0.182773
Vermont	328214.0	0.818155	0.181845
North Dakota	427461.0	0.838287	0.161713
Connecticut	1809929.0	0.842953	0.157047
Mississippi	1230289.0	0.844524	0.155476
Kentucky	2001827.0	0.844710	0.155290
Virginia	3957518.0	0.845036	0.154964
Missouri	2935029.0	0.852597	0.147403
Wyoming	278620.0	0.871976	0.128024
Iowa	1641167.0	0.877136	0.122864
Massachusetts	3653350.0	0.878616	0.121384
South Carolina	2133140.0	0.878973	0.121027
Idaho	732118.0	0.880070	0.119930
Indiana	3197494.0	0.881589	0.118411
Pennsylvania	6098930.0	0.881943	0.118057
New York	9501484.0	0.884425	0.115575
Tennessee	3003119.0	0.891773	0.108227
Nebraska	1000596.0	0.894702	0.105298
Arkansas	1246375.0	0.896230	0.103770
South Dakota	428624.0	0.898174	0.101826

Table B.2: Table B.1 continued

Name	Total	Intrastate flows	Interstate flows
Oregon	1877248.0	0.902754	0.097246
New Mexico	832344.0	0.905638	0.094362
Alabama	1998093.0	0.907275	0.092725
Illinois	6139657.0	0.911857	0.088143
Wisconsin	2974121.0	0.916772	0.083228
Nevada	1308896.0	0.922300	0.077700
Maine	614770.0	0.922472	0.077528
Ohio	5431966.0	0.926383	0.073617
North Carolina	4392883.0	0.926953	0.073047
Georgia	4328510.0	0.929859	0.070141
Washington	3249909.0	0.931439	0.068561
Louisiana	1944373.0	0.931859	0.068141
Minnesota	2905656.0	0.932087	0.067913
Oklahoma	1615079.0	0.932413	0.067587
Montana	458437.0	0.948641	0.051359
Michigan	4303332.0	0.954900	0.045100
Utah	1376095.0	0.959478	0.040522
Arizona	2706420.0	0.960954	0.039046
Colorado	2526124.0	0.970203	0.029797
Texas	11827331.0	0.973138	0.026862
Florida	8346751.0	0.979197	0.020803
California	16639025.0	0.982980	0.017020

C

Building area clustering

In Figure C.1 the HDBSCAN algorithm does not take into account the building areas, only the distances between them. Since the blocks in Manhattan are treated as individual buildings, there are around 5000 of them and they are at a larger distance to each other, relative to areas such as the Bronx where the median detected building has a smaller area and distance to its nearest neighbouring building on average. This leads to the analysis procedure treating the most densely populated of the five boroughs of New York City as a sparsely populated area, separating the Bronx, Brooklyn and Queens. In Figure C.2, the sizes based on the area replace the raw counts of the buildings and the algorithmic procedure delineates all of the boroughs together.

Figure C.1: HDBSCAN boundaries for New York

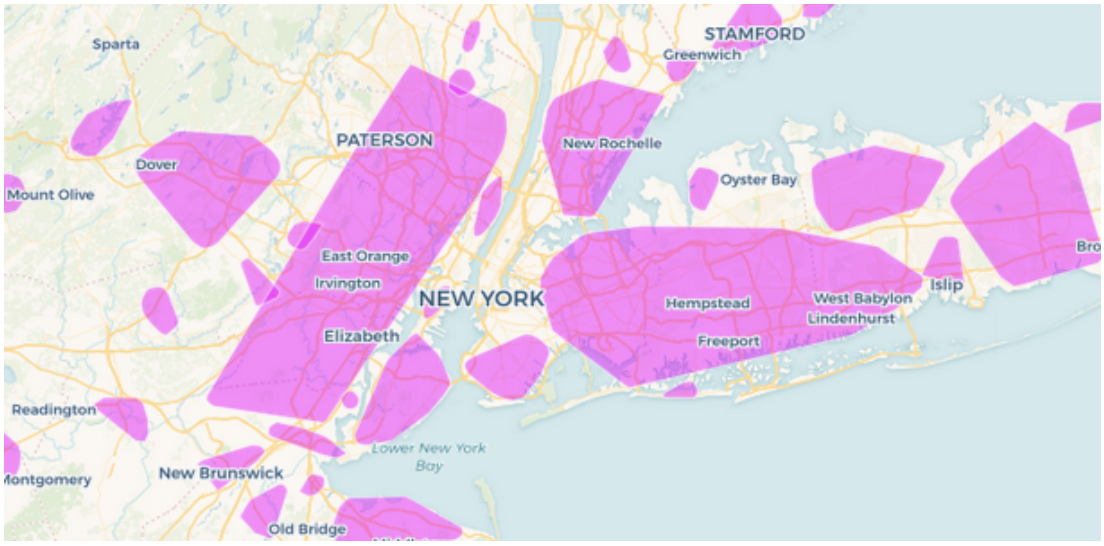


Figure C.2: HDBSCAN boundaries when area is incorporated into the methodology for New York

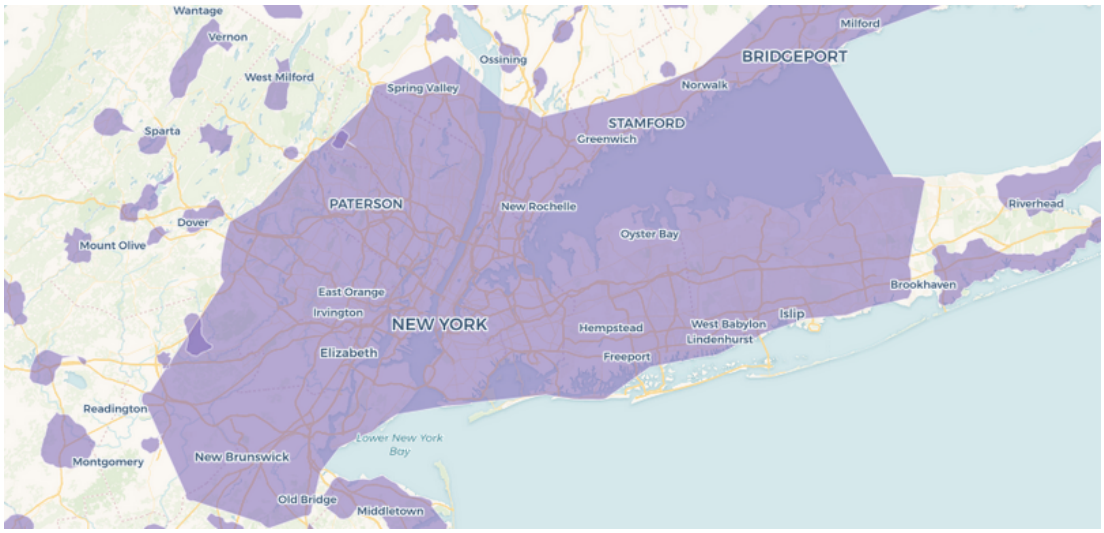
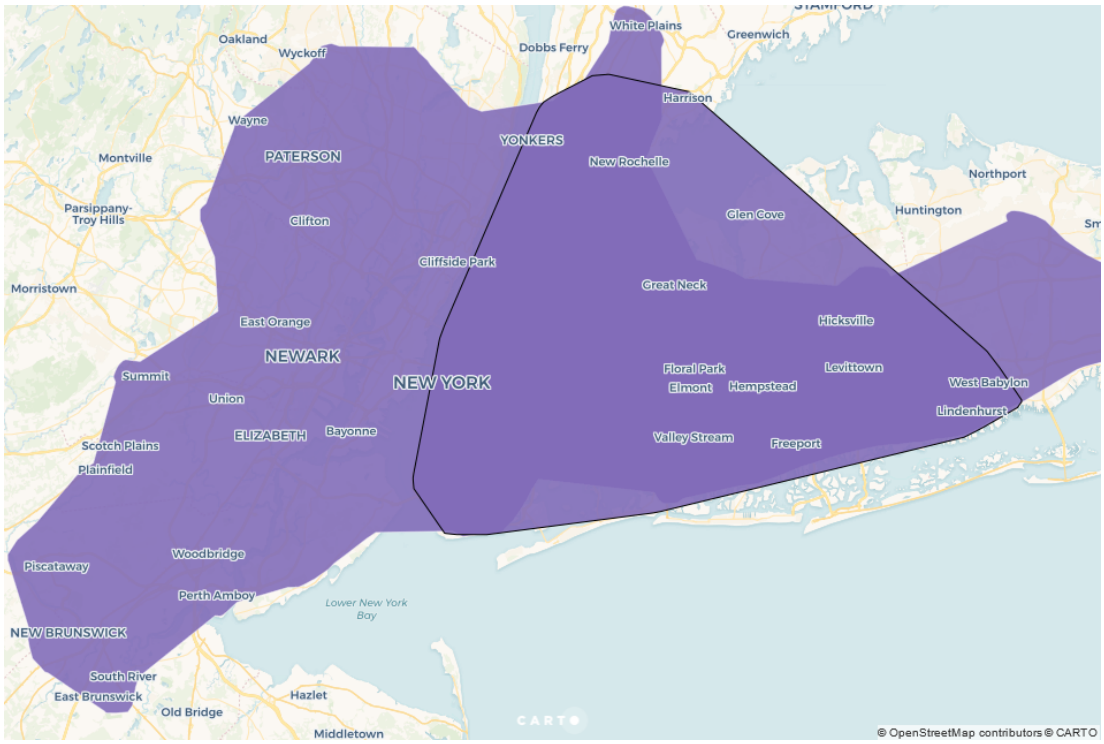


