NETWORK CENTRALITIES IN POLYCENTRIC URBAN REGIONS:

Methods for the Measurement of Spatial Metrics

A thesis submitted for the degree of Ph.D. in Urban Space and Computation

Space Syntax Laboratory

Bartlett Faculty of the Built Environment
University College London

By

Kimon Krenz

April
2018

DECLARATION

I, Kimon Krenz confirm that the work presented in this thesis is my own. Where information has been derived from other sources, I confirm that this has been indicated in the thesis.

Signature

ABSTRACT

The primary aim of this thesis is to explain the complex spatial organisations of polycentric urban regions (PURs). PURs are a form of regional morphology that often evolves from post-industrial structures and describe a subnational area featuring a plurality of urban centres. As of today, the analysis of the spatial organisation of PURs constitutes a hitherto uncharted territory. This is due to PURs' inherent complexity that poses challenges for their conceptualisation. In this context, this thesis reviews theories on the spatial organisation of regions and cities and seeks to make a foundational methodological contribution by joining space syntax and central place theory in the conceptualisation of polycentric urban regions. It takes into account human agency embedded in the physical space, as well as the reciprocal effect of the spatial organisation for the emergence of centralities and demonstrates how these concepts can give insights into the fundamental regional functioning. The thesis scrutinises the role that the spatial organisation plays in such regions, in terms of organising flows of goods and people, ordering locational occupation and fostering centres of commercial activity. It proposes a series of novel measurements and techniques to analyse large and messy datasets. This includes a method for the application of large-scale volunteered geographic information in street network analysis. This is done, in the context of two post-industrial regions: the German Ruhr Valley and the British Nottinghamshire, Derbyshire and Yorkshire region. The thesis' contribution to the understanding of regional spatial organisation and the study of regional morphology lies in the identification of spatial structural features of socio-economic potentials of regions and particular areas within them. It constitutes the first comparative study of comprehensive large-scale regional spatial networks and presents a framework for the analysis of regions and the evaluation of the predictive potential of spatial networks for socio-economic patterns and the location of centres in regional contexts.

IMPACT STATEMENT

This research into network centralities in polycentric urban regions (PURs), and their relationship to socio-economic patterns of density and location has enhanced understanding of spatial structures and geometric relations in cities and regions.

This project benefits inside academia in the disciplines of urban planning and design, mathematical graph theory, sociology and computer science, by bringing together central place theory and space syntax; by developing methodologically and theoretically the approach of configurational analysis; by providing methods for the analysis and evaluation of regional spatial networks; by proposing new ways for the generation of randomised regional spatial networks and contributing to the knowledge of their statistical properties; and by establishing a database of street-level population data for potential use in the fields of epidemiology, ecology, geoscience or communication.

Results of this research have been presented at international conferences and disseminated in a series of peer-reviewed proceedings and journal articles. The knowledge resulting from this research has impacted the teaching curriculum at UCL in the form of series of workshops and seminars in the MSc Space Syntax: Architecture and Cities, MRes Inter-disciplinary Urban Design, MA Architecture and Historic Urban Environments, as well as the MArch Urban Design at the Bartlett School of Architecture. The methods presented in this project have the potential to contribute to space syntax and urban planning curricula towards a stronger emphasis on the importance of regional relationships.

This research can benefit outside academia by informing the development of tools for the assessment of the impact of large-scale infrastructural developments and urban expansion, as well as for the management of PURs. The methods and workflows presented in this thesis have informed the development of a publicly accessible tool for the simplification of street network data for spatial network analysis. Moreover, the knowledge resulting from this project can inform the identification of areas for development investment and potentially aid policy strategies aiding economic growth.

ACKNOWLEDGMENT

TABLE OF CONTENTS

TABLE OF FIGURES

TABLE OF ABBREVIATIONS

AADF .... Annual average daily flow
AADF_All .... Average annual daily traffic for all motor vehicles
AADF_FR .... Average annual daily traffic for all freight traffic
AADF_HV .... Average annual daily traffic for all heavy goods vehicles
AADF_PA .... Average annual daily traffic for all passenger traffic
AADT .... Annual average daily traffic
AB .... AddressBase Premium dataset
AbsH2 .... Relative building height
AbsHMax .... Maximum building height
AbsHMin .... Minimum building height
ALKIS .... Authoritative real estate cadastre information system
ASA .... Angular segment analysis
BC .... Betweenness centrality
CC .... Closeness centrality
CDF .... Cumulative distribution function
CPT .... Central Place Theory
CRS .... Complete spatial randomness
DPA .... Douglas-Peuker algorithm
EDA .... Exploratory data analysis
EFA .... Exploratory factor analysis
EFA BC 1 - I .... Inter-regional latent centrality structure for betweenness centrality
EFA BC 2 - II .... Intra-regional latent centrality structure for betweenness centrality
EFA BC 3 - III .... Metropolitan latent centrality structure for betweenness centrality
EFA BC 4 - IV .... City latent centrality structure for betweenness centrality
EFA BC 5 - V .... Neighbourhood latent centrality structure for betweenness centrality
EFA CC 1 - I .... Inter-regional latent centrality structure for closeness centrality
EFA CC 2 - II .... Metropolitan latent centrality structure for closeness centrality
EFA CC 3 - III .... City latent centrality structure for closeness centrality
EFA CC 4 - IV .... Neighbourhood latent centrality structure for closeness centrality
ER .... Erdős-Rényi random graph model
ERPG .... Erdős-Rényi random planar graph
ERPGr .... Erdös-Renyi random planar graph with radius restriction
ESSnet .... European Statistical System
EU .... European Union
FUR .... Functional urban region
GE .... Germany
GIS .... Geographic Information Systems
ITN .... Integrated Transport Network
KSL .... Lillifors test
LCF .... Latent centrality factor/structure
LoD 0-4 .... 3D level of detail classification
MSD .... Metric step depth
NACH .... Normalised choice i.e. NBC
NBC .... Normalised betweenness centrality

19

NCB .... National Coal Board
NCC .... Normalised closeness centrality
NDY .... Nottinghamshire, Derbyshire and Yorkshire region
NRW .... North-Rhine Westphalia
OS .... Ordnance Survey
OSM .... OpenStreetMap
PA .... Parallel analysis
POI .... Points of interest
PUR .... Polycentric urban region
RANN .... Fast nearest neighbourhood search algorithm
RGG .... Random geometric graph
RTA .... Regional trade agreements
SIMP .... Simplified OSM network
SLW .... Segment length weighted
SM .... Segment map
TCB .... Town centre boundaries
TD .... Total depth
UK .... United Kingdom
VGI .... Volunteered Geographic Information
VPGr .... Variance gamma planar graph with radius restriction
WTO .... World Trade Organisation

CHAPTER 1

FROM THE CITY TO THE REGION

1 CHAPTER

1.1 INTRODUCTION

This thesis is concerned with the spatial organisation of regions, cities and the urban areas within them. More particularly, it focuses on the role that centralities play in the way society configures regional spaces. It seeks to make a foundational methodological contribution by bringing together space syntax and central place theory in the conceptualisation of polycentric urban regions, proposing methods for the use of complex and highly messy datasets and exposing the inadequacy of existing polycentric models. The object is to pinpoint the methodological and theoretical shortcomings in the current regional economics and urban planning literature. It will be argued that economic and planning theories focused on the distribution and organisation of regions neither take sufficiently into account human agency embedded in the physical space, nor the reciprocal effect of the spatial organisation for the emergence of centralities, which means relative accessibility.

The case studies selected for this research are two former coal-mining regions: the German Ruhr Valley and the British Nottinghamshire, Derbyshire and Yorkshire region. Such regions are now referred to as 'post-industrial' in an acknowledgement of their transitional process from an economy primarily based on the manufacturing sector to an economy based on the service sector. The spatial organisation of both regions is highly influenced by developments during the industrialisation that produced a complex spatial organisation, which is so often referred to as polycentric urban region (PUR). Both regions are comparable in their historical and socio-economic development as well as their form of spatial organisation, which differs substantially from other regional types. Post-industrial regions face major transitional challenges, for instance, a decline in manual labour (e.g. factory worker) and an increase in professional workers (e.g. creative-industry, research). The ability to adapt to such transitions is linked to the past developments of these regions and their particular physical realities, such as spatial mismatches describing locational disparities between supply and demand.

Theories dealing with the spatial organisation or urban economies are often either focused on individual cities, thus neglecting the importance of inter-city and regional relationships, or deal with the regional continuum and its regional relationships by simplifying cities to abstract nodes. In this context, this thesis will reflect on two main bodies of work: Walter Christaller's (1933) 'central place theory' which explains the spatial arrangement, size and number of cities and their centres by mechanisms of

market size and economic distance; and Bill Hillier's (2004) notion of cities as 'movement economies' which emphasizes the role of the spatial configuration for movement and accessibility. Both theories are consistent within their conceptual frameworks, the region and the city, and make fundamental contributions to the way we can understand the spatial organisation of both. A comprehensive epistemological and methodological discussion of these two schools of thought is presented in Chapter 2. It will be argued that neither of these approaches, however, fully captures the complexities and spatial challenges of post-industrial urban clusters.

Research undertaken for this dissertation tests both the notion of central places and spatial configuration to the problems of two post-industrial polycentric urban regions, i.e. their particular spatial organisation. I advocate a hybrid approach, for which I develop new methodologies to pursue this. The primary aim is to grasp the complex spatial organisation of PURs. What are the main drivers in the spatial development of PURs? Related to this primary question are subsequent questions such as: what role does the spatial organisation play in fostering socio-economic activities in PURs? To what extent are microeconomic mechanisms, which explain the spatial organisation of traditional regions of any relevance for PURs?

In order to respond to these questions, I develop and propose a series of novel measurements and techniques (Chapter 4–6). This work informs the first comparative study of comprehensive large-scale regional spatial networks (Chapter 8). It contributes by extending the field of space syntax to the regional continuum, developing new methods to overcome radii selection in space syntax analysis and providing a conceptualisation for hidden spatial scales (Chapter 7). It presents a framework for the analysis of regions and the evaluation of the predictive potential of spatial networks for socio-economic patterns and the location of centres in regional contexts (Chapter 8). The thesis establishes a novel methodology to employ volunteered geographic information enabling efficient large-scale spatial network analyses (Chapter 5), as well as a new model for randomising regional spatial networks for the purpose of comparisons of large-scale intra-regional spatial networks (Chapter 6). One of the outcomes of this endeavour is a new database of precise population estimates at the street level.

1.1.1 FROM THE CITY TO THE REGION

The United Nations Population Division ascertains that the urban population represents the biggest share of the world's inhabitants since the new millennium (United Nations 2013). As for today and even more so in the future, cities are the major human habitat with a rapid growth in population. Globalisation is certainly one major

driver of this process. Acknowledging the suggestion of some that globalisation is not an entirely new phenomenon (Hirst et al. 2009; Steger 2013) and the question of whether this process can be seen as entirely global, it still seems difficult to misjudge the pace and the extent of growing transnational interactions and trade. The number of Free Trade Areas (such as the European Economic Area), for instance, has grown significantly over the last decades. While a total of 124 of Regional Trade Agreement (RTA) notifications have been made in the period of 1948-1994, over 400 of such arrangements were notified since then (WTO Secretariat 2016). Almost all members of the World Trade Organisation have agreed on at least one RTA (ibid.). Numerous authors have linked market liberalisation with increasing productivity and economic performance. There are a number of cities, such as New York or London, that can be considered as benefiting massively from these developments, by producing growing population and wealth generation during the recent decades.

A corresponding body of work has been produced to examine the spatial implications of globalisation on cities and their metropolitan areas under the notion of 'global cities' or 'world cities' (Castells 2000; Friedmann 1986; Sassen 1991). Global cities are particularly characterised by their primary economy, i.e. finance and service sector (ibid. p. 126). These cities feature an extensive range of internationally operating corporate service firms and an advanced producer service sector. While for some examples, like London or Tokyo, their distinction as independent cities is relatively clear, others have – fuelled by their intensive growth – merged with surrounding cities and blurred their distinctive geographical boundaries such as the cities Nagoya, Osaka, Kyoto and Kobe in Japan or Hong Kong, Shenzhen and Guangzhou in China. Where the early 21st century city begins and ends has become increasingly fuzzy (MacLeod and Jones 2011). It is this fuzziness of newly emerged metropolitan areas that causes new challenges to traditional representations and presents difficulties for the theoretical and analytical conceptualisation (Sieverts 1997).

Other scholars have argued that, in fact, it is not so much the individual city that should be at the focus to understand spatial implications of globalisation but instead the region as a whole (Meijers 2007; Soja 2011). The tendency towards a more regional focus can be understood as an awareness that cities cannot be studied independently. Any form of human settlement exists in relation to other settlements, whether this is in the form of exchanges of goods, capital, information or human interaction (see Figure 1 for an example of the complex relationship of commuting patterns in the Bay Area, US). Exchanges of such flows can assume different forms and take place on different networks, spatial and non-spatial, but with urban areas at the core of their functioning. Rather than individual cities, it has been argued that a new type of polycentric agglomeration is linked to this phenomenon, which highlighted the rise of a new type of

urban morphology, the 'polycentric urban region' (Agnew 1993; Parr 2004, 2008; Scott 2002; Soja 2011).

The study of the spatial organisation of regions can look back on a long history of studies such as Johann Heinrich von Thünen's 'The Isolated State' in the early 19th century and Walter Christaller's 'central place theory' in the 1930s. The study of the spatial organisation of regions reached a peak with Peter Haggett's 'locational analysis' in the mid 1960s. After this peak, the field ceased to be of greater importance, which was partly caused by the concerns about environmental determinism.

PURs are considered to be a spatial manifestation of a globalising world. This puts regions back at the focus of the research agenda and sees them as important elements to understanding socio-economic trajectories, particularly in the light of globalisation. This has led to claims such as that 'regions are once again emerging as important foci of production and as repositories of specialised know-how of technological capability, even as the globalisation of economic relationships proceeds apace' (Scott 1999 p. 9).
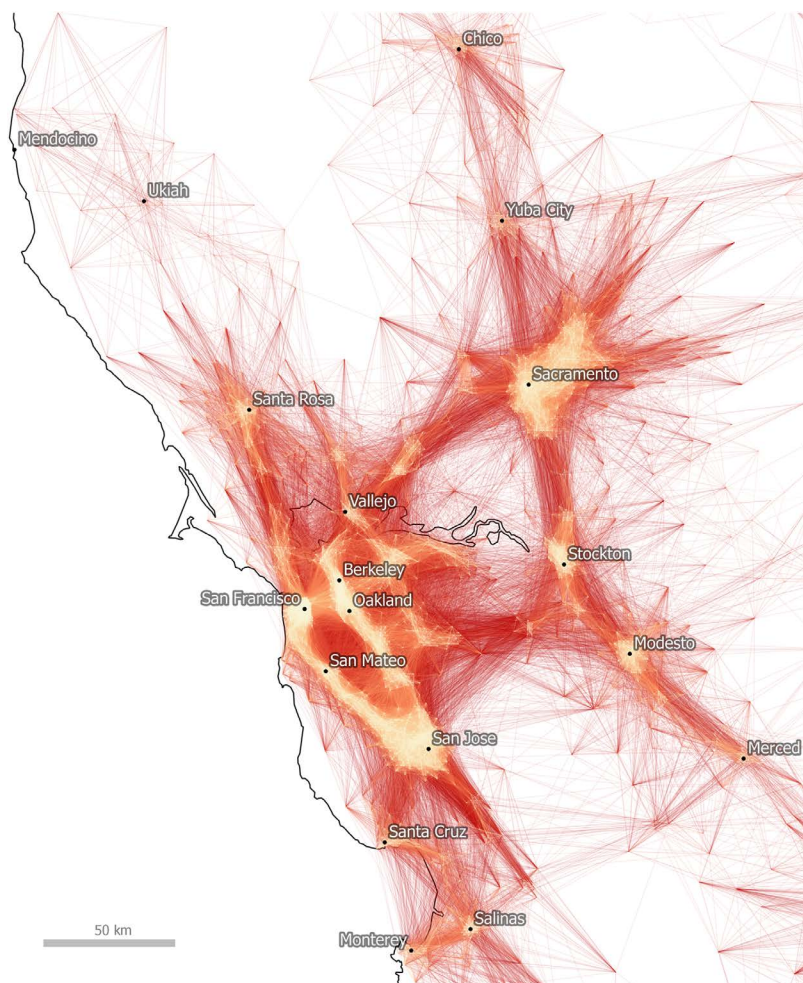
Figure 1: Tract-to-Tract commutes of 80km or less in the Bay Area, United States of America. (Figure by Dash Nelson and Rae 2016)

It remains a question what *exactly* makes regions in general and PURs in particular such important drivers for innovation and economic success? Apart from the role PURs play in a globalised economy, it might be their particular spatial configuration that holds a pivotal role for productivity and the generation of innovation. However, very little is known in terms of their spatial organisation. This is because, on the one hand, traditional theories on the spatial organisation of regions and regional urban space lack the concepts for such complex entities, and, on the other hand, because there is a lack of appropriate methods for the analysis of large spatial entities.

In recent years, it has become clearer that in order to arrive at meaningful approaches to the study of cities and regions, there needs to be more of a common ground. This becomes even more pronounced with the increase in interdisciplinarity, the new magnitude of computational power, sophisticated quantitative approaches and big data. Recently Michael Batty (2013) has emphasised what such a common ground could be and even called for the establishment of an independent science of cities. Batty synthesises concepts from complexity theory and links them to network science. His notion draws on the connection of defining cities as places of flows that are organised in networks. Batty argues that in order to understand cities, one needs to understand their networks (ibid. p. 3). This is not only the case for cities but should be extended to the regional continuum, because if we want to understand PURs, we need to understand their networks. Network-based approaches offer an incorporation of an entire set of simultaneous relationships allowing us to decipher complex interdependences from the local neighbourhood to the overall region. In the context of PURs, this implies that one can investigate, on the one hand, relationships between cities as systems of cities, and, on the other hand, also relationships between local urban areas as systems of urban areas in a regional embedding. Employing network-based approaches allows examining the regional continuum on its interrelation of urban areas, on interaction, connection and as spaces of flows, or as Sieverts puts it on the 'in-between' (Sieverts 1997). It is this complex 'in-between' which makes the spatial organisation of PURs particularly difficult to decipher and where the network can provide valuable insights into visible and hidden relationships.

1.2 RESEARCH OBJECTIVES

From a general point of view, this thesis aspires to contribute, first, by bridging the gap between local and regional theories on the spatial organisation of society. The research sets out with the juxtaposition of two general points of view on how society organises itself in space; Bill Hillier's (1996, 2004) notion of 'cities as movement economies' focuses on urban space and Walter Christaller's (1933) 'central place theory' aims to explain the regional construct. This is discussed in the context of polycentric urban

regions, an emerging regional morphology whose spatial configuration constitutes a hitherto uncharted territory. The two regions selected are particularly interesting in this context as they are both post-industrial and coal-mining regions that do not adhere to the traditional growth developments due to their extensive and rapid growth during industrialisation. Their particular economic histories lead them to develop a different hierarchy of centres that is not comparable with traditional regional continua. Both regions are, therefore, expected to exhibit different patterns of spatial organisation than traditional regions. The aim is to investigate the role the spatial organisation plays in such regions in organising flows of goods and people, ordering locational occupation, i.e. places of residency and fostering centres of commercial activity. This research responds to Batty's (2013) definition of the 'new science' of cities by examining the polycentric urban region as a system of networks and flows transferring the notion to the regional continuum and highlighting the role that regional embedding plays for urban areas.

The second contribution is a proposition on how space syntax can be employed to the region, revitalising the tradition of regional analysis from the early 20th century and combining it with dimensions of human agency that space syntax incorporates. This study employs Bill Hillier's (2005) concepts of to- and through-movement, which builds on the two network centrality measures betweenness and closeness centrality, as proxies for human movement behaviour and socio-economic patterns on a regional scope. This study proposes methods and tests ways of studying regions via their network and their fundamental physical domains: the building and the street. This thesis contributes by proposing a novel method for the use of volunteered geographic information for the construction of comprehensive spatial network representations of regions. As a consequence of the lack of existing comparable findings in systematic regional morphological investigations, this thesis develops a method of generating randomised regional street networks as a tool for comparison.

Ultimately, the thesis seeks to contribute to the field of regional network analysis and to engage future researchers in the field by providing not only the tools but also the concepts for potential future work in the area of regional morphological analysis. Finally, this thesis aspires to highlight the importance of considering the reciprocal relationship of urban space and its regional embedding. Such an embedding is particularly important in the context of transitional challenges of PURs because it highlights the need to consider regional relationships for policies and planning interventions aimed at the local scale.

1.3 REGIONAL SPATIAL ORGANISATIONS

1.3.1 DEFINING POLYCENTRIC URBAN REGIONS

Before elaborating on the theoretical implications of the proposed shift from a city-based to an inter-city/regional perspective, it is necessary to define fundamental terms and their origin. This section seeks to clarify what is meant when the term 'region' is used. This is necessary as the term is used with reference to a variety of geographic contexts, sizes and scales in the academic literature. A focus will be placed on what has been termed the 'polycentric urban region' (PUR), as it constitutes a very particular regional morphology that often applies to post-industrial regions, such as the two selected cases of the Ruhr Valley and the Nottinghamshire, Derbyshire and Yorkshire region.

In geography the term, region, is broadly applied to three fundamental territories, *trans-, supra-* and *sub-national* (Trippl et al. 2012 p. 13). While supra-national territories are regions that constitute several nations (e.g. Latin-America, south-east Asia) and trans-national territories are regions that stretch across two or more adjacent states (e.g. EUREGIO[1]), this study's interest is in the third kind that can be described as sub-national territory (ibid. p. 14). Sub-national regions are to be found within one independent nation and share, contrary to trans- and supra-national territories, the same political and socio-economic systems. There are numerous definitions of different types of regions, and an analytical description of these would go beyond this discussion, rather it should be noted that the object of this research is a particular type of region within the category of sub-national territories, namely polycentric urban regions. The concept of polycentrism is often used to describe a hierarchical organisation of many different centres of any kind.

Hierarchy in polycentrism is usually referred to from political, social or economic perspectives, and centres are defined on spatial as well as non-spatial properties. To employ the notion of polycentrism in the context of PURs it is necessary to give a more precise definition. PURs are regions that consist of a number of historically distinct, politically and administratively independent cities in close proximity to each other (Kloosterman and Lambregts 2001 p. 718). Important is the lack of a dominating central city and a rather even distribution of a smaller number of similar-sized cities of equal economic importance and an even greater number of smaller cities (ibid. p. 719). There are other terms used to describe such structures, such as 'Zwischenstadt' (i.e. in-between city) (Sieverts 1997), 'city-region' (Scott 2002), 'city networks' (Camagni and

[1] Cross-border region between the Netherlands and Germany and the first Euroregion.

Salone 1993), 'network cities' (Batten 1995), or Hall and Pain's (2006) concept of the 'polycentric mega-city region'. The latter entails the inclusion of metropolitan cities with multiple cores and does, hence, not apply to the given definition above. This is also the case for concepts such as the 'multi-core metropolis' (Hall 1999) that is considered to be polycentric but relates to morphologies that are governed by a primary urban centre, rather than equivalent centres such as in PURs. The particular interest of this thesis is on the spatial organisation of polycentric urban regions and the intra-regional relationships between parts within such regions. The reason for why the focus is particularly on the intraregional relationship is because, as it will be argued, it is this internal relationship between parts and the larger continuum that constitutes challenges in transitional processes.

URBAN HIERARCHY AND THE RELATIONSHIP OF CITIES. Relationships between cities are usually defined as concepts of hierarchy (Meijers 2007). Such Hierarchies help to define the relative importance between parts and distinguish different systems by their hierarchical structures (i.e. linear, branching, flat, overlapping, nested hierarchy). On a city scale, most notably the notion of 'a city is not a tree' by Christopher Alexander (1966) describes how the urban form is hierarchical, intrinsically network-based and multi-relational. On a regional scale, definitions of hierarchical forms underwent an extensive development. It developed from Walter Christaller's 'Central Place Theory' (CPT) (1933) and August Lösch's (1940) 'Economics of Location' concept. Both authors argue that urban functions are organised hierarchically with distinctive levels. The development further stretched to contemporary notions such as Saskia Sassen's 'New Economy' and the 'Global City' (1991, 2005) emphasizing that urban functions depend on networks. In alignment with Sassen's notion is also Manuel Castells' 'Network of Flows' (2000). Castells focuses on the importance of global flows of capital, people and information as relational factors of urban functions.[2] The general regional debate is highly influenced by differences in scale, which have a strong influence on the concepts and their transferability to alternative scales. The allocated hierarchical level of a city in the global economy network does not necessarily relate to a hierarchy a city holds within the regional context. Moreover, the concept of hierarchies is not strictly defined, affecting the ability to connect concepts across different geographies.

The traditional concept of hierarchy refers to a natural order (or rank) that emerges between cities and urban settlements, which share the same socio-economic system. In the past, such hierarchy was mainly operationalised as independent cities and their size, i.e. a city's total population (see also Zipf's rank-size distribution (1949)). In this sense,

---

[2] See Neal (2011) for a comprehensive review.

hierarchy is a concept that is determined by single entities in network-based systems, rather than by their network relationships. For Michael Batty (2006) this kind of size-based hierarchy can be described as 'natural ordering' of urban systems. I argue that this traditional operationalisation of hierarchy derived from a city's size becomes inapplicable for the context of PURs, where boundaries are blurred, and the independent city dissolves into an agglomeration of urban spaces.

Many scholars from fields such geography, urban studies and planning have argued that this 'simple' operationalisation of hierarchy alone cannot account for the complexity encountered in cities and regions (Alexander 1966; Batty 2006; Meijers 2007). Furthermore, Batty (2006) shows that hierarchy is manifold and can occur in many domains aside of the distribution of population sizes, such as in retail activities or network connectivity. He emphasises the need to loosen the strict idea of hierarchies and instead focus on the exploration of models that can account for overlaps of different hierarchies. Batty's point of departure is an evolutionary perspective, where 'natural hierarchies' are seen as an intrinsic part of cities, as a result of bottom-up evolutionary processes that grow organically over time, rather than being introduced by top-down planning interventions (ibid. p. 166). Even though, early notions such as Zipf's rank-size distribution, or Christaller's CPT fit into this line of thought, i.e. that hierarchies develop as the result of bottom-up processes, this has not been expressed as clearly as in Batty's work.

An impediment to the enhancement of the notion of hierarchies is the focus of the single city. The work of Sassen (1991) and Castells (2000) shows how in the context of globalised economies a focus on the network relationship and the flows within them can help to reveal the global hierarchy of cities. How such a shift can also inform our understanding of PURs, however, remains an open question. The current academic discourse on regions has already begun to shift the focus of the concept of hierarchy from definitions through the 'within' (i.e. the individual city) and instead focuses on the 'between' (i.e. the relationship of cities). Most notably Meijers (2007) demonstrates how an emphasis of the 'between' can provide substantial insights into the urban functioning of PURs. Her work on synergies (a notion where cities relate to each other in such a way that their network creates more than the sum of its parts) in PURs, highlights how cities and urban areas within PURs can benefit from their internal relationships and even create knowledge agglomerations that can be compared to cities of the global hierarchy from an economic point of view. Nevertheless, Meijer's approach also operationalised hierarchy in terms of the network of cities in PURs, with independent administratively defined cities.

POLYCENTRIC URBAN REGIONS: A DEFINITION. The term polycentric urban region is, however, contentious and often used ambiguously in the literature. The increased academic interest in polycentric regions of urban character is largely due to their alleged superior economic potential (Parr 2004 p. 232). Although the PUR is believed to be potentially economically successful due to a less hierarchical flow of goods and people, there is very little rigorous, empirical and comparative evidence on the spatial organisation that fosters such flows and whether this causally produces economic success. Simultaneously, post-industrial regions which often are considered to be PURs often face an initial process of economic decline, sometimes combined with a population decline, which can lead to a spiral of urban deterioration. It is hence, important to emphasise that the concept of PURs is first of all one of a description for a sub-national territorial spatial organisation.

The social theorist John Parr provides a maximum definition that is based on seven conditions that are most often defined as crucial components of PURs by the majority of scholars (2004). The maximum definition precedes a somewhat basic approach towards PURs that could be described as a minimum definition. According to this minimum definition, a PUR must be a polycentric region of urban character, where the 'polycentrality' denotes a plurality of centres, a 'region' refers to a subnational territory and 'urban' means that the vast majority of the population, as well as employment, must be located in a specified set of urban centres (ibid.). As this definition provides us with only little concrete information, Parr provides a set of seven more concrete conditions that must be in place for an area to be defined as PUR. This maximum definition is comparative in nature, as it requires the benchmark of a putative non-polycentric region, preferably within the same nation state or larger supra-national context, against which another potential PUR can be evaluated. The benchmark region can be either an existing region, or a constructed entity based on the average characteristics of the reference nation, yet it should have comparable levels of population, urbanisation and aggregate income.

The first condition refers to the clustering of centres. In a PUR there must be urban centres that are spatially divided by open, vacant or agricultural space. The centres form a cluster that is neither random nor regular, while it is trivial whether they are positioned in a linear, circular or polygonal shape. This criterion refers to the difference between spaces which are built-up and those that are not evenly settled. The second requirement refers to the upper limit on centre separation. Centres must lie within a maximum level of separation of another as otherwise a PUR could comprehend an entire nation. Here, scholars have varying different definitions as to how far this distance can be, yet usually, the travel time to the nearest centre should not exceed one hour. This time might be used as a benchmark to determine whether a neighbouring city is part of

the PUR. Parr criticises that these distances are inevitably arbitrary, yet if they are coherently employed, they prove useful (ibid.). The third criterion focuses on the distance between the respective centres in the PUR by including a lower limit on centre separation to ensure that the region considered as a PUR is in fact not a conurbation or a multi-centred metropolitan area. Again, the distance is arbitrary but should be employed reasonably. The fourth condition refers to the size and spacing of centres. Overall, the centres in a PUR must be more closely located to one another than in the benchmark region, and the sizes of the centres must be greater, meaning that the urban sprawl must be greater than in cities of a comparable absolute population and economic capacity. The fifth condition is concerned with the size and distribution of the centres and demands that there must not be a marked difference in size between the largest centres of the PUR's region (ibid.). Parr distils a sixth criterion, dealing with the interaction among centres: the amount of economic interaction between all centres within the PUR, such as overlapping labour markets and commute, must be larger than in a comparable region. The domain of this exchange can span over retail and a large range of business services. Parr emphasises that the trade in a PUR must be less hierarchical and can exist between two equally proportioned centres or can even be mutual between a smaller and larger centre. This strong interrelatedness leads to equal economic trading levels of the centres. The seventh criterion looks at the centre specialisation. The criterion demands that a centre of a PUR must have a higher specialisation compared with a centre in a benchmark region of the same size. The underlying idea is that there is a spatial division of labour across centres (Massey 1984). Parr emphasises that due to its maximalist extent, the definition might possibly be too narrow and exclude too many regions that, if not conceptualised as PUR, would hinder a fruitful analysis of this type. Still, he does not specify which of the criteria should be of lesser importance or whether a minimum number of the seven criteria must be fulfilled (Parr 2004).

Whereas Parr employs a series of spatial characteristics, such as distances between centres and the spatial pattern of settlements, there is no established method on how to statistically evaluate these factors. The PUR debate is mostly driven by semantics and debates over the very definition itself, rather than on an actual spatial manifestation or how polycentric regions are spatially configured and what this configuration implies for the socio-economic success of these regions. This situation is partly caused by the complexity PURs exhibit and partly by the lack of comparative methods to grasp specific patterns in these complexities. Moreover, what can be considered as a 'centre' or at which degree a centre becomes of importance for the relevance of PURs is not clarified.

1.4 RESEARCH OUTLINE

This introduction highlighted that contemporary globalisation processes have pushed approaches to regional scale studies back up the research and policy agenda. This return is caused by the new emergence of large urban regions in general and polycentric urban regions in particular. PURs are a new form of regional morphology that often evolves from post-industrial structures and poses transitional challenges linked to their industrially shaped spatial organisation. How this particular spatial organisation can be understood has largely remained unanswered, and it is one of the aims of this thesis to propose methods and concepts to establish such knowledge. The thesis is outlined as follows:

Chapter 2 offers an epistemological and methodological discussion of central place theory and space syntax, which are two major theories on the spatial organisation of regions and cities. This discussion is the foundation for setting up a dialogue between the two schools of thought. It will be argued that both schools have not been linked so far due to differences in their definition of hierarchy, scale and their conceptualisation of distance. I argue that in the light of the fresh impetus to revitalise the tradition of regional scale that Walter Christaller and August Lösch established almost a century ago, space syntax has an unexpected contribution to make in this arena. Space syntax adds new dimensions such as the incorporation of human agency and the conceptualisation of space as relational entity. The chapter concludes with a proposition on how these notions can be applied in the context of polycentric urban regions. Moreover, it will be argued that space syntax has an unexpected contribution to make in this arena, beyond the added dimension of human agency.

Chapter 3 introduces both of the selected case studies, the German Ruhr Valley and the British Yorkshire, Derbyshire and Nottinghamshire region and argues that both regions can be considered prime examples of PURs. By analysing their historical spatial development, it will be shown that both regions feature a highly interrelated and dense network of railways, canals and streets, which led to a fragmented polycentric settlement structure without a dominating city. It will be argued that both regions form prime examples of polycentric urban regions in alignment with Parr's definition of PURs. However, this analysis also exposes the inadequacy of the PUR definition in providing a framework that allows understanding the complex regional organisation that PURs exhibit.