Figure C.3 shows the difference between the New York delineation based on a parameter choice of 2000 versus 19000. The split occurs along the Hudson bay, since that is an area with no buildings.

Figure C.3: New York delineation based on a minimum parameter size of 2000 and 19000



Figures C.4, C.5, C.6 show the the delineations for three of the top 15 cities for the HDBSCAN, GHSL and CUC delineations. GHSL urban delineations are in blue, CUC in green, our HDBSCAN in red.

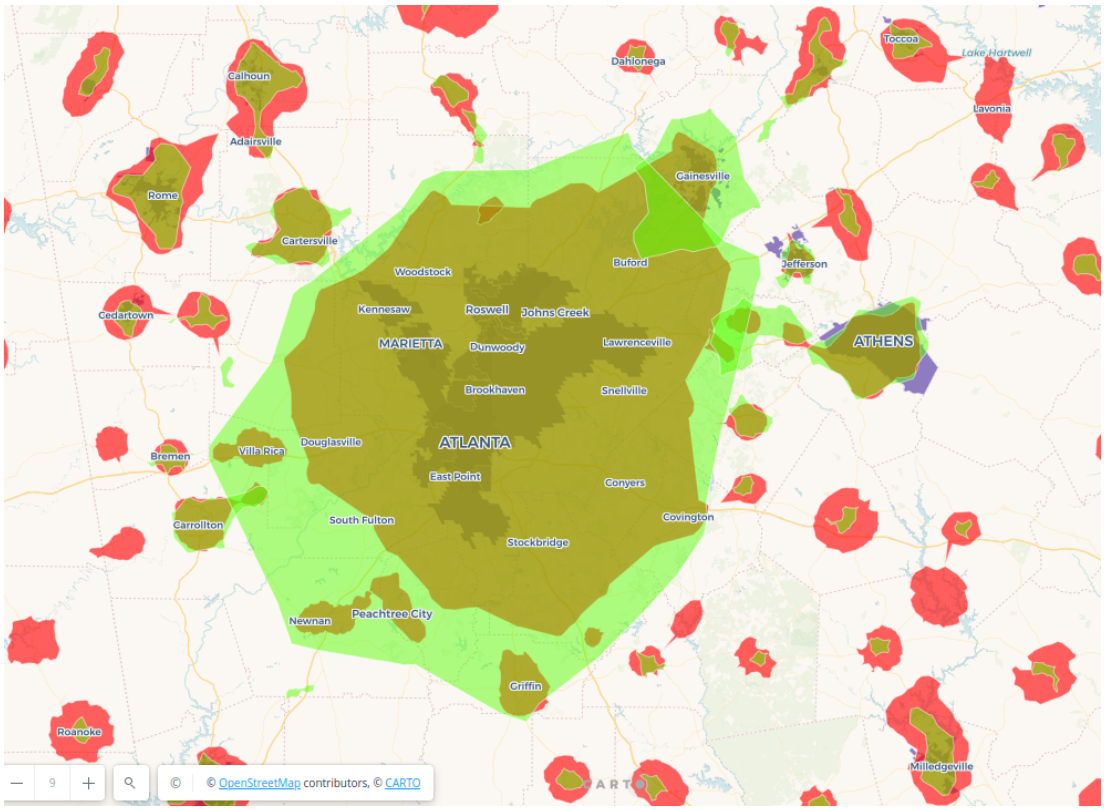


Figure C.4: Atlanta extent for each of the three methods

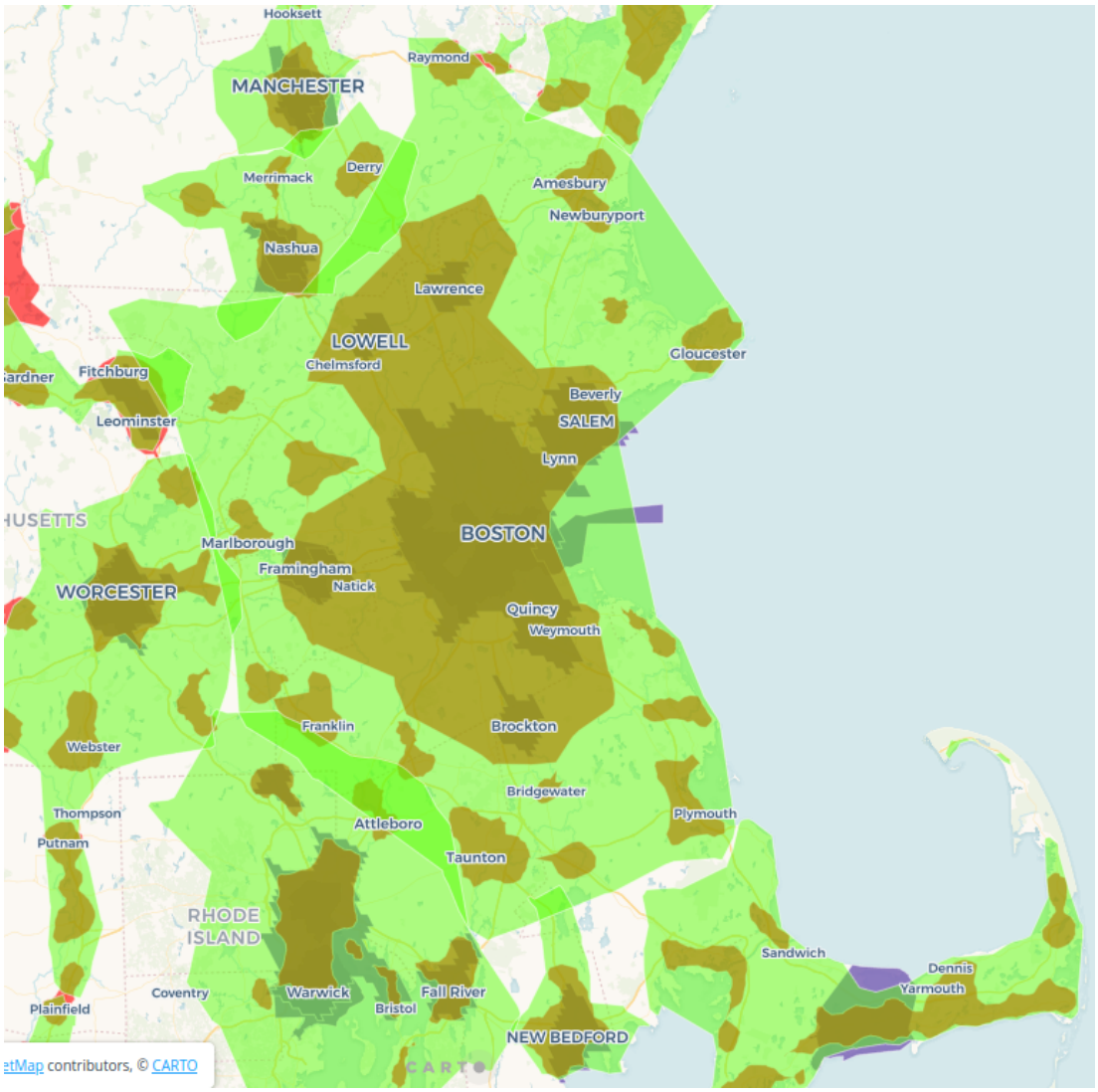


Figure C.5: Boston extent for each of the three methods

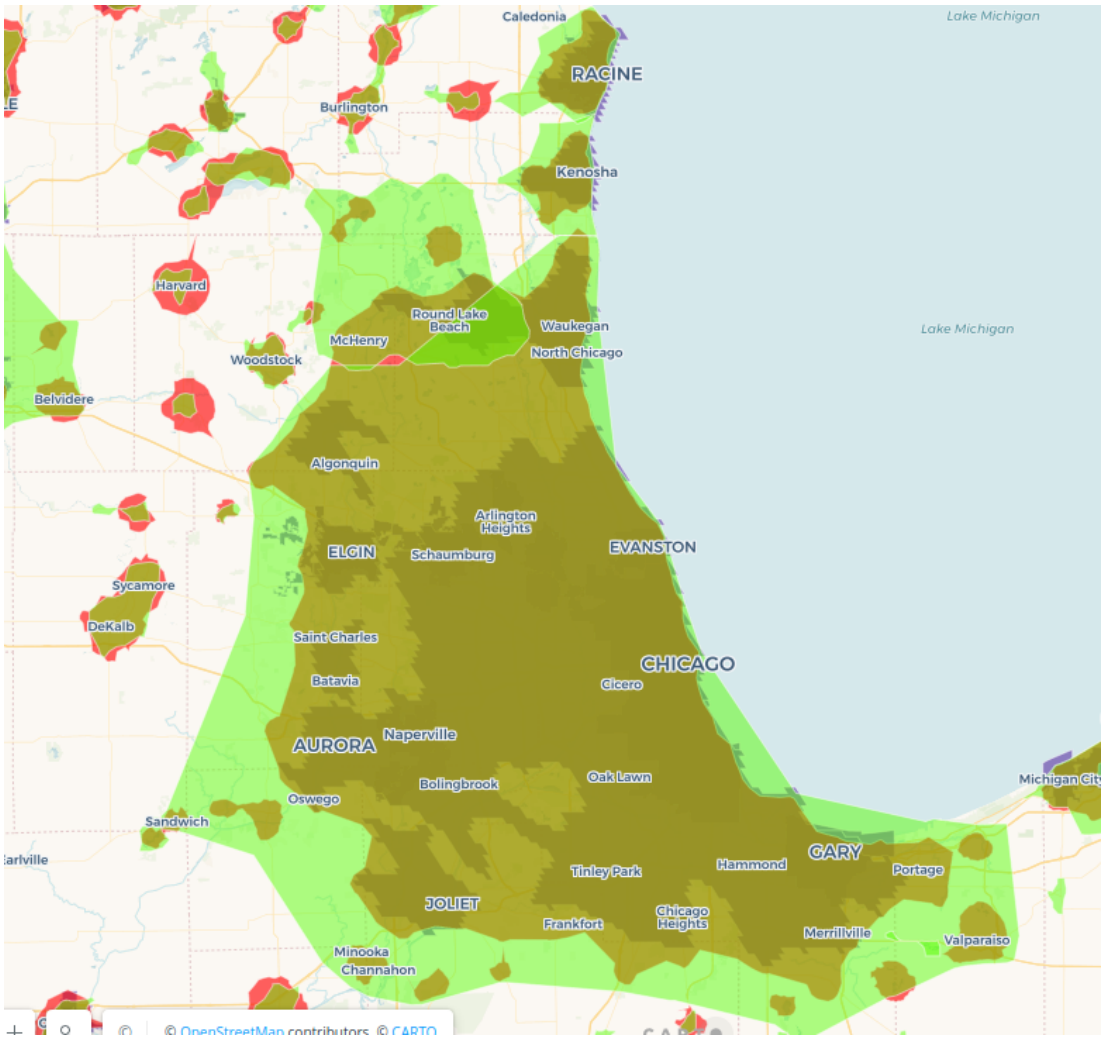


Figure C.6: Chicago extent for each of the three methods

Bibliography

Adolphson, M. (2009). Estimating a Polycentric Urban Structure. Case Study: Urban Changes in the Stockholm Region 1991–2004. *Journal of Urban Planning and Development*, 135(1):19–30.

Aghabozorgi, S., Seyed Shirkhorshidi, A., and Ying Wah, T. (2015). Time-series clustering - A decade review. *Information Systems*, 53:16–38.

Ahuallachain, B. (2012). Inventive Megaregions of the United States :. *Economic Geography*, 88(2):165–195.

Alexander, C. (2017). *A city is not a tree*. Sustasis Press/Off The Common Books.

Alonso, W. (1960). *A model of the urban land market: Location and densities of dwellings and businesses*. University of Pennsylvania.

Amekudzi, A. A., Banerjee, T., Barringer, J., Cmapbell, S., Contant, C. K., Doyle, J. L. H., Ankner, W., Fainstein, N., Fainstein, S. S., Faludi, A. K., et al. (2012). *Megaregions: Planning for global competitiveness*. Island Press.

Angel, S. and Blei, A. M. (2016). The spatial structure of american cities: The great majority of workplaces are no longer in cbds, employment sub-centers, or live-work communities. *Cities*, 51:21–35. Current Research on Cities.

Arcaute, E., Hatna, E., Ferguson, P., Youn, H., Johansson, A., and Batty, M. (2015). Constructing cities, deconstructing scaling laws. *Journal of the Royal Society Interface*, 12(102).

Arcaute, E., Molinero, C., Hatna, E., Murcio, R., Vargas-Ruiz, C., Masucci, A. P., and Batty, M. (2016). Cities and regions in Britain through hierarchical percolation. *Royal Society Open Science*, 3(4).

Arribas-Bel, D. (2014a). Accidental, open and everywhere: Emerging data sources for the understanding of cities. *Applied Geography*, 49:45–53.

Arribas-Bel, D. (2014b). Accidental, open and everywhere: Emerging data sources for the understanding of cities. *Applied Geography*, 49:45–53.

Arribas-Bel, D., Garcia-López, M.-À., and Viladecans-Marsal, E. (2021a). Building(s and) cities: Delineating urban areas with a machine learning algorithm. *Journal of Urban Economics*, 125(September 2018):103217.

Arribas-Bel, D., Green, M., Rowe, F., and Singleton, A. (2021b). Open data products-A framework for creating valuable analysis ready data. *Journal of Geographical Systems*, 23(4):497–514.

Arribas-Bel, D. and Sanz-Gracia, F. (2014). The validity of the monocentric city model in a polycentric age: US metropolitan areas in 1990, 2000 and 2010. *Urban Geography*, 35(7):980–997.

Arribas-Bel, D. and Schmidt, C. R. (2013). Self-organizing maps and the US urban spatial structure. *Environment and Planning B: Planning and Design*, 40(2):362–371.

Arribas-Bel, D. and Tranos, E. (2018). Characterizing the Spatial Structure(s) of Cities "on the fly": The Space-Time Calendar. *Geographical Analysis*, 50(2):162–181.