Chapter 4 gives an extensive account of the data used in this study as well as the methodological approach chosen to scrutinise the research questions discussed. It functions as an introduction to the difficulties large messy datasets bring, as well as how

such data can be operationalised in the context of regional network analysis. The selected data is based on the theoretical positioning of CPT and space syntax theory and explained in Chapter 3. The data selection aims to reflect three fundamental characteristics of human activity in space: i) where humans move in space, ii) where humans occupy space, and iii) where humans commercially interact in space. Human spatial occupation is physically represented in the form of 3D-building information and population data, human movement is operationalised by regional traffic flow data, and human microeconomic activity is represented in the form of semantic information on the location service and trade functions. Moreover, this chapter introduces challenges in relation to scale, resolution and precision across different datasets and introduces a series of data disaggregation and aggregation methods to bridge differences in spatial representations from the large statistical sample grid, to the level of buildings and streets.

Chapter 5 (and 6) constitutes a methodological extension of Chapter 4, with a specific focus on spatial network analysis. It introduces networks as a form of spatial representation employed in space syntax analysis and elaborates on the historical development of this method. The chapter includes a critical reflection on the implications and challenges posed by transferring this method to a regional context. I propose to make use of volunteered geographic information, i.e. OpenStreetMap (OSM) data, as a globally comparable data source for spatial network analysis. I present and test a method developed for this thesis, which enables the use of such data in large-scale network analysis applications.

Chapter 6 elaborates on the properties and statistical characteristics that regional street networks exhibit. It reflects on the methodological problem that arises due to the lack of comparable research in the field of regional morphology and introduces methods for a randomised regional street network generation. It includes a critical reflection of these methods and emphasises the shortcomings in comparability of these methods when applied to real-world models. Finally, I propose a novel method for random regional street networks (i.e. Variance Gamma Planar Graph with radius restriction) that offers comparability to real-world street networks.

Chapter 7 presents the concept of latent centrality structures for hidden scale structures in regional spatial networks, as well as the exploratory factor analysis (EFA) as a method to reveal such fundamental hidden structures. The chapter presents the results of an EFA building on a large set of 400,000,000 centrality values of betweenness and closeness centrality of the two case study regions (introduced in Chapter 5) and the two randomised regional street network models (introduced in Chapter 6). These results are compared with each other and inferences on the impact of human action on regional

spatial organisation are presented. The chapter also includes the proposition of a novel method for the visualisation of such latent centrality structures as well as a new combined multi-scalar model that resolves bias issues in radius selection.

Chapter 8 presents a comprehensive analysis of latent centrality structures with socio-economic variables (introduced in Chapter 4). More precisely, this chapter presents the results of the prediction of regional movement by spatial metrics, followed by regional population predictions and finally by an identification of the relationship between service and trade centres and latent centrality structures, i.e. scales. The tests show that spatial network metrics and latent centrality structures hold substantial explanatory power for the prediction of regional movement and the location of service and trade centres on the level of the spatial network segment.

Finally, Chapter 9 constitutes the concluding discussion. It presents a detailed account of how this research contributes to knowledge, the novelties and limitations, as well as potential prospects for future research.

CHAPTER 2

SPATIAL CONFIGURATION OF PURs

2 CHAPTER

This chapter presents a discussion on methodological and epistemological challenges in understanding polycentric urban regions. It reflects upon two fundamental theories on the spatial organisation of cities and regions, Christaller's central place theory and Hillier's theory on movement economy. I highlight the strength and limitations of both theories in the context of polycentric urban regions and explain the implications for a methodological operationalisation in the context of PURs. This is done by elaborating on two fundamental concepts of hierarchy and scale within both theories. This chapter proposes to move beyond the notion of the individual cities within the CPT and space syntax towards a concept of urban space and its regional embedding. The analysis of relative centralities for each urban space is a novelty in the investigation of centres and their accessibility on a regional scale. It will be argued that the concepts derived from both theories form valuable starting points, but a combinatory approach is necessary in order to conceptualise the spatial organisation of polycentric urban regions. In this approach, the centres in PURs are expected to neither follow a strict CPT hierarchy, nor a space syntax dichotomy of local and global networks. Instead, it argues that it is necessary to investigate to which particular scale a centre relates when analysing the spatial configuration of PURs and its distribution of commercial centres. Scale, it will be argued, is the fundamental hidden (i.e. latent) spatial structure formed by human activity. Finally, methodological implications are provided as well as an outlook on the following chapter.

2.1 THE SPATIAL CONFIGURATION OF PURS

2.1.1 SPACE SYNTAX AND THE STUDY OF REGIONS

Compared to the previously discussed authors in Chapter 1 space syntax is a fundamentally different approach to the concept of the spatial organisation of cities. Space syntax is a theory and method for the description of buildings and cities and their relation to patterns of human interaction, movement and socio-economic activity (Hillier 2004; Hillier and Hanson 1984), in this regard it is situated in the field of urban morphology, the study of the form of human settlements and the process of their formation and transformation. Space syntax theory evolved from the study of the spatial arrangement of spaces and small settlement patterns in the late 1970s and developed into a larger conceptual framework on the reciprocal relationship of society and space (Hillier 2008). The fundament of the theory is the concept of space as a relational entity. Space in space syntax refers in a broader sense to any form of void (i.e. rooms, streets, plazas) defined by its surrounding enclosure. Spaces are shaped by human action and

thus also influence the future potential of human interaction. The combination of individual spaces together form a network of spaces, which can be analysed by their network properties to determine their spatial configuration, or pattern (Hillier and Vaughan 2007). In an urban context this network of spaces is operationalised through a representation of space by so-called axial lines (Hillier and Hanson 1984) (see Chapter 5 for an extensive discussion). Axial lines are defined as 'the longest line that can be drawn through an arbitrary point in the spatial configuration' (Turner et al. 2005), as such it forms a representation of the visibility and accessibility between individual spaces. Visibility and accessibility highly influence the way humans move through space.

Movement plays a fundamental role in space syntax theory, as the spatial configuration of urban spaces is seen as a mediator of the so-called 'natural movement', which is the proportion of movement that is governed by the configuration of space itself rather than functional attractors (Hillier et al. 1993). Every spatial configuration holds such natural movement potential, as it is the proportion of movement that is solely governed by the configuration itself. In the last decade, several authors demonstrated that this proportion of natural movement strongly correlates with actual human behaviour, i.e. movement in space (Barros et al. 2007; Gao et al. 2013; Hillier and Iida 2005; Jayasinghe et al. 2015; Jiang and Liu 2009; Patterson and Jones 2016; Penn 2003; Penn et al. 1998a; Serra et al. 2015).

Out of these enquires that employed such network-based approaches to cities arose the concept of the 'movement economy' (Hillier 2004 pp. 125–7). The theory of the movement economy proposes that in the evolutionary process of settlements, spaces are configured in such a way that they organise natural movement in quieter and busier patterns of flows (ibid. pp. 125–7). This process is linked to an evolutionary process that is based on human movement and increases efficiency, but it is also linked to principles that naturally govern social interaction. Movement flows arising out of this dual principle go on to influence the location of land use, and given that a specific functional threshold is reached, generate in turn 'multiplier effects' on movement creating further attraction to such locations. This multiplier effect then leads to a feedback loop on land-use choices and the spatial organisation triggering additional spatial developments (ibid. pp. 125–7). A crucial achievement of space syntax has been the fundamental observation of a strong relationship between the natural movement estimated through the network measurement of closeness and betweenness centrality and the location of economic activities (Hillier 1996a, 1996b, 1999; Hillier et al. 1993; Penn et al. 1998b). The core observation is that economic activity is highly influenced by network accessibility (or one might argue by its degree of being central), more specifically by the potential to be a destination (to-movement, i.e. closeness centrality) and by the potential to be on the way of journeys (through-movement, i.e. betweenness centrality).

Closeness centrality and betweenness centrality are two different measurements and can be used as a proxy for two different modes of travel; the former being a proxy for the likelihood of choosing a space as a destination, the latter being a proxy for the random chance of being encountered on the way of a journey. The former reflects, in this sense, a deliberate human decision to go to a place, while the latter reflects a by-product of journeys, rather than a deliberate decision. In a hypothetical situation where one would need to choose a location in a city in order to be found by a friend wandering through the system, the two measurements would, hence, lead to two fundamentally different answers on which location would be preferred. The notion goes even further (largely based on the measurement of betweenness centrality) and argues for the existence of a generic form of the city, a foreground network of linked centres and a background network of largely residential spaces (Hillier 2004 p. vii; Hillier et al. 2012). These observations are all made in the context of the independent city, and it remains unanswered whether the generic form of the city can also refer to a generic form of the region and PURs in particular.

Only very few space syntax studies have set the region at their focus of analysis. Turner's (2009) study into the linkage of the local to regional continuum is a pioneering piece of work for the field. His study is a novelty, not only due to his methodological proposition to make use of road-centre line data for the analysis, but by focusing on a collection of cities in the regional context. The application of network analysis in the field of regional studies opens up the possibility of new understandings of spatial relations.

Space syntax, applied to a regional scale, however also introduces some epistemological and methodological challenges. As such, space syntax is theorised in a fundamentally local context, the human body in space, which has not been explored so far in a regional context. Epistemological challenges evolve also from questions about the region as an entity itself: can knowledge about the region derive from local information in the same way as it is the case for cities? Concepts of cities differ substantially to concepts of polycentric urban regions, and it is not clear whether a part to whole relationship observed in cities can also be applied to PURs. Are regions the sum of urban space, or do they form a particular spatial organisation itself, and if so what role does the city play in this organisation? I argue that space syntax methods provide the necessary tools for such questions, despite its historic development in a city context. However, by doing so some methodological challenges become apparent, such as that axial line generation is theorised in an urban context posting the question: can this method be transferred to rural and agricultural space where spaces are not clearly enclosed. Other methodological challenges are the creation of appropriate spatial models, the definition of the model boundary, or the selection of appropriate radii. The following sections will elaborate on these questions in the context of PURs.

Hitherto there are no common concepts for the region in space syntax. The few space syntax studies that scrutinise regions and large metropolitan areas develop dissimilar concepts. An overview of all space syntax studies published between 2007 and 2015 can be found in Table 1. The most apparent problem is a lack of a coherent definition of the term 'region'. This is rooted in the difficulties with the very definition of the entity itself, and most obviously in the difference of model sizes, which range from 20 to 950km. Particularly the term metropolitan area seems to be often used synonymously with that of the region, which makes the comparative application of findings problematic. Additionally, to the differences in model sizes comes a variety of different model types. These types vary from manually drawn axial lines by researchers, over models based on governmental data to models based on voluntarily produced geographic information. Also, the level of detail and resolution within each model differs from an inclusion of all open spaces to analyses based only on upper tier highway systems.

Adding to this variation within each approach, the inconsistent use of space syntax measures and their respective scales of analysis pose a further problem. All of this is, on the one hand, due to the constant development of analytical procedures and technologies in the field and, on the other, due to the developmental stage of regional studies in the field of space syntax. However, this situation presents us with difficulties in the comparability of findings. None of the studies addresses specifically, how the method needs to be altered or adjusted in order to account for the differences between regions and cities. Instead, space syntax studies working on the field of regions use the same methods and approaches derived from the city and employ them assuming a general validity across scales. Whether this is actually the case has not been the focus of analysis so far.

Table 1: Overview of space syntax studies dealing with the regional and metropolitan scale, from 2007 – 2015.

Author	Year	Studie Area	Model type	Level of Detail	Model size	Model form	Analysed Scales (in km)	Measures*
van Nes	2007	Leiden metropolitan are, Netherlands	Axial line	all spaces	20km	irregular	topological n	CC
Turner	2009	Manchester and Leeds-Bradford metropolitan area, UK	ITN road-centre line	vehicular network	120km	administrative border	metric 0.5, 1, 1.5, 10, 20, 30, 40	ASA BC SLW
van Nes	2009	Randstad region, Netherlands	Axial line	motorway, main roads & local streets	90km	irregular	topological 1, 2, 3, n	BC & CC
Serra and Pinho	2013	Oporto metropolitan area, Portugal	Axial line	all spaces	60km	circular	metric 0.4, 1.2, 2.1, 3, 3.8, 6, 8.2, 10.4, … 23.5, 25.6, 30	ASA BC & CC SLW
Cermasi and Psarra	2013	Emilia Romagna region, Italy	OSM road-centre line	vehicular network	120km	irregular	metric 10, 20, 30	ASA BC & CC
Gil	2013	Randstad region, Netherlands	OSM road-centre line	vehicular network and rail	130km	administrative border	n	ASA BC & CC
Hanna, Serras and Varoudis	2013	Europe	NIMA Road-centre line	highways and rail	2000km	geographic border	metric 1, 50, 100, 500, 1000	ASA CC
Mohamed et al.	2013	Greater Cairo metropolitan area	Road-centre line	vehicular network	40km	irregular	metric 2, 5, 10, n	ASA BC & CC
Psarra, Kickert and Pluviano	2013	Detroit metropolitan area	Road-centre line	vehicular network	450km	administrative border	metric 1, 10, n	BC & CC
Karimi, Parham and Acharya	2015	Jeddah sub-region, Saudi Arabia	Axial	all spaces	60km	Irregular	n	CC
Krenz et al.	2015	Rio de Janeiro, Brazil	OSM road-centre line	all spaces	60km	administrative border	metric 0.4, 0.8, 1.2, 2.1, 3, 3.8, … 23.5, 25.6, 27.8, 30, n	ASA NBC & NCC
Krenz	2015	Ruhr Valley and Halle–Leipzig region, Germany	OSM road-centre line	all spaces	120km	administrative border	metric 2, 10, 30, n	ASA BC & NCC
Serra, Hillier and Karimi	2015	Great Britain, United Kingdom	Meridian 2 road-centre line	motorway, main roads (A-level)	930km	geographic border	metric, 1.2, 2, 2.4, 5, 10, 20, 30, 40, 50, 60, … 800, 1000, n	ASA BC & CC SLW
Serra, Hillier and Karimi	2015	Great South Eastern region, United Kingdom	ITN Road-centre line	vehicular network	250km	administrative border	metric 0.8, 1.2, 2.4, 5, 10, 50, 100	ASA BC & CC SLW
Ugalde et al.	2015	Rio Grande do Sul, Brazil	Axial line	urban spaces & regional roads	40km	irregular	topological n	BC & CC
Law and Versluis	2015	Great Britain, United Kingdom	Meridian 2 road-centre line	motorway, A-roads, B-roads and minor roads	950km	geographic border	metric 0.8, 2, 5, 10, 20, 50, 100, 250, 500	CC

*CC: closeness centrality, BC: betweenness centrality, NCC: normalised closeness centrality, NBC: normalised betweenness centrality, ASA: angular segment analysis, SLW: segment length weighted.

Space syntax studies exploring metropolitan and regional forms are very scarce. The majority of research applies network analytical approaches to the scale of the 'city', the city, in these studies is mostly defined by natural or administrative boundaries. Each model consists of one independent city. These investigations have led to a series of cross-country comparisons of cities and their morphological structures and give valuable insights into their socio-economic functioning (Figueiredo and Amorim 2007; Hanna 2009; Peponis et al. 2007). Most recently Hillier et al. (2012) have pointed out, in their study of 50 different cities that there is a globally occurring dual relation between a foreground and background structure of cities. This dual relation has been theorised by Hillier as the generic city (Hillier 2014). The 50 different cities, Hillier et al. compared in their 2012 study vary in size significantly. The three smallest networks in their list of cities, Mytiline, Nicosia and Venice, measure approximately 1km, 1,5km and 5km in width, whereas the largest networks include Istanbul, Beijing and London with approximately 26km, 34km and 64km. The largest system is hence 64 times larger than the smallest system. For Hillier et al. (2012 p. 164) such a comparison is nevertheless appropriate, because they developed a method to normalise betweenness centrality, which allows to arrive at a range of comparable values, which they argue 'permit[s] direct comparison of radii within and across cases' from 'local to global'. They argue that their analytical approach allows a comparison across different sizes as the systems under investigation feature the same unit, namely streets and hence 'share the same scale and mean the same thing' (ibid. p. 167). The definition of scale, however, remains uncertain.

What is referred to as 'scale', here, could be better described as 'resolution' and does not sufficiently account on scale as a whole. While Hillier et al. (ibid. p. 164) do not specify what they refer to as 'local', they proceed in their analysis to investigate the 'global pattern' of each city comparing the radius n, or in other words all segments with each other for each case. Yet, in fact these comparisons are rather problematic due to the following reasons:

a) The boundary selection has a strong impact on the observed structure. This impact has been termed 'edge effect'. The model of the city of Tokyo and Beijing for example are cut-outs of larger continuous metropolitan agglomerations and areas at the border of the model, thus, do represent a fragmented network of the real-world situation. A study by Gil (2015 p. 2), demonstrated that 'centrality measures are affected differently by the "edge effect" and that the same centrality measure is affected differently depending on the type of distance used'. This effect is stronger the larger the applied radius is and consequently effects radius n the most.

b) Radius *n* is not a distance-free measure. Rather it is the distance necessary to capture the two segments in the graph that are the farthest away from each other. In other words, radius *n* has a precise distance: it is the longest shortest path (or the network geodesic) of the system. One can assume that for example for the model of Mytiline, radius *n* is slightly larger than the geographic distance of the model boundary ≥ 1km, whereas for the model of London the radius *n* must approximate ≥ 64km.

When comparing these two betweenness centrality structures, the comparison is hence based on one structure that exhibits movement on a very small radius (some might refer to as 'local') and another structure of a very large radius (some might refer to as 'global'). This difference between scales becomes apparent, as depicted in Figure 2. The figure shows a model of two hypothetical cities, a small town and a larger city that fundamentally consist of a continuous structure of 9 small towns. Both cities are analysed on betweenness centrality on radius *n*, which in the case of the small town is equivalent to a radius of 1000 metres. For the larger city radius *n* is equivalent to a radius of 3000 metres. Moreover, the figure also includes the result of betweenness centrality for the larger city on radius 1000 metres (to compare the pattern against radius *n* of the small town). This simple scenario highlights that the structures of the radius 1000 metres are much more comparable than those of radius *n* for both models. The question must therefore not be whether different structures exist, as it can be demonstrated that they do, but rather whether a comparison of structures derived from a global radius are meaningful when models of different sizes are compared. This issue becomes even more pressing when the size of the model, or it's arbitrarily chosen boundary influencing radius n, and subsequently the 'global' structure, are much harder to define as is the case in PURs.
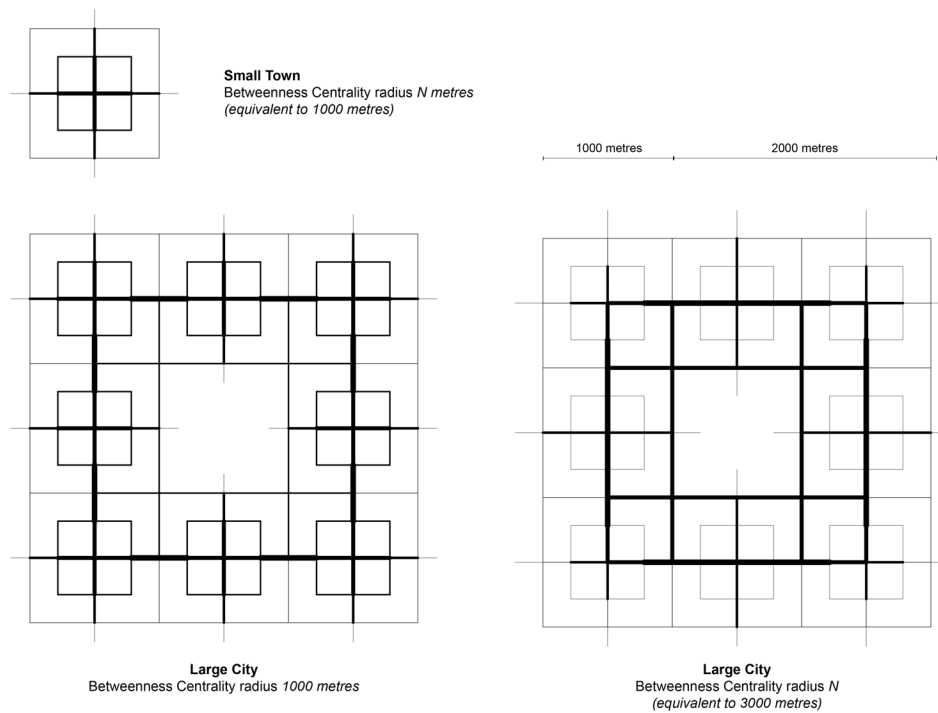
Figure 2: Comparison of different betweenness centrality structures on radius N and 1000 metres for two hypothetical cities.

Initial tests show that the proposed normalisation method by Hillier et al. (2012) practically eradicate differences between larger radii (>3,200 metres) in regional models. Figure 3 is a correlation matrix of a regional spatial network model (see Chapter 6 for a detailed explanation of the model employed and the reasoning behind the radii selection). The model has been analysed on angular segment analysis segment length weighted betweenness centrality (ASA SLW BC), from these values a corresponding value of normalised least angle choice (NACH) has been calculated following the method proposed in Hillier et al. (ibid.). Each pixel corresponds to a specific radius pair and the colour indicates the degree of correlation value r^2. The upper left grid shows correlations between the 49 radii of NACH and their respective correlation with each other. Effectively, every radius after 3,200 metres correlates to 0.95 with each other. That such a strong correlation does not reflect the initial network centralities can be observed in the lower right grid, where ASA SLW BC radii are compared against each other. Here a more complex pattern of potentially three distinctive correlation groups can be observed: first, from 100 to 3,200, then from 3,200 to 39,200 and finally from 39,200 to 110,500 metres. The results show that NACH leads to a substantial loss of information when applied to regional spatial networks and raise questions about the appropriateness of this method for regional network analysis.
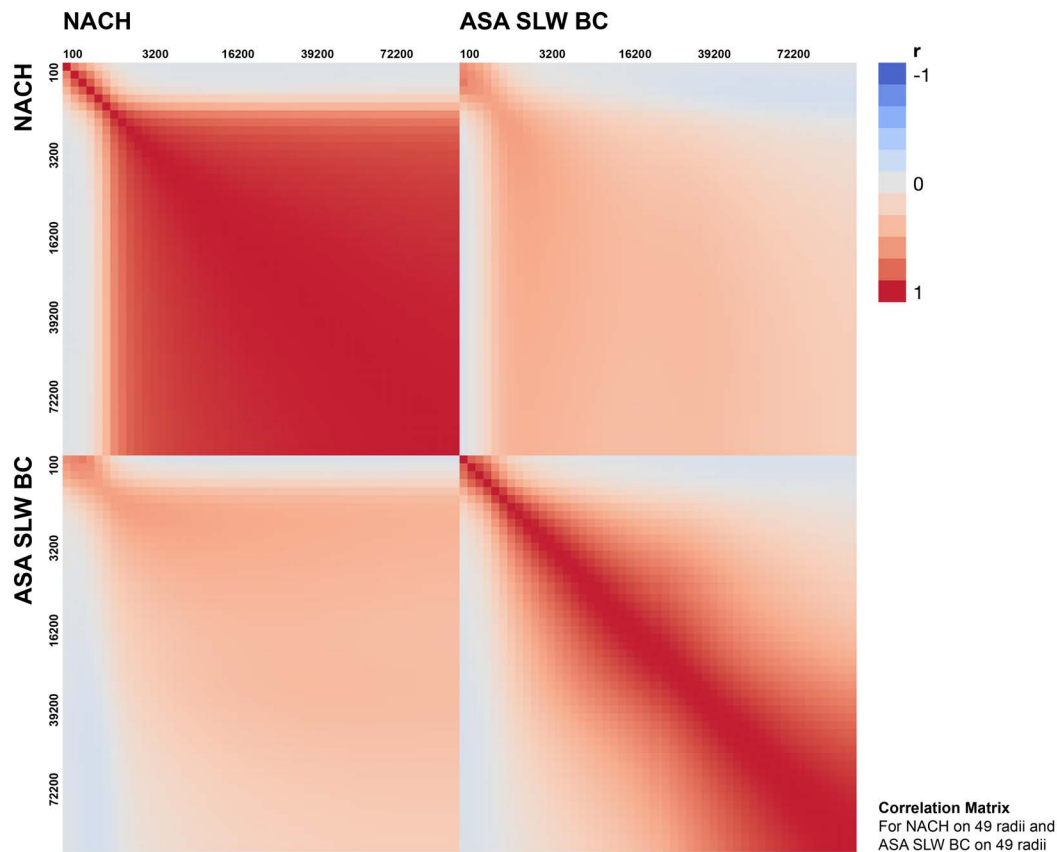
Figure 3: Correlation matrix for normalised least angle choice and angular segment analysis for segment length weighted betweenness centrality on 49 different radii from 100 metres to 110500 metres of a spatial network model of a UK Region.

Both difficulties, the boundary selection and the usage of radius n, are rooted in the lack of theorising scale in space syntax and the fact that the radius of what is considered 'local' and 'global' changes dramatically throughout the body of space syntax literature depending on the object under investigation. The general use of the term is initially derived from cellular spaces and graph theory terminology, but was introduced to the context of society above and beyond network relationships at later stages. First referred to by Hillier et al. (1976 p. 153), 'local' and 'global' was used in a descriptive context of cellular agglomeration patterns derived from a simple rule sets. Here, 'local' refers to an individual cell and its rule, while 'global' describes the agglomerated object as a whole, that is all individual cells together, and their global structure that emerges from that. For Hillier et al. (1976) it is not of particular importance at which scale 'global structure' emerges, or if there are other structures in-between; rather their focus is on the theoretical positioning that it emerges at all and its subsequent implication for the observed entity. While it is clear in the context of cellular spaces what is meant when the term 'local' and 'global' is used, the terminology becomes vague when the authors convey their concept to real world examples, where they argue that scale becomes an intrinsic aspect of any analytical endeavour. Cellular agglomerations are theoretical

constructs and ultimately non-spatial and therefore do not feature spatial scales. What differentiates them is their topological relationship. When network principles are applied to real world spaces, scale becomes an important factor. This is because, when leaving the theoretical sphere of non-spatiality, geometrical characteristics, such as the metric distance become an important factor of differentiation (Salheen and Forsyth 2001). If 'global' relates to the agglomeration of all human journeys in space, and as a product generates a spatial configuration that is shaping movement, then all journeys can only refer to those taking place within the model and hence exclude any inter-city relationships. A similar point has been made by Griffiths in the context of temporality (2011 p. 79).

A large body of work in the field of mathematical methods of spatial analysis dealing with the spatial organisation of society on inter-city and regional relationships was already established at the time when Hillier et al. (1976) first formulated their notion of space syntax, but the authors decided to not engage with these strands due to the fundamental differences in their conception of distance and space. The ramifications of this decision become particularly apparent when Hillier et al. (ibid.) transfer their theoretical models on real world examples. When the scaleless model becomes spatialised – and hence starts to incorporate scales – in forms of buildings, neighbourhoods and settlements of ranging size, the term 'local' and 'global' starts to refer to entities of entirely different sizes. The authors bridge these differences with the terminology of 'small' or 'large' scales, or synonymously with that of 'levels' (ibid. p. 183), while simultaneously describing 'local' and 'global' characteristics of the respective system. What is here considered as 'global', however, needs to be seen in the context of each respective spatial scale.

The reasoning behind this can be found in The Social Logic of *Space*, where Hiller and Hanson state that they deliberately excluded notions of 'distance' and 'location' in their theory, arguing that space syntax is ultimately distance free and that the notion of location can be replaced by the notion of morphology, enabling the incorporation of an entire set of simultaneous relationships (1984 p. xii). They further argue, it is the analysis of these simultaneous relationships and 'the global properties of such complexes of relations' that reveal hidden structures, which prior approaches building on distance notions, have failed to provide (ibid. p. xii). Such global properties indeed exhibit hidden structures, but, as argued earlier, the comparability of these properties across systems remains debatable and becomes difficult in regional applications. Hiller and Hanson's decision to exclude the notions of distance and location from their theory prevented a potential convergence of developments of mathematical methods of spatial analysis in quantitative geography. Particularly the work of Peter Haggett (1965) and his colleagues Richard Chorley (1967, 1969), Richard Morrill (1970) as well as Abler et al.

(1971) focused on finding patterns of spatial relations and their geometric network properties, as well as stressing the importance of distance in human spatial organisation. This is why it comes as a surprise that this body of thought was not incorporated and might be the reason for the vague concept of scale in space syntax literature. Yet, although not incorporating these ideas in the general theory, Hillier and Hanson have expressed their appreciation for the theories of von Thünen (1826), Christaller (1933a) and Lösch (1940). All of these authors played an important role for the development of the field of quantitative geography and specifically deal with the notions of distance and location. A quote by Peter Haggett from 1965 exemplifies the proximity of his thinking to that of Hillier and Hanson.

"One of the difficulties we face in trying to analyse integrated regional systems is that there is no obvious or single point of entry. Indeed the more integrated the system, the harder it is to crack. Thus in the case of nodal regions, it is just as logical to begin with the study of settlement as with the study of routes. As Isard comments: "the maze of interdependencies in reality is indeed formidable, its tale unending, its circularity unquestionable. Yet, its dissection is imperative. … At some point we must cut into its circumference." We chose to make that cut with movement." (Haggett 1965 p. 31)

Both authors see the entry point of analytical ventures in understanding human spatial organisation in the study of streets with the focus on movement at its core, opening up points of contact. With the developments in the field of space syntax during the last decade, particularly the development of angular segment analysis and the introduction of metric distance radii (Hillier and Iida 2005; Turner 2001), the possibility of a point of connection has been established. While the majority of space syntax studies put the focus of their research on the city, quantitative geography departed towards an understanding of regions as integrated systems of different settlements from early on. This is particularly the case for the geographic strand of economic theories, which started with a one-city theory (von Thünen 1826) and moved to a system of different hierarchically ordered cities (Christaller 1933a) into what has now been coined as a more complex, network-based relationship of cities and their hinterland (Sassen 1991; Taylor 2004). This thesis will work at bringing these two traditions into a new and productive dialogue. The following section will elaborate on Christaller's central place theory to point out its fundamental notion of space and further identify points of contact with space syntax theory.

2.1.2 CENTRAL PLACE THEORY

Christaller's central place theory (CPT) (1933a) explains the spatial dispersion of economic and social activities through spatial market competition and centrality in

regions. I argue that Walter Christaller's notion of hierarchical order can bring valuable insights into the spatial organisation of polycentric urban regions. This is because, up to now CPT has proven to be a self-consistent theory of economically driven human spatial organisation that is able to inform urban and regional economic development and planning strategies (Mulligan et al. 2012 p. 407). This still holds even though, several investigations and practical applications on real world examples have shown that the regional distribution of urban areas can follow a more complex relationship (Arthur 1994; Batty 2007; Blumenfeld 2007; Eaton and Lipsey 1979; Fik 1988; Fik and Mulligan 1990; Fotheringham 1983; Glaeser and Maré 2001; Haining 1983; Krugman 1991; Thill 1986; West and Von Hohenbalken 1984; White 1974, 1977, 1978).

Walter Christaller developed CPT in the early 1930s. The theory employs a spatial competition principle to describe the location of the tertiary sector (i.e. all urban functions of services and trade that serve face-to-face end user demands). Christaller's aim was to describe the location, size and spatial organisation of settlements with functions of the tertiary sector. Christaller analysed and categorised differently sized urban areas and their relationship, based on commercial services to their surrounding rural area (1933a). This notion is based on the idea that cities are points of economic exchange. This economic exchange follows a hierarchical order in such a way that specific economic trades occupy particular areas of potential distribution and spatially compete with trades of the same kind. This leads to an economically even spatial distribution with an efficient accessibility for each of the trades. Settlements that are centrally located offer more goods and services and have larger populations. For Christaller 'central' means accessibility from an economic point of view and implies that distance is intrinsically linked to the cost of transport. Relative locational centrality is the fundamental determinant for this notion.

The CPT underlies the following simplifying assumptions: the market is competitive in nature; humans act consistently rational, maximise their profit, have infinite access to information (based on the principle of the *homo economicus*), and strive to minimise travel cost; space and resources follow a homogeneous flat distribution (Christaller 1933b p. 35).

These assumptions have often been the reason for criticism. I would like to specifically focus on Christaller's assumption of a homogeneously distributed and flat space and elaborate on the role of space in CPT. Christaller's theorisation of space has a strong impact on the actual real world applicability of his theory. I will argue that Christaller neglected the inherent complexity of space and that the effect that this complexity has on the affordance and accessibility of locations, and ultimately on what makes a space central. Not taking into account these factors hinders a comprehensive understanding

of the spatial distribution of cities, their size and population, while also preventing an effective real world application of his theory.

THE ROLE OF SPACE IN CENTRAL PLACE THEORY. For Christaller the conceptualisation of the city is one of abstract nodes within a networked economy. This abstraction comes as a surprise as Christaller's main interest is the hierarchical spatial distribution and size of cities, which is a fundamental spatial phenomenon. The terminology 'central place' does not describe a spatial entity as such. Christaller makes clear that it is not the 'appearance of a town', or its spatial configuration, on which the focus of his theory lies, but on the 'function in human community life' such a place provides (ibid. p. 15). He further specifies, that a 'central place' does not refer to a town or settlement as the entity in their entire spatial manifestation, instead, for Christaller a 'central place' is the 'localization of the functions of a centre at the geometrical location' within a settlement (ibid. p. 17). These are all functions that serve the needs and demands of the surrounding population. His work fits into the general framework often found in economic and micro-economic theories interested in the study of spatial dispersion, where cities are simplified to abstract entities. Nevertheless, the CPT system does not come without any spatiality at all. In Christaller's framework, spatiality is thought of as distance and market area. Christaller's use of Euclidean distance might be one of the reasons why his contribution has been largely neglected by those that emphasize non-Euclidean spatialities. For Christaller, however, distance means not only Euclidean distance, but economic distance, 'the cost of freight, insurance, and storage; time and loss of weight or space in transit; and, as regards passenger travel, the cost of transportation, the time required, and the discomfort of travel' (ibid. p. 22). Proximity is hence what is economically accessible, rather than what is physically close, which is still the dominant definition of distance in urban economics.