Atkinson, P. M. and Tate, N. J. (2000). Spatial scale problems and geostatistical solutions: A review. *Professional Geographer*, 52:607–623.

Balk, D., Leyk, S., Jones, B., Montgomery, M. R., and Clark, A. (2018). Understanding urbanization: A study of census and satellite-derived urban classes in the United States, 1990-2010. *PLoS ONE*, 13(12):1–20.

Balk, D., Leyk, S., Montgomery, M. R., and Engin, H. (2021). Global harmonization of urbanization measures: Proceed with care. *Remote Sensing*, 13(24):1–26.

Ballantyne, P., Singleton, A., Dolega, L., and Macdonald, J. (2022). Integrating the who, what, and where of us retail center geographies. *Annals of the American Association of Geographers*, pages 1–23.

Barrington-Leigh, C. and Millard-Ball, A. (2015). A century of sprawl in the United States. *Proceedings of the National Academy of Sciences of the United States of America*, 112(27):8244–8249.

Batten, D. F. (1995). Network cities: creative urban agglomerations for the 21st century. *Urban Studies*, 32(2):313–327.

Batty, M. (2006). Rank clocks. *Nature*, 444(7119):592–596.

Batty, M. (2012). Building a science of cities. *Cities*, 29(SUPPL. 1):S9–S16.

Batty, M. (2018). *Inventing future cities*. Mit Press.

Bellisario, J., Weinberg, M., and Mena, C. (2016). The Northern California Megaregion: Innovative, Connected, and Growing. Technical Report June, Bay Area Council Economic Institute.

Bergmann, L. and O'Sullivan, D. (2018). Reimagining GIScience for relational spaces. *Canadian Geographer*, 62(1):7–14.

Bertaud, A. (2018). *Order without design: How markets shape cities*. MIT Press, London, England.

Blondel, V. D., Guillaume, J. L., Lambiotte, R., and Lefebvre, E. (2008). Fast unfolding of communities in large networks. *Journal of Statistical Mechanics: Theory and Experiment*, 2008(10):P10008.

Bosker, M., Park, J., and Roberts, M. (2018). Definition Matters: Metropolitan Areas And Agglomeration Economies In A Large Developing Country. *Definition Matters: Metropolitan Areas And Agglomeration Economies In A Large Developing Country*, (September).

Brambilla, G., Confalonieri, C., and Benocci, R. (2019). Application of the intermittency ratio metric for the classification of urban sites based on road traffic noise events. *Sensors (Switzerland)*, 19(23).

Bruns, A. and Liang, Y. E. (2012). Tools and methods for capturing Twitter data during natural disasters. *First Monday*, 17(4).

Burger, M. and Meijers, E. (2012). Form Follows Function? Linking Morphological and Functional Polycentricity. *Urban Studies*, 49(5):1127–1149.

Calafiore, A. (2021). Workshop on mobility data in urban science. *Alan Turing Institute*, (October).

Calafiore, A., Palmer, G., Comber, S., Arribas-Bel, D., and Singleton, A. (2021). A geographic data science framework for the functional and contextual analysis of human dynamics within global cities. *Computers, Environment and Urban Systems*, 85(November 2020):101539.

Campbell, C. J. (2018). Space, place and scale: Human geography and spatial history in pastand present. *Past and Present*, 239(1):e24–e45.

Campello, R. J., Kröger, P., Sander, J., and Zimek, A. (2020). Density-based clustering. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 10(2):1–15.

Campello, R. J. G. B., Moulavi, D., Zimek, A., and Sander, J. (2015). Hierarchical Density Estimates for Data Clustering, Visualization, and Outlier Detection. *ACM Transactions on Knowledge Discovery from Data*, 10(1):1–51.

Chaudhry, O. and Mackaness, W. A. (2008). Automatic identification of urban settlement boundaries for multiple representation databases. *Computers, Environment and Urban Systems*, 32(2):95–109.

Chazal, F. and Michel, B. (2017). An introduction to topological data analysis: fundamental and practical aspects for data scientists.

Chen, M., Arribas-Bel, D., and Singleton, A. (2019). Understanding the dynamics of urban areas of interest through volunteered geographic information. *Journal of Geographical Systems*, 21(1):89–109.

Christaller, W. (1980). *Die zentralen Orte in Süddeutschland: eine ökonomisch-geographische Untersuchung über die Gesetzmässigkeit der Verbreitung und Entwicklung der Siedlungen mit städtischen Funktionen*. Wissenschaftliche Buchgesellschaft.

Christopher, A. (1965). A city is not a tree. In *Architectural forum*, volume 122, pages 58–62.

Cici, B. (2015). *Mobile Data Analysis For Smart City Applications*. PhD thesis.

Cici, B., Gjoka, M., Markopoulou, A., and Butts, C. T. (2015). On the Decomposition of Cell Phone Activity Patterns and their Connection with Urban Ecology. In *Proceedings of the 16th ACM International Symposium on Mobile Ad Hoc Networking and Computing - MobiHoc '15*, volume 2015-June, pages 317–326, New York, New York, USA. ACM Press.

Cohen, B. (2006). Urbanization in developing countries: Current trends, future projections, and key challenges for sustainability. *Technology in Society*, 28(1):63–80. Sustainable Cities.

Cohen-Steiner, D., Edelsbrunner, H., and Harer, J. (2007). Stability of persistence diagrams. *Discrete and Computational Geometry*, 37(1):103–120.

Coombes, M. (2014). From City-region Concept to Boundaries for Governance: The English Case. *Urban Studies*, 51(11):2426–2443.

Costa, J. P. and Škraba, P. (2014). A topological data analysis approach to epidemiology. *European Conference on Complexity Science 2014*, (September 2014).

Couclelis, H. (1996). The death of distance.

Craig, S. G., Kohlhase, J. E., and Perdue, A. W. (2016). Empirical polycentricity: The complex relationship between employment centers. *Journal of Regional Science*, 56(1):25–52.

Csáji, B. C., Browet, A., Traag, V. A., Delvenne, J. C., Huens, E., Van Dooren, P., Smoreda, Z., and Blondel, V. D. (2013). Exploring the mobility of mobile phone users. *Physica A: Statistical Mechanics and its Applications*, 392(6):1459–1473.

Curlin, T., Jaković, B., and Miloloža, I. (2019). Twitter usage in Tourism: Literature Review. *Business Systems Research*, 10(1):102–119.

Dadashpoor, H. and Malekzadeh, N. (2020). Driving factors of formation, development, and change of spatial structure in metropolitan areas: A systematic review. *Journal of Urban Management*, 9(3):286–297.

Dadashpoor, H. and Malekzadeh, N. (2021). Evolving spatial structure of metropolitan areas at a global scale: a context-sensitive review. *GeoJournal*, 0.

Davies, A., Green, M. A., and Singleton, A. D. (2018). Using machine learning to investigate self-medication purchasing in England via high street retailer loyalty card data. *PLoS ONE*, 13(11):1–14.

de Bellefon, M.-P., Combes, P.-P., Duranton, G., Gobillon, L., and Gorin, C. (2019). Delineating urban areas using building density. *Journal of Urban Economics*, (October 2018):103226.

De Nadai, M., Staiano, J., Larcher, R., Sebe, N., Quercia, D., and Lepri, B. (2016). The Death and Life of Great Italian Cities. In *Proceedings of the 25th International Conference on World Wide Web - WWW '16*, pages 413–423, New York, New York, USA. ACM Press.

Deng, Y., Liu, J., Liu, Y., and Luo, A. (2019). Detecting urban polycentric structure from POI data. *ISPRS International Journal of Geo-Information*, 8(6).

Derudder, B., Liu, X., Wang, M., Zhang, W., Wu, K., and Caset, F. (2021). Measuring polycentric urban development: The importance of accurately determining the 'balance' between 'centers'. *Cities*, 111(November 2020).

Dijkstra, L., Florczyk, A. J., Freire, S., Kemper, T., Melchiorri, M., Pesaresi, M., and Schiavina, M. (2021). Applying the degree of urbanisation to the globe: A new harmonised definition reveals a different picture of global urbanisation. *Journal of Urban Economics*, 125:103312. Delineation of Urban Areas.

Dolega, L. and Celińska-Janowicz, D. (2015). Retail resilience: A theoretical framework for understanding town centre dynamics.

Dolega, L. and Lord, A. (2020). Exploring the geography of retail success and decline: A case study of the liverpool city region. *Cities*, 96:102456.

Dolega, L., Pavlis, M., and Singleton, A. (2016). Estimating attractiveness, hierarchy and catchment area extents for a national set of retail centre agglomerations. *Journal of Retailing and Consumer Services*, 28:78–90.

Dolega, L., Reynolds, J., Singleton, A., and Pavlis, M. (2021). Beyond retail: New ways of classifying uk shopping and consumption spaces. *Environment and Planning B: Urban Analytics and City Science*, 48(1):132–150.

Duranton, G. (2021). Classifying locations and delineating space: An introduction. *Journal of Urban Economics*, 125(April).

Edelsbrunner, H., Kirkpatrick, D., and Seidel, R. (1983). On the shape of a set of points in the plane. *IEEE Transactions on Information Theory*, 29(4):551–559.

Emrani, S., Chintakunta, H., and Krim, H. (2014). Real time detection of harmonic structure: A case for topological signal analysis. *ICASSP, IEEE International Conference on Acoustics, Speech and Signal Processing - Proceedings*, (3):3445–3449.

Ewing, R. and Hamidi, S. (2015). Compactness versus Sprawl: A Review of Recent Evidence from the United States. *Journal of Planning Literature*, 30(4):413–432.

Feng, Y., Wu, S., Wu, P., Su, S., Weng, M., and Bian, M. (2018). Spatiotemporal characterization of megaregional poly-centrality: Evidence for new urban hypotheses and implications for polycentric policies. *Land Use Policy*, 77(129):712–731.

Florczyk, A. J., Melchiorri, M., Orbane, C., Schiavina, M., Maffenini, M., Politis, P., Sabo, S., Freire, S., Ehrlich, D., Kemper, T., Tommasi, P., Airaghi, D., and Zanchetta, L. (2019). Description of the GHS Urban Centre Database 2015. Technical report, Publications Office of the European Union.

Florida, R., Gulden, T., and Mellander, C. (2008). The rise of the mega-region. *Cambridge Journal of Regions, Economy and Society*, 1(3):459–476.

Frias-Martinez, V. and Frias-Martinez, E. (2014). Spectral clustering for sensing urban land use using Twitter activity. *Engineering Applications of Artificial Intelligence*, 35:237–245.

Frias-Martinez, V., Soguero-Ruiz, C., Frias-Martinez, E., and Josephidou, M. (2013). Forecasting socioeconomic trends with cell phone records. In *Proceedings of the 3rd ACM Symposium on Computing for Development - ACM DEV '13*, page 1, New York, New York, USA. ACM Press.

Frias-Martinez, V., Soto, V., Hohwald, H., and Frias-Martinez, E. (2012). Characterizing urban landscapes using geolocated tweets. *Proceedings - 2012 ASE/IEEE International Conference on Privacy, Security, Risk and Trust and 2012 ASE/IEEE International Conference on Social Computing, SocialCom/PASSAT 2012*, pages 239–248.

Friedmann, J. (2019). Thinking about complexity and planning. *International Planning Studies*, 24(1):13–22.

Furno, A., Fiore, M., Stanica, R., Ziemlicki, C., and Smoreda, Z. (2017). A Tale of Ten Cities: Characterizing Signatures of Mobile Traffic in Urban Areas. *IEEE Transactions on Mobile Computing*, 16(10):2682–2696.

Furno, A., Stanica, R., and Fiore, M. (2015). A comparative evaluation of urban fabric detection techniques based on mobile traffic data. In *Proceedings of the 2015 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining, ASONAM 2015*, pages 689–696, New York, New York, USA. ACM Press.

Galdo, V., Li, Y., and Rama, M. (2021). Identifying urban areas by combining human judgment and machine learning: An application to india. *Journal of Urban Economics*, 125:103229.

Garcia-Pulido, A. L. and Samardzhiev, K. P. (2022). Geometric reconstructions of density based clusterings.

Geddes, P. (1915). *Cities in evolution: an introduction to the town planning movement and to the study of civics*. London, Williams.

Georg, I., Blaschke, T., and Taubenböck, H. (2018). Are we in boswash yet? A multisource geodata approach to spatially delimit urban corridors. *ISPRS International Journal of Geo-Information*, 7(1):15.

Gibadullina, A., Bergmann, L., and O'Sullivan, D. (2021). For Geographical Network Analysis. *Tijdschrift voor Economische en Sociale Geografie*, 112(4):482–487.

Gidea, M. and Katz, Y. (2018). Topological data analysis of financial time series: Landscapes of crashes. *Physica A: Statistical Mechanics and its Applications*, 491:820–834.

Glaeser, E. L. (2010). *Agglomeration economics*. University of Chicago Press.

Glaeser, E. L. (2011). *Triumph of the City*. Macmillan.

Glass, M. R. (2014). Conflicting spaces of governance in the imagined Great Lakes megaregion. *Megaregions: Globalization's New Urban Form?*, pages 119–145.

Glocker, D. (2018). The Rise of Megaregions:Delineating a new scale ofeconomic geography. Technical report, Organisation for Economic Co-operation and Development.

Gottmann, J. (1957). Megalopolis or the urbanization of the northeastern seaboard. *Economic geography*, 33(3):189–200.

Graham, M. R., Kutzbach, M. J., and McKenzie, B. (2014). DESIGN COMPARISON OF LODES AND ACS COMMUTING DATA PRODUCTS.

Grauwin, S., Sobolevsky, S., Moritz, S., Gódor, I., and Ratti, C. (2015). Towards a comparative science of cities: Using mobile traffic records in New York, London, and Hong Kong. In *Computational Approaches for Urban Environments*, pages 363–387. Springer International Publishing, Cham.

Groos, J. C. and Ritter, J. R. (2009). Time domain classification and quantification of seismic noise in an urban environment. *Geophysical Journal International*, 179(2):1213–1231.

Habitat, U. (2013). *State of the world's cities 2012/2013*. Routledge.

Hagler, Y. (2009). Defining US Megaregions. Technical Report November, Regional Plan Association.

Hall, P. G. and Pain, K. (2006). *The polycentric metropolis: learning from mega-city regions in Europe*. Routledge.

Hamilton, R. and Rae, A. (2018). Regions from the ground up: a network partitioning approach to regional delineation. *Environment and Planning B: Urban Analytics and City Science*, page 239980831880422.

Harrison, J. (2013). Configuring the New 'Regional World': On being Caught between Territory and Networks. *Regional Studies*, 47(1):55–74.

Harrison, J. and Hoyler, M. (2015a). *Megaregions : globalization's new urban form?* Loughborough University.

Harrison, J. and Hoyler, M. (2015b). *Megaregions : globalization's new urban form?* Loughborough University.

Hartshorne, R. (1939). The Nature of Geography: A Critical Survey of Current Thought in the Light of the Past (Conclusion). *Annals of the Association of American Geographers*, 29(4):413.

Henderson, J. V., Nigmatulina, D., and Kriticos, S. (2021). Measuring urban economic density. *Journal of Urban Economics*, 125(August 2019):103188.

Hennig, C. (2015). What are the true clusters? *Pattern Recognition Letters*, 64:53–62.

Heris, M. P., Foks, N. L., Bagstad, K. J., Troy, A., and Ancona, Z. H. (2020). A rasterized building footprint dataset for the United States. *Scientific Data*, 7(1):1–10.