However, since space in CPT is of an isotropic nature, i.e. uniform in all orientations, this notion of distance has nevertheless a Euclidean real world realisation. Starting with the premise of a homogeneous distribution of space, Christaller employs two fundamental concepts: 'threshold population' and 'distance decay', which exemplify what economic distance means. Threshold population refers to the total population necessary at a given location to make a good or service economically feasible, making it interchangeable with demand. Distance decay describes the effect of distance on a spatial interaction, or in other words, the farther two entities are away from each other, the less likely is their potential of interaction. This implies that if a good or service is too far away from a customer the cost of travel outweighs the need for the respective good. In this sense distance decay defines the range in which a service or good can be traded. Figure 4, shows how the relationship between these concepts defines economic distance, and thus space. Cost is in a negative linear relationship to demand, the higher the

demand the lower the cost (a). Whereas cost and distance to the place of supply remains in the inverse, namely positive linear relationship; the farther away from the place of supply, the higher the cost of the good or service (b). The combination of these, leads to a decline in demand with an increase of distance from the place of supply (c).
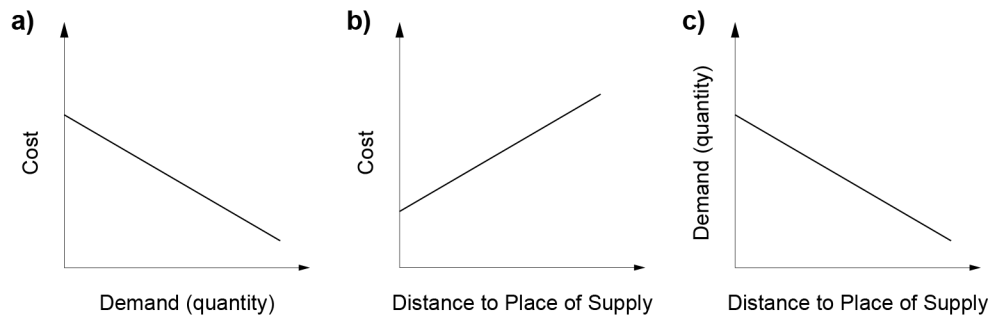
Figure 4: Graphs showing the relationship of cost, demand and distance to a place of supply.

The threshold population causes a service and good to emerge, whereas distance decay defines the probability of interaction with the respective service or good within the resulting economic distance. Given the assumption of an isotropic space this relationship leads to a central place, surrounded by a circular market area, with a minimum and maximum distance (Figure 5:a, b). The size of such centres and the market area are both reciprocally dependent on the threshold population and the respective effect of the distance decay. Given an even population distribution, these market areas will emerge into a closely fitted pattern covering the entire region. However, because circular geometries are not able to cover a surface with no overlaps and no gaps, Christaller made use of a hexagon tessellation. Figure 5:c shows how the pattern emerges over time.
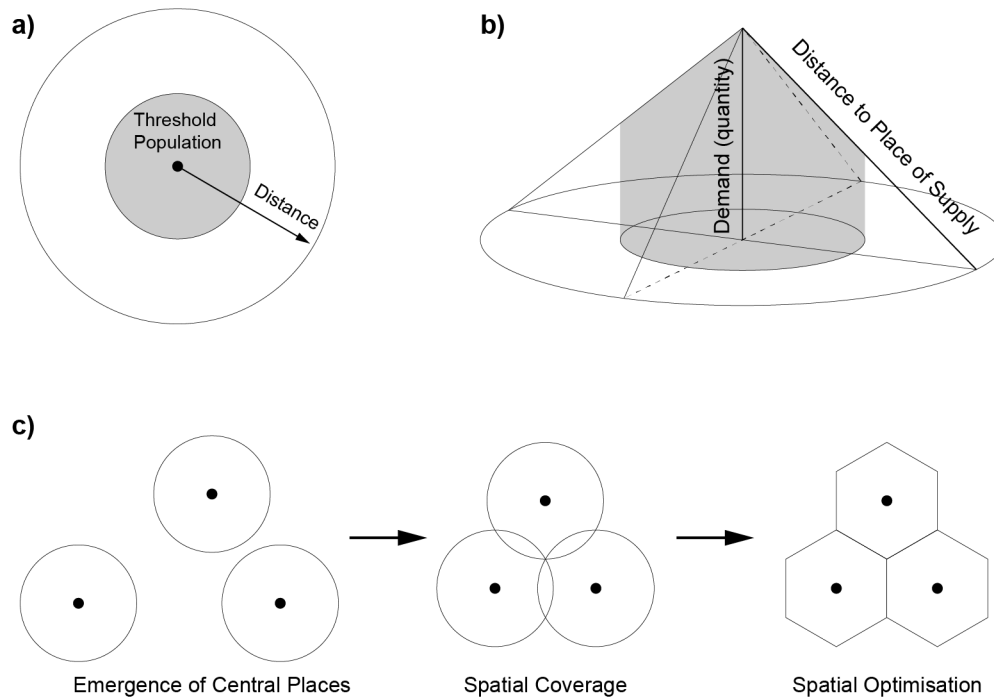
Figure 5: Threshold population and distance decay model (a, b). Process resulting in optimised spatial coverage.

As some goods and services require larger threshold populations, or have a lower distance decay their market areas vary substantially in size. One can think of the market area of a small local tailor serving a neighbourhood, in comparison to a large department store selling mass products. Both market areas are different in threshold population and range. Such differences lead to a further subdivision and the emergence of hierarchy within the central place system. Retail trade and service activities with similar population threshold and range tend to cluster at preferred locations forming urban agglomerations and, thus, central places of different hierarchies.

a)

Central Place System
Principle of supply

- **G** Place
- **B** Place
- **K** Place
- **A** Place
- **M** Place

- **B** Ring
- **K** Ring
- **A** Ring
- **M** Ring

b)

Central Place System
Market areas

- **G** Area Border
- **B** Area Border
- **K** Area Border
- **A** Area Border
- **M** Area Border
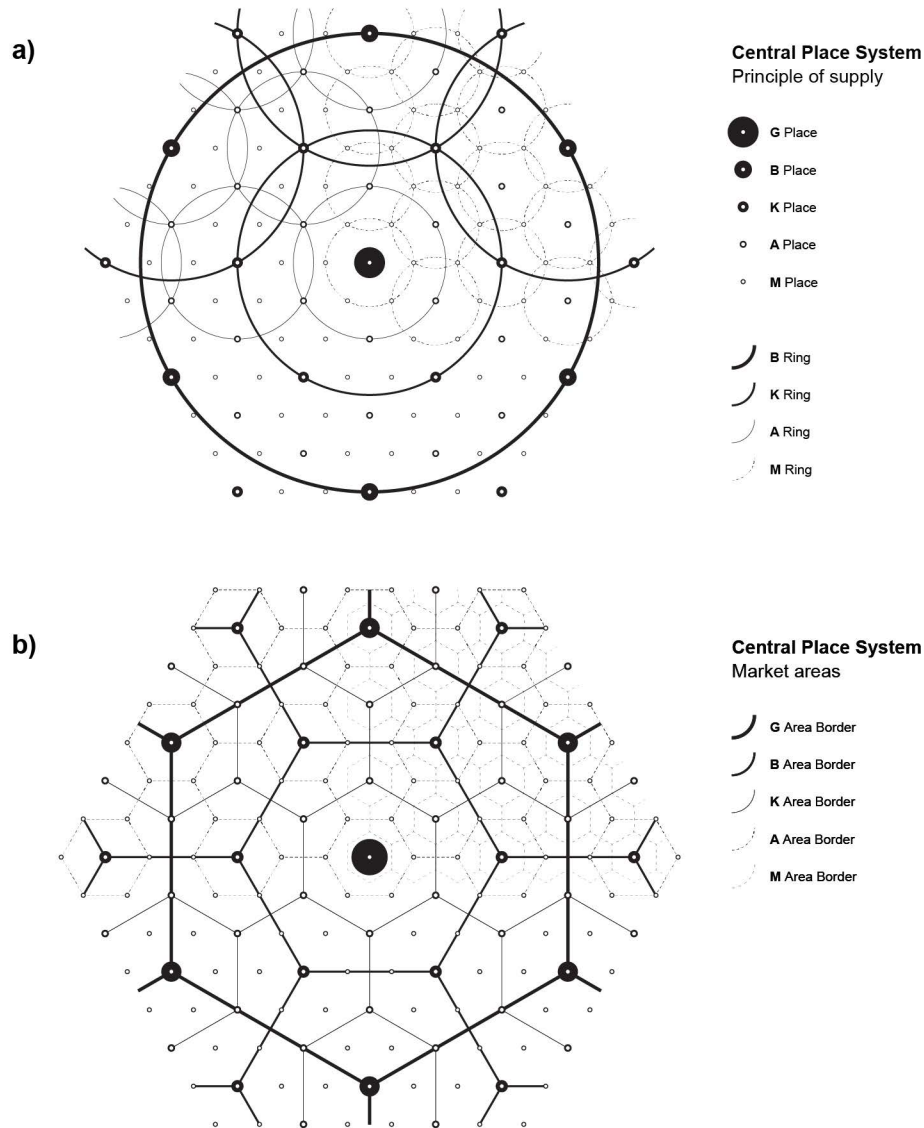
Figure 6: Central Place Systems and the principle of supply (a) and market areas (b) (based on Christaller, 1933b, p. 338).

Christaller's central place system is divided into seven hierarchical levels of urban forms (Table 2), ranging from a small town, the *Marktort*, with a population of a 1,000 up to large scale cities *Landstadt* with populations larger than 500,000. Each hierarchy features a potential market population (i.e. threshold population) as well as a given market radius.

Table 2: Christaller's central place hierarchy and market radii (based on Christaller 1933a p. 72).

Type	Urban Population	Market Population	Market Radius (m)
Marktort (M)	1,000	3,500	4,000
Amtsort (A)	2,000	11,000	6,900
Kreisstadt (K)	4,000	35,000	12,000
Bezirkstadt (B)	10,000	100,000	20,700
Gaustadt (G)	30,000	350,000	36,000
Provinzstadt (P)	100,000	1,000,000	62,100
Landstadt (L)	500,000	3,500,000	108,000

Figure 7:c shows how his theory manifests itself if mapped for the case of Southern Germany. Here L centralities form the upper tier of interconnected centres. In the order of P, G, B and K centres are then cluster around the respective next upper level. Christaller tested his theory empirically by counting the number of landlines per city and subtracted the total population. His assumption was that those cities that have a higher surplus of landlines exhibit a higher degree of commercial activity.

Particular for his model is that relationships are inherently one-directional, this means that each lower class depends on the level above. Since each level is characterised by a, for the hierarchy relevant, cluster of particular economies; horizontal interdependencies are considered as redundant and, hence, non-existent. This implies that interregional relationships do only exist on the level of large metropolitan cities, which is an assumption that does specifically not hold for PURs (Meijers 2005).

a)

Distribution of Central Places in Southern Germany

- ◎ L-Ort
- ◉ P-Ort
- ◉ G-Ort
- ◉ B-Ort
- • K-Ort
- • A-Ort
- · M-Ort

- •••••• 21 km-K-Ring (schematisch)
- ——— Ring der B-Orte (normal 36 km)
- ┼┼┼┼┼┼ Grenzen der L-Systeme
- ═══ L-Richtungen 1. Grades
- ══ ══ L-Richtungen 2. Grades

b)

Central Place System
Network of movement

- ╱ Primary Route
- ╱ Secondary Route
- ╱ Local Route
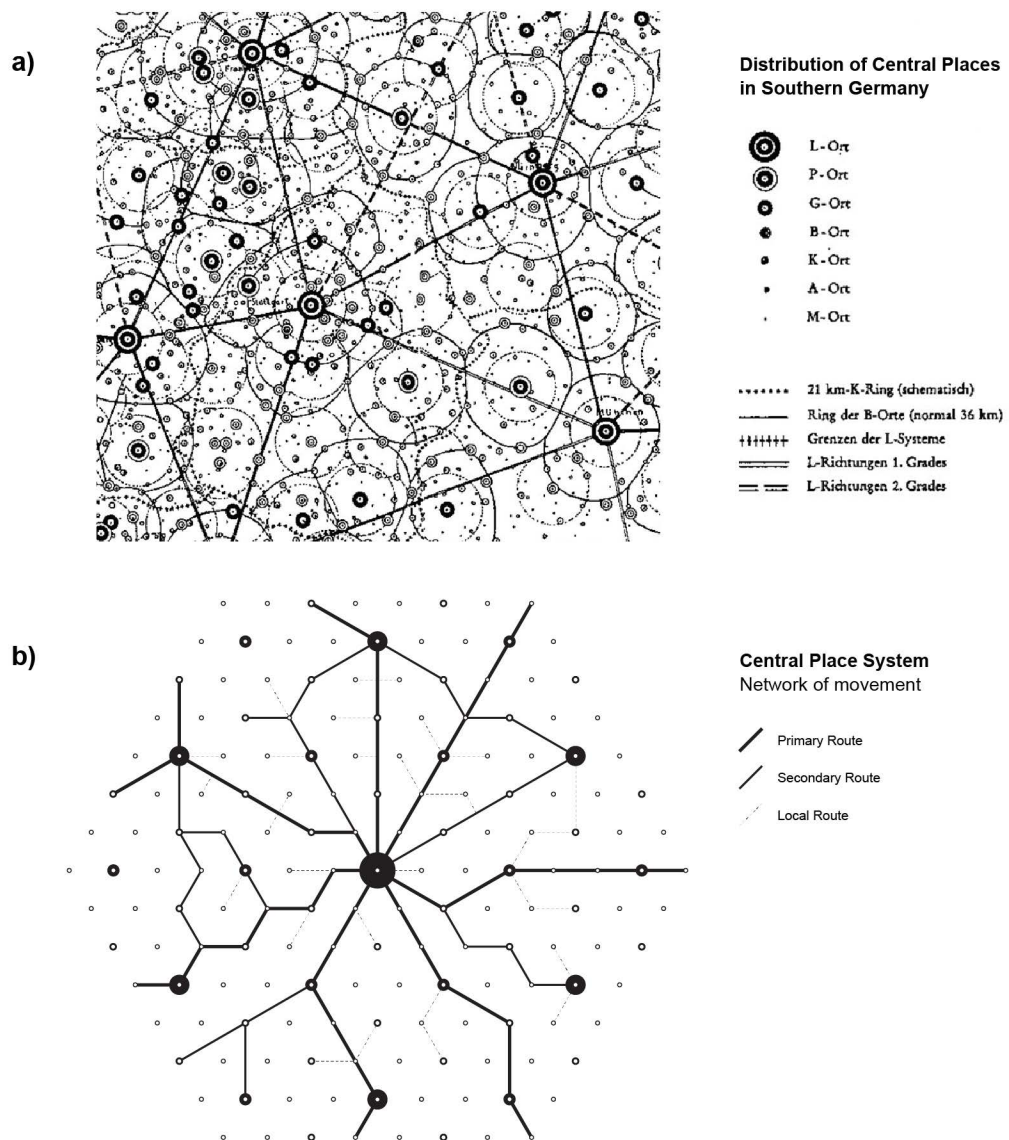
Figure 7: Central Place Systems with resulting networks of movement (a). Also, mapping of the theoretic central place system in a real world example of Southern Germany (b) (based on Christaller 1933a p. 338).

2.1.3 CRITIQUE AND DEVELOPMENTS OF CENTRAL PLACE THEORY

Several authors have highlighted cases in which CPT cannot fully account for observed phenomena, these observations have led to criticism and further developments of the theory, such as Berry (1961), Bourne et al. (1978) and Haggett (1969) among others (see Coffey (1998) for an extensive review).The following section presents the core strands of critique and extensions as well as a reflection on their importance on the spatial organisation. This section builds on a comprehensive review of contemporary publications on CPT by Mulligan et al. (2012).

One of CPT's strength is to accommodate different theorems that explain organisation in space. Two of the most influential ones are firm agglomeration and functional hierarchy (ibid.). One of the premises, with which CPT explains the mechanisms by which centres evolve and finally establish in space, is based on the assumption that firms choose their location on two conditions: first, they seek to optimise the price-quantity combination (supply and demand relationship). Second, they strive to settle at a location where the distance to their closest rivals in the industry is largest in order to secure the biggest share of potential customers possible (competitive accessibility). Yet, at the same time, firms cluster with great proximity to other firms from the same sector, following the idea of an economy of scale (ibid.). At first sight CPT fails to account for such firm or urban agglomerations, as competitors are normally expected to search for the maximum distance between each other, yet urban agglomeration is a common phenomenon.

2.1.4 URBAN AGGLOMERATION

The term of urban agglomeration stems from the field of urban economics, and is of special importance for the concepts from economies of scale. Urban agglomeration denotes the aspect in urban economics in which firms are spatially clustered closely together allowing them to save costs from jointly used infrastructure but also profit through shared suppliers, greater specialisation of employees and division of labour (Glaeser and Maré 2001). This can entail that the production cost of each firm decreases, which might attract not only more firms in the same industry but also more suppliers (Jacobs 1961). For this phenomenon, spatial proximity between firms is not only one of the driving forces of this process but it is also the process that changes the spatial configuration within the city. Alternatively, in the case of a growing city, this effect might also change the importance of the centre in the regional context as well as the regional embedding into the street network.

Urban economists distinguish two positive kinds of externalities, i.e. advantages of an exogenous nature that help a firm thrive. The first is the localisation economy, which means an increase of profit for a firm that follows when it increases the number of produced units and can thus reduce the production cost needed for each unit. The second is the urbanisation economy, which means that a firm profits from a larger pool of skilled labour to which they have access due to their location in urban areas (Glaeser and Maré 2001). At the same time, those firms also profit from collectively used goods, such as transport infrastructure. Both, localisation and urbanisation economy, are intrinsic properties of cities.

Especially the advantages of urbanisation economies can stimulate firms to cluster and thus create centres. Scholars such as Haining (1983), West and Hohenbalken (1984) as well as Fik (1988) provide empirical evidence for yet another additional mechanism that leads to further growth of spatially stimulated economies. If the customer demand at a location is large enough, corresponding with the increased population threshold, even rivalling firms from the same segment of an industry might cluster, which might entail the formation of location economies. Other advantageous effects that proximity to rivalling firms could have which could eventually contribute to the establishment of centres, is the idea that firms profit from a creative exchange of ideas and talent that could bring about new developments and innovation (Glaeser and Maré 2001).

Mulligan et al., who refer to this phenomenon as geography clustering, regard this effect of growth of a centre as compatible with CPT purely because of its advantageous circumstances for business (Mulligan et al. 2012). This is, however, not entirely coherent with Christaller's theory, which posits that rivalling firms will occupy a location based on the maximal distance to their rivals while staying in a radius that encompasses enough potential customers. In the case of a big market potential, demand could, however, be sufficiently high so that two firms could settle right next to one another, without fearing that their rival were to serve all of the available customers. Taking into account the additional advantages of spatial agglomerations, this situation would in fact be more profitable for a firm, than distancing itself spatially.

Theories and empirical evidence from economics identify similar mechanisms that explain how agglomeration could bring about centres (Krugman 1991). The interaction produced by spatial proximity of firms leads to spill over effects in knowledge, leading to success, which leads to more innovation and knowledge in a perpetuating cycle. Simultaneously, firms can profit from common relationships to suppliers, which might again attract more firms in this segment. These positive effects for very similar, as well as different businesses in the same industry segment offers an explanation for why large centres can emerge, based on other dynamics than suggested by the CPT framework.

Other authors (Eaton and Lipsey 1979; Thill 1986) emphasise the advantages that consumers and buyers gain from an agglomeration of firms which again spur further agglomeration tendencies. The advantage from a consumer's perspective is the possibility of multistep or multipurpose shopping, that allows customers to diminish their cost of transportation by buying multiple goods in one or multiple places in close vicinity of another, while also allowing them to compare options and prices of goods. Both of these drivers were found to contribute to the competing destinations model, which embodied hierarchical organisation (Fik and Mulligan 1990; Fotheringham 1983; Mulligan et al. 2012). In such scenarios, it might become of importance whether a

shop lies on the way of a shopping journey rather than being the destination as such. Another factor contributing to the survival of agglomeration dynamics might be that new industries can relatively easily take over from gaps that out-dated industries leave open (Blumenfeld 2007). This leads to influences of historic developments on the future potential of areas to become places of agglomeration, in the light of post-industrial regions such gaps are imminent after the main industry moves out.

According to Arthur (1994), in terms of industrial agglomeration two different dynamics could be at work in such a case. In the path dependency approach (a notion emphasising the impact of past decision on the possibilities in the future) new industries could be attracted by older ones. This can happen largely irrespective of the quality of the location. Alternatively, in the deterministic option, the locational choice of firms could be a function of the geographic characteristics, such whether there is an access to a port or not (Mulligan et al. 2012). This take-over of gaps would entail that firms could still create a thriving centre over time despite of the centres' relatively unpractical location. Agglomeration could therefore be a dynamic that interferes in the restrictions on where large centres could emerge in the hierarchy and introduce a certain imbalance or impurity of the observed pattern. Such mechanisms might be of particularly relevance for post-industrial regions that face physical relatives that are largely shaped by historic industrial developments.

2.1.5 FUNCTIONAL HIERARCHY

It has been argued that CPT lacks a comprehensive explanation on why hierarchies appear as a result of systemic self-organisation processes (Batty 2007). There are two strands of theories that could bridge this gap. The first one is the theory of spatial evolution, the second, is a microeconomic approach. Both concepts could be summarised under the term functional hierarchy (Mulligan et al. 2012). According to this strand, hierarchies emerge as products of systemic self-organisation that strive to become more efficient (Batty 2005). According to Christian (2004) societies can only climb higher levels of complexity by virtue of these feedback loops.

The first approach to explain the generation of hierarchies in space was initiated by White (1974, 1977, 1978) who employed computational simulation methods, in which he varied different parameters, such as population, household transportation rates etc. These simulations yielded different results for different baselines. First, if transportation costs were low, central location beat other variables so that locations high up in the spatial hierarchy flourished. If, however, transportation was costlier, the hierarchy became flatter as transportation distance was kept very low. The overall findings of these simulations are that a hierarchical structure, resembling that of

Christaller can emerge, yet the population magnitude of the different centres is clearly dependent on indivisible production costs, as well as transportation costs (ibid.). Interestingly, the structure seems to be largely independent of the locations chosen at the start or the sizes of the centres; and what is more, positive feedback loops endogenously enforce growth.

The second and more integrated approach to the evolution of functional hierarchies comes from Allen and Sanglier (1981a, 1981b). These scholars can show how feedback loops change the logic of centre survival if a third good is added to a situation in which previously only two goods existed. The centres cannot grow or decline anymore along those lines of their initial adaptive capacities, but survive or decline according to the logics laid out about spatial agglomeration. A centre prevails only if it has the right balance of local population and central location (i.e. accessibility).

The economist Hotelling (1929) delivers a proof for the idea that differently sized centres are based on microeconomic theorems that postulate that firms change their locations based on the new customer densities. Beckmann (1986) showed that the market demand could come in the three shapes monopoly, competition and super-competition. This entails that firms attempt to settle right between their rivals to protect their monopolies. In a scenario where all firms are equally distributed in space and offer the same product, price is the deciding factor, thus establishing a stable equilibrium over time. In a different scenario, firms are already in different locations and adapt their price and location decisions (Eaton 1972) so that a stable equilibrium can only be established if no firm changes their price-output-location decision. However, many disturbances and inequalities can corrupt this balance, so that dominating firms are able to emerge. An example of this would be that actors are tied to a certain place due to special obligations or requirements, therefore establishing asymmetry.

Eaton and Lipsey (1976) adopted a more realistic approach by holding transport rates constant, allowing for the indivisibilities of goods as well as multipurpose shopping. What resulted were different hierarchies of two levels of which Christaller's system was one of multiple spatial equilibria. This is often explained with the idea that given multi- as well as single-shopping, there will always be a higher flow to higher places, so that spatial asymmetry will emerge. This is an important observation with regards to network centralities, as it implies a) that Christaller's CPT is not the only spatial realisation of microeconomic decision making processes, and b) that human movement behaviour has a stronger impact on commercial viabilities.

Mulligan et al. (2012) emphasise a third approach, that is the two-good cost-minimization model by McLafferty and Ghosh (1984). Here, the authors examine

different locations for economic actors given multipurpose shopping for dispersed households taking into account each household's decision calculus. This showed that economic actors in higher levels would often trump those in lower positions.

What these different approaches to functional hierarchy show is that the hierarchies can occur through evolutionary processes thriving to improve spatial efficiency. Moreover, functional hierarchies do not have to follow a CPT logic, but CPT can form one of the potential outcomes. If a hierarchy has grown over time based on a bottom-up process aiming to improve spatial efficiency, this might also affect the overall spatial configuration. In the case of post-industrial regions, this also implies that what has been an efficient configuration for a region based on coal and steel mono industries, can be inefficient for industries based on service and trade. For PURs in transitional processes it is then important to understand the implications of such spatial configuration so that planning policies can react in alignment with the spatial structure, rather than against the physical reality. Such a reaction might be to identify those industry gaps at locations that are spatially beneficial, even after transitional processes.

2.1.6 IMPLICATIONS FOR THE UNDERSTANDING OF THE SPATIAL ORGANISATION OF PURS

What differentiates Christaller's CPT theory from a space syntax approach is that distance is abstractly defined as a connecting line between nodes, rather than considering distance through the human-shaped configuration. Christaller's theory is a strong simplification that envisages a place in an ideal planar space. His aim is to understand the distribution of cities, rather than the immediate morphology that such a process produces. Moreover, Christalller's notion is spatial in the sense that it defines centrality through spatial accessibility, while replacing physical distance through economic distance. Spatial accessibility refers, here, to accessibility in a plane homogenous space. Christaller's concept of accessibility i.e. economic distance, however, results inevitably in physical distance. In fact, any human action in space has to overcome the same physical distance to begin with, which thus affects economic distance. This 'overcoming' ultimately influences the fundamental potential of any interaction in space. It remains questionable to which degree closeness centrality, as a spatial accessibility measure, conforms with the hierarchical order of centres that CPT describes. Nevertheless, following CPT's core assumption that the result of microeconomic processes is indeed a hierarchical order of different centres, created by human decisions making processes in space, the sum of these decisions must lead to a particular spatial realisation of such hierarchical order. This hierarchical order must be retraceable in space and detectable through its spatial configuration. This poses questions to whether the spatial organisation of PURs, that is largely shaped by

industrial processes also reflect such hierarchical order, or whether these follow different or no order principles at all.

If Christaller's CPT is transferred into an actual physical space and expressed in network terminology, his notion of economic accessibility becomes one of closeness centrality. Every space has a specific 'neareness' (a relative distance) to each other and some spaces are thus closer than to other spaces. In network terminology, this spatial accessibility is best described through closeness centrality and according to Hillier related to to-movement potential, the probability that a space is selected as a destination. Christaller's central places, would, if economic distance were transferred into physical distance, be located at those configurationally beneficial locations that exhibit higher values of closeness centrality.

Space syntax, on the other hand provides methods and tools to explore the morphology of spatial configurations in cities. The findings reporting strong relationships between the pattern produced by spatial network analysis such as betweenness centrality have only been made in the context of the city. Until today, no other study has systematically and in a comparative fashion investigated whether this relationship is also valid in regions. If Christaller's theory has validity then there is more human activity on scales of market spaces than on other scales, which implies that over time such pattern forms a spatial organisation, which fosters efficient movement and is hence traceable through the spatial configuration itself. Since post-industrial regions are strongly influenced by alternative factors, such as the location of settlements based on geological locations, the pattern might be significantly skewed. This assumption is based on the fact that movement during the industrialisation within the region must have taken place on a fundamentally different origin-destination relationship, namely from settlements to work place, i.e. coal mines and steel factories. Whether this development had an impact on the current relationship between microeconomic activity and the spatial organisation can be derived from a comprehensive spatial network analysis.

2.1.7 NETWORK CENTRALITY MEASUREMENTS

This section will introduce different network measures in the context of space syntax analysis. Space syntax's measures for the analysis of spatial configurations are based on graph theory. A number of researchers have focused on statistical properties of networked systems. For spatial analysis and relations in space proximity plays a crucial role. In this section, investigations in network centrality form the core interest. Notions for such statistical properties were founded in social network science in the late 1960s, of which the most relevant forms are Sabidussi's closeness centrality (1966) and Freeman's notion of betweenness centrality for social networks (1977). Sabidussi and

Freeman were both interested in understanding the role of network characteristics for social and psychological disciplines. Sabidussis's (1966) closeness centrality is based on the length of the average shortest path between a vertex and all vertices, calculated as show in formular (1).

$$C_C(i) = \left[\sum_{j=1}^{N} d(i,j) \right]^{-1}$$

(1)

Freeman's (1977) betweenness centrality gives an indicator of a vertice's centrality in a network. This is calculated by counting the number of shortest paths from all vertices to all others that pass through that vertice, where g_{jk} (i) equals the number of shortest paths connecting jk passing through i and g_{jk} equals the total number of shortest paths (2).

$$C_B(i) = \sum_{j \neq k} g_{jk}(i) \, / \, g_{jk}$$

(2)

Both measures were developed in a non-spatial context of social networks. Spatial networks, on the other hand, are different to social networks in the way that they can incorporate geometric information, such as the distance between nodes or the angular relation between them. The question arose, to whether the proposed measures would be suitable to provide insights into the properties of spatial configurations and the effect to human behaviour and movement (Hillier and Hanson 1984).

Building on these fundamental concepts of closeness and betweenness centrality, Hillier and Iida (2005) formulated theoretical and practical developments for their spatial application. Their proposition is based on the notion that cognitive information can be retrieved from spatial networks thorough statistical centrality measures (ibid. p. 481). The conceptualisation of distance plays a crucial role in the way humans are moving in space, rather than the distance itself (ibid. p. 476). As results in cognitive sciences have indicated (Javadi et al. 2017), route decision making is influenced by minimisation of directional changes and the tendency to create linear routes (Hillier and Iida 2005 p. 476). In order to be able to investigate these geometric-spatial properties Hillier and Iida further advanced the axial line representation to a segment model (Figure 8, (c)). Their advancement consists of the introduction of three fundamental concepts of distance costs. In addition to the already established centrality measures, which make use of steps between network nodes in the graph, Hillier and Iida propose to analyse the dual graph instead (ibid. p. 481): "[t]he distance cost between two line segments is measured by taking a 'shortest' path from one to the other, so the cost of travel between S and a can be given as $w(\pi - \theta) + w(\phi)$, while the cost between S and b can be $w(\vartheta) + w(\pi - \varphi)$" (ibid. p. 481).
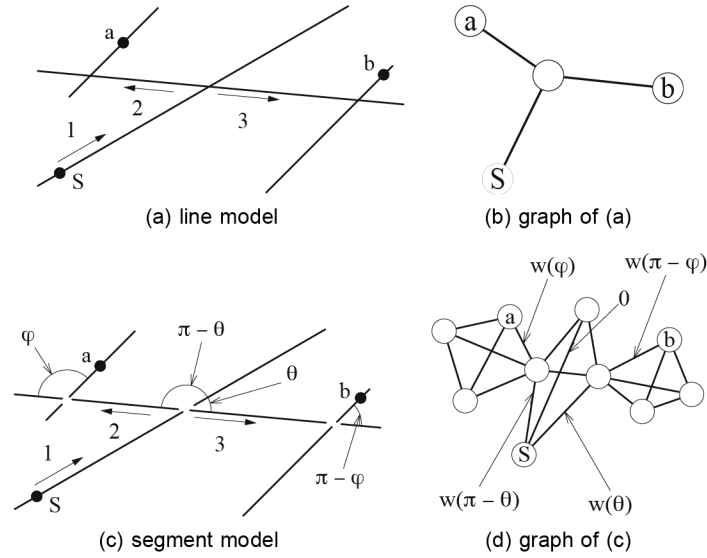
Figure 8: Line-network model and graph representation. Source: Hillier & Iida (ibid. p. 482).

These three distances costs are defined as:

i) topological, which describes geometrical properties unaffected by the continuous change of the shape or the size of objects (in Figure 8), for segment model (c) and graph and (d), $w(\vartheta) = w(\pi - \vartheta) = w(\varphi) = w(\pi - \varphi) = 1$ (While, $w(0) = 0$)) (ibid. p. 482).

ii) angular, which references differences in the angle of one object to another. This is calculated through a weighting of each intersection, where a straight-line connection results in a value of 0 and a right angular turn will be 1. $w(\vartheta) \propto \vartheta$ $(0 \leq \vartheta < \pi), w(0) = 0, w(\pi/2) = 1$ (ibid. p. 481).

iii) metric, incorporating the metric distance within the network from one object to another. Here the starting point is not the end of the segment, but the mid-point. The distance between two segments, equals hence the sum of half their length (ibid. p. 481).