Hermosilla, T., Palomar-Vázquez, J., Balaguer-Beser, Á., Balsa-Barreiro, J., and Ruiz, L. A. (2014). Using street based metrics to characterize urban typologies. *Computers, Environment and Urban Systems*, 44:68–79.

Holliman, N., Turner, M., Dowsland, S., Cloete, R., and Picton, T. (2017). Designing a cloud-based 3d visualization engine for smart cities. *Electronic Imaging*, 2017(5):173–178.

Hu, Q. (2019). Twitter data in public administration: a review of recent scholarship. *International Journal of Organization Theory and Behavior*, 22(2):209–222.

Huang, X., Wang, C., and Li, Z. (2019). High-resolution population grid in the Conus using Microsoft building footprints: A feasibility study. *Proceedings of the 3rd ACM SIGSPATIAL International Workshop on Geospatial Humanities, GeoHumanities 2019*.

Hupeng, W., Kang, J., and Hong, J. (2019). Effects of urban street spatial parameters on sound propagation. *Environment and Planning B: Urban Analytics and City Science*, 46(2):341–358.

Italia, T. (2015). Telecommunications - sms, call, internet - mi.

Jacobs, J. (1961a). *1993, The Death and Life of Great American Cities, Modern Library ed*. Random House, New York.

Jacobs, J. (1961b). Jane jacobs. *The Death and Life of Great American Cities*, 21(1):13–25.

James, P., Dawson, R., Harris, N., and Joncyzk, J. (2014). Urban observatory environment.

Jochem, W. C., Leasure, D. R., Pannell, O., Chamberlain, H. R., Jones, P., and Tatem, A. J. (2020). Classifying settlement types from multi-scale spatial patterns of building footprints. *Environment and Planning B: Urban Analytics and City Science*, 0(0):1–19.

Jonczyk, J. C., Quinn, P. F., Heidrich, O., James, P., Harris, N., Dawson, R. J., and Pearson, D. J. (2016). Demonstration of a green-blue approach for a strategic management of urban runoff. *AGUFM*, 2016:H13M–1597.

Khiali-Miab, A., Van Strien, M. J., Axhausen, K. W., and Grêt-Regamey, A. (2019). Combining urban scaling and polycentricity to explain socio-economic status of urban regions. *PLoS ONE*, 14(6):1–23.

Knyazeva, I. S., Makarenko, N. G., Kuperin, Y. A., and Dmitrieva, L. A. (2016). The new approach for dynamical regimes detection in geomagnetic time series. *Journal of Physics: Conference Series*, 675(3):032028.

Kockelman, K., Huang, Y., and Quarles, N. (2019). The Rise of Long-Distance Trips , in a World of Self-Driving Cars : Anticipating Trip Counts and Evolving Travel Patterns Across the Texas Triangle Megaregion. Technical Report April, The University of Texas at Austin.

Krehl, A. (2015). Urban spatial structure: An interaction between employment and built-up volumes. *Regional Studies, Regional Science*, 2(1):290–308.

Kropf, K. (2009). Aspects of urban form. *Urban Morphology*, 13(2):105–120.

Kwon, K. and Seo, M. (2018). Does the polycentric urban region contribute to economic performance? The case of Korea. *Sustainability (Switzerland)*, 10(11).

Lacy, K. (2016). The New Sociology of Suburbs: A Research Agenda for Analysis of Emerging Trends. *Annual Review of Sociology*, 42:369–384.

Lancichinetti, A. and Fortunato, S. (2009). Community detection algorithms: A comparative analysis. *Physical Review E - Statistical, Nonlinear, and Soft Matter Physics*, 80(5):1–11.

Lang, R. E. and LeFurgy, J. (2003). Edgeless cities: Examining the noncentered metropolis. *Housing Policy Debate*, 14(3):427–460.

Lang, R. E., Lim, J., and Danielsen, K. A. (2020). The origin, evolution, and application of the megapolitan area concept. *International Journal of Urban Sciences*, 24(1):1–12.

Lau, B. P. L., Wijerathne, N., Ng, B. K. K., and Yuen, C. (2018). Sensor Fusion for Public Space Utilization Monitoring in a Smart City. *IEEE Internet of Things Journal*, 5(2):473–481.

Lenormand, M., Picornell, M., Cantú-Ros, O. G., Louail, T., Herranz, R., Barthelemy, M., Frías-Martínez, E., Miguel, M. S., and Ramasco, J. J. (2015). Comparing and modelling land use organization in cities. *Royal Society Open Science*, 2(12).

Leyk, S., Leyk, S., Leyk, S., Uhl, J. H., Uhl, J. H., Connor, D. S., Braswell, A. E., Braswell, A. E., Mietkiewicz, N., Mietkiewicz, N., Balch, J. K., Balch, J. K., Balch, J. K., Gutmann, M., and Gutmann, M. (2020). Two centuries of settlement and urban development in the United States. *Science Advances*, 6(23):1–13.

Liu, X., Derudder, B., and Wang, M. (2018). Polycentric urban development in China: A multi-scale analysis. *Environment and Planning B: Urban Analytics and City Science*, 45(5):953–972.

Lobo, J., Alberti, M., Allen-Dumas, M., Arcaute, E., Barthelemy, M., Bojorquez Tapia, L. A., Brail, S., Bettencourt, L., Beukes, A., Chen, W., Florida, R., Gonzalez, M., Grimm, N., Hamilton, M., Kempes, C., Kontokosta, C. E., Mellander, C., Neal, Z. P., Ortman, S., Pfeiffer, D., Price, M., Revi, A., Rozenblat, C., Rybski, D., Siemiatycki, M., Shutters, S. T., Smith, M. E., Stokes, E. C., Strumsky, D., West, G., White, D., Wu, J., Yang, V. C., York, A., and Youn, H. (2020). Urban Science: Integrated Theory from the First Cities to Sustainable Metropolises. *SSRN Electronic Journal*.

Lugomer, K. and Longley, P. (2018). Towards a comprehensive temporal classification of footfall patterns in the cities of Great Britain. *Leibniz International Proceedings in Informatics, LIPIcs*, 114(43):1–6.

Lynch, K. (1984). *Good city form*. MIT press.

Malzer, C. and Baum, M. (2020). A hybrid approach to hierarchical density-based cluster selection. In *2020 IEEE International Conference on Multisensor Fusion and Integration for Intelligent Systems (MFI)*, pages 223–228.

Manduca, R. (2020). The spatial structure of US metropolitan employment: New insights from administrative data. *Environment and Planning B: Urban Analytics and City Science*, 0(0):1–16.

Marquet, O. and Miralles-Guasch, C. (2015). The Walkable city and the importance of the proximity environments for Barcelona's everyday mobility. *Cities*, 42(PB):258–266.

Marull, J., Galletto, V., Domene, E., and Trullén, J. (2013). Emerging megaregions: A new spatial scale to explore urban sustainability. *Land Use Policy*, 34:353–366.

McInnes, L. and Healy, J. (2017). Accelerated Hierarchical Density Based Clustering. *IEEE International Conference on Data Mining Workshops, ICDMW*, 2017-Novem:33–42.

Meijers, E. (2008). Summing small cities does not make a large city: Polycentric urban regions and the provision of cultural, leisure and sports amenities. *Urban Studies*, 45(11):2323–2342.

Miao, Q., Qiao, Y., and Yang, J. (2018). Research of urban land use and regional functions based on mobile data traffic. In *Proceedings - 2018 IEEE 3rd International Conference on Data Science in Cyberspace, DSC 2018*, pages 333–338. IEEE.

Milego, R., Michelet, J. F., Arévalo, J., Jupova, K., and Larrea, E. (2019). ESPON FUORE: functional urban areas and regions in Europe.

Möck, M. and Küpper, P. (2020). Polycentricity at its boundaries: consistent or ambiguous? *European Planning Studies*, 28(4):830–849.

Morillas, J. M. B., Gozalo, G. R., González, D. M., Moraga, P. A., and Vílchez-Gómez, R. (2018). Noise pollution and urban planning. *Current Pollution Reports*, 4(3):208–219.