Of these three distances, the authors highlight particularly the importance of angular and topological properties for the human concept of distance (ibid. p. 488). Each distance can then be assigned to a particular scale of analysis. Scales are defined as a radius from a root segment and form as a cut-off mechanism for the analysis. Rather than investigating the network as a whole (each node to all other nodes) the cut off allows investigations of the network in precise scales from the local to a global level. These different scales can then simultaneously be evaluated and give insights into the networks structure, such as the frequency of usage of different parts of the network based on the chosen distance concept. The result is a potential multi-scale analysis of spatial configurations, which allows understanding spatial networks for two measures

(closeness and betweenness centrality) via the three fundamentally different distances as discussed above (geometric, angular and metric) on multiple scales (geometric, angular and metric). Of these three distance concepts, particularly angular distance has been proven to be a reliable indicator for predicting human movement flows. Moreover, Turner (2007) has proposed to combine angular segment analysis (ASA) with metric scales allowing an analysis which differentiates between scales. This combination allows the researchers to investigate network structures of spatial configurations of any size at any scale.

This approach is important, because such network structures are changing depending on the scales of analysis. It forms one of the core strengths of a space syntax approach to provide a methodology that enables investigations of scale-dependent relations. However, this approach also poses an additional challenge, as of today there has been no method to overcome bias or arbitrary radii selection in the application of space syntax to spatial networks (Serra and Pinho 2013). While the common procedure is suitable to define specific journeys, lengths based on the mode of transport or distances to urban functions, such as ideal spacing of bus stops of approximately 400 metres, the procedure does not overcome the problem of possibly hidden, missing patterns within the network. Most recent space syntax studies build either on approaches by Bill Hillier (2009), proposing the existence of a distinct local and global structure of cities. Here local, depending on the interpretation, indicates a scale of 500 metres to 2000 metres and global is considered as n (all-to-all). Alternatively, scales are based on distances between certain areas, or different modes of transport. Particularly in the context of regional analysis, the notion of a 'global structure' becomes fuzzy. This is because it is based on the paradigm of the city as an enclosed entity, something that can be defined by a precise border and size. As elaborated earlier such clear urban boundaries do not exist in PURs, posing the need for new methods to overcome this issue in a regional context.

2.1.8 GEOGRAPHY AND THE NOTION OF SCALE

The notion of scale has long been a major concern for the discipline of geography in general. Over the last decades several authors have stressed the importance of scale (Harvey 1969; Meentemeyer 1989; Watson 1978), particularly so in physical geography and remote sensing in GIS (Quattrochi and Goodchild 1997). When dealing with geographic data, it is essential to specify the respective scale of investigation. Lam and Quattrochi (1992) summarise the core notions on spatial scale and their related difficulties when dealing with spatial processes in a triad of *cartographic*, *geographic* and *operational scale*. Here, *cartographic scale* refers to the ratio between the mapped representation and the real world. While *geographic scale* relates to the spatial extent or the scope of the analysis, *operational scale* gives an account on the level at which the

respective process operates. In addition to this, scale can also be interpreted as the level of detail, or resolution (ibid. p. 89). For the authors one of the core reasons for the importance of dealing with precise definitions when talking of scale is that spatial patterns are usually related to a precise scale and different processes might lead to similar spatial patterns (ibid. p. 89). This conundrum makes it necessary to define the spatial extent and the spatial resolution of any data and its analysis in order to determine at which scale processes operate. This is particularly the case since the advancement of GIS allowed cross comparisons of different scales, albeit their potential incomparability.

Apart from these data-related aspects of scale in physical geography, a large body of work in human geography has dealt with scale on a theoretical level. Here, the focus is on understanding how 'the production of *scale* is implicated in the production of *space*' (Marston 2000)[3]. Erik Swyngedouw (2004 p. 129) argues that the social and physical transformation of the world is taking place in an 'interlocked and nested geographic scale'. For Swyngedouw social life is process-based, and constantly iterates, transforms and reconfigures itself. This process stays in a reciprocal relation to nature and produces in its appropriation and transformation 'historically specific and physical natures that are infused by a myriad of social power relationships' (ibid. p. 130). These '[s]ocio-spatial relations operate over a certain distance and produce scalar configurations' (ibid. p. 131). Swyngedouw's notion of environmental transformations as integral parts of the social and material production of scale allows us to perceive such scalar relationships through space.

What Swyngedouw describes are scales that manifest themselves in space (see also Soja (1989)). Since these scales are in their generation dynamic and process-based, Swyngedouw argues '[s]tarting [the] analysis from a given geographic scale seems [...] to be deeply antagonistic' (Swyngedouw 2004 p. 132). For him, contrary to Christaller's CPT, scales are not primarily shaped by economic activity, but by human activity and the very nature of social life. Swyngedouw further emphasizes that scales incorporate complex power structures that govern social relations (ibid. p. 131). This is because scales generate geometries of power that produce advantages and disadvantages in their very existence. In space syntax terminology one can speak of integrated and segregated locations, or as Stephen Read (2013 p. 10) has put it in his typology of urban levels of 'being in or out of the network'.

---

[3] See Marston et al. (2005)–albeit their contested criticism on the existence of scales in human geography–for a comprehensive review on scale-related literature of the past 20 years.

Based on these two notions, one can conceptualise scale in the context of space syntax. Scale, is hence the structure shaped and constantly reshaped by socio-spatial and economic processes, operating over a certain distance and time. While the process shaping the scale structure can stay in a fast and constant transformation, the spatial scale, it is argued here, is changing in a rather inert way. Still, spatial scales are, in Swyngedouw's words, 'never fixed, but perpetually redefined, contested and restructured in terms of their extent, content, relative importance and interrelations' (2004 p. 133). When analysing spatial networks on different centrality radii the patterns that one can observe are constantly influenced by this underlying or latent scale structure that is manifested in the very configuration of the network. It is proposed that in order to understand the fundamental morphology of a region, one needs to empirically demonstrate this hidden or latent scale structure, or in other words, this multi-levelled, interrelated system of spatial scales that cause certain centrality patterns to emerge. Instead of starting from the dichotomy of the 'local' and 'global' radii in the analysis, the spatial configuration needs to be understood through an extensive collection of different metric radii.

2.2 SUMMARY

The two theories, CPT and movement economy, provide valuable insights into the emergence of hierarchies within regions and cities. On the one hand, Christaller's CPT, which is based on microeconomic mechanisms and the presupposition that human action is based on spatial efficiency, is demonstrably able to explain the hierarchical distribution of cities and their centres in organically grown regions. Hillier's movement economy, on the other hand, provides valuable insights into the relationship of economic activity and the location within cities. His theory is based on the concept that evolutionary processes govern movement within cities into a hierarchy of two fundamental networks, i.e. foreground and background.

Both theories, however, cannot sufficiently explain the spatial organisation of PURs. This is because PURs are not comparable to traditional regions due to their historical emergence. Their development is highly influenced by industrial growth, which influences the total population at locations that follow the principle of geological availability within these regions. This violates Christaller's first assumption, namely that of a homogenous population distribution. For Christaller, however, this total population, i.e. the population threshold is one of the key mechanisms to cause centres to emerge. It is therefore of interest to understand whether the hierarchical order of centres still emerges in PURs and if so, what the influencing factors governing this evolutionary process are. Moreover, urban areas within PURs do not feature clear boundaries and existing centres cannot clearly be related to a particular 'city' or even a

particular population. For this endeavour, it is necessary to leave behind the notion of the individual city and re-phrase the definition of location to one that is fundamentally local, namely the individual space according to the space syntax definition mentioned earlier. Thus, it is possible to investigate potential hierarchical relationships of each location in relation to the overall region, which is a fundamental necessity to understand the spatial organisation of PURs and aid with knowledge in the processes of transition.

Hillier's notion of the movement economy, provides concepts for such relationships in the context of the city, i.e. local to global, and foreground and background network. As mentioned earlier, several studies employing a space syntax approach have shown that cities feature multiple centres, which contrasts the operationalisation of CPT, based on a simplification of cities to a single centre. Such simplification makes it difficult to understand a potentially complex distribution of centres observable in PURs. However, it remains questionable whether the concept of the movement economy is transferable to the regional continuum in the form in which it has been developed. As it has been theoretically and analytically demonstrated in conjunction to the CPT, rather than a dichotomy of local to global, it is more reasonable that human action in a regional context leads to multiple hierarchies in space (i.e. hidden spatial structures). It is, hence, expected to find different spatial scale patterns in relation to a potential central place hierarchy in traditional regions. Whether such a hierarchical organisation of spatial-scale patterns can also be observed in PURs is unclear. In either way, a network-based approach will provide valuable insights into the relationship of local centres to the overall regional system in PURs and potentially reveals i) a hidden centrality structure and ii) the relation of socio-economic processes and such a structure.

Space syntax provides the necessary tools for such an analysis for the context of the city. As has been demonstrated, these methods and concepts for the interpretation of network centralities are based on the individual city. It is hence necessary to establish a set of methods and concepts that enable the application of a network-based approach to the regional continuum. One of the core difficulties in the application of the space syntax method to the regional space is the selection of appropriate radii and the definition of what global means or to be precise about whether specific scales exist and how one can reveal them. It has been shown, that the majority of space syntax studies employs incomplete networks, which make the comparison of findings highly problematic. In order to overcome the lack of comparable networks, a randomised regional spatial network will be employed to compare the effect of the spatial configuration on the resulting spatial metrics. Such a comparison produces valuable insights into the role the spatial configuration plays for the presumed hidden centrality structures.

In order to understand the spatial organisation of PURs, and as a result of the above reflection of the existing literature, the aim of this thesis is a) to develop the necessary tools to make investigations into the regional continuum possible. This includes the development of methods for the application of large scale spatial networks (see chapter 5 for an extensive discussion), the development of randomised street networks as a tool for comparison (see chapter 6 for the reasoning and discussion behind this approach), the development of a technique to overcome arbitrary radii selection and a method to reveal hidden centrality structures in regional models (see chapter 7 for a discussion an proposition of such method) and finally methods for the comparison of socio-economic data with such hidden structures, or precisely the comparison of centralities with information on movement, spatial occupation and commercial activity (see chapter 8 for a presentation of this strand of inquiries). Finally, the thesis aims to provide b) a conceptualisation of what an analysis of the spatial configuration should focus on. This is proposed through the concept of scale i.e. latent centrality structures.

The following chapter will introduce the two selected case studies and evaluates to what extent both regions can be considered to be polycentric urban regions.

CHAPTER 3

POLYCENTRIC URBAN REGIONS CASE SELECTION

3 CHAPTER

Where the previous chapter elaborated theories and their underlying logic of what produces a hierarchical structure of centres, this chapter will look at the spatial configuration of centres and street networks in two case studies, one in the UK and one in Germany. Christaller's theory rests upon the premise that agents shape the location and growth of settlement structures in an evolutionary manner over time. This assumption requires agents to be able to move around and freely choose their ideal professional location, as their resources are mobile. Yet, what happens to the configuration of centres in a region where the entire economy rests purely on the extraction of raw material resources that are locally fixed and immobile? How do regional patterns of centres and settlement clusters evolve within hundreds of years when they are determined by the accessibility and the existence of an exhaustible resource that sets clear limitations to where cities and clusters of settlements can be established?

This chapter will exemplify the emergence and development of settlement structures relying on pre-determined resources for an economy by investigating the cases of the English and German former coal and steel region, the NDY region and the Ruhr valley. Both regions are exceptional in comparison with other polycentric regions, as they both grew purely based on an industry based on coal and steel production. The following chapter will explain how these regions' structure emerged historically and spatially and which factors were most influential in shaping the configuration of centres. By initially evaluating the necessary criteria that must be met by a region to qualify as a PUR the chapter will evaluate whether the NDY region as well as the Ruhr valley can be regarded as prime examples of polycentric urban regions and what their defining characteristics look like in detail.

The chapter, starts by introducing the case selection and potential generalizability of the findings in the later analyses. It will present both of the cases in detail, beginning with the NDY region, followed by the Ruhr Valley. It will then explain to which extent the NDY region qualifies as a PUR under a minimum and maximum definition. The part to follow will focus on the historical pathways and processes that brought about the characteristic spatial structure of the region. The section after this will focus more precisely on the systems of transport and formation of infrastructure. After this, the Ruhr Valley will be scrutinised following the same steps. The conclusion of the chapter will summarise the most important points and aspects.

3.1 CASE STUDIES: THE SELECTION PARAMETRES

This study builds on two different case studies: the Ruhr Valley in West Germany and the former coal fields in central England, comprising the regions of the Yorkshire, Derbyshire and Nottinghamshire (NDY) coalfields. This selection was made, as both regions have been often been referred to as PURs in the literature (Hall and Pain 2012; Münter 2011; Parr 2004; Reicher et al. 2015): both regions feature large-scale fragmentations of urban areas with independent mid-size cities. This similar regional structure stems from their comparable development during industrialisation and post-industrialisation. The regions are also very similar in other regards: both form the largest coal-mining regions in their respective country as well as in Central Europe and are highly comparable regarding their trajectories, path dependences and regional scale. They have undergone extremely similar historical formation processes and produced similar settlement patterns based on the extraction and processing of coal and ore. Both regions faced periods of decline and shrinkage and underwent similar socio-economic and infrastructural developments based on the demise of the mining industry.

Due to their large size, both regions provide a lot of intra-regional variation of the structure of PURs. Yet, as they are regions whose structure emerged in different countries, cultures and industrial periods they also provide us with inter-regional variation. Only if we find the same patterns in PURs with intra- and interregional variation can we generalise the findings and tentatively infer them to other PURs. Both cases allow us to draw conclusions about the structure and features of industrialised PURs in Western Europe but also about other PURs that developed based on the occurrence and mining of natural resources. The results drawn from the former mining regions in Germany and the United Kingdom are of special importance to all regions depending on coal deposits depicted in Figure 9 but especially to the vastly expanding coal regions in China and India.
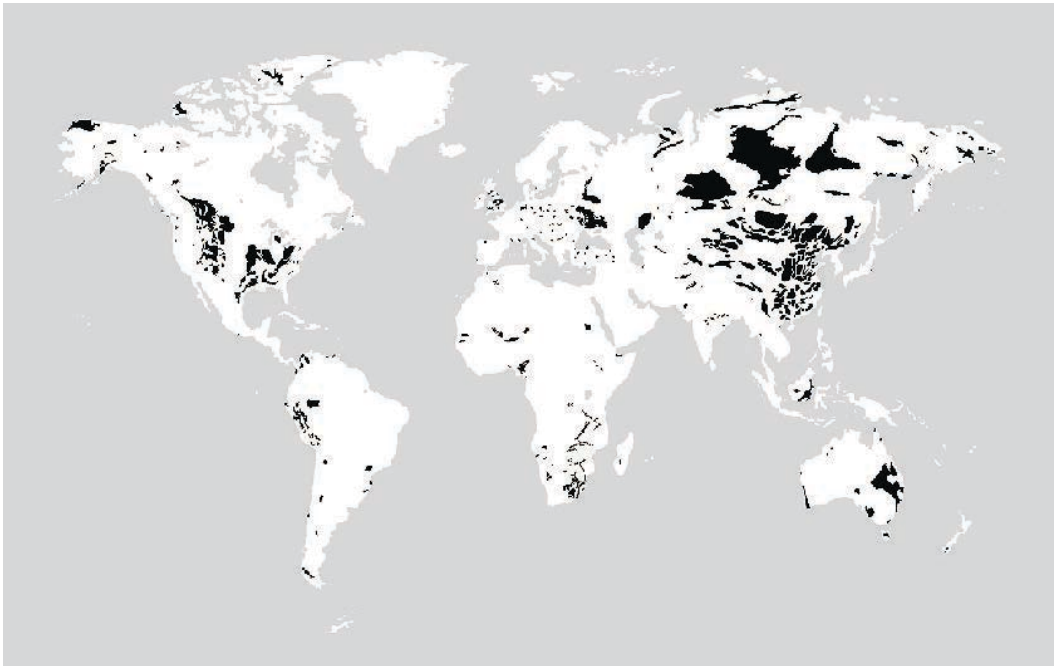
Figure 9: Location of the worlds coal reserves. World Coal Association © 2014.

The findings and methods of this study will also be of importance for other non-coal industrial PURs such as the Pearl River Delta and the Yangtze River delta in China, Great Jakarta or the Tokaido corridor in Japan. More immediate European examples of industrialised PURs for the research of which this study will be of interest are the South East England region, the Randstad in the Netherlands, Central Belgium, the Rhine Main region in Germany, the European Metropolitan Region, Northern Switzerland, the Paris region and Greater Dublin (Hall and Pain 2012).

3.2 THE YORKSHIRE, DERBYSHIRE AND NOTTINGHAMSHIRE

The recent configuration of street networks, towns and cities of the NDY region reach back to the century-long mining of coal and iron ore. The region stretches from the former coalfields of West and South Yorkshire down to the Nottinghamshire coalfields and over to the Eastern parts of Derbyshire. In the South the area analysed covers the Eastern half of the Derbyshire county including the area south of Sheffield, including Chesterfield reaching all the way down to Derby where the Southeast demarcation of the area ends. Eastwards, the area includes Nottingham, Sutton-in-Ashfield, Worksop stretching further north up to Doncaster. Even further northwards the area covers the heart of South Yorkshire ending in the East with the city of Doncaster and in the West with the Peak District National Park. Encompassed by this area are towns such as Barnsley and Rotherham. Northwards, the area covers most of West Yorkshire, including Huddersfield, Halifax and Bradford, including Keighley and Ilkley.

Figure 10: Area of investigation Nottinghamshire, Derbyshire & Yorkshire region, United Kingdom with highlighted areas of coal mining. Background showing European Urban Atlas land monitoring.

In its structure and growth the region's development has always been dependent on the coal deposits, so that the locations of the first larger settlements were entirely determined by the location of the coal seams as well as their accessibility.

3.2.1 THE NDY REGION AS A POLYCENTRIC URBAN REGION

Drawing on the most common criteria of all PURs that have been discussed in the theory section, the following discussion will argue that the NDY region can be considered a PUR. Although not explicitly presented as such, Parr provides us with a minimum and maximum definition of the PUR (Parr 2004). A PUR must encompass a plurality of centres, must be a subnational territory and must be urban in character as the vast majority of population as well as the employment are located in the urban centres. As this minimum definition proves to be too wide and would identify too many regions, entailing that the definition were to lose its analytical precision, the NDY region will be scrutinised as regards to seven more narrow criteria. Not all of these criteria must apply for every PUR although the majority should be met for a justified classification as a PUR.

For the first condition to be fulfilled the clustering of the centres in the region must be positioned neither randomly nor regularly, meaning that there must be open or agriculturally used land separating the centres. For this condition it is not important whether the centres are positioned in a linear, circular or polygonal shape. The pattern of centres in the NDY qualifies as highly polycentric in this regard as the structure is not only very decentralised, but also highly fragmented, due to the fact that the position of new settlements was entirely determined by the location of the coal seams that are at times irregular. The early coal working in the West led to a pattern of many small centres following the coal seams, with agricultural and open land separating the small clusters. Further to the East, where coal lies in lower segments and could only be produced after the 1860's, the settlements are fewer and larger, leaving vast stretches of open or farmland in between them (North and Spooner 1978 p. 265).

The second condition refers to the upper limit of centre separation so as to prevent the classification of an entire country as a PUR. Although scholars are not consistent on this issue, Parr mentions a travel time of one hour between two neighbouring centres as a maximum (2004). The definition of when a city, town or settlement cluster can be defined as a centre is often ambiguous and will also be challenged by this study, yet, for the sake of succinctness, this section will refer to a centre as a city that has more than 100,000 inhabitants. In the NDY region these cities are Bradford, Leeds, Huddersfield, Doncaster, Rotherham, Sheffield, Chesterfield, Nottingham and Derby. Table 3 displays how much time in minutes it takes to travel from one inner city centre to that of each immediate neighbouring city centre. The time shown is in minutes by car and has been calculated conservatively so that only the longest route in time of all shortest-distance routes was counted. The table also shows how much time the longest car journey for each city to its neighbouring city takes. Under the second condition the NDY region clearly qualifies as a densely settled PUR as the longest route, which is between Chesterfield and Sheffield takes only 43 minutes. Figure 11 shows the commuting pattern for a selection of cities. The pattern revealed highlights the complex intra-regional relationships between cities within the NDY region.

Table 3: Neighbouring cities and travel time between these and cities in the NDY region.

City	Neighbouring cities, (max. travel time in min.)	Max. travel time by car in min.
Bradford	Leeds (27), Huddersfield (33)	33
Leeds	Bradford (27), Huddersfield (36)	36
Huddersfield	Leeds (36), Bradford (33)	36
Doncaster	Rotherham (32), Sheffield (43)	43
Rotherham	Sheffield (22), Doncaster (32), Chesterfield (37)	37
Sheffield	Rotherham (22), Doncaster (43), Chesterfield (38)	43
Chesterfield	Rotherham (37), Sheffield (38)	38
Nottingham	Derby (35)	35
Derby	Nottingham (35)	35
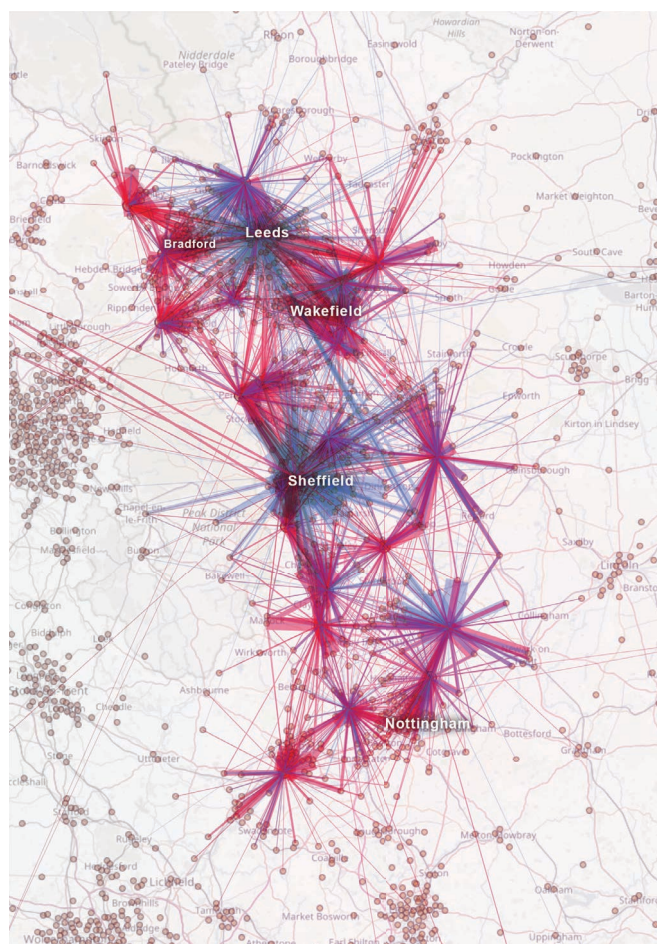
Figure 11: Commuting flows between selected cities of the NDY region. Blue lines indicate net gain, red lines indicate net loss, and the line thickness represents the relative number of flows. ©
http://commute.datashine.org.uk/

The third criterion demands a lower limit between the centres of the PUR that must not be exceeded to avoid that cities, which coalesce into another in the form of a conurbation or a multi-centred metropolitan area would be wrongly qualified as PURs. This is easily fulfilled by the NDY region, as the vast majority of centre structure formed in the 18th, 19th and early 20th century. The oldest centres did not coalesce, as they were usually too far apart as well as too small in size and additionally had to accommodate farmland in close vicinity to the houses of miners to secure their subsistence. The more recent settlements and mines lay by design within quite a distance to another, as it was more efficient to drill only one shaft and extend the coal working horizontally in an underground system of tunnels.

The fourth condition requires that the centre sizes in the region given the same spacing of a comparable region must be larger, meaning that the PUR's centres are more stretched, resulting in an overall closer location to another (Hill 2000). This criterion is fulfilled best in the East of the NDY region where the coal could only be worked towards the end of the 19th century so that smaller towns dramatically expanded with the development of coal mining. An illustrative example for this process is Doncaster that incorporated multiple villages and former suburbs during the peak of its expansion, today forming the Metropolitan Borough of Doncaster. Stimulated by the mining boom in Doncaster, the steel towns Rotherham and Sheffield expanded spatially as the coal produced in Doncaster could now be transported directly to the steel production centres via the river Don.

The fifth condition demands that there must be no prominent difference between the sizes of the larger centres in the region, i.e. there should not be a markedly dominant centre in the region (Parr 2004). The NDY region has also a polycentric character in this regard: Leeds is with its 188 km² the largest urban area in the NDY region, yet is closely followed by Nottingham with 176 km², followed by Sheffield with 167 km². The cities of the second largest category of urban area are Derby and Bradford with each 64 km² and Barnsley with 60 km². The third category constitutes the cities Doncaster with 44 km², Chesterfield with 35 km² and Harrogate with 23 km² (Demographia 2018). Although Leeds is clearly the largest city within the area, it does not dominate the region.

For the sixth condition to be fulfilled, there must be an exchange of economic goods, services and commutes between the regional centres of a magnitude larger than in a non-polycentric benchmark region. This criterion entails that there is also considerable trade between two equally sized centres in the same region – an idea that would not conform to Christaller's theory of hierarchical trade relations. Historically, there has always been trade of iron and coal between the centres in the NDY region. The close relationship between the four counties that make the former coal region today can be

illustrated best in terms of commuter flows: in 2011 the vast majority of the in- and out-commuter flows from Derbyshire were to Nottinghamshire and South Yorkshire. While 4 per cent of the out-commuters and 52 per cent of the in-commuters are with South Yorkshire and Nottinghamshire, only 1 to 11 per cent is with all of the other bordering counties[4] (County Council Derbyshire 2011). The in- as well out-commuting to West Yorkshire is with 1 per cent each very low, which however could most likely be explained by the fact that the border between the two counties is minimal. Overall it must be noted that the industry and labour market is still far from recovery since the mine closure following the 1960s, so that due to the low increase of non-coal jobs out-migration from the former coalfields was dominating the commuting patterns until it turned in 2007 to become moderately positive (Beatty et al. 2007).

The seventh condition refers to the centre specialisation, i.e. the specialisation of the centres in the PUR must be higher than that of a centre in a benchmark region. The NDY region does not fulfil this criterion for most of its history as the entire region's economy was based on a mono-economy of coal and steel. The demise of the coal industry caused 95 to 99 per cent of the overall male job losses in Yorkshire, Derbyshire and Nottinghamshire in the period from 1981 to 2004. This job loss could only be replaced by an increase of little more than 9 per cent of non-coal jobs in the same period (ibid.). The socio-economic situation in the coalfields can still be described as comparably precarious, although a certain degree of specification can be noted: Nottingham's biggest sector in the economy has become the service sector with the business services as the largest sub-sector being responsible for approximately one fifth of output in the city (Nottingham Economic Strategy Research Bureau 2012 p. 12). Derby's biggest economy is in retail, while Leeds has become the UK's largest centre for financial and business services as well as digital, creative and broadcasting after London. Doncaster's economy is very different from that of Leeds as its major focus lies on retail, wholesale and construction (Doncaster Council 2013 p. 9). In Sheffield, a considerable share of the economy still relies on metallurgy and steel making, whilst almost 50 per cent of jobs are in the public sector. The economy in Rotherham is largely based on manufacturing and lower-tech service industries although it is currently diversifying due to large investments (Rotherham Council 2008). Current projections point towards an increasing specification of the centres in the NDY region although the majority of the economy in the former coalfields is still based largely on manufacturing and low-skilled labour sectors.

---

[4] This statistic also includes 20 per cent of the in- and 17 per cent oft he out-population into Derby from Derbyshire.

The NDY region can be clearly classified as a PUR in six out of seven criteria. Whether the NDY region will become more polycentric in terms of centre-specification is yet uncertain. As the region in all other points qualifies unequivocally as decidedly polycentric, it can be classified as a PUR.

3.2.2 HISTORIC ANALYSIS: THE DEVELOPMENT OF THE NDY REGION TO A POLYCENTRIC URBAN REGION

The South and West Yorkshire coalfields were formed by two distinctly different coal-swamps that were separated by a barrier of land when they formed during the Carboniferous period (Hill 2000). The location and quality of those varying coal deposits would almost fully determine the industrial and settlement structure throughout the last number centuries.

The coalfields in Yorkshire spread from the Lower to Middle Coal measures. The best quality of coal in in this area can be found in the area between Silkstone and Barnsley. In the north-south direction a scarp terminates the coalfield. In this area instead of coal measures large deposits of Magnesium limestone that reach until the West of Doncaster can be found. In the West the coal seams are cropping out, which made them easily accessible even in pre-medieval periods; towards the central region of Yorkshire, the coal is of best and densest quality. In the east the seams are thinning and are finally interrupted by the "Don Faults", a structural disturbance between Sheffield and Mexborough which can be held accountable for the relatively late economic and structural development of this very region (ibid.). The larger geological coal measures stretch from Yorkshire down through Derbyshire over Nottinghamshire.

The line parting the two coalfields is located north of Barnsley and Doncaster and stretches from Woolley to Askern. At this line the main coal seam in Yorkshire changes from over 6ft in thickness to one of inferior quality. The coal seams further eastwards descend until they are covered by Permian and Triassic rock that is hard to work through giving the area the name "concealed coalfield" (ibid.). Based on this irregular pattern of location, quality and accessibility of coal, the earlier settlements relying on coal formed in the West. With the technical development of mining tools the settlement and industry could tap the more high-quality coal seams in the East, halting extensive settlement and industrial development until the early 20th century.

Although coal was worked already under the Roman occupation and throughout the Middle Ages, it did not have a notable effect on the local economy or infrastructure. In the 13th century coal was worked in the East of the South Yorkshire are only, reaching from Silkstone to West of Barnsley and Masborough. The output did not exceed a few

thousand tons per year and mining was only a seasonal occupation after the harvest (ibid.).

In the central-east region of the area coal was worked around Barnsley, Rotherham and Sheffield during the 16th and 17th century. Between 1550 and 1650 a minor industrial revolution caused the output of coal to rise. Up until the 18th century the working of coal and ironstone increased and with it the places where collieries were sunk. This drilling process started moving away from the Western outcrops in an eastwards direction as drift bell-pits and shallow shafts were introduced which reached deeper lying measures of coal (ibid.). At that time, miners lived in immediate proximity to the shafts.

Up to the 17th century the mining was confined to the land-sale industry, which meant that coal was consumed only in a small area around the place it was worked, as there was a lack of navigable rivers, canals and seaports to transport the coal over long distances (ibid.). Still, the working of iron ore gradually lead to an increasingly bigger iron and steel trade, with the town of Ripley in Derbyshire producing 25 per cent of the English steel (Natural England 2014).

Due to the industrialisation, the 18th and 19th century saw a quickly growing demand for coal. In 1788 shaft guides were developed which simplified coal winding and the steam engine multiplied the output that could be worked. The industry grew and with it the numbers of workers and settlements around the collieries. During the Turnpike era from 1740 to 1826 the road and canal infrastructure was improved significantly, which facilitated trade and operating procedures (ibid.). As depicted in Figure 12, the improved and expanded transport system stimulated the sinking of more mines, so that a dense pattern of small collieries covered the entire East of the region until the 1850s.
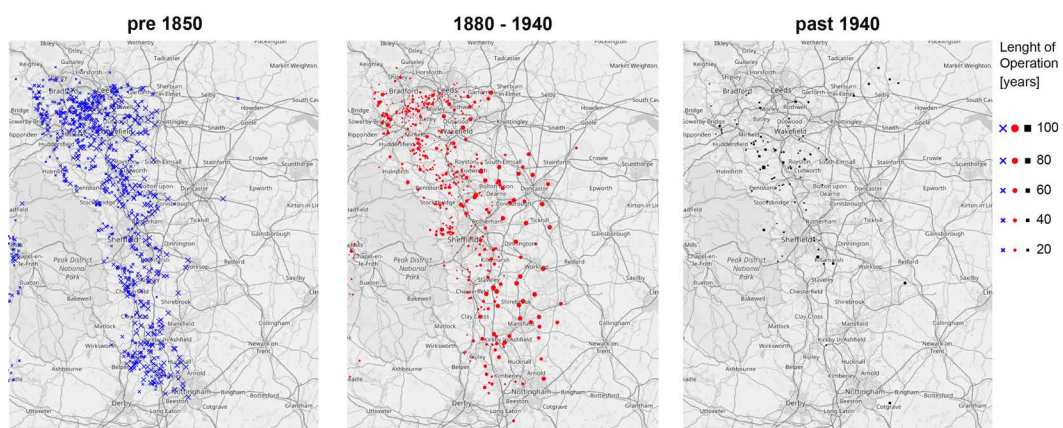
Figure 12: UK Colliery locations divided by date of startup and differently sized by length of years in operation. Three distinctive pattern are identified: pre-1850, 1880 to 1940 and past 1940.

In the 19th century, the seams in the West of the Yorkshire coal region were quickly exhausted, which caused the majority of companies to move further Eastwards on the search for yielding coal seams, leaving behind the settlements that had established around the collieries. The mining primarily focused on some remaining parts in the west of the region where the coal measures were easier to reach. As the methods of working coal were still limited it proved more efficient to sink pits in close vicinity and work coal around them until moving on. The costs of mining rose as the working depth increased which could only be solved by producing greater outputs. This lead to change of the configuration of settlement and industry: instead of many and smaller patches of collieries and settlement, now, fewer and larger structures were established. An exception to this posed the central coalfields in South Yorkshire that yielded by far the coal of highest quality and additionally and exhibited rich measures of coal. This is why most of the collieries that were sunk here between 1851 and 1875 were being worked until the nationalisation of the collieries in 1947 (ibid.) and the infrastructure there could evolve, was expanded and consolidated.