Moro, E., Calacci, D., Dong, X., and Pentland, A. (2021). Mobility patterns are associated with experienced income segregation in large us cities. *Nature communications*, 12(1):4633.

Mumford, L. (1937). What is a city?

Naboulsi, D., Fiore, M., Ribot, S., and Stanica, R. (2016). Large-scale mobile traffic analysis: A survey.

Nelson, A. and Lang, R. (2018). *Megapolitan america*. Routledge.

Nelson, A. C. (2017). Megaregion Projections 2015 to 2045 with Transportation Policy Implications. *Transportation Research Record: Journal of the Transportation Research Board*, 2654(1):11–19.

Nelson, G. D. (2020). Communities, Complexity, and the 'Conchoration': Network Analysis and the Ontology of Geographic Units. *Tijdschrift voor Economische en Sociale Geografie*, 0(0):1–19.

Nelson, G. D. and Rae, A. (2016). An economic geography of the United States: From commutes to megaregions. *PLoS ONE*, 11(11):e0166083.

Newman, M. E. (2006). Modularity and community structure in networks. *Proceedings of the national academy of sciences*, 103(23):8577–8582.

Oden, M. and Sciara, G. C. (2020). The salience of megaregional geographies for inter-metropolitan transportation planning and policy making. *Transportation Research Part D: Transport and Environment*, 80(August 2019):102262.

Onda, K., Sinha, P., Gaughan, A. E., Stevens, F. R., and Kaza, N. (2019). Missing millions: undercounting urbanization in India. *Population and Environment*, 41(2):126–150.

Openshaw, S. (1979). A million or so correlation coefficients: Threeexperiments on the modifiable areal unitproblem. *Statistical Applications in the Spatial Sciences*, pages 127–144.

OpenStreetMap contributors (2017). Planet dump retrieved from https://planet.osm.org . `https://www.openstreetmap.org`.

Orga, F., Socoró, J. C., Alías, F., Alsina-Pagès, R. M., Zambon, G., Benocci, R., and Bisceglie, A. (2017). Anomalous noise events considerations for the computation of road traffic noise levels: The DYNAMAP's Milan case study. *24th International Congress on Sound and Vibration, ICSV 2017*.

O'Sullivan, A. (2011). *Urban Economics*. McGraw-Hill Education.

Paasi, A. and Zimmerbauer, K. (2016). Penumbral borders and planning paradoxes: Relational thinking and the question of borders in spatial planning. *Environment and Planning A*, 48(1):75–93.

Parr, J. B. (2007). Spatial definitions of the City: Four perspectives. *Urban Studies*, 44(2):381–392.

Pei, T., Sobolevsky, S., Ratti, C., Shaw, S. L., Li, T., and Zhou, C. (2014). A new insight into land use classification based on aggregated mobile phone data. *International Journal of Geographical Information Science*, 28(9):1988–2007.

Perea, J. A., Deckard, A., Haase, S. B., and Harer, J. (2015). SW1PerS: Sliding windows and 1-persistence scoring; discovering periodicity in gene expression time series data. *BMC Bioinformatics*, 16(1):1–12.

Pereira, C. M. and de Mello, R. F. (2015). Persistent homology for time series and spatial data clustering. *Expert Systems with Applications*, 42(15-16):6026–6038.

Petrović, A., Manley, D., and van Ham, M. (2020). Freedom from the tyranny of neighbourhood: Rethinking sociospatial context effects. *Progress in Human Geography*, 44(6):1103–1123.

Piangerelli, M., Rucco, M., Tesei, L., and Merelli, E. (2018). Topological classifier for detecting the emergence of epileptic seizures. *BMC Research Notes*, 11(1):392.

Purkarthofer, E., Sielker, F., and Stead, D. (2021). Soft planning in macro-regions and megaregions: creating toothless spatial imaginaries or new forces for change? *International Planning Studies*, 0(0):1–19.

Rae, A. (2015). Mapping the american commute.

Rahiminejad, S., Maurya, M. R., and Subramaniam, S. (2019). Topological and functional comparison of community detection algorithms in biological networks. *BMC Bioinformatics*, 20(1):212.

Rains, T. and Longley, P. (2021). The provenance of loyalty card data for urban and retail analytics. *Journal of Retailing and Consumer Services*, 63(November 2020):102650.

Rappaport, J. and Humann, M. (2021). The Size of U.S. Metropolitan Areas. *SSRN Electronic Journal*, (May).

Re Calegari, G., Carlino, E., Peroni, D., and Celino, I. (2015). Extracting urban land use from linked open geospatial data. *ISPRS International Journal of Geo-Information*, 4(4):2109–2130.

Reba, M. and Seto, K. C. (2020). A systematic review and assessment of algorithms to detect, characterize, and monitor urban land change. *Remote Sensing of Environment*, 242(May 2019):111739.

Roberts, M., Blankespoor, B., Deuskar, C., and Stewart, B. (2017). Urbanization and Development: Is Latin America and the Caribbean Different from the Rest of the World? *Urbanization and Development: Is Latin America and the Caribbean Different from the Rest of the World?*, (March).

Ross, C., Barringer, J., Yang, J., Woo, M., Danner, A., West, H., Amekudzi, A., and Meyer, M. (2009). Megaregions: Delineating Existing and Emerging Megaregions. Technical report, Federal Highway Administration.

Ross, C., Woo, M., and Wang, F. (2016). Megaregions and regional sustainability. *International Journal of Urban Sciences*, 20(3):299–317.

Rowe, F., Calafiore, A., Arribas-Bel, D., Samardzhiev, K., and Fleischmann, M. (2023). Urban exodus? understanding human mobility in britain during the covid-19 pandemic using meta-facebook data. *Population, Space and Place*, 29(1):e2637.

Roy Chowdhury, P. K., Bhaduri, B. L., and McKee, J. J. (2018). Estimating urban areas: New insights from very high-resolution human settlement data. *Remote Sensing Applications: Society and Environment*, 10(October 2017):93–103.

Rozenfeld, H. D., Rybski, D., Andrade, J. S., Batty, M., Stanley, H. E., and Makse, H. A. (2008). Laws of population growth. *Proceedings of the National Academy of Sciences of the United States of America*, 105(48):18702–18707.

Rykwert, J. (1988). *The idea of a town: the anthropology of urban form in Rome, Italy and the ancient world*. Mit Press.

Schiavina, M., Moreno-Monroy, A., Maffenini, L., and Veneri, P. (2019). GHSL-OECD Functional Urban Areas. *European Commission, Joint Research Centre (JRC) Technical Report - Public Release of GHS-FUA*.

Schneider, C. M., Belik, V., Couronné, T., Smoreda, Z., and González, M. C. (2013). Unraveling daily human mobility motifs. *J R Soc Interface*, 10.

Schubert, E., Sander, J., Ester, M., Kriegel, H. P., and Xu, X. (2017). DBSCAN Revisited, Revisited. *ACM Transactions on Database Systems*, 42(3):1–21.

Secchi, P., Vantini, S., and Vitelli, V. (2015). Analysis of spatio-temporal mobile phone data: a case study in the metropolitan area of Milan. *Statistical Methods and Applications*, 24(2):279–300.

Shen, L. and Stopher, P. R. (2014). Review of GPS Travel Survey and GPS Data-Processing Methods. *Transport Reviews*, 34(3):316–334.

Shen, Y. and Batty, M. (2019). Delineating the perceived functional regions of London from commuting flows. *Environment and Planning A*, 51(3):547–550.

Singleton, A. and Arribas-Bel, D. (2021). Geographic Data Science. *Geographical Analysis*, 53(1):61–75.

Singleton, A. D. and Spielman, S. E. (2014). The Past, Present, and Future of Geodemographic Research in the United States and United Kingdom. *Professional Geographer*, 66(4):558–567.

Smith, L. and Turner, M. (2019). Building the urban observatory: Engineering the largest set of publicly available real-time environmental urban data in the uk. In *Geophysical Research Abstracts*, volume 21.