The movement further eastwards becomes evident in Figure 12, which allows us to trace the pattern of coalmines that can account for the settlements in the NDY region today. The progress in mining technology and the depletion of easily accessible coal in the West did not only cause the establishment of new settlements further to the East of the region, but also influenced the configuration of these new settlements. The pattern of active coalmines recorded in the map for the period of 1880 to 1940 shows that the spacing between the collieries was considerably larger than in the East and that mines were continuously operated for almost 100 years. This could explain why we find comparatively larger settlements such as Sheffield and Doncaster in the East of the NDY region that also lie at a comparably larger distance to another. The only region that defies the eastwards trend is the Northern NDY region of Yorkshire, in which the mining continued throughout the last 200 years. The exceptionally rich deposits and high quality of the coal seams in South Yorkshire that proved most profitable until the nationalisation in 1947 could be the reason why the area around Leeds has the largest settlement cluster until today and still is economically the most viable sub-region.

The introduction of powerful winding machines in the process in Doncaster and the central area led to a vast expansion of the coalfield and settlements in an eastwards manner in the period from 1890 to 1930 (ibid.). Up until the first half of the 19th century the major customers of the coalfields were located in the greater region. Customers were households and railways, but also the iron and steel industry in Sheffield and Rotherham as well as the textile industry west of Riding, all of which were fuelled by the coal from the coalfields (ibid.) and had to be connected to the region. As the South Yorkshire coal in particular was of such high quality a new market outside of the region

opened up and the coal was being exported so that in 1875 the South Yorkshire coalfields alone produced six per cent of the national output. Now, customers were steam lines, railway companies and the manufacturing industry in the Eastern and Southern counties of England, which brought the expansion of the infrastructure to accommodate the new markets. The export even increased with the development of the larger collieries of the Doncaster area, which were located closer to the ports and coast (ibid.).

In the Doncaster area, the coal production started relatively late but did so at such a remarkable pace that in between the years of 1905 and 1913 the agricultural town of Doncaster had developed into a large industrial city providing 85 per cent of the Yorkshire coal output in 1924 (ibid.). By 1913 the export trade of the South Yorkshire coalfield comprised 20 per cent of the output with the Baltic states being the most important customers (ibid.). At the same time further settlements around Wakefield and Barnsley after 1908 such as Fitzwilliam and Upton were established (Natural England 2014)

As a reaction to a newly increased competition on the market, the Coal Miners Act of 1930 led to an amalgamation of mining companies and meant to increase the efficiency and productivity of the mines. In 1938 the state expropriated and compensated the landowners of the coal deposits in the region in order to optimise the working of coal in the region (Hill 2000). Small industrial sites and settlements were given up or made more efficient, and the organisation in space moved from a decentralised structure to multiple and bigger centres.

After the nationalisation of all remaining collieries in the UK, the National Coal Board (NCB) managed the entire coal industry and as well as its infrastructure comprising of 225,000 acres of farmland, 141,000 houses, offices, shops and hospitals etc. At this time, four per cent of the British workforce was employed by the NCB corresponding to 796,000 of work staff. Once more the organisation of the coal industry was reorganised administratively and the NCB attempted to modernize the entire industry (ibid. pp. 37-40): since the productivity and output suffered in some less efficient collieries in 1957, the NCB organised 30 reconstructions at a cost of around £65,161,000 (ibid.). Also, the coal deposits were now worked in a holistic manner meaning that the coal should be worked in the most effective way from the collieries in the area. The strategy also included out-of date collieries to be closed, development schemes were established and new collieries were built (ibid.). The third map in Figure 12 shows the drastic reduction in numbers of coalmines after 1940 and how this once again led to much more clustering around fewer but larger collieries, bringing with it more settlement in the larger centres and abandonment of the older structure in the smaller collieries. Again, the Yorkshire

coalfields were the only ones that outlasted the majority of negative trends in the NDY region.

After their production high in 1957 the mines had reached their limits and in the following years, cheap and plentiful oil began to dominate the markets (ibid.). The resulting shrinkage for the industry lasted from 1957 to 1973. When in 1956 there were 840 collieries in Britain employing 700,000 men, in 1971 there were only 292 collieries left employing 290,000 men. In the 1960's half of the industry was discarded as inefficient, and by the mid nineties the government had ended all subsidised to the industry. Again, the focus lay on the most productive coalfields only and the workforce was reduced from 700,000 to 300,000 workers in 1969 (ibid.).

Between 1979 and 1984 the biggest reconstruction programme in British mining history was launched: In the Yorkshire coalfield, three areas were reconstructed and three major complexes were created; also a higher output for open cast mines was decided (ibid.). Now, each site had a high capacity computerized coal preparation plant with Grimethorpe being one of the largest in Europe (ibid.). Unfortunately the major customer, the steel industry in the South Yorkshire Coalfield, collapsed and again collieries were closed, infrastructure abandoned, while the population declined.

In the early 1980s the government had decided to regenerate the coalfields and launched several aid programmes in order to create new jobs (Beatty et al. 2007). The regions of Yorkshire and Nottingham received assisted area status and the programmes had a positive effect until the end of the 1980s (ibid.). Since 1989 the EU started subsidising the coalfields of which most of the financial aid was used to improve business support and restore and adapt infrastructure (Beatty et al. 2007).

In 1992, only 53 collieries in all of the United Kingdom were left. In the following years most of the industry was privatised (Hill 2000). In 2005, the by then privatised coal industry employed fewer than 4,000 in only 8 remaining collieries (Beatty et al. 2007). In 1990, the EU started the RECHAR programme of aid for coalfields so that in 1996 the redevelopment government agency "English Partnerships" was extended and came to public sector expenditure of £600 million. The South Yorkshire and Nottinghamshire regions that had until 2000 received "Objective 2" support by the EU, were then classified as an "Objective 1" qualifying them for even more financial support (ibid.).

3.2.3 SYSTEMS OF TRANSPORT AND FORMATION OF INFRASTRUCTURE

The structural development of the region can be explained best by the depth of the coal measures, that were surfacing in the West of the region, making possible early workings of coal without refined techniques, whereas further Eastwards the coal seams ran much

lower requiring much more advanced mining skills and equipment. The better the production methods became, the further moved the mining and with it the settlement expansion towards the East. The different depth and quality of coal measures can therefore explain the late settlement and industrial development of the Eastern parts of the NDY region in comparison to the early-developed Western parts (Hill 2000).

In the Middle Ages, the clearance of forests led to a landscape that was characterised by farm land, villages, hamlets and individual farms (Natural England 2014). After the 1650s, the last remaining open fields were progressively enclosed - first privately, then by the Parliamentary Enclosure Acts - resulting in a more regularly shaped field structure. Before the 17th century the coal mining was confined to the land-sale industry due to the lack of navigable rivers, canals and seaports so that coal products could only be consumed in the area close to where the coal was being worked (Hill 2000), so that there were hardly any connections or much trading between singular towns or settlement clusters. Until the 17th century the settlement patterns were characterised by a high density of dispersal with levels of nucleation rising to the south whereas hamlets to the north were typically set around commons, greens and farmsteads along track ways. Typically, the entire NDY region was built of settlements around hall-churches (Natural England 2014). Starting in 1734 the river Don was used for the transportation of coal. The further construction of canals in the South Yorkshire coalfield until 1819 created a network of transport that provided an access to the centre of Sheffield. In 1838, the first processes begun that initiated the polycentric character of the NDY region: the first railway between Rotherham and Sheffield but also a route between Derby, Rotherham, Chesterfield and Leeds was established (Hill 2000), so that slowly a network of canals, railway and track ways was established that would interconnect the larger region. The creation of better roads throughout the area characterised the Turnpike era of 1740 to 1826 and accelerated the process of interconnection; especially in the years before 1750 the navigation of the Don had been improved as far as to the outskirts of Sheffield. Before the 1850s the Dearne and Dove Canal had been built which meant that with the further developing canal and road infrastructure more collieries could be built. After 1850, the development of the railway system led to a further boost of the coal economy (ibid.) bringing with it more infrastructure. Due to the industrialisation in the late 18th century a process of further nucleation of settlements as well as their expansion could be observed over the region (Natural England 2014). Throughout the first half of the 19th century the Yorkshire region was characterised by shallow workings of coal, which required the companies to move on to new areas as soon as the levels of coal they had reached, were exhausted. The second half of the 19th century, however, saw a change in structure of the collieries' and settlement location. Due to the extension of the railway network, most of the collieries

and settlements clustered around main lines. Railway routes were built to criss-cross with other major routes entailing a boost for all collieries, which were positioned on major routes or connected to the network through short branches (Hill 2000). It was at this time when the sinking of collieries moved eastwards. The newly sunk collieries worked deeper and mostly clustered around the rivers Dearne and Don. Most of the collieries sunk after 1851 were worked until the nationalisation of the industry almost 100 years later (ibid.) leading to a stabilisation and further development of the infrastructure built in the mid 19th century.

Until the first half of the 19th century the major turnover of the coalfields was made through trade with the regional industry only (ibid.), which built a network structure that would serve only the inter-regional connectedness. The very high quality of the South Yorkshire coal soon led to the expansion of the market outside of the region and the coal from the South Yorkshire coalfield was being exported. Now, the new customers were located in the Eastern and Southern counties of England, outside of the region and the infrastructure was extended to that the region would also be connected well to out-of-region destinations. In the following years, the export increased with the development of the larger collieries in Doncaster as the area was closer to the ports and coast (ibid.). The years from 1876 to the 1900 saw a consolidation of the central region of the Yorkshire coalfield, which had the largest deposits regionally (ibid.).

In the first decade of the 20th century the large-scale development of the "concealed coalfield" in the east of the region began and the mining shifted towards the Doncaster area where first collieries were sunk (ibid.). The year from 1911 to 1920 saw the greatest expansion of the Doncaster coalfield despite of Second World War (ibid.); but not only the Doncaster area but also other regions in the northwest and east had become industrialised in the last years and were expanding such as the villages of Fitzwilliam and Upton. Further East in an area to the south and southwest of Barnbow mining did not start until 1942 (Natural England 2014). At the same time, some collieries were sunk in Nottinghamshire, which is why the infrastructure there is still rather young and built to accommodate only few and large mines (Hill 2000).

The period after the nationalization in 1947 saw a change from shaft to drift mines. An underground network made it possible for coal to travel directly to the factories where it was processed without even entering the surface once. Additionally, the network made it possible for different units to share the same surface facilities. Over the years more and more facilities introduced this full retreat mining technology (ibid.). This process entailed that the infrastructure below the surface was not further extended and was in character different from that of the much older coalfields in the West.

The region of the former coalfields is dominated by extensive urban influences in the north and the central parts of the region, whereas the South remains comparably rural with fewer large urban centres. Throughout the entire region, the size of fields and field patterns are irregular and inconstant: This is due to medieval clearances of woodland, the piecemeal enclosure of strip fields and the land that was farmed by miners and weavers to secure their subsistence. In the 18th and 19th century the commons were enclosed. Due to the expansion of farms following the 1950s and the emergence of urban fringes, in some areas the old field patterns gave way to new structures, whereas in some areas the old patterns remained intact (Natural England 2014). Most of the settlements in the coalfield region, however, do not date back as long as that, as most mining villages were built anew due to the rapid industrial expansion in the 19th century (ibid.). Due to the fragmentation of land, farming remains difficult in the area until today (ibid.).

Today, the region of the former coalfields of Nottinghamshire, Yorkshire and Derbyshire has strong transport features linking the region with other areas. For one, the M1 running from North to South connects Leeds and London. The A1 is the main route of transport north of Leeds. Additionally, there a number of railway lines passing through the whole area of which most of them are former mineral lines which have now been reclaimed to constitute new multi-user trails (ibid.). In the last years recent engineering, manufacturing and light industrial use, commercial and retail sites have reached the urban periphery taking over from the coal industry. The networks that resulted from this development are motorways such as the M1, which was opened in the late 1950s and 1960s and continues to be improved, as well as the M62 and more railways and canals. Recently, more warehousing development around motorway junctions can be observed. Today, new and improved roads generate a growing demand for warehouses but also for housing development, commerce and industry; in some places suburbanisation processes can be observed. Wherever spoil heaps and other reclaimed land could be used for light industry and warehousing, an economic surge can be made out, especially in the vicinity of motorways (ibid.).

A variety of rivers flow through the region connecting urban with rural areas mostly from the West to the East. Due to urban fringe pressure the landscape takes on a rather fragmented pattern: Built-up areas, industrial land, smaller areas of dereliction and farmed open country are adjacent to substantial stretches of intact agricultural land of arable and pastoral use (ibid.). Among the most influential rivers are the Aire, the Calder, the Dearne, the Rother, the Don and the Erewash, which were central for the development of the area. Especially the areas around Leeds and Sheffield can be described as some of far-reaching urbanisation, extensive housing and industrial development: Both cities exhibit terraced and back-to-back housing and grand 19th century municipal buildings. Generally, cities in the coalfield region are structured

around centrally positioned churches. Canals, roads and railways are complemented by an extensive network of multi-user trails, canal towpaths and long-distance food paths such as the Trans Pennine Trail and the Ebor Way (ibid.)

Concluding, the region has a highly-fragmented character still. The structure to be found in the West is very different from that in the East due to the much later and spatially extensive coal-mining development in the East. The East has fewer but larger centres as well as fewer but larger roads that were established to connect locations farther away. The West is connected by a more fine-grained and decentralised structure with multiple smaller centres as coal had to be transported above the surface and the simple machinery did not allow for long-term and deep drilling in only one location.

3.3 THE RUHR VALLEY

Geologically the Ruhr zone is part of a bigger northwest European coal belt of mining industry that stretches over the South of Poland and Belgium over the North of France and South Wales to the Midlands in England. The Ruhr region is an area located in the very west of Germany in the Land North Rhine Westphalia. The Ruhr valley as a cohesive industrial region formed gradually and compared to the NDY relatively late in the beginning of the 19th century. Just as the NDY region, the Ruhr region's infrastructure has largely been determined by the working of coal, ore and its industry. The Ruhr region is neither based on historical borders nor natural landscapes. Starting in the South of the Ruhr zone, the region expands further north to the Hellweg zone enclosing the cities Duisburg, Mülheim, Essen, Bochum and Dortmund. Further north, the Emscher lowlands and even further north the areas of Recklinghäusen and Höhenrücken are part of the region. North of this, the area stretches over the Haard and the Lower Rhine to the Hohe Mark, which lies north of where the river Lippe flows.

The region of the Ruhr can be split into three different natural landscapes that determined the different usage and settlement of the land: the lower Rhine Bight, the southern Rheinische Schiefergebierge and the Westphalian Bight in the North. The region is named after the lowland of the river Ruhr, which flows through the Schiefergebierge. It was also this river that first disclosed the carboniferous measures in the rocks that lead to early coal mining and open diggings. In between the north of the Ruhr valley and the river Emscher lowland lays the Hellweg. The strata of Emscher marl made the area around the river rather marshy, which made it unattractive for settlements up until industrialisation in the 19th century. Early settlements could instead be found in the mountainous area of the Recklinghäuser and Vestische Höhenrücken further north. Much alike the Emscher river zone, the area around the river Lippe remained unattractive to settlers due to its wet soils. The three landscapes of

the Ruhr are structured along a line from the west to the east. This eastwards change is matched by the geological structures in the region: The highest measure exhibits Upper Cretaceous deposits which is often called the 'upper burden' as they cover the lower lying carboniferous measures (Keil & Wetterau 2013 p. 4). As the coal seams were easier to access, the first settlements and coal working started here. In the south of the Ruhr valley the coal deposits crop out, further north they lie two to three times lower. The quality of the coal increases systematically from the south to the north towards the Westphalian Bight throughout the entire region with the coal of higher quality usually lying in higher measures and the coal of higher quality usually to be found in deeper strata (ibid. p. 5). Whereas in the English midlands the coal measures lie rather high up, in the Ruhr regions they lie under hard segments of marlstone are interrupted by faults and generally lie rather deep (Kersting 2009).
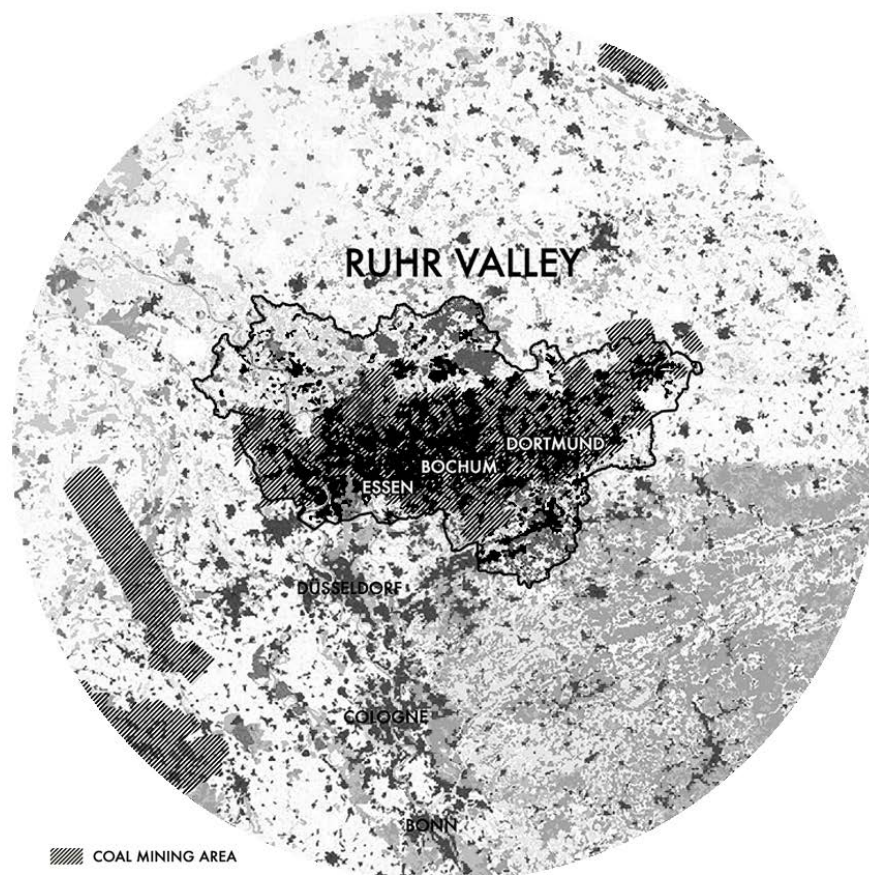
Figure 13: Area of investigation Ruhr Valley region, Germany with highlighted areas of coal mining. Background showing European Urban Atlas land monitoring.

This is why in the North the industry and settlement developed only in the 19th century when powerful machinery became available, whereas in the South in the higher lying seams settlers already produced coal in a small industry in the 18th century. Although

most of the collieries closed in the 1960's and 1970's, there are only two mines that produce coal in 2018.

3.3.1 THE RUHR REGION AS A POLYCENTRIC URBAN REGION

The Ruhr Valley is commonly identified as a prime example by the literature (Hall and Pain 2012; Keil and Wetterau 2013; Kersting 2009; Münter 2011; Reicher et al. 2015; Strohmeier et al. 2002, 2015; Strohmeier and Kersting 1992). Under the minimum definition it easily qualifies as a PUR: With its 53 municipalities and eleven larger cities it encompasses a plurality of centres, in the state of North Rhine-Westphalia and has a distinctively urban character throughout (Reicher et al. 2015). The Ruhr agglomeration has a population of approximately five million people and is the fifth largest conurbation in Europe and the most densely populated region of Germany (Keil and Wetterau 2013).

For the Ruhr region to qualify as a PUR, irregular patches of open or agricultural land must separate the clustering of centres. The Ruhr region easily fulfils this condition, as there are plenty of wide spaces, green sites, meadows and fields between settlements that can be reached from almost anywhere in the region in under less than 30 minutes of a car journey (Reicher et al. 2015). This structure dates back to the scattered and non-central locations of settlements that were determined by the accessibility and quality of the coal seams.

The second condition for a PUR requires an upper limit of centre separation for which the analysis of the NDY region employed a maximum travel time with a car of one hour to all its immediately neighbouring cities and only focused on those centres with more than 100.000 inhabitants. In the Ruhr Valley, these are the fourteen cities that meet this criteria: Moers, Duisburg, Oberhausen, Mühlheim an der Ruhr, Bottrop, Essen, Gelsenkirchen, Recklinghausen, Herne, Bochum, Witten, Dortmund, Hagen and Hamm. Table 4 shows how much time it takes to travel from one inner city centre to that of the immediate neighbouring city (in minutes). All of the journeys have been calculated to be made by car with the recommended speed for each respective street. The Table 4 shows the longest travel time possible for each city's neighbouring city. The criterion of the upper limit is clearly fulfilled as the longest journey, which is between Hamm and Dortmund, takes only 45 minutes. Figure 14 shows how these distances are also expressed in the complex interrelated pattern of daily commuting flows for the Ruhr Valley.

Table 4: Neighbouring cities and travel time between these and cities in the Ruhr Valley region.

City	Neighbouring cities, (max. travel time in min.)	Max. travel time by car in min.
Moers	Duisburg (15), Oberhausen (24), Mühlheim a. d. Ruhr (31)	31
Duisburg	Moers (15), Oberhausen (17), Mühlheim a. d. Ruhr (24)	24
Oberhausen	Mühlheim a. d. Ruhr (24), Bottrop (21), Essen (26), Duisburg (17), Moers (24)	26
Mühlheim a. d. Ruhr	Oberhausen (24), Bottrop (32), Essen (25), Duisburg (24), Moers (31)	31
Bottrop	Oberhausen (21), Essen (33), Gelsenkirchen (18), Mühlheim a. d. Ruhr (32)	33
Essen	Gelsenkirchen (24), Mühlheim a. d. Ruhr (25), Bottrop (33), Oberhausen (26), Bochum (28)	33
Gelsenkirchen	Bottrop (18), Essen (24), Bochum (25), Herne (20), Recklinghausen (28)	28
Recklinghausen	Herne (24), Gelsenkirchen (28), Bochum (28)	28
Herne	Gelsenkirchen (20), Bochum (18), Recklinghausen (24)	24
Bochum	Essen (28), Witten (21), Herne (18), Gelsenkirchen (25)	28
Witten	Bochum (21), Dortmund (21), Hagen (29)	29
Dortmund	Witten (21), Hagen (28), Hamm (45)	45
Hagen	Witten (29), Dortmund (28)	29
Hamm	Dortmund (45)	45
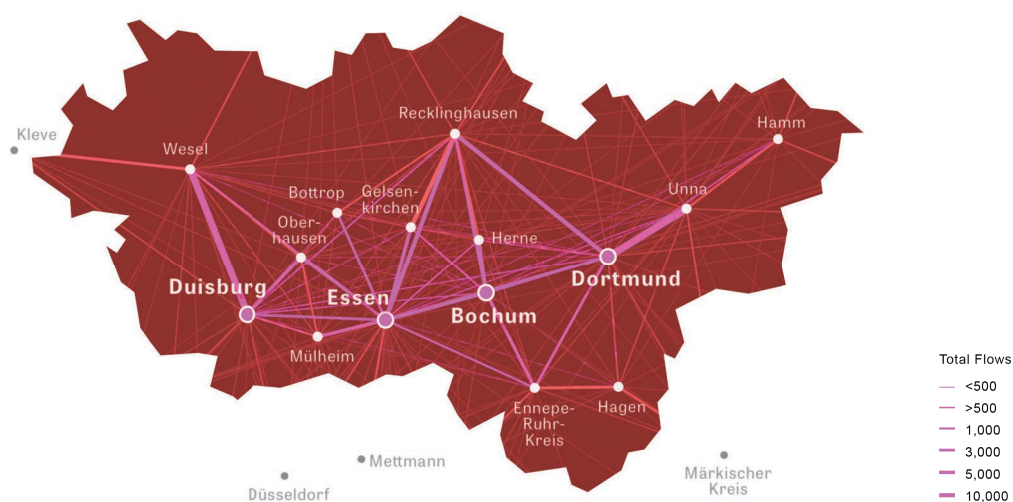
Figure 14: Commuting flows between cities of the Ruhr Valley. © Zeit Verlag
http://www.zeit.de/feature/pendeln-stau-arbeit-verkehr-wohnort-arbeitsweg-ballungsraeume

The third criterion requires that a certain lower limit between the centres of the PUR must be maintained, to prevent that a region of coalesced cities is identified as a PUR. In the Ruhr Valley, the settlement structures that lie outside of the cities consist of a combination of residential and industrial areas, as well as storage facilities and energy supplying infrastructure so that it is oftentimes difficult to determine the end of a centre (ibid.). It is also true that the cities and towns of the Ruhr Valley are exceptionally evenly populated compared to other regions commonly identified as PURs, such as South Central Belgium. The city centres across the region are usually not much more densely populated than the fringes of the city, yet there is no coalescence of cities or towns between any of the cities or towns in the region, nor are there any corridors where centres merge (Münter 2011).

For the fourth condition to be fulfilled, the city centres must be comparably large in size compared to other regions, resulting in an overall closer location to another. In the Ruhr region, the settlement structures in the North are plentiful in absolute numbers but smaller in size, as they largely date back to collieries established in the early 19th century when machinery was neither powerful nor refined, making it more efficient to maintain many smaller and superficially dug mines. With the technical development towards the end of the 19th century, the approach to build fewer but deep shafts dominated the industry and initiated a coal mining boom in the North of the Ruhr. This led to fewer but spatially larger settlements. With the arrival of the car the process of urban sprawl across the entire region was actively promoted by the construction and extension of new motorways and arterial roads (Keil and Wetterau 2013).

According to the fifth criterion, PURs cannot have one prominently larger centre that dominates the network. In this regard, the Ruhr Valley is most likely one of the world's prime example due to its relative lack of a hierarchical regional structure. Hall and Pain scrutinise the region's largest four cities Essen, Bochum, Dortmund and Duisburg and find that they are extremely similar the terms of employment, economic output, commuter flows and size (Hall and Pain 2012). For the analysis of hierarchical settlement structures they employ the spatial entity of the functional urban region (FUR) that consists of a core, defined in terms of employment, size and density as well as a ring described in terms of regular daily journeys commuting to the core (ibid.). Fewer FURs in a region denote a much more decentralised structure. Against the background of the extremely large population of five million of the Ruhr Valley, the region has only seven FURs, with only the megacity region of Dublin having fewer FURs in West Europe, while South East England has 51, Paris 30, Randstad in the Netherlands 25 and Central Belgium and EMR Northern Switzerland 8 (ibid.).

The sixth condition demands that the exchange of economic goods, services and commute between equally proportioned centres in the region must be larger than in non-polycentric regions. Hall and Pain measure polycentrality based on commuter flows between the centres of a regions. They find that there are particularly high commuting flows between the equally-sized centres of Essen, Bochum and Dortmund (ibid.). Their analyses of regional commuting patterns show that the Ruhr compared to other PURs in Western Europe, such as Central Belgium, Randstad, South East England or the German Frankfurt metropolitan area, is much more de-centralised and interconnected (ibid.). In the larger centres, such as Dortmund, Bochum, Essen, Hagen or Duisburg, more than 60 per cent of the population live and work in the cities, while 10 to 12 per cent are classified as daily out-commuters, whereas 15 to 18 per cent are commuting into the cities on a daily basis (ibid.). Münter finds that the commuter zone of the Rhein-Main region that encompasses the Ruhr has the largest daily commuter flows of all German regions (Münter 2011). Hall and Pain provide a measure of relative connectivity of the centres in a region, which measure the gradients of connectivity from the highest- to the lowest-linked city pertaining to the relative importance of business networks to other cities. Their analyses show that the Ruhr Valley has an enhanced and very un-hierarchical connectivity structure between the cities Dortmund, Essen and Duisburg (Hall and Pain 2012). The authors conclude their comparative study with the finding that the Ruhr Valley is among the most polycentric regions in terms of business exchanges in Europe, and the most polycentric region, when the bordering nodes of Düsseldorf and Cologne are included (ibid.).

The seventh condition requires a high specialisation of the centres in a PUR. Before the decline of the coal industry in the 1960s an entire century was characterised by an extremely uniform economic structure across the region with the exception of smaller industries producing chemicals, stone and sand that outlasted the coal industry. In the early 1970s a new development programme brought the Opel car production to Bochum but failed in attracting other industries until the late 1970s (Kersting 2009). The new industries that gradually diversified the economy in the Ruhr were mostly focused on information, communication and biomedical technologies. A small industry in Dortmund specialised on technology research in cooperation with the Technical University, but it was not until the 2000s that any notable economic specialisation developed when a regional structural policy programme of the Project Ruhr aimed to strengthen the local businesses in twelve sectors. The energy sector and new energy technologies in Recklinghausen, Oberhausen, Essen and Gelsenkirchen were promoted; Dortmund, Essen, Bochum and Hagen specialised in Information Technology; while especially the towns closer to the border of the regions focused their development on medical and health technologies. Other sectors that the initiative

successfully promoted locally are microstructure technologies and microelectronics, water and sewage technologies, new chemistry, mining technology, material management, mechanical engineering, logistics, design and tourism and leisure (Keil and Wetterau 2013). Although these industries have registered steadily increasing turnover, the degree of specialisation in the Ruhr region does not exceed that of other non-polycentric regions in Germany yet. Although, the Ruhr valley does currently not qualify as a PUR in this regard, more specialisation can be expected to develop in the years to come.

The Ruhr region qualifies as a PUR in six out of seven criteria, although the seventh criteria is likely to be fulfilled in the years to come. As the Ruhr turns out to be markedly polycentric with regards to the remaining six criteria, it can be concluded that it is a highly suitable example for the study of polycentric urban regions.

3.3.2 HISTORIC ANALYSIS: THE DEVELOPMENT OF THE NDY REGION TO A POLYCENTRIC URBAN REGION

During the 8th century towns such as Duisburg, Dortmund and Recklinghausen became important centres for trade and business. Next to its use as a military route, the Hellweg primarily served as a trading route with cities such as Mülheim, Essen, Bochum, Castrop, Dortmund and Unna and stretched from the mouth of the river Ruhr to the river Weser and to the river Elbe. Some of the cities located on this route developed into trading cities, some of them even joining the Hanseatic League in the 12th century, which profited their further development. The following centuries of wars disrupted the economic development of the area and the population mainly relied on agriculture and handcraft (ibid. pp. 6–10).

Coal was worked as early as the 14th century around the easily accessible outcrops in the south where the coal production did not require refined technology or skill. The 'lean coal' from this area was used for domestic purposes only. This meant that settlements in the area could profit from the coal but did not establish any trade or infrastructure with other parts of the region (ibid. p. 6).

With the 1780 establishment of sluices between Fröndenberg-Ruhrort and Duisburg the Ruhr could be used in large parts as a means of transport for the still very small coal industry. This new route of transport facilitated trade between towns (ibid. p. 7). The following years until 1870 can be described as the early industrial period of the Ruhr valley.

In 1803, the first iron smelting plants of the Ruhr region were located at Gutehoffnungshütte where they today still operate under the name of MAN Turbo AG.

Unlike in the English Midlands, coal could only be worked in large quantities and comparatively late (ibid. p. 12). Once the Coalition wars had ceased in the second half of the century, trade started to flourish once more through a new transport route that connected the Lower Rhine area with the Netherlands so that especially the trade with regional crafts thrived (ibid. p. 11). When the region became part of the Prussian Kingdom in 1815 the entire German federation opened up as a new large market so that trade could develop further, and the networks were expanded to accommodate the new markets and the process of industrialisation begun. It was over the course of this century that a distinctively polycentric and urban structure developed in the region (Reicher et al. 2015). By the 1850s a dense pattern of small collieries scattered across the Southern half of the Ruhr Valley (Figure 15).
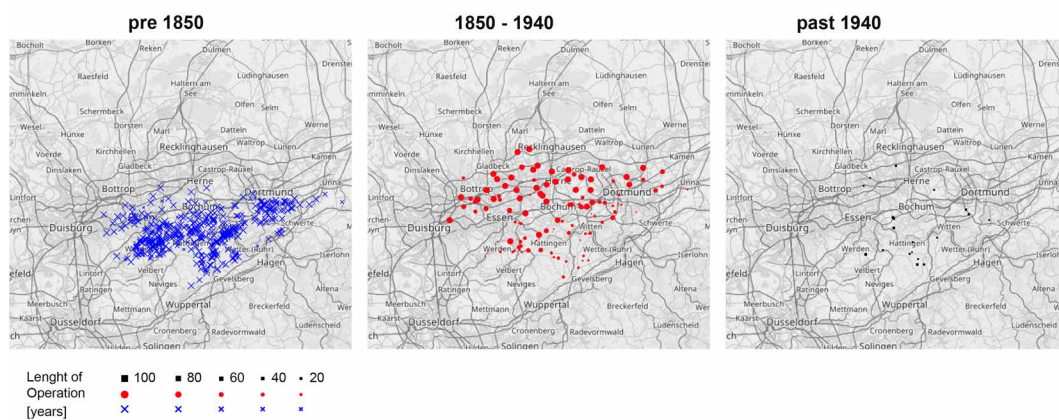
Figure 15: German Colliery locations divided by date of startup and differently sized by length of years in operation. Three distinctive pattern are identified: pre-1850, 1850 to 1940 and past 1940.

For the first time in the Ruhr coking coal was accessible in large quantities, which defined the starting point of the age of coal mining in the entire region. The high quality of the coal did also revolutionise the smelting of iron ore, which entailed not only the surge of the coal but also of the iron industry. In the course of the 19th century, the development of the coal industry proceeded further north: in the 1840s the technology of deep mining had spread to the Northern parts of the Ruhr region, most notably the cities of Duisburg, Essen, Bochum, and Dortmund in the Hellweg zone. Yet, as the demand for coking coal remained unfulfilled, the development progressed further northwards so that in the 1860's coal mining was initiated in the Emscher zone in the cities of Oberhausen, Bottrop, Herne and Castrop (Kersting 2009). Around the river Lippe, the major coal chemistry industry formed where most coal-fired power stations were run entailing a slight growth of this area. It was also around this time that the first iron smelting plants were built in the Emscher zone where bog iron was produced in opencast mines.

The middle of the 19th century saw a surge in industrial growth as most of the transportation could be transferred to the railway system. Through the connection of railway networks to the world's largest inland port and inland container transhipment port the first step was taken for the region to partake in the global market and become PUR connected to the international market (ibid. pp. 5–7). Another catalyser for the regional industry and its growth was the improvement and expansion of the canal system. As the industry expanded further East- and Northwards the establishment of canal structures followed. The Emscher and Hellweg zones of all the areas in the region developed quickly in terms of industrialisation after 1850. The Lippe and Lower Rhine area however would stay a development area until gas-flame coal was worked and used in the coal chemistry and power generation (ibid. p. 14). The period after 1895 until the beginning of the First World War can be regarded as the peak of industrialisation and booming period for the industry (ibid. p. 13). For the period between the 1850s and the two World Wars, the pattern of coal mining in the Ruhr resembles the trends of the NDY region. Figure 15 allows us to trace how the pattern of collieries moved to the North-East of the Ruhr and changed to a structure of fewer but larger collieries with much more space in between them. Structurally, this could explain how the four cities Dortmund, Essen, Duisburg and Bochum profited from this shift. The settlement structure became more clustered around the areas causing these four centres to expand significantly in a short time making them the largest cities of the region today.

It was due to the coal working and industry that the Ruhr Valley until the beginning of the Second World War had become the most influential industrial conurbation in Europe. The period before and between the First and Second World War can be regarded as the last industrial phase of the Ruhr region. At war times, the primacy of the industry was to support the war economy so that established and traditional structures and processes were maintained which prevented any innovation or change. In the 1950s the coal industry was the major pillar of the economic reconstruction of Germany: it was the energy source for the transport systems, iron and smelting industry and it was primarily used in electricity generation, the chemical industry and in households. The first coal crisis of 1958 was initally regarded as a mere temporary crisis and the signs of growing energy diversification on a European market were widely discounted. Yet, as Germany had then become part of the international market, the competition with cheap coal from abroad and the unfavourable mining conditions in the Ruhr valley made the decline of the Ruhr's coal industry inevitable. Additionally, Germany was obliged to import coal from the United States. German coal lost most of its market share as natural gas and oil constituted reasonable alternatives. Moreover, drastic efficiency reductions in the iron, steel, shipping and railway industries contributed their parts in the coal industries decline. Towards the end of the 1950's the coal measures south of the Emscher were

exhausted and the industry withdrew from the area closing most of the mines and with it the settlement structures were either abandoned or shrunk drastically (see the panel for the period after 1940 in Figure 15). In the 1960s most of the coal production was situated North of the river Emscher (ibid. pp. 13–5). In the 1980s the last coal mines around Dortmund and in the Eastern regions were discontinued. Since the crisis more than 130,000 million Euros were invested as subsidies into the coal mining industry, which were decided to end in 2007 and entirely ceased in 2018. When in 1960 in the Ruhr alone there were almost 264,000 jobs in the coal and steel industry, this number plummeted to 44,000 in 2006; in 2018, there will only be two mining sites, one for anthracite, one for coal left in Ruhr region (ibid. p. 39).

The coal production as the main pillar of the Ruhr's industrial development was complemented by the steel and iron industry and, in later years, by a smaller industry working chemical and other natural resources, such as bog iron, that contributed to the growth of the area. As the iron industry still consumed large amount of coal it was best located very close to coal deposits. As it was cheaper to transport the iron ore to the coal plants than the other way around, an extensive transportation network was built in between iron and coal plants. In a short period of approximately 60 years a wide range of interconnections between the production units had developed and spread all over the Hellweg zone.

Unlike the coal industry, the steel industry did not follow the northwards development. Due to the post-war economic crisis, the steel and iron industry formed large-scale enterprises to better their chances on the world market but were soon to be de-centralised by the allies through the Ruhr Statute with the aim of dismantling the industry. The closing down of old production sites first seemed to harm the economic growth of the area but at the same time presented the industry with the opportunity to modernize and built on new technologies. This restructuring led to an economic growth of the crude steel industry from 1950 to 1972. The global competitive market that had already hit the coal industry severely, had now started to affect the steel industry too: New materials such as ceramics and plastics as well as the transition to lightweight constructions, led to a gradual decline of the steel industry in the Hellweg zone. Whereas British and other competing steel producers obtained large subsidies, the Ruhr's steel industry had to do without this support. The decline led to a significant restructuring of the Ruhr zone (ibid. p. 16).

The second much smaller industry on to which the Ruhr region still relies, produced chemical, stone and sand: Since 1925 iron ore salt is worked in the area around the Wesel rural district, the products of which are still used in the chemical industry and for road salt (ibid. p. 6). One of the resources still worked in the region until today are sandstones

and sands. Close to the town of Hagen compact limestone and dolomite is worked which is still used in by the steel and concrete industry. More sandstone is excavated near Witten and Herdecke; in the Northwest of the Ruhr gravel and clay are produced still; In Haltern, quartz is worked. Even though these relatively smaller industries did not reach the size and influence of the coal industry they still contributed their share to the industrial development and growth of the region.

3.3.3 SYSTEMS OF TRANSPORT AND FORMATION OF INFRASTRUCTURE

Compared to the NDY region the regional structure of the Ruhr developed in an interconnected manner in a much shorter time. The settlements and coal mining were much short-lived than the English midlands as the often-occurring faults, the marlstone layer as well as the relatively flat coal seams did not allow for the mines to be worked as long, which could explain the relatively small settlement clusters across the South of the region.

Also, the often-changing administrative and political affiliation of the different areas comprising the Ruhr valley today had an impact on economic structures and settlement patterns and the integration into the larger region: After the reorganisation after the Congress of Vienna in 1815 changed the structure as smaller administrative territories such as ecclesiastical territories disappeared and other territorial claims were abandoned as the Ruhr became part of Prussia. The region was split in threes different administrative divisions: The parts to the West were split into the Rhine-province (Rheinpreußen, Rheinland), the administration of Düsseldorf whereas the larger eastern part was incorporated into the provinces of Westphalia. In each respective province, the authorities declared larger cities as independent, such as Essen (1873), Duisburg (1874), Dortmund (1875), Bochum (1875) and Hagen (1887). As happened with the example of Bochum in 1885, once cities became too large, parts were detached and established as new towns, as happened with the districts of Gelsenkirchen and Hattingen in Bochum or Buer, Hörde, Osterfeld and Hamborn. A constant influx of migrants and economic growth along with the practice to split off districts led to an extensive urban sprawl in the region (ibid. p. 18). The second reorganisation took place in 1926 and 1929 when several cities were incorporated into others and boundaries changed. In 1946, the Prussian provinces were dissolved and the Ruhr became part of the new state Northrhine-Westphalia. In the mid seventies, the region was restructured into much larger administrative units to make administration more efficient. With this step also the urban planning competence was transferred onto lager administrative bodies where before smaller authorities had planned the districts (Reicher et al. 2015).

The Ruhr region's typical polycentric structure was brought about by the working of coal. Before 1840 the region was only sparsely populated with only 240,000 people living in the area. Today most of the continuous building-up is done along the Hellweg and the river Emscher. The earliest bigger mining complexes started in the South of Ruhr where mostly lean and forge coal could be worked. The first bigger settlements were called cottar settlements and lay in the vicinity of the mines. The smallholdings that the miners had for agricultural purposes led to the characteristically dispersed settlement pattern in the area. On the contrary, due to its fertile soils the Hellweg zone to the north was an important agricultural zone and exhibited nucleated villages next to the town of Duisburg, Mülheim, Essen, Bochum and Dortmund. When coal was tapped in this region the population increased quickly. North of the Hellweg zone, the Emscher lowlands are characterised by thick layers of marl and prevented almost any settlement. Even further up north, the areas of Recklinghäuser and Höhenrücken are fertile zones that favoured settlement. Again, further north, the Haard and the Lower Rhine as well as Hohe Mark which lies north of the river Lippe had always been only moderately attractive to settlers (ibid. p. 20).

In the mid 19th century the railway network had been extended significantly and was used for the majority of transportation. It was also at this time, that the railways were connected to the biggest inland port. With the Cologne-Minden railway in 1847 and the Bergisch-Märkische in 1862, the two major railway lines of the Ruhr valley were established. The former connected Duisburg, Oberhausen, Dortmund and Hamm, the latter connected the cities of Duisburg, Mülheim an der Ruhr, Essen, Bochum and Dortmund. Not only did the two lines facilitate working processes and trade but also were themselves customers for the coal.

When the era of industrialisation was at its peak in the second half of the 19th century, early findings of inferior coal caused the early-dispersed structure of settlement. This area, called the Ruhr zone, has been changed least by industrialisation as the newer more promising coal findings and new technology caused a northwards movement. In contrast to this, the cities further of Duisburg, Mühlheim, Essen, Bochum and Dortmund quickly turned into compact and industrialised cities that exhibited large iron and coking plants and drew migrants during the higher phase of development towards the end of the 19th century. It was due to the new establishment of the first Cologne-Minden railway that the sparsely settled Emscher zone gained of locational value and a great number of large mines and smelting work were established. The industry in this area grew so quickly that workers from East Prussia were recruited which lived in newly erected colonies in close vicinity to the plants and mines. The colonies were characterised by their compact and homogenous building structure. Due to the industrial boom the cities along the Cologne-Minden railway were mostly build

only to accommodate the workers in close proximity to their workplace. The increase in industry and population made the Emscher and Hellweg zone the centre of the Ruhr. Whereas the typical structure of colonies erected for the miners around the 1860s were characterised by their small building along side streets, the later building structured tried to imitate much more naturally grown villages with spacious places and winded streets. The area of expansion yet to develop were the Lippe zone in the North and the Lower Rhine area in the west (ibid. pp. 21–2). In the 1920s and 1930s the urban planning and the Settlement Association of the Coal District (Siedlungsverband Ruhrkohlenbezirk, SVR) attempted to promote a non-industrial urban development of the cities in the region: The cities' peripheries in the south were supposed to be free of industrial buildings and should be used for middle class housing and recreational purposes; the cities northern perimeters were subject to massive industrial development. After the destruction of the Second World War the reconstruction of the cities followed the industrial early patterns, as one tried imitate to the old structures and success as soon as possible. Due to large in-migration the cities in the Hellweg zone was extended swallowing smaller towns south of the Emscher (ibid. pp. 23–4).

Despite the industrial decline of the second half of the 20th century, the urban development continued. The newer structures exhibited more and extended traffic spaces, open and semi-opened settlement forms and large-scale sited for trade and industry; also seven green belts were decided to prevent further uncontrolled expansion of the cities (ibid. p. 25). This containment of further expansion of the cities could not be achieved in the West of the Ruhr, where the built-up areas of Duisburg, Oberhausen and Mülheim are partly interconnected in many ways.

The aim of the 1950s reconstruction and planning of the cities was one of structure and dense concentration. Accordingly, plot sizes were increased and traffic areas enlarged. In the 1960s the cities were primarily adapted to cars: national roads, ring roads and motorways were planned such as the A-level roads A40, A42, A59. A new planning principle was also 'urbanity through density'. In all larger cities in the Ruhr area densely built new housing districts were erected, and as they were created outside of the city structure they furthered the trend of suburbanisation. At the same time the old city centres were dismantled, the inhabitants resettled in new suburban areas and the centres restructured for retailing, administration, banks and the like.

At the same time, there have been attempts to upgrade and integrate the abandoned mining sites into the region, e.g. the steel plant Phoenix West, which has been transformed into a recreational area with a large lake. The Krupp works that occupy a space of one by three km between the city centre of Essen and districts in the west have been an obstacle of urban development. Since the 1990 there have been attempts to

integrate them into the city structure, by means such as the conversion of factory buildings into musical theatres, car parks and furniture department stores and young and active housing area including shops and restaurants. The Essen city centre has therefore been gradually moved to the West (ibid. pp. 28–30).

Until the middle of the 20th century, the railway, trams and bicycles were the primary transport for goods and people. Yet, when cars became accessible to more and more people the transport was reorganised entailing that the highway system was expanded to become one of the densest one in Europe. Most of the cities underwent large rebuilding plans to favour the mobility with cars. This added network stratum of mobility made the region exceedingly interconnected. After the demise of the mining industry a large share of the industrial and commercial infrastructure was rebuilt to initiate economic diversity and attract new investors and industries. Other infrastructure was abandoned or turned into cycle tracks or green areas (ibid.).

3.4 SUMMARY

The NDY region and the Ruhr Valley can both be classified as PURs. Both regions developed a highly interrelated and dense network of railways, canals and streets that led to a highly fragmented and polycentric settlement structure. In both regions, there is no clearly dominant centre. In each region, the immediately neighbouring larger centres can be reached in under 45 minutes travel time and there is overall no coalescence of centres although the urban sprawl caused the centres to become spatially relatively large. Ample patches of brownfields, green sites and agricultural land separate the settlement clusters. The configuration of clusters in each of the regions does not follow any regular shape. Even though both PURs went through decades of economic decline caused by the demise of the coal industry in the second half of the 20th century, processes of economic diversification and recovery are progressing, although very slowly. Today there is a comparably high percentage of commute between the centres within each region. The economic and employment structure in both regions is at different stages: while the economy in the Ruhr is slowly diversifying and a specialisation of the centre evolves since the early 2000's, the economy of the NDY region with the exception of Leeds, still largely relies on low-skilled and precarious work in the manufacturing sector while visible and invisible unemployment are still the highest in the UK (Beatty et al. 2007; Beatty and Fothergill 1996).

The entirety of the settlement structure in both regions has most likely been determined by the position, quality, magnitude and accessibility of coal deposits causing the earliest settlement structures to establish in the West of the NDY region and the South of the Ruhr Valley. The development of the settlement and colliery structure in both regions

resembles each other. Whilst the clusters were relatively small and lay with little distance to one another before the first half of the 19th century, the mining strategies in the period from the 1850s to 1950s created a centre structure of fewer but much larger towns and cities that lay within a larger distance. The requirements of the coal and steel industries created a dense structure of seemingly unhierarchical transport networks that increased the polycentrality of both regions and might stimulate the economic recovery in the decades to come.

CHAPTER 4

DATA AND METHODOLOGY

4 CHAPTER

This chapter introduces the datasets and the methodology chosen to evaluate if regional spatial network metrics hold valuable information for the estimation of socio-economic variables. This study makes use of a regional population dataset provided by the European Statistical System (ESSnet), movement data, 3-dimensional building information as well as semantic information about buildings functions and the residential occupational status on a regional scope. Each of these datasets is of a fundamentally different geographic representation, making it necessary to develop appropriate tools to bridge the differences in precision, geometry and resolution. I present methods for each of these procedures and also provide methods to verify the results. In this chapter, I will also introduce methods of data disaggregation to bridge different levels of geographic resolution, from the statistical grid to the building- and street-level. The chapter will begin with an introduction of the data types used. Followed by this, I will introduce the methodologies chosen and elaborate on their respective limitations.

4.1 DATASETS

4.1.1 MOVEMENT DATA

Measuring movement on a regional extent is a non-trivial task as movement takes place on a variety of different scales and for multiple purposes. Several authors have argued that traffic choices are affected by socio-demographic factors, such as age and gender (Arentze et al. 2004), as well as different trip purposes, destinations and modes. The nature of trips that can be made in a region varies a lot. There are journeys made by heavy goods vehicles transporting products across the country, inter-city car journeys of home-to-work travels, inner-city journeys for shopping purposes, cycling trips to the closest park or a short walk to a neighbourhood located restaurant to name only some. If all of these attributes that define this large variety of trips are combined into a single dataset, we are facing the issue of multi-scale complexity if we want to make comparisons.

This becomes clearer with regard to a simple vehicular traffic count dataset on a regional scope. Generally, such datasets are collected by counting actual vehicular flows on a given number of sampled streets, by means of so-called manual or automated gate counts. If a vehicle passes through one of the gates, it is added to the total number of counts. In addition to this, different types of vehicles are classified. These classifications differ from country to country, which is an issue that needs to be taken

into consideration when doing cross-country comparisons. The classifications are mostly based on a combination of more specific vehicular differentiations (Table 5). The purpose of these detailed vehicular differentiations is to be able to infer different types of journeys. Final counting classifications divide vehicles into broader classifications such as pedestrian-based movement, light goods vehicles and heavy goods vehicles. Furthermore, these broader classifications are recorded to aid transport planning purposes, such as which roads and highways need structural or spatial reinforcement or where to locate traffic management control systems.

Table 5: Vehicular classification scheme for cross-country comparability. (Table based on (Kathmann et al. 2009 p. 13))

Class	2 Two-wheeled motor vehicles (e.g. motorcycles etc)	3 Cars, taxis, cars with trailer, vans up to 9 seats	4 Buses and coaches	5 Lorries,,, light goods vans ≤ 3.5 tonnes	6 Lorries without trailer, heavy goods vehicle ≥ 3.5 tonnes	7 Lorries with trailer, heavy goods vehicle ≥ 3.5 tonnes
Pictogram						

Vehicular counts are either collected and recorded separately for both movement directions of one road or they are collected as a single value. The resulting data can give a representative account of how movement is distributed within a state or region, as the gate counts are done on several weekdays as well as weekends at specific time periods. In a next step, the annual average daily traffic (AADT) is calculated. The AADT, sometimes referred to as annual average daily flow (AADF), is a well-established measurement in the field of transport engineering and gives the number of vehicles that drive on a specific stretch of road on an average day of the year (Department for Transport 2016).

The dataset, hence, combines journeys of different modes and purposes on multiple scales. This implies that direct correlations with network metrics such as betweenness centralities of specific radii become questionable as these metrics describe explicit distances rather than the necessary multi-scalarity. A recent study focusing on the relationship between network metrics and the estimation of movement flows showed that correlations differ significantly between different metric radii (Serra et al. 2015), which means that alternative methods are necessary to understand the phenomenon. AADT datasets are the only available source of reliable and robust information on human movement on a regional scope. This study will draw on the comparison of two large-scale datasets from Germany and the United Kingdom while elaborating on the issue of scalarity in the analysis.

GERMAN MOVEMENT DATA. For the German case, the dataset employed is provided by the Landesbetrieb Straßenbau Nordrhein-Westfalen (Straßen.NRW) (State Office for Road Management North-Rhine Westphalia). It comprises a regional road traffic count dataset that the Straßen.NRW provides in georeferenced form since 2015, covering the entire state of North-Rhine Westphalia, including 9086 individual gate counts. The street network of North-Rhine Westphalia is one of the most heavily used in Europe. The Straßen.NRW operates two different systems for traffic monitoring purposes. An automated counting system is updated monthly, next to a quinquennial manual gate count. The core difference between the two systems is the number of gate counts and the location of counting points. The automated counting system operates 339 individual counting points with the majority located on national highways (171) and federal highways (112), while the quinquennial count features a more comprehensive approach covering all street levels and uses up to 9000 individual counting points. Both methods are combined in the published dataset for 2015. Figure 16 gives an overview of the count position distribution with increased gate numbers in urban agglomerations and an even distribution of counts in rural areas. The selected data consists of 3779 count locations. Of these 3779 counts 3421 are manual counts, 110 automatic traffic counts and 19 are traffic count estimates.
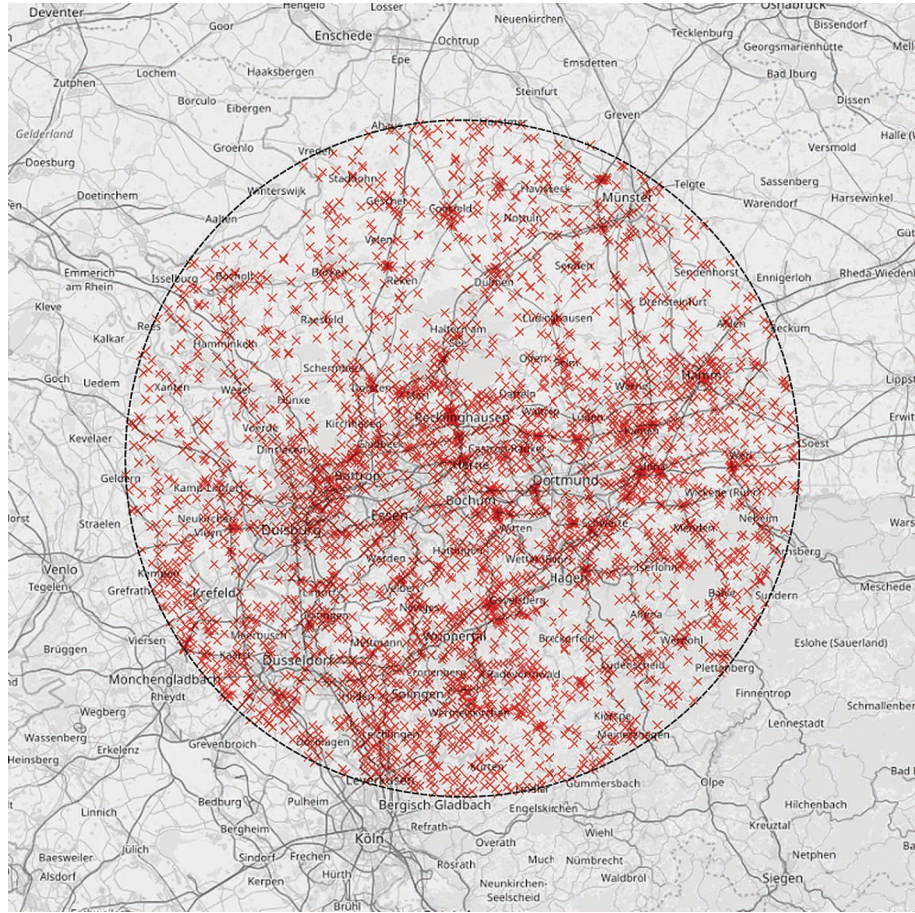
Figure 16: Location of all traffic count positions within the German case study area of a 120km diameter. Red crosses indicate gate locations. Background map © OpenStreetMap contributors.

The German gate count data is divided into four main classifications for AADT (Table 6). DTVKFZA accounts for all vehicles passing through the respective gates. DTVPVA gives insights into passenger traffic, as it combines all vehicles with passenger purposes, such as motorbikes, cars, vans or busses. DTVGVA is the number of all vehicles that are classified as freight traffic. This includes lorries and light goods vans below and above 3.5 tonnes, as well as heavy-weight lorries with trailers. The last class, DTVSVA, only contains those vehicles, which are above 3.5 tonnes and are therefore considered as heavy good vehicles.[5]

---

[5] For a comprehensive account on the traffic count methodology and the underlying mathematical models the reader is referred to the official methodology guide (Lensing 2013).

Table 6: AADT classification scheme for the German dataset.

Class	Primary Description (translation)	Vehicle Type (see Table 5)
ZSTNR	Gate count ID	
ZSTART	Count method (automatic, manual or estimate)	
DTVKFZA	Average annual daily traffic for all motor vehicles	2, 3, 4, 5, 6, 7
DTVPVA	Average annual daily traffic for all passenger traffic	2, 3, 4
DTVGVA	Average annual daily traffic for all freight traffic	5, 6, 7
DTVSVA	Average annual daily traffic for all heavy goods vehicles	4, 6, 7

BRITISH MOVEMENT DATA. The British dataset is provided by the British Department for Transport (DfT) and extracted from the countrywide AADT dataset for 2015. Comparable to the German traffic count system, the traffic in the United Kingdom is monitored by an automatic and manual system. While the automatic system uses 180 automatic traffic counters providing continuous data from the national network, the final annual road traffic counts are based on approximately ten thousand manual counts. The final count contains a series of traffic flow estimates next to manual and automatic traffic count data. The DfT traffic estimates are mainly calculated for minor roads and their calculations are based on growth rates extracted from the manual counts of the major road system and AADTs of previous years (Department for Transport 2016). The traffic census in the United Kingdom is, unlike that of Germany, carried out annually. The data is gathered on motorways and A roads (major roads), as well as B, C and unclassified roads (minor roads). For this study, both major and minor roads are combined into a single dataset. An overview of the count position distribution within the 120km selection diameter of the British case study area can be found in Figure 17. Similar to the German case, the number of gates in urban agglomerations increases and is rather evenly distributed in rural areas. The selected data consists of 2773 count locations. Of these 2773 counts 815 are manual counts, 51 derived from a neighbouring counted gate, 138 estimated through a neighbouring link, while the majority of counts (1769 data points) is estimated through AADF values from previous years. The British dataset is therefore respectively smaller in terms of the empirically gathered data.
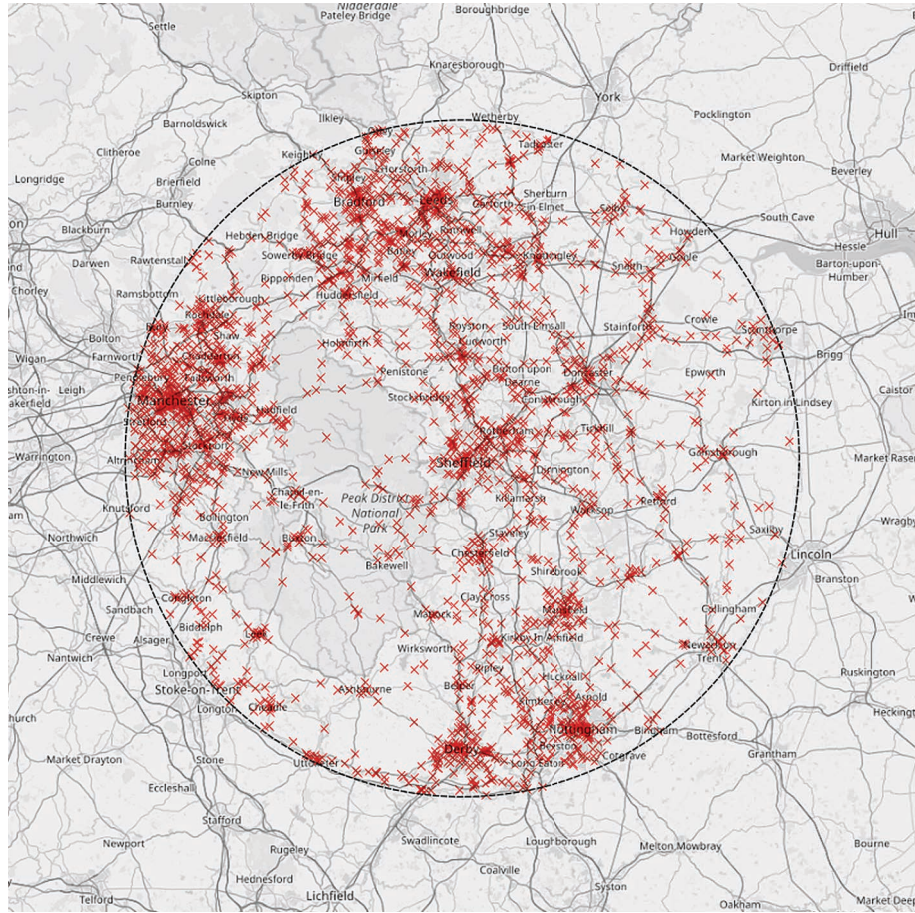
Figure 17: Location of all traffic count positions within the British case study area of a 120km diameter. Red crosses indicate gate locations. Background map © OpenStreetMap contributors.

The British dataset is divided into 11 different classes (Table 7). These classes are much in alignment with the differentiation classification of Table 5. In fact, the British classification scheme features a higher level of detail and provides a variable for every of the vehicle types (2 - 7), as well as wider subcategories for heavy-weight vehicles. Classes Fd2WMV, FdCar, FdBus and FdLGV, each describes an individual vehicle class, while FdHGVR2 to FdHGVR6 differentiates heavyweight vehicles based on their number of axles. FdAll_MV is comparable with the German DTVKFZA and describes the average annual daily traffic for all vehicles for each gate. This detailed classification allows the creation of a set of new variables to match the more simplified German dataset.

Table 7: AADT classification scheme for the British dataset.

Class	Primary Description (translation)	Vehicle Type (see Table 5)
CP	Gate count ID	
Estimation	Count method (automatic, manual or estimate)	
Fd2WMV	Average annual daily traffic for all two-wheeled motor vehicles	2
FdCar	Average annual daily traffic for all cars and taxis	3
FdBus	Average annual daily traffic for all busses	4
FdLGV	Average annual daily traffic for all light goods vehicles	5
FdHGVR2	Average annual daily traffic for all two-articulated axle heavy goods vehicles	6
FdHGVR3	Average annual daily traffic for all three-articulated axle heavy goods vehicles	6
FdHGVR4	Average annual daily traffic for all four-articulated axle heavy goods vehicles	7
FdHGVR5	Average annual daily traffic for all five-articulated axle heavy goods vehicles	7
FdHGVR6	Average annual daily traffic for all six-articulated axle heavy goods vehicles	7
FdHGV	Average annual daily traffic for all heavy goods vehicles	6, 7
FdAll_MV	Average annual daily traffic for all motor vehicles	2, 3, 4, 5, 6, 7

Four new variables are derived from the classifications scheme (Table 8). Of these four variables, particularly AADF_PA holds valuable information for the estimation of human movement. The average annual daily traffic for all passenger vehicles accounts for every trip taken by individuals and groups of people making use of the same mode. The variable is adjusted for goods and industrial transportations, as these are traditionally operated through heavy-weight vehicles. In this regard, AADF_FR accounts for any movement generated through industrial activity. AADF_All is the totalled value of all measured vehicular movements within the region.

Table 8: AADT variable table for cross-country comparison. Including different variable combinations of the British dataset to guarantee comparability with the German data.

Class	Primary Description (translation)	GE Variable	UK Variable
G_ID	Gate count ID	_	_
COUNT	Count method (automatic, manual or estimate)	_	_
AADF_All	Average annual daily traffic for all motor vehicles	DTVKFZA	FdAll_MV
AADF_PA	Average annual daily traffic for all passenger traffic	DTVPVA	Fd2WMV + FdCar + FdBus
AADF_FR	Average annual daily traffic for all freight traffic	DTVGVA	FdLGV + FdHGV
AADF_HV	Average annual daily traffic for all heavy goods vehicles	DTVSVA	FdBus + FdHGV

4.1.2 POPULATION DATA

Population data, or precisely the total number of people resident at a spatial location, is available in a large variety of different often not comparable geographic boundaries. This difference in size, form and distribution poses a challenge for the statistical analysis, because datasets are either not comparable, or methods of data transformation can negatively influence the statistical procedures of the analysis. This challenge is particularly grave where statistical boundaries do not match the geographical level of observations, or where the small-scale application is of interest (e.g. building- or street-level analysis). In this study, both challenges are present and demand a solution to methodologically bridge the gap between the different geographical scales and physical representations, e.g. from the administrative boundary to the street network, to enable a statistical comparison.

The population dataset used in this study is provided by the GEOSTAT project, which is an initiative by the ESSnet in collaboration with the European Forum for Geography and Statistics (EFGS). The GEOSTAT is an ongoing project aiming at providing comparable population data for the entire surface area of all member states of the European Union. Specifically, this research project is based on the GEOSTAT 2006 and 2011 population grid. The GEOSTAT grid is a unique dataset allowing cross-country comparisons of population and housing Census within the European Union. The dataset is based on a 1km by 1km grid following the infrastructure for spatial information in Europe (INSPIRE) framework. Grid-based datasets feature a series of advantages over conventional administrative boundary-based datasets, as they are comparable in size, stable over time and can be used to construct hierarchical models by combining smaller

grid cells to their next larger squared aggregation (4, 9, 16, 25, etc.) (Bloch Holst 2011 p. 49). The GEOSTAT dataset consists of Census information, which is joined using the GIS 'reaggregation' methods, to produce an evenly distributed statistical grid of 1km. 'Reaggregation' is a collection of statistical and spatial GIS methodologies used to increase or decrease spatial resolution between different datasets and can further be divided into the subcategories of aggregation, disaggregation and methods of mixed uses of the former. The GEOSTAT dataset makes use of all three of these approaches, depending on the respective country.