Sorensen, A. and Labbé, D. (2020). Megacities, megacity-regions, and the endgame of urbanization. In *Handbook of Megacities and Megacity-Regions*, pages 1–19. Edward Elgar Publishing.

Soto, V. and Frías-Martínez, E. (2011). Automated land use identification using cell-phone records. In *Proceedings of the 3rd ACM international workshop on MobiArch - HotPlanet '11*, page 17, New York, New York, USA. ACM Press.

Statham, T., Fox, S., and John, L. (2021). Identifying urban areas : A new approach and comparison of national urban metrics with gridded population data. (0):1–28.

Statham, T., Wolf, L., and Fox, S. (2020). Applications of Gridded Population Datasets : Delineating Urban Areas.

Stich, B. M. and Webb, P. (2019). Disconnects in Megaregional Freight Planning Are Holding Back the Louisiana Gulf Coast. *Public Works Management and Policy*, 24(2):149–159.

Sulis, P., Manley, E., Zhong, C., and Batty, M. (2018). Using mobility data as proxy for measuring urban vitality. *Journal of Spatial Information Science*, 16:137–162.

Sung, H. G., Go, D. H., and Choi, C. G. (2013). Evidence of Jacobs's street life in the great Seoul city: Identifying the association of physical environment with walking activity on streets. *Cities*, 35:164–173.

Taubenböck, H., Standfuß, I., Wurm, M., Krehl, A., and Siedentop, S. (2017). Measuring morphological polycentricity - A comparative analysis of urban mass concentrations using remote sensing data. *Computers, Environment and Urban Systems*, 64:42–56.

Taubenböck, H., Weigand, M., Esch, T., Staab, J., Wurm, M., Mast, J., and Dech, S. (2019). A new ranking of the world's largest cities—Do administrative units obscure morphological realities? *Remote Sensing of Environment*, 232(July):111353.

Toch, E., Lerner, B., Ben-Zion, E., and Ben-Gal, I. (2019). Analyzing large-scale human mobility data: a survey of machine learning methods and applications. *Knowledge and Information Systems*, 58(3):501–523.

Toole, J. L., Ulm, M., Bauer, D., and Gonzalez, M. C. (2012). Inferring land use from mobile phone activity.

Torija, A. J., Ruiz, D. P., and Ramos-Ridao, A. (2013). Application of a methodology for categorizing and differentiating urban soundscapes using acoustical descriptors and semantic-differential attributes. *The Journal of the Acoustical Society of America*, 134(1):791–802.

Umeda, Y. (2017). Time Series Classification via Topological Data Analysis. *Transactions of the Japanese Society for Artificial Intelligence*, 32(3):D–G72_1–12.

UN (2020). World Cities Report: Unpacking the Value of Sustainable Urbanization. pages 43–74.

Usui, H. (2019). A bottom-up approach for delineating urban areas minimizing the connection cost of built clusters: Comparison with top-down-based densely inhabited districts. *Computers, Environment and Urban Systems*, 77(July):101363.

van Oort, F., Burger, M., and Raspe, O. (2010). On the economic foundation of the Urban network paradigm: Spatial integration, functional integration and economic complementarities within the Dutch Randstad. *Urban Studies*, 47(4):725–748.

Virtanen, T., Plumbley, M. D., and Ellis, D. (2018). *Computational Analysis of Sound Scenes and Events*. Springer International Publishing, Cham.

Wang, M. and Debbage, N. (2021). Urban morphology and traffic congestion: Longitudinal evidence from us cities. *Computers, Environment and Urban Systems*, 89.

Wang, M., Derudder, B., and Liu, X. (2019). Polycentric urban development and economic productivity in China: A multiscalar analysis. *Environment and Planning A*, 51(8):1622–1643.

Wang, Q., Phillips, N. E., Small, M. L., and Sampson, R. J. (2018). Urban mobility and neighborhood isolation in america's 50 largest cities. *Proceedings of the National Academy of Sciences*, 115:7735–7740.

Wardrop, N., Jochem, W., Bird, T., Chamberlain, H., Clarke, D., Kerr, D., Bengtsson, L., Juran, S., Seaman, V., and Tatem, A. (2018). Spatially disaggregated population estimates in the absence of national population and housing census data. *Proceedings of the National Academy of Sciences*, 115(14):3529–3537.

Wasserman, L. (2011). Topological data analysis. *Inverse Problems*, 27(12):120201.

Wei, L., Luo, Y., Wang, M., Cai, Y., Su, S., Li, B., and Ji, H. (2020). Multiscale identification of urban functional polycentricity for planning implications: An integrated approach using geo-big transport data and complex network modeling. *Habitat International*, 97(129):102134.

Wheeler, S. M. (2015). *Five reasons why megaregional planning works against sustainability*. Number January.

Williams, A. (2012). What is a city? *Architectural Design*, 82(1):66–69.

Wolf, L. J., Fox, S., Harris, R., Johnston, R., Jones, K., Manley, D., Tranos, E., and Wang, W. W. (2020). Quantitative geography III: Future challenges and challenging futures. *Progress in Human Geography*.

Wu, K., Tang, J., and Long, Y. (2019). Delineating the regional economic geography of China by the approach of community detection. *Sustainability (Switzerland)*, 11(21).

Yang, J. and Zhou, P. (2020). The obesity epidemic and the metropolitan-scale built environment: Examining the health effects of polycentric development. *Urban Studies*, 57(1):39–55.

Yang, P. P. and Yamagata, Y. (2020). *Urban systems design*. Elsevier Inc.

Yildirim, Y., Allen, D. J., and Albright, A. (2019). The relationship between sound and amenities of transit-oriented developments. *International Journal of Environmental Research and Public Health*, 16(13).

Yu, M. and Fan, W. (2018). Accessibility impact of future high speed rail corridor on the piedmont Atlantic megaregion. *Journal of Transport Geography*, 73(October):1–12.

Zambon, G., Benocci, R., Bisceglie, A., Roman, H. E., and Bellucci, P. (2017). The LIFE DYNAMAP project: Towards a procedure for dynamic noise mapping in urban areas. *Applied Acoustics*, 124:52–60.

Zambon, G., Benocci, R., and Brambilla, G. (2016). Statistical road classification applied to stratified spatial sampling of road traffic noise in urban areas. *International Journal of Environmental Research*, 10(3):411–420.

Zanella, A., Bui, N., Castellani, A., Vangelista, L., and Zorzi, M. (2014). Internet of things for smart cities. *IEEE Internet of Things Journal*, 1(1):22–32.

Zanganeh Shahraki, S., Sauri, D., Serra, P., Modugno, S., Seifolddini, F., and Pourahmad, A. (2011). Urban sprawl pattern and land-use change detection in Yazd, Iran. *Habitat International*, 35(4):521–528.

Zhan, X., Ukkusuri, S. V., and Zhu, F. (2014). Inferring Urban Land Use Using Large-Scale Social Media Check-in Data. *Networks and Spatial Economics*, 14(3-4):647–667.

Zhang, F., Wu, L., Zhu, D., and Liu, Y. (2019a). Social sensing from street-level imagery: A case study in learning spatio-temporal urban mobility patterns. *ISPRS Journal of Photogrammetry and Remote Sensing*, 153(December 2018):48–58.

Zhang, M., Fu, H., Li, Y., and Chen, S. (2019b). Understanding Urban Dynamics From Massive Mobile Traffic Data. *IEEE Transactions on Big Data*, 5(2):266–278.

Zhang, X. Q. (2016). The trends, promises and challenges of urbanisation in the world. *Habitat international*, 54:241–252.

Zuo, F., Li, Y., Johnson, S., Johnson, J., Varughese, S., Copes, R., Liu, F., Wu, H. J., Hou, R., and Chen, H. (2014). Temporal and spatial variability of traffic-related noise in the City of Toronto, Canada. *Science of the Total Environment*, 472:1100–1107.