AGGREGATION refers to a decrease in spatial resolution. As such aggregation is used in the GEOSTAT project to count/aggregate the value of geo-referenced Census microdata (such as address points, parcel locations, building and building part information or Census areas), whose location falls within each respective grid cell (ibid. p. 60). The aggregated data values are then joined onto the respective INSPIRE grid (Figure 18 and Figure 19) and, thus, form the base for the final GEOSTAT population grid dataset. The aggregation methodology is used in those countries, which provide geo-referenced Census microdata, as it is the case in Germany and for urbanised parts of the United Kingdom. Aggregated population databased on micro-scale datasets provides the highest quality possible, while avoiding issues of privacy disclosure. In European countries and in areas in the UK where such micro-scale datasets are not available, alternative methods of data disaggregation are employed.
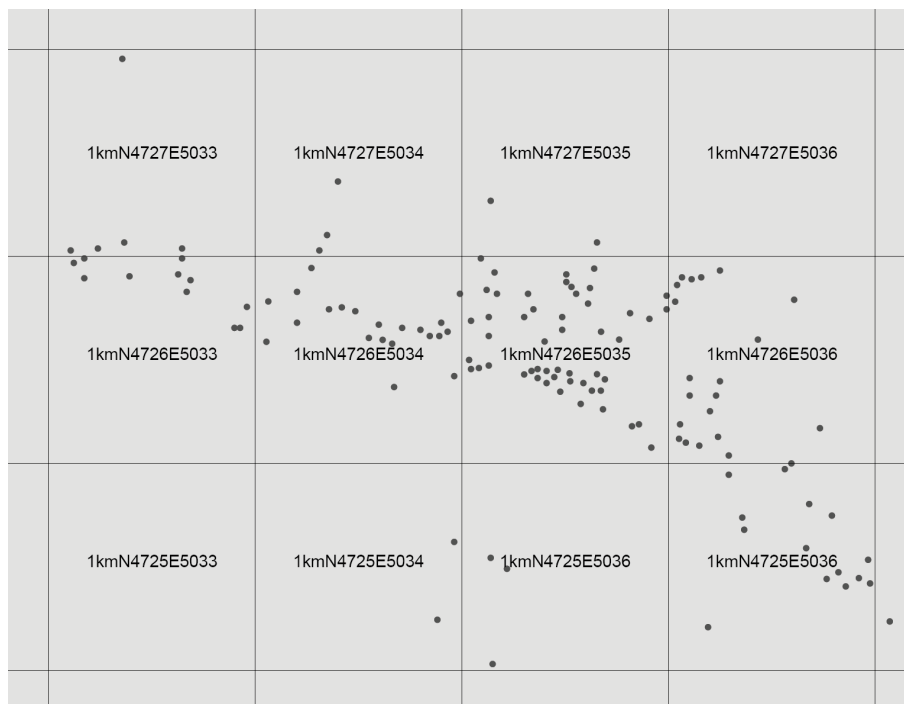
Figure 18: Location pattern of the source micro-dataset and ETRS89-LAEA projected population grid (figure based on (ibid. p. 76)).
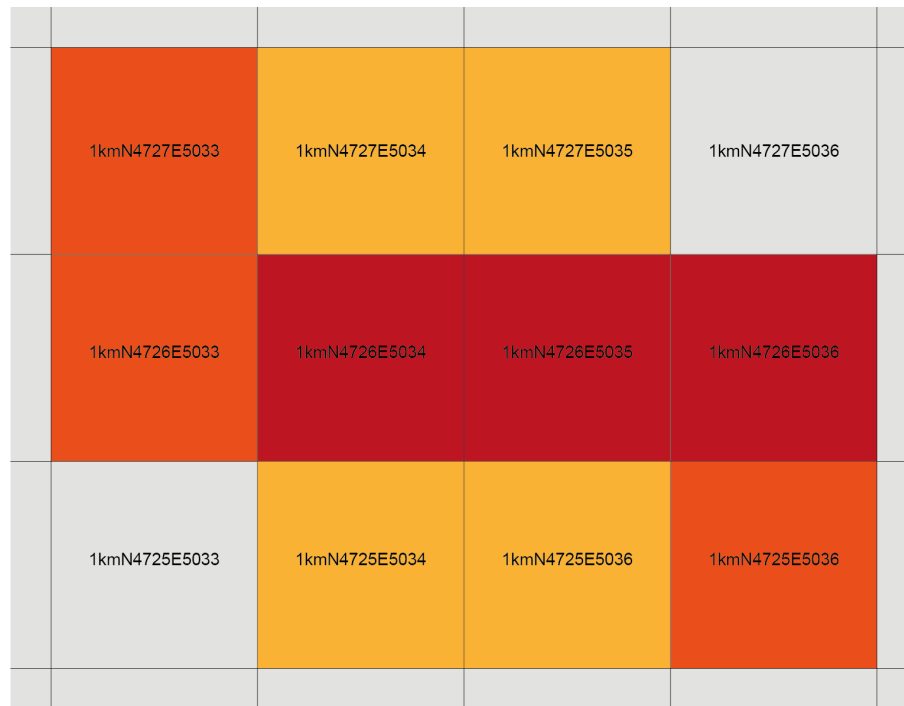
Figure 19: Grid cells with aggregated point information. Red indicating higher total population (figure based on (ibid. p. 76)).

DISAGGREGATION opposed to aggregation is a method in which low-resolution data is distributed to geographic entities of higher resolution. This is, for example, necessary when population data is only available on a city-administrative level, yet population estimates are needed on a neighbourhood level. The GEOSTAT project makes use of disaggregation for the GEOSTAT 2006, as well as for selected non-urbanised areas in the UK of the GEOSTAT 2011 where Census microdata is not available.

Due to its cross-country comparability, the GEOSTAT population grid is an appropriate source for researchers interested in different population densities across Europe and is of great value for endeavours such as in this study. The uniform grid can further be used as a form of bias reduction in urban data sampling because it does not deal with size differences between urban to rural administrative boundaries. Moreover, the uniform grid enables the creation of hierarchical modelling, which allows the testing of the importance of scalar effects in urban sampling. Figure 20 and Figure 21 present a visualisation of the dataset previously mentioned for both models at the extent of 230 km. Comparisons with alternative data of the built environment will be based on those grids, which fully intersect with all datasets employed. Both maps are produced with the same colour breakdown, enabling a direct comparison. The model regions differ in their average density. While the German case features fewer grid cells of a low density, or no population at all, the British case features substantially larger and more cells of low density, while areas of high urbanisation appear more clustered and larger in their extent.
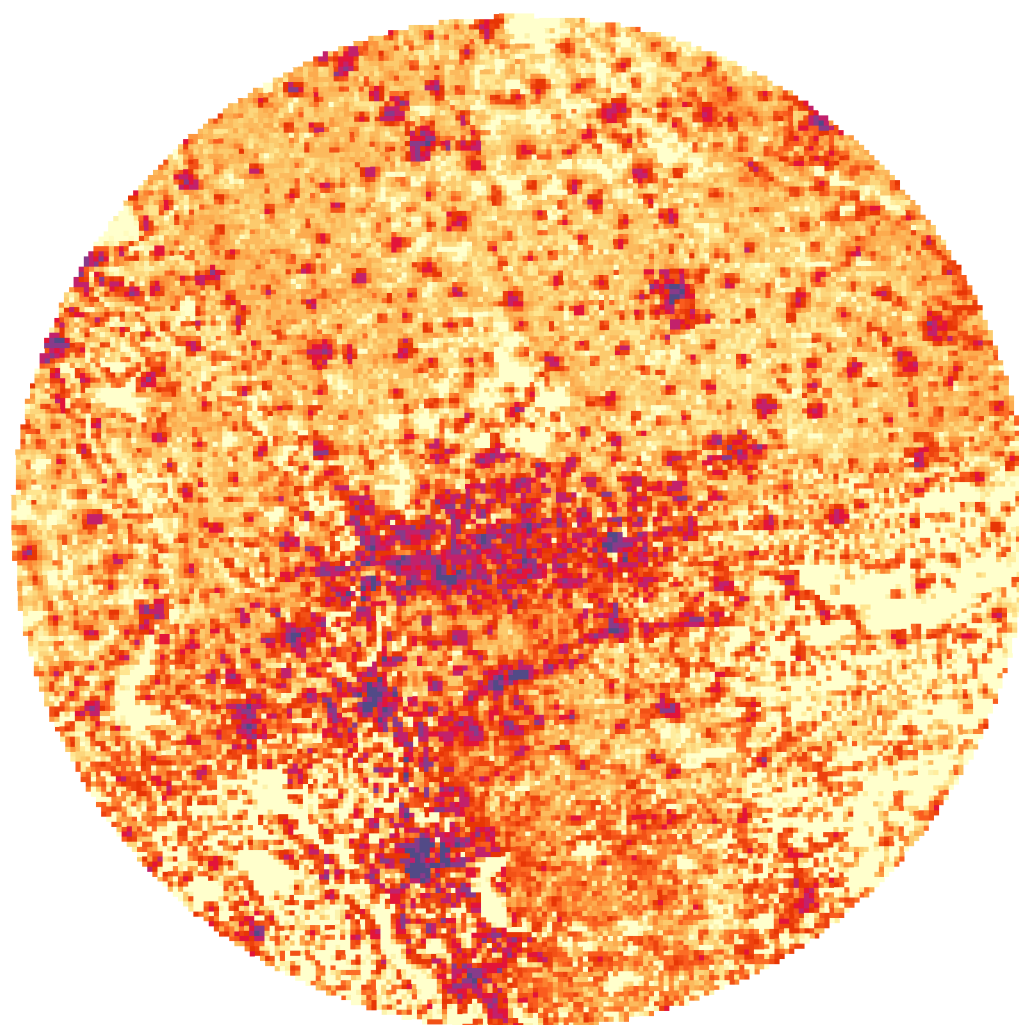
GE: Total Population in 2011

min 0   max 18401

Figure 20: Visualisation of the GEOSTAT 2011 dataset, total population per 1 x 1 km square grid for the German model.

UK: Total Population in 2011

min 0   max 14754
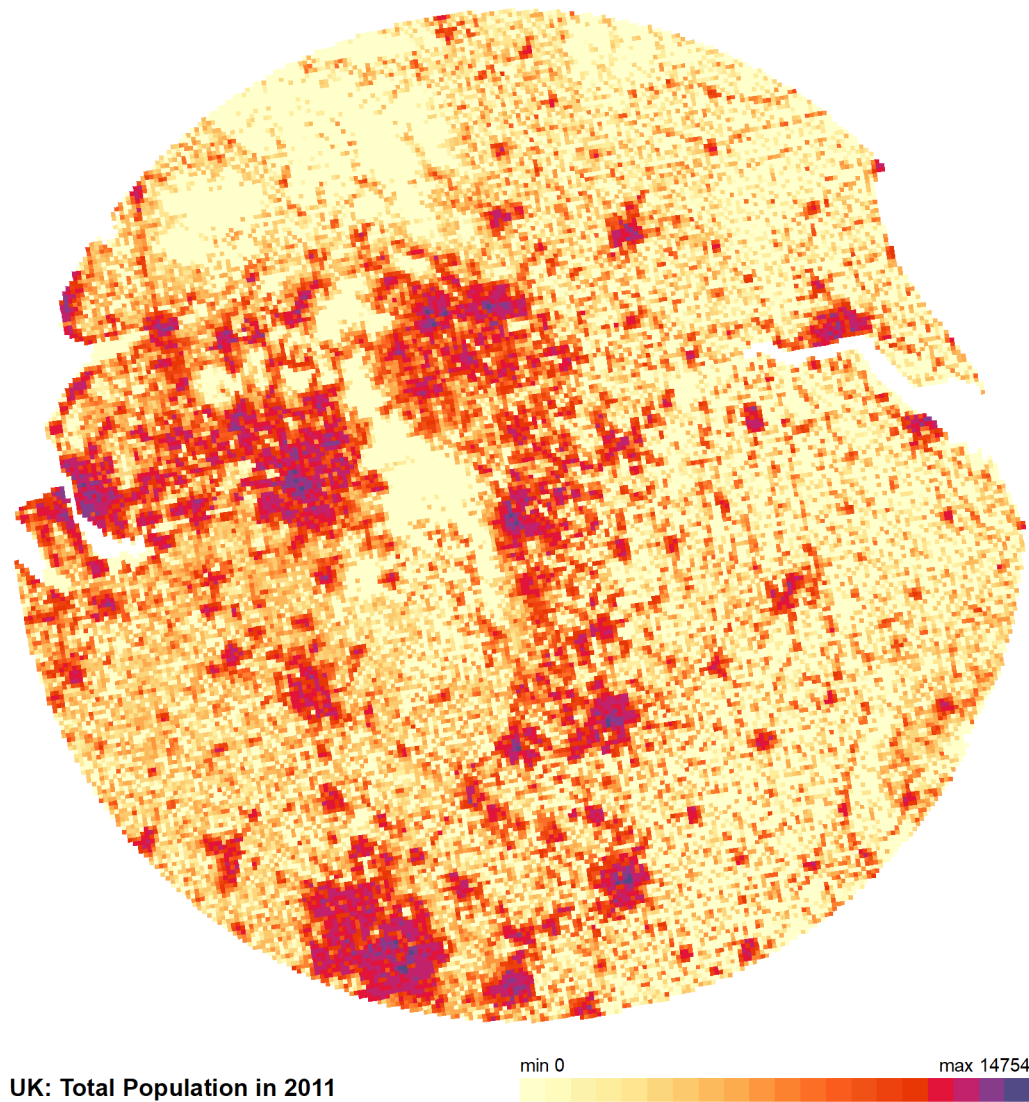
Figure 21: Visualisation of the GEOSTAT 2011 dataset, total population per 1 x 1 km square grid for the British model.

4.1.3 3D-BUILDING INFORMATION DATA

This study makes use of a large 3-dimensional (3D) building information model (CityGML) for both study areas. CityGML is an open source data model based on the XML format for the comparability and exchange of 3D city models (Gröger et al. 2012 p. 9). One of the core obstacles in cross-country 3D model comparisons are differences in the level of detail. The CityGML was specifically developed to 'reach a common definition of the basic entities, attributes, and relations' of 3D models (ibid. p. 9). CityGML 3D city models are available in different, hierarchically structured components, varying in scale and precision. In general, such datasets are divided into 5 different levels of detail (LoD) providing building information at different spatial resolutions ranging from 2-dimensional footprint representations (LoD0) to fully

113

detailed 3-dimensional models including information about the interior of buildings (LoD4) (Figure 22) (ibid. p. 11).
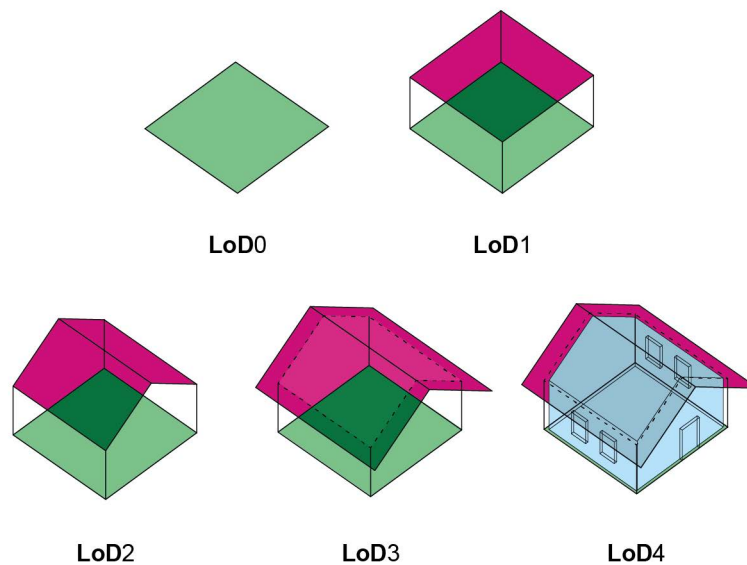
Figure 22: Visualisation of the CityGML level of detail classification LoD0-4.

The LoD concept provides specific guidance on the level of detail and their potential scale of application (Table 9). LoD1 models are particularly suitable for regional and city-scale applications, as the model consists of extruded building footprints and does not lead to excessive file sizes. Within LoD1 datasets, each building is represented by a flat roof structure. Such simple building block representations, as in the LoD1, do not incorporate other roof types. Buildings with complex geometries and multi-level heights are simplified to plain block representations and the modelled geometry can potentially be significantly different to the real-world geometry. Such complex buildings are generally found in areas of high building densities. Following the recommendation of the Open Geospatial Consortium (ibid. p. 12) I make use of LoD1 models for the following regional analysis. LoD1 models provide sufficient information on building volume and shape on a regional scale.

It should, however, be noted that the consistency of the LoD concept has been challenged recently, due to the fact that the five LoD classifications (LoD0 - 4) do not provide sufficient possibilities for differentiation. LoD levels can feature differences within each LoD level, which has raised concerns pertaining to comparability (Biljecki et al. 2014 p. 9; Biljecki, Ledoux, et al. 2016). In this context, Biljecki, Ledoux and Stoter (2016) propose 11 additional LoD subcategories and arrive at a stricter specification, thus, improving consistency between datasets of similar LoDs. The problem of consistency has a stronger impact when the level of detail increases (such as it is the case from LoD2 to 4). For this study, these difficulties of consistency can be neglected, as the

methodologies for the creation of the German as well as the British datasets are similar in their construction. The specific methodology of both model constructions is explained in the following sections.

Table 9: CityGML level of detail classification LOD0-4.

Class	Primary Description (translation)	Model scale description
LoD 0	2-dimensional building footprints	Region
LoD 1	3D-building blocks (footprint extrusion)	City, region
LoD 2	3D-building blocks with classified roof structure	City, city districts, projects
LoD 3	Detailed 3D-building model including façade elements	City districts, architectural models (exterior), landmarks
LoD 4	Detailed 3D-building modelling including interior information	Architectural models (interior), landmark

Both countries base their LoD1 model on cadastre-based building footprints and use building height information generated through remote sensing technologies. In general, such a process consists of the extraction of building height information from a 3Dlaser scan of regional surface areas. By a combination of digital terrain information, several building heights reference points are extracted (Figure 23). These extracted points are: the minimum building height, i.e. the lowest point of the building in reference to a digital terrain model (AbsHMin), the maximum building height, i.e. the highest measured point of the roof structure (AbsHMax) as well as a relative building height (AbsH2). AbsH2 is a result of the lowest reference point of the roof structure or fascia and the relative distance to AbsHMin. Any information regarding the geometric building boundary is extracted from the building footprint or cadastre information. The actual building height (RelH2) can be used in isolation of additional digital terrain information and provides specific building height data forming the base of the 3D city model.
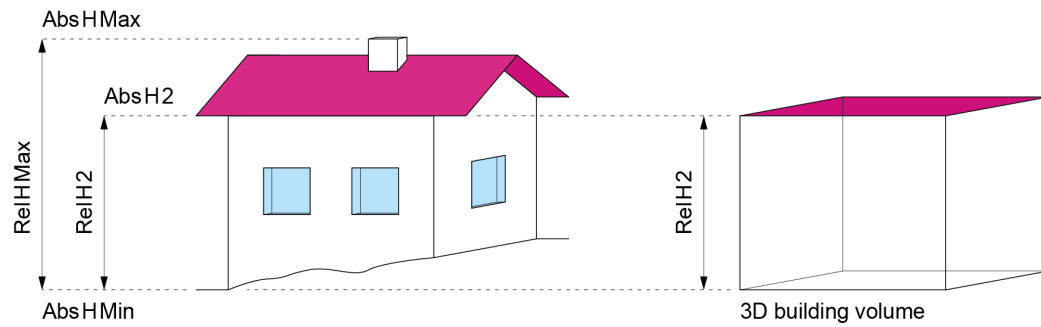
Figure 23: Illustration depicts different height information within the 3D city model, as well as the resulting building volume. 'Abs' refers to the absolute distance, whereas 'Rel' refers to the relative height.

A visual representation of such data can be found in Figure 24. Variations in height are visible not only for different types, but also within similar building types.

0.0   39.7 m

Figure 24: Visualisation of LoD1 data for a randomly selected area in the UK. Building heights ranging from 0 to 39.7 metres. © Ordnance Survey.

GERMAN BUILDING INFORMATION. The German LoD1 dataset is provided by the Bezirksregierung Köln (District Council Cologne) and covers the area of the entire North-Rhine Westphalian state. The building information consists of more than 10 million individual buildings and is the largest administrated dataset by a federal state in this field. The dataset has been released in 2017 and is the only German, publicly available dataset of this kind. Automated remote sensing methods are employed to generate building heights, due to the size and extent of the state. The model is constructed by a combination of governmental cadastre information as the source for building footprints, digital surface and terrain models and airborne laser scanning methods for building height measurements. Additionally, aerial imagery is employed as

a complementary method, where the combination of cadastre and laser scanning produced insufficient results. This is, for example, the case for buildings with a specific importance for the greater landscape such as churches, museums and castles. These buildings are successively updated through methods of stereoscopy. Buildings are interactively improved with the use of areal imagery of different spatial angles. Moreover, the same method is applied where cadastre information is inconsistent, or buildings were not erected in alignment with the local building codes.
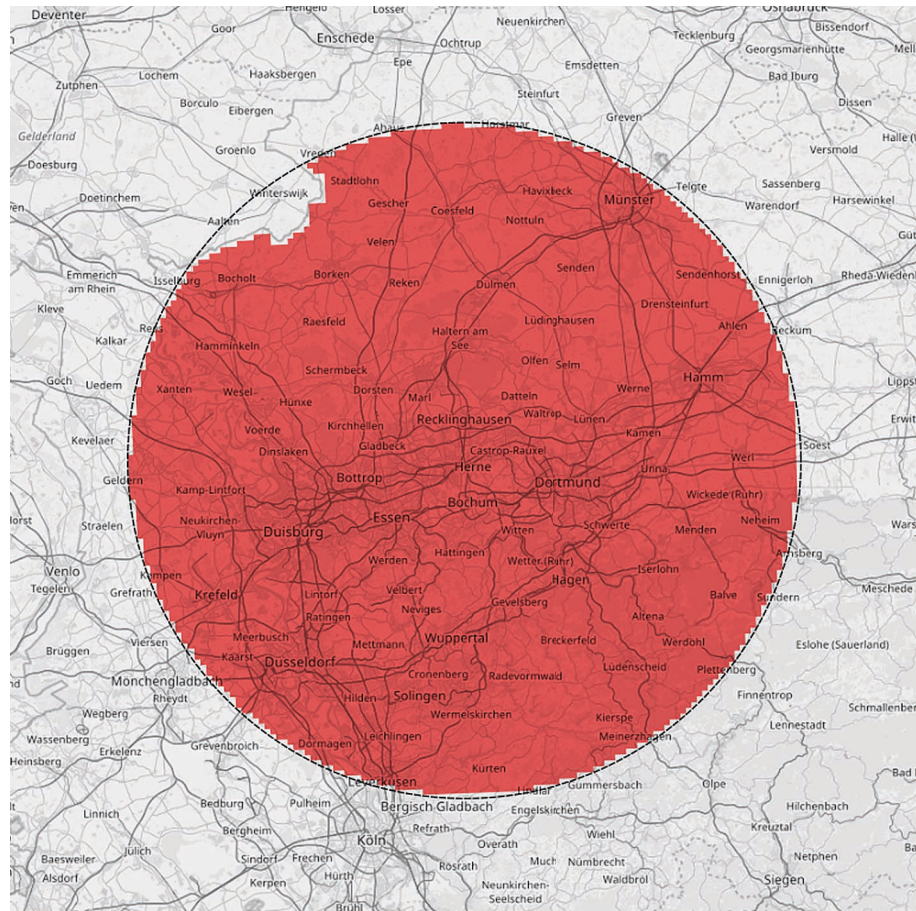
Figure 25: Data coverage for the German 3D-building information within a 120km diameter model radius (dashed line), red indicates areas fully covered by 3D information. Background map © OpenStreetMap contributors.

The German dataset is compared with the spatial location of the available GEOSTAT population data grid and the 120km circular cutout of the street network model. The final selection is based on all population grids that are fully contained by the 120km street network boundary and fully covered by the 3D-building information. This step of pre-selection is necessary in order to remove all buildings that are not fully covered by either of the two other datasets (GEOSTAT and street network), as well as to remove all GEOSTAT grid units that are not fully covered by 3D-building information. Figure 25, shows the final surface area for the German model that is covered by the final selection

of 3D-building information. The final data consists of 10,928,000.00 km² covered surface area and 5,436,915 individual buildings.

BRITISH BUILDING INFORMATION. The British LoD1 dataset is provided by the Ordnance Survey and includes 3D-building information for large parts of the United Kingdom. The data coverage focuses mostly on urbanised areas and highly populated agglomerations, with an underrepresentation of rural and agricultural landscapes. Comprising of 23 million buildings, the UK dataset is an alpha release and cannot be considered as complete. Moreover, the Ordnance Survey does not indicate further quality enhancements, which are comparable to landmark buildings in the German dataset. Still, most parts of the case study area are fully covered. Similar to the German case, the 3D-building information is generated through a fully automated process, drawing on remotely sensed data, including aerial imagery and digital surface models. The Ordnance Survey provides no further level of detail beyond the LoD1 level. The footprint information is based on the OS MasterMap Topography Layer, which is a highly accurate large-scale digital database of detailed surface features and of comparable quality to the German governmental cadastre information.

Finally, the British dataset resembles the German dataset, compared to the spatial location of the GEOSTAT population data grid and the 120km circular cutout of the street network model. The final data consist of 5,148,000.00 km² of the covered surface area and a total of 4,312,347 buildings. Figure 26, shows how the coverage is distributed within the model area. All major urban agglomerations are fully covered, while large parts of the rural and non-urbanised areas are excluded.
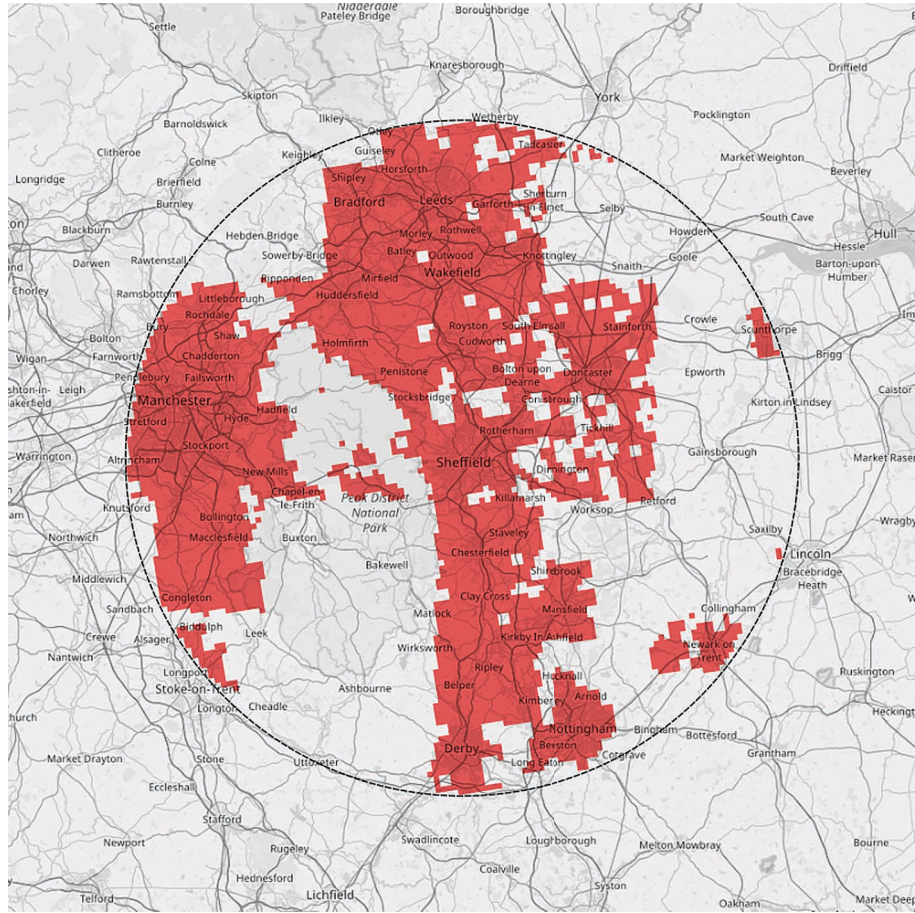
Figure 26. Data coverage for the British 3D-building information within a 120km diameter model radius (dashed line), red indicates areas fully covered by 3D information. Background map © OpenStreetMap contributors.

4.1.4 SEMANTIC BUILDING INFORMATION

Semantic information plays a pivotal role for detail and quality enrichment for urban and regional modelling. This is particularly the case for small-scale applications, where traditional governmental data does not provide the necessary level of detail, sufficient information or where time periods between dataset updates are too lengthy. Several authors have investigated potential sources for such successful building classifications. Approaches can be divided into two main categories, a) the estimation of the building function via building morphology and b) the estimation of the building function through complementary data. Authors working in the former category make several propositions for building function estimation, such as identifying the building form and inferring their residential or industrial function through an analysis of volunteered geographic information (Fan et al. 2014). Fan, Zipf and Fu reach an accuracy of more than 90% for the detection of residential buildings. Other authors employ LIDAR and multispectral imagery to identify differences between building geometries and can detect urban residential buildings (Awrangjeb et al. 2010). Morphological investigations with the purpose of building classifications have been applied to 3D-

building information. Henn et al. (2012) propose a machine learning approach to classify 3D-building information. Their core methodology is the identification of building types (e.g. terraced, detached or apartment buildings etc.). While these approaches are meaningful and promising, their accuracy level remains insufficient for reliable population estimation. This lack of accuracy is particularly grave for areas of high urban density. For the purpose of this study, a more differentiated classification is needed where classifications are providing higher levels of detail beyond simplified group descriptions such as 'residential' or 'industrial'.

In the second category, authors employ complementary data and make substantial use of point data information to estimate the building function. Such point data can either be based on governmental or proprietary address databases or is a part of volunteered geographic information. Orford and Radcliff (2007) are the first to propose employing point-based address data for the classification of 2D residential building footprints with the aim of enhancing the quality of census data. Smith et al. (2010) have advanced this approach and introduce a residential building classification of 2D building footprints. In recent years, only very few authors have explored the potential of semantic enrichment in the context of 3D-building information. In this context, Kunz and Hecht (2015) demonstrate how OSM data can be used to refine governmental building classification. Their approach makes use of small-scale non-residential point information to estimate floor area usage of buildings that are pre-classified as residential in governmental cadastre datasets. Their approach follows a two-step process, where first the semantic information is matched to the building geometry. The matching is based on the spatial relation of the OSM points of interest datasets and 3D-building information. In a second step, the average floor area of the respective function is estimated based on the type of usage. This allows an estimation of the floor area usage per building and function. In the context of regional semantic enrichment of 3D-building information, none of the proposed approaches can produce satisfying results.

While there are no address databases available for the German region, the regional government provides detailed cadastre information including building classifications for every building since the recent launch of an open data initiative. For the British case, such cadastre information is not available. Instead the British Ordnance Survey provides a high-quality address database that was updated recently. Voluntary geographic information in the form of OSM points of interest data is available for both cases. In this context, I will draw on three specific datasets for the semantic enrichment of 3D-building information. For the German case, sufficient building classification is available from the Amtliches Liegenschaftskataster-informationssystem (ALKIS) (Authoritative Real Estate Cadastre Information System), which can be linked directly to individual 3D-building elements in the LoD1 3D model by a specific ID layer. For the

British case study, however, no such governmental classification scheme is available for a large-scale application. As an alternative, the Ordnance Survey AddressBase will be used to generate comparable building classifications. Finally, I will employ OSM points of interest data to enrich both dataset with small-scale non-residential information.

In relation to the aims previously set out to explore the potential of network metrics for the estimation of socio-economic variables, I have selected specific semantic building information that indicate residential usage, as well as building functions, which hold information on economic activity. The following section presents the differences of these two datasets, the methods of selection and the final selection variables.

GERMAN SEMANTIC BUILDING INFORMATION. The ALKIS is the official real estate cadastre information system developed by the Arbeitsgemeinschaft der Vermessungs-verwaltungen der Länder der Bundesrepublik Deutschland (AdV) (Working Committee of the Surveying Authorities of the Laender of the Federal Republic of Germany). ALKIS is a combination of the real estate register and the real estate field mapping and, by now, the nationwide standard for the management of official geospatial reference data that takes into account the international standards of ISO / TC 211. The dataset contains entries for each individual building describing its function in a land-use classification system. While, the building function/type of land-use is defined by the predominant functional significance of the building at the time of the survey (also referred to as 'principle of dominance'), land-use classes are only assigned to entire buildings, rather than to building parts or storeys. This implies that buildings of mixed-use are assigned the class of their predominant function. A combined class is applied in cases where such a predominant function cannot be evaluated precisely (such as 'mixed-used building with housing'). Building functions are divided into 501 different land-use classes, including a substantial differentiation of building functions for housing, business, industry or agricultural and many others (AdV 2015 pp. 211–33).

GERMAN RESIDENTIAL LAND-USES. Table 10 shows a comprehensive selection of all residential building classes of the ALKIS classification that indicate buildings with the residential function of permanent residences. The selection comprises of 23 classes including solely residential functions, such as 'residential buildings', 'residential houses', 'residential accommodation', as well as mixed-use buildings were residential and alternative functions coexist such as 'residential and office buildings', 'residential building with retail and commercial usage' or 'building for trade and services with housing'. The selection contains 2,618,827 classified buildings of residential land-use, the majority of which are 'residential houses' with 2,360,516 buildings.

Table 10: AdV ALKIS/ALK: residential land-use classification and number of building units.

Class	Primary Description (translation)	Units
31001_1000	Residential Building	64,806
31001_1010	Residential House	2,360,516
31001_1020	Residential Accommodation	2,061
31001_1021	Children's Home	449
31001_1022	Home for the Elderly	7,140
31001_1023	Nurse Home	216
31001_1024	Student Accommodation	572
31001_1100	Mixed-used Building with Housing	45,791
31001_1110	Residential Building with Public Facilities	16,131
31001_1120	Residential Building with Retail and Commercial Usage	43,557
31001_1121	Residential and Administrative Buildings	48
31001_1122	Residential and Office Buildings	682
31001_1123	Residential and Retail Buildings	2,621
31001_1130	Residential building with Commercial and Industrial Usage	36,884
31001_1131	Residential and Operational Buildings	198
31001_1210	Land and Forestry Residential Building	18,880
31001_1220	Land and Forestry Residential and Operational Buildings	10,721
31001_1221	Cottage	5,296
31001_1222	Residential and Agricultural Buildings	353
31001_1223	Forester's Lodge	83
31001_2310	Building for Trade and Service with Housing	787
31001_2320	Buildings for Business and Industry with Housing	261
31001_3100	Building for Public Use with Housing	774

GERMAN COMMERCIAL LAND-USES. Following the same selection approach as presented in the previous section, all buildings with a commercial function have been selected from the ALKIS classification scheme. This selection has a specific focus on trade and service land-uses (including functions of mixed-use). All of these are end-user oriented retail and service functions, such as markets, kiosks, shops as well as pharmacies, post offices and restaurants. All of these functions rely on face-to-face relations and are included in the central place spatial demand model. There are 14 classes that fulfil this criterion and the selection comprises of 110,778 buildings in total. The largest class are the mixed-use buildings that combine residential with commercial, retail or service

functions ('31001_1100' and '31001_1120') with 45,791 and 43,557 buildings, followed by buildings with sole retail and commercial usage '31001_2050' with 5,311 entries.

Table 11: AdV ALKIS/ALK: trade and service land-use classification and number of building units.

Class	Primary Description (translation)	Units
31001_1100	Mixed-used Building with Housing	45,791
31001_1120	Residential Building with Retail and Commercial Usage	43,557
31001_1123	Residential and Retail Buildings	2,621
31001_2050	Building with Retail and Commercial Usage	5,311
31001_2051	Department Store	1,129
31001_2052	Shopping Centre	551
31001_2053	Market Hall	131
31001_2054	Shop	4,344
31001_2055	Kiosk	871
31001_2056	Pharmacy	17
31001_2080	Building for Service and Hospitality	2,026
31001_2081	Restaurant	3,263
31001_2310	Building for Trade and Service with Housing	787
31001_3013	Post	379

In the next step, the commercial land-use selection is enhanced by the use of OSM POI data. All POIs indicating service and trade land-uses are selected (see Figure 25 for a full list). There are 16,837 additional buildings, which can be classified as land-use of service or trade by this method. Most of the buildings (2,862) that have previously not been classified with their respective land-use function are buildings with 'restaurants' usage. This is followed by buildings with a 'fast-food' (1,396) and a 'bakery' (1,045) function. The employment of volunteered geographic information reveals that particularly food service functions are not sufficiently covered by the German cadastre system. This is especially the case in smaller towns, where service and trade functions make up only small parts of the overall building geometry and the building classification follows the principle of dominance. In such cases, buildings are often classified as 'residential', as this is the predominant function, so that the database lacks crucial information about retail and service functions.

Table 12: OSM points of interest classes of service and trade for the German model.

Primary Description	Buildings	Primary Description	Buildings
bakery	1045	hairdresser	1152
bank	859	jeweller	163
bar	173	kiosk	940
beauty_shop	264	laundry	169
beverages	244	mall	7
bicycle_shop	218	mobile_phone_shop	118
biergarten	24	newsagent	46
bookshop	168	nightclub	84
butcher	217	optician	219
cafe	976	outdoor_shop	19
chemist	143	pharmacy	972
clothes	703	post_office	269
computer_shop	91	pub	920
convenience	293	restaurant	2862
department_store	101	shoe_shop	208
doityourself	163	sports_shop	67
fast_food	1396	stationery	84
florist	376	supermarket	778
gift_shop	3	toy_shop	49
greengrocer	58	travel_agent	171
		video_shop	25

Figure 27 shows this misclassification at the example of the city centre of Bochum and Witten. In Bochum's centre, which is an area of high urban density, almost all buildings are correctly classified with regard to service and trade land-uses with only a few exceptions that are misclassified. Compared to the city centre of Bochum, the city centre of Witten does not have any service and trade classifications in the ALKIS cadastre information. The superposition of OSM POI data, however, highlights that this is a significant underrepresentation of the real functional usage.
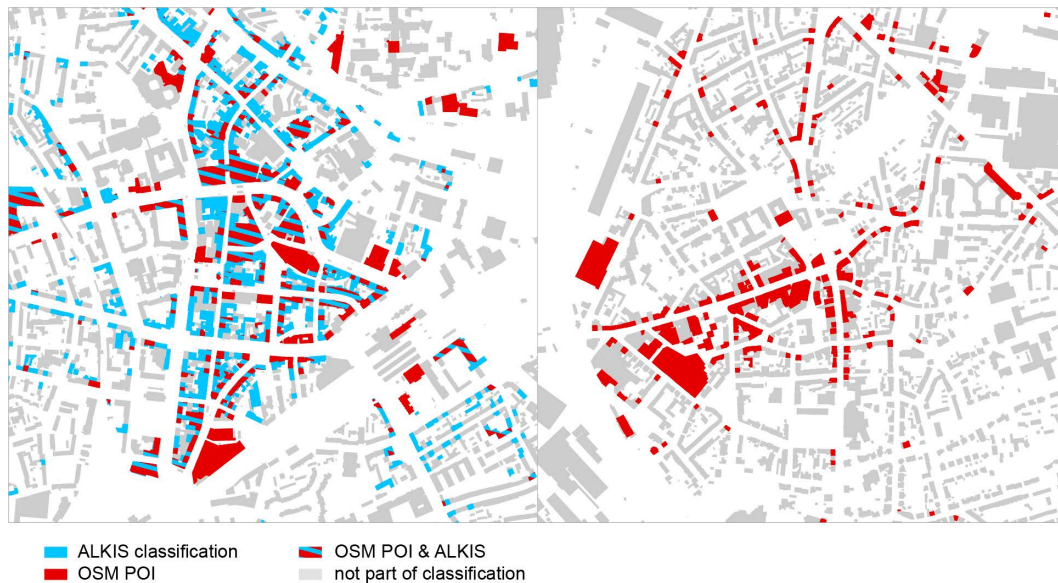
Figure 27: Superposition of the ALKIS cadastre information and the OSM points of interest building classifications. Buildings hatched in red and blue are covered by both datasets. On the left, the city centre of Bochum, an area of high urbanisation. On the right, the city centre of Witten.

The combined service and trade data of the German cadastre and OSM POI, leads to a final classification of 127,684 buildings for the German model.

BRITISH SEMANTIC BUILDING INFORMATION. The British semantic building information is generated by using the AddressBase Premium dataset provided by the Ordnance Survey. Unlike in Germany, the British government provides no official real estate cadastre information. For this reason, the OS AddressBase Premium (AB) is employed as a way to create semantic building information. The AB is a database that contains more than 34 million postal addresses and classifies them into different land-uses. This dataset includes the most accurate list of georeferenced postal addresses in Great Britain. Each property is referenced with a unique 12-digit property reference number (UPRN) and represented by a georeferenced point. These georeferenced points allow us to link the dataset through a spatial join with figure-ground polygons of the British 3D-model. There are a few difficulties that arise, due to the nature of the data, in the form of multiple address points at the same location. The AB data model has more data points than existing building polygons in the 3D-model, a phenomenon, which is also referred to as 'one-to-many'. This means a building polygon can have hundreds of different address points of similar or different classes at the same spatial location – a problem, which demands a methodological approach to overcome this issue.

The AB Premium dataset provides specific address classifications. These classifications are comparable with the German official real estate cadastre information. The classification scheme holds 563 different classifications, ranging from residential, commercial to industrial functions (Ordnance Survey 2013) and provides detailed

tertiary descriptions, next to the primary and secondary descriptions on the class function. An address point with the combinatory class 'RD03', for example, describes an address of semi-detached residential buildings. RD03 is of the primary class 'residential', secondary class 'dwelling' and of tertiary class 'semi-detached'. This level of detail enables the selection of those classes that are similar to the German cadastre information and facilitates the construction of a comparable classification scheme.

BRITISH RESIDENTIAL LAND-USES. Smith and Crooks (2010 p. 24) proposed to classify residential types by pursuing a dwelling address and geometry-based functional classification method. Their approach makes use of address point densities by defining the building class depending on the number of residential address points that share a spatial location with a building polygon. The authors propose 6 building categories ranging from detached houses to flats in contiguous blocks. Single building polygons that are only covered by one address point are hence classified as 'detached house', whereas building polygons that are covered by more than two address points are classified as 'flats in a single block'. Since the work of Smith and Crooks (2010), the Ordnance Survey has made significant improvements to the AB database, which is now released as AB Premium, including the incorporation of property types as well as improvements to the data coverage and completeness.[6] These improvements have resolved the issues pointed out by Smith and Crooks regarding general residential building classification.

The building geometry information employed in this study is 3-dimensional, which causes alternative issues with regard to the classification. Smith and Crooks' (ibid.) model use a 2D building footprint dataset and their classification is limited to the modelled footprints only. If 3-dimensional data is employed the model has a higher level of detail including potential numbers of stories and the total volume of the building, allowing inferences to the total habitable volume. An increase in the level of detail via 3-dimensional building information brings up additional classification issues. Such issues occur where the total geometric volume of a building does not correspond with the number of residential addresses. This is especially the case with multi-storey mixed-used buildings, as well as buildings of predominantly industrial or commercial function, where residential addresses are present but the total number of residential addresses is insignificant in comparison to the number of addresses of the predominant function.

---

[6] Smith and Crooks (2010 p. 19) pointed out at the example of London that a significant number of buildings is not covered by the AB resulting in unclassified building polygons. The completeness of the AB has substantially increased since 2010 resolving issues of non-classified building polygons.

The average habitable floor space in the UK in 2015 is 90.18 m² (Jennings and Lewis 2017), so that taking an approximate average floor height of 3 meters leads to an average habitable volume of 270.54 m³. Buildings that are classified as residential according to the AB Premium database, but exhibit significantly higher building volumes, might be examples of misclassification. These misclassifications occurs especially if the building polygon also features address points of the primary descriptor 'education', the secondary descriptor 'medical' or 'industrial applicable to manufacturing, engineering, maintenance, storage/wholesale distribution and extraction sites'. The primary descriptor 'education' leads to a misclassification of school buildings as residential buildings. The secondary descriptor 'medical' misclassifies hospitals, while the secondary descriptor 'industrial applicable' misclassifies warehouses and large industrial buildings. In all these cases, a misclassification can be detected by the significant difference between the expected habitable volume and the actual measured geometric volume of the building. The strongest outlier is a hospital building with an expected habitable volume of 270.54 m³ and an actually measured habitable volume of 1,555,840.88 m³. Following this observation, all buildings that have a measured habitable volume of 5,000.00 m³ above the expected habitable volume have been excluded from the database. This applies to 1240 of the 2,648,920 buildings.

The results of this selection can be found in Table 13, where all classes of the OS AB Premium database of a residential function that indicate a permanent address and the number of units in the covered area are listed. There are 14 classes that fall under this category, with a total of 3,107,523 address point entries. The labels 'residential', 'detached', 'semi-detached' and 'terraced' make up the largest part of the address points and have the sole function of residential usage. 'Self-contained flat' and 'house in multiple occupation' can aside of their sole residential function also describe flats in buildings with predominantly different functions. The majority of the data points is 'dwelling' with 1,164,117, followed by 'semi-detached' with 630,250, so that both of these labels make up more than half of all entries. This address point selection can then be used to classify all building geometries in a one-to-many spatial join geo-process. The column 'units' of Table 13 gives an account of the differences between number of address points and number of classified building geometries and allows direct comparisons with the German model. The largest class is 'dwelling' with 1,070,345 buildings followed by 'semi-detached' with 625,662. The biggest difference between address points and classified buildings occurs in the class 'self-contained flat' with 320,655 fewer classified buildings then there were address point entries. This numeral difference is due to multi-dwelling housings, such as high-rise buildings that contain multiple flats that are covered by multiple address points. In total the British model consists of 2,648,920 buildings of residential usage.

Table 13: Ordnance Survey AddressBase Premium: residential land-use classification, number of address points and resulting building units.

Class	Primary Description	Address Points	Units
R	Residential	19,406	8,453
RB	Ancillary Building	157	78
RD	Dwelling	1,164,117	1,070,345
RD01	Caravan	2,861	2,527
RD02	Detached	331,862	329,172
RD03	Semi-Detached	630,250	625,662
RD04	Terraced	513,226	505,875
RD06	Self-Contained Flat	423,552	102,897
RD07	House Boat	52	11
RD08	Sheltered Accommodation	1,551	250
RH	House in Multiple Occupation	567	99
RH01	HMO Parent	2,793	396
RH02	Non Self Contained Accommodation	14,828	1,073
RH03	HMO Not Further Divided	2,301	2,082

BRITISH COMMERCIAL LAND-USES. A comparable service and trade classification has been established in alignment with the AB classification process for residential buildings. The selection of address point classes is based on the selection criteria for trade and services of the German cadastre information (Table 14). There are 14 different address base classes that have been identified as comparable with the German dataset. These classes cover commercial functions such as retail, shops, or markets as well as services such as agents, banks or restaurants. The class with the largest number of address points is 'shop/showroom' with 39,384 addresses, followed by 'retail' with 6,714 addresses. These 66,296 address points are used to select and classify all buildings that intersect with points. The resulting number of classified buildings is smaller than the number of address points, with almost half as many buildings classified as 'retail' as address points in the dataset. This is caused by the many-to-one relationship with multiple addresses being covered by single building geometries.

Table 14: Ordnance Survey AddressBase Premium: trade and service land-use classification, number of address points and resulting building units.

Class	Primary Description	Address Points	Units
CR	Retail	6,714	3,931
CR01	Bank / Financial Service	1,313	976
CR02	Retail Service Agent	2,079	1,371
CR02PO	Post Office	198	165
CR04	Market (Indoor / Outdoor)	3831	294
CR06	Public House / Bar / Nightclub	4,157	3,411
CR07	Restaurant / Cafeteria	4,853	3143
CR08	Shop / Showroom	39,384	27,847
CR09	Other Licensed Premise / Vendor	1,075	821
CR10	Fast Food Outlet / Takeaway (Hot / Cold)	2,692	2,208

The land-use classification based on the AB dataset can further be enhanced through small-scale VGI information of the OSM points of interest data. For this purpose, all POIs with trade and service usage have been selected and used to identify additional buildings that hold important commercial land-uses. Table 15, shows the number of identified buildings as well as their function. The largest category of land-uses that has not been already classified by an address point are 'pubs' with 1660, and 'fast-food' with 1368 buildings. Overall, the use of volunteered geographic information form of OSM points of interest data has unveiled additional 4653 buildings with service and trade based land-use, that have formerly not been covered by the address point data.

Table 15: OSM points of interest classes of service and trade for the British model.

Primary Description	Buildings	Primary Description	Buildings
bakery	171	hairdresser	712
bank	320	jeweller	67
bar	232	kiosk	9
beauty_shop	153	laundry	73
beverages	90	mall	3
bicycle_shop	54	mobile_phone_shop	74
biergarten	1	newsagent	144
bookshop	48	nightclub	34
butcher	92	optician	99
cafe	726	outdoor_shop	20
chemist	25	pharmacy	381
clothes	341	post_office	392
computer_shop	45	pub	1660
convenience	797	restaurant	808
department_store	24	shoe_shop	52
doityourself	111	sports_shop	24
fast_food	1368	stationery	28
florist	72	supermarket	307
gift_shop	58	toy_shop	17
greengrocer	34	travel_agent	70
		video_shop	2

Only those buildings that have not already been classified are added to the final building class. A substantial number of buildings (5085) exists in both datasets (OS AB premium and OSM points of interest) and is excluded from the final OS POI selection. There are no cases of substantial misclassification or lack of coverage, as it is the case in the German ALKIS cadastre dataset.

Figure 28: Superposition of the OS AddressBase and the OSM points of interest building classifications. Red and blue hatched buildings are covered by both datasets.

The combined service and trade data of the OS AddressBase and OSM POI, leads to a final classification of 48,844 buildings for the British model.

4.2 METHODOLOGIES

4.2.1 REGIONAL MOVEMENT PREDICTIONS THROUGH NETWORK CENTRALITIES

Several authors have investigated the relationship and predictive power of measurements of centrality for human movement or traffic flows (Barros et al. 2007; Gao et al. 2013; Hillier and Iida 2005; Jayasinghe et al. 2015; Jiang and Liu 2009; Patterson and Jones 2016; Penn 2003; Penn et al. 1998; Serra et al. 2015). These studies are all comparable in terms of their methodological approach. First, network centralities are generated for the respective spatial, street or traffic network. Secondly, gate count data of pedestrian, vehicular or alternative modes of transport are correlated with the coinciding network segment. The majority of previous research compares betweenness centrality, closeness centrality or a comparison of both. The results reported indicate linear relationships with different degrees of explanatory power for the data variance (R^2=0.30 to 0.80) highlighting the importance of the relative location of a street for the expected pedestrian or vehicular flows. In this context, it has been suggested that particularly the location of a street in relation to the overall city is a determining factor for the mean level of vehicular movement (Penn et al. 1998 p. 77). The role of inter- and intra-regional relationships have largely been neglected by the researchers interested in cities.

As previously laid out, the city as entity brings a series of disadvantages, such as edge effects and issues of boundary selection, into the context of network analytics. In this context Patterson and Jones (2016) highlight the issues related to low traffic prediction rates in urban settings, artificially selected areas within cities and isolated towns. The authors present a method of 'boundary weighting' to increase prediction in those cases where edge effects had a negative effect on the prediction accuracy. As shown previously, a simplified OSM model is comparable to initial axial line models, and model extensions do not constitute a limitation. Rather than to adjust centrality measures through a specific boundary-based weighting, I propose that these issues can be substantially excluded by employing the simplified OSM model proposed.

The work of Serra et al. (Serra et al. 2015) explores the relationship between traffic flows and large-scale network metrics. The authors correlate traffic data to centrality measures on 27 different analysis radii of the A-level vehicular street network (most importantly routes, i.e. motorways and trunk roads connecting major cities) and report the highest correlations at a radius of 20 km (for the entire dataset with r=0.69 for closeness centrality and 0.56 for betweenness centrality). Following these findings Serra et al. conclude that a 20km radius must indicate a 'main scale for movement at the national level' and that spatial network metrics can be employed to predict traffic flows.

These findings present a relevant contribution to the understanding of regional traffic flows for A-level street networks. What cannot be answered form this initial investigation is the interplay of different spatial scales. A-level roads are limited to a very specific mode of transport, i.e. vehicular motorway traffic, and can therefore not account for relationships that go beyond these distances. Moreover, the reported low correlation scores at small radii are most likely caused by the A-level road network morphology, as distances between intersections in A-level networks are larger than in the citywide street networks. A-level roads account for only 12.7% of the entire 395,000 km road network (Department for Transport 2017) and feature a rather incomplete representation of the region under investigation. This study is, however, particularly interested in the relationship between small-scale urban areas and the relative location in the region, independent of the theoretical scale, model size or analysis radius.

To answer the question of the importance of intra- and inter-regional factors for the prediction of vehicular flows and to account for the mulit-scalarity of flow data, I will compare country wide traffic count datasets with closeness and betweenness centrality measures of latent centrality factors as well as the combinatory model proposed.

4.2.2 SAMPLING THROUGH HIERARCHICAL GRID MODELLING

Sampling through hierarchical grid models, describes a sampling system with different hierarchical levels. Each of these levels is of a different scalar dimension. For a seamless comparison, the dimensions of each layer needs to be congruent with the next dimension and their spatial locations have to be aligned. Such a hierarchical model allows understanding the influence of sample sizes on the predictive potential of network metrics on the spatial variables, and allows us to draw conclusions about scalar effects on the phenomenon observed.

As mentioned in section 4.1.2, the population data used is stored in a 1 x 1 km grid cell format. This grid cell format enables the creation of a hierarchical grid cell model. A grid cell plays a pivotal role in geographic analysis and is in the context of GIS raster files also referred to as pixel. Grid cells are used for the purpose of sampling, observation and modelling. For the purpose of sampling, grids need to be of regular tessellation, where each cell is formed of congruent regular polygons. Regular tessellations exist only in three different forms: equilateral triangles, squares and regular hexagons. Squares, or regular rectangles are the most frequently used form of grids in geographic analysis. The effect of the form and size of sample polygons is a much-discussed issue within spatial modelling (Birch et al. 2007; Hengl 2006). While there are good reasons for the application of hexagonal grids in urban sampling, such as minimisation of polygon-edge effects or more symmetric nearest neighbourhoods (Birch et al. 2007 p. 353)(Figure 30),

hexagonal grids do not allow seamless hierarchical modelling. Rectangular grids, in contrast, have an advantage with regard to hierarchical modelling. Each rectangular grid cell can be divided into an integer number of smaller squares maintaining congruency and alignment with the previous scale (ibid. p. 355).
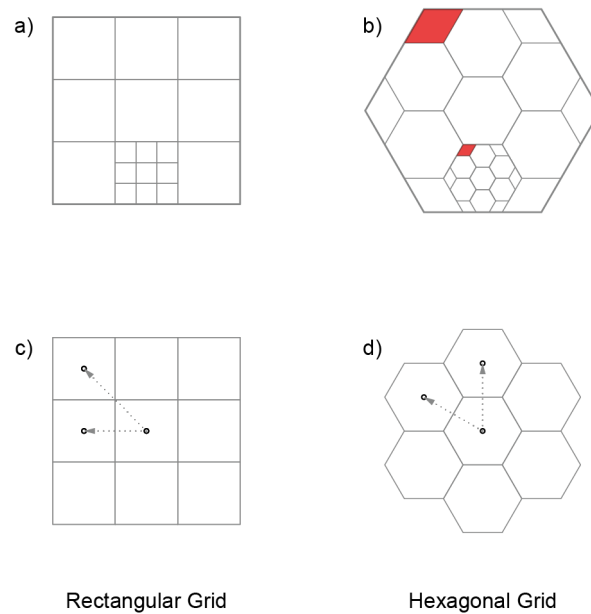
Figure 29: Rectangular and hexagonal grids and their potential congruency and alignment a) and b). Difference in nearest neighbourhood between rectangular and hexagonal grids c) and d).

Hierarchical modelling is of particular interest when the observed phenomenon underlies a specific scalar relationship, or in other words, if a phenomenon becomes only present at specific scales. This could be the case when spatial processes cluster only at a particular distance or when spatial variables correlate only when compared at specific area sizes. This raises issues related to the respective size of the sample polygon, a problem that is inherent to any regular tessellation. According to Florinsky and Kurkyakova (2000) a selection process of grid sizes, should be based on a statistical method performing a correlation analysis of a series of grid sizes with the variable and their respective spatial objects. The adequate size can be determined by a smoothed plot of the correlation coefficients of each series. While this approach confidently arrives at an adequate grid size, it proves to be time-intensive when multiple spatial variables are employed. In this context, Hengl (2006 p. 1296) investigates the effect of different grid sizes on terrain data and formulates three generally recommended grid sizes that can be translated to cases of urban data. According to Hengl, the sample polygon size should be either of the three following: a) the coarsest legible grid, with respect to the positional accuracy and size of objects, b) the finest legible grid, covering at least 95% of spatial objects, and c) a compromise between the two.

Following Hengl's (ibid. p. 1296) recommendations for the selection of grid cell sizes, three sample grids have been selected based on the initial data source, the spatial objects and a compromise between the two. The core spatial components are the spatial network and 3D-building geometries. The crucial spatial network element is the segment, of which 90% have a length of up to 203.5 metres (GE) and 213.5 (UK). Whether a 3D-building geometry falls within a sample grid is defined by the building's footprint, of which 97.5% are of up to 433.05 m² (GE) and 205.78 m² (UK). With a minimum building depth of 3 metres, these square metre values imply that none of the 97.5% of all buildings will be longer than 150 metres. A satisfactory coverage of 90% of all spatial objects can hence be achieved with a sample grid cell that is at least 250 metres in both dimensions.

The largest or coarsest grid possible is defined by the GEOSTAT dataset and is a 1 x 1 km grid. The finest legible grid can be defined by the size of the spatial objects (segments, buildings); in both cases, these objects do not exceed 250 metres, which leads to a finest-legible grid size of 250 x 250 metres. As a compromise, a thrid sample size has been selected based on a size that is in between the coarsest and finest grid. Coinciding with the power of two of the resulting grid cells, the compromising grid measures 500 x 500 metres. A 500 x 500 metres grid is the squared integer of 1, which allows a direct scalar embedding between the finest and coarsest grid. Figure 30, exemplifies how these three scales are interlinked through their squared dimensions at the example of the coarsest grid cell downwards. Through the method of reaggregation, spatial variables can be collected and correlated at any of the three spatial scales. The analysis will build on the hierarchical model with three sampling levels of 1 x 1 km, 500 x 500 metres and 250 x 250 metres.

1km x 1km

500m x 500m
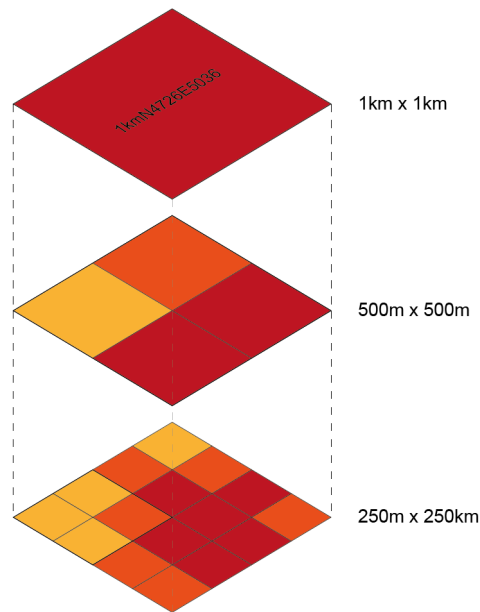
250m x 250km

Figure 30: Hierarchical grid model with three scale levels: 1km x 1km, 500 m x 500 m and 250 m x 250m.

4.2.3 ESTIMATING REGIONAL POPULATION DENSITIES

In this section, I present a model for the estimation of population on the level of 3D-building information and street segments. Following a brief introduction of different methods of population estimation, I introduce the methodology developed and the tests for their appropriateness.

Employing precise small-scale population data is not only of interest for this study but for a large variety of research applications, such as evaluations of environmental hazards, health and epidemiology, or transport and planning applications. Yet, the availability of comparable small-scale population data of regional scope is limited across the globe. Despite the potentially broad applications of such small-scale datasets, the access to reliable population estimates at the necessary spatial resolution is scarce. The access to building precise population datasets, for example, is on one hand limited due to concerns of privacy disclosure, a problem that increases when datasets are combined with population related variables (e.g. age class, gender, housing, among others) and, on the other, due to the cost-intensity of data collection. While in exceptional cases population data is regularly updated and provided on the small-scale statistical unit of the building block (such as the Netherlands), the vast majority of countries provide Census based population data on the geographic level of larger administrative boundaries such as the scope of the city or commune. These administrative boundaries come with the drawback of arbitrarily drawn contours and significant differences in geographic size within a single dataset.

Figure 31 exemplifies how the boundaries of the smallest statistical unit in the UK (lower super output areas) arbitrarily cut through settlements and differ in area size. One of the main reasons for this is that municipalities and urban administrations try to generate administrative boundary areas that are equal in their number of captured residents rather than according to the comparability of their spatial make-up. The average size of such boundaries grows with a decrease in urbanity and an increase in rurality of the captured area due to the difference in population density. Arbitrary or purposely defined administrative boundaries affect not only statistical comparisons, but are regularly used as a political tool in gerrymandering, which is a practice aiming to establish a political advantage by manipulating district boundaries. This makes such datasets particularly problematic for applications in small-scale, block, street or building levels, because the discrepancy between detail levels undermines potential correlations with local variables. Specifically, the issue of the ecological fallacy plays a pivotal role in this process. The ecological fallacy is committed where an inference about the nature of individuals is derived from the statistics fathered at the level of the group of which the individuals are part. In the context of population data, this would imply that the population is evenly distributed over the observed space and that correlations with spatial variables at the level of an administrative boundary hold the potential to infer to the level of the individual object and ultimately to the individual behaviour. Additionally, Census data, on which population estimates are traditionally based, can be inconsistent or incomplete, particularly in developing countries (Alahmadi et al. 2013; Anderson et al. 2014).
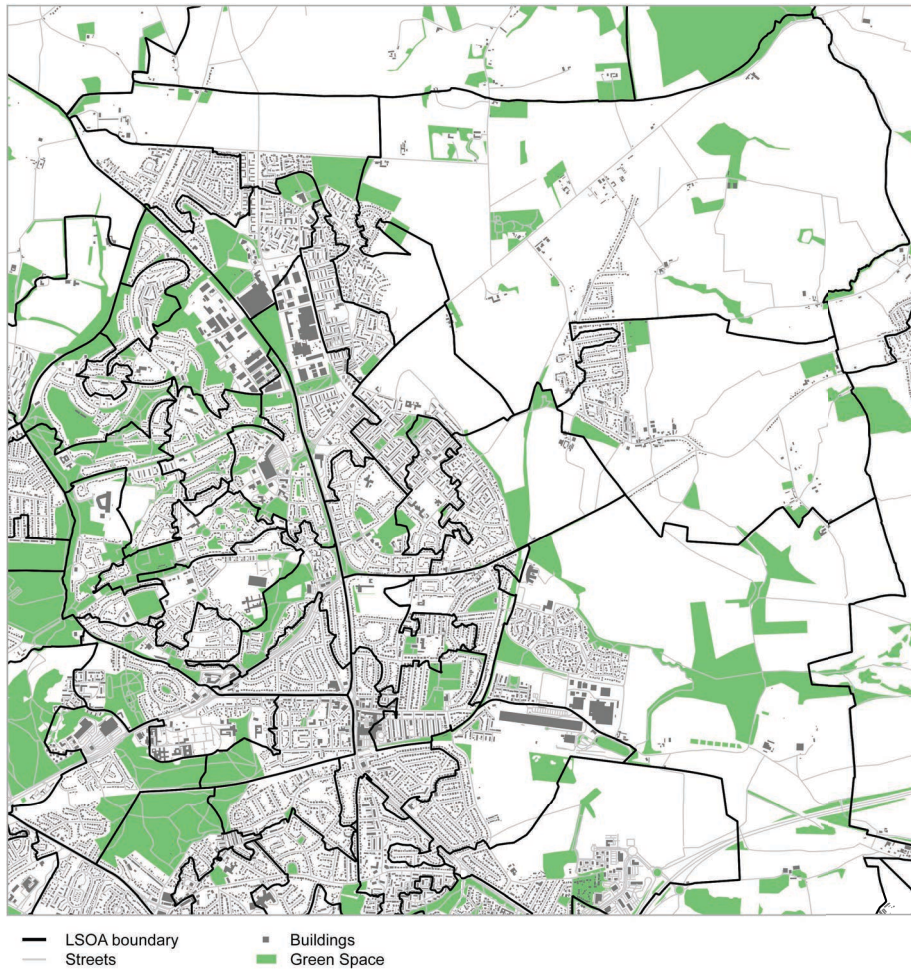
Figure 31: Small scale statistical boundaries (Lower super output areas) for an exemplary area in the United Kingdom.

POPULATION ESTIMATION USING GEO-STATISTICAL METHODS. A common solution to the issues of differently sized statistical units and differences in scale is the estimation of population density for small-scale resolutions with the help of geographical information. Population estimation for fine-grained spatial resolutions has a long-standing interest in sciences in geography and demographics (Kraus et al. 1974), with methods of population mapping such as 'dasymetric mapping' dating back to the early 20th century (Petrov 2012). In the past, such estimations have widely used population data in combination with additional ancillary information, such as 2D geographic information, satellite imagery, or spatio-temporal data (e.g. twitter information). In general, population estimation approaches can be divided into two categories: areal interpolation methods and statistical modelling (see Wu, Qui and Wang (2005) for a comprehensive review).

AREAL INTERPOLATION is applied where population data is available but needs to be transferred between varying source areas and target areas. In such cases, the main obstacle is that the initial data source (e.g. Census information) is spatially incongruent

with the target zone, e.g. areas of interest, such as market catchment areas or postal zones (Qiu et al. 2012 p. 645). Areal interpolation is an appropriate geo-statistical procedure for those cases, where population data can be taken from an existing set of spatial polygons to predict the population of a new set of spatial polygons. In this sense, coverage is not the issue but the challenge lies in bridging different geographic scales. A specific focus in areal interpolation is to maintain the coherence of the total population. The aforementioned disaggregation (see 4.1.2) is a type of areal interpolation and results of disaggregation processes can be enhanced through the employment of ancillary information. This process is also referred to as dasymetric mapping and commonly considered as the approach producing the highest level of accuracy (Wu et al. 2005 p. 70).

STATISTICAL MODELLING describes the methodological approach where population data is estimated through a correlation of ancillary information. This is necessary where census data is either scarce or incomplete. In such cases, the existing population information is correlated with alternative parameters, e.g. urban areas, land uses blocks, dwelling units and other socio-economic characteristics, so that population predictions can then inferred from these correlations (ibid. p. 66).

For the purpose of this study, specifically the approaches of statistical modelling that makes use of land-use and dwelling information are of interest. This strand of research has a long tradition in the estimation of population densities (Green 1956; Kraus et al. 1974). While researchers have focused on manual counts of buildings and their classification in the fields' early years, since recently the access to large scale classification databases has increased, opening up possibilities of large-scale applications. To estimate population at a given location the authors classify each building of potential habitation as 'residential', and count the total number of residential buildings per statistical area. This count can then be used to calculate a person-per-dwelling unit ratio; leading to reliable estimations (Green 1956; Wu et al. 2005 p. 67). In later approaches the classification is improved by introducing differentiations of the building types (single-family, multi-family, trailer parks, commercial) (Kraus et al. 1974). These investigations make substantial use of aerial imagery and 2D building information. Several authors have pointed out that besides incorporating semantic information, also the building geometry can play a pivotal role in enhancing population estimations (Bakillah et al. 2014; Wu et al. 2008).

Very recently, with the access of city wide 3D-building data, this strand of research has branched into employing 3D information derived from aerial imagery (i.e. LIDAR), as well as complex manually produced 3D-building geometries for the precise estimation of population (Biljecki, Ohori, et al. 2016; Tomás et al. 2016). The general assumption

is that the building volume provides a higher level of detail in the physical description of the lived environment, capturing differences between identical buildings types (such as number of floors, or differences in floor plan efficiencies) at a much more fine-grained level and arise, thus, positively impacting the estimation accuracy. Such geometric information is of higher importance in built environments that feature higher degrees of spatial complexity (i.e. multi-storey and multi-level buildings or mixed used functions) often found in urban conurbation and urban centres.

Specifically the seminal work of Biljecki et al. (2016) shows how the use of 3D-building information together with semantic information enrichments can produce reliable population predictions, superior to traditional 2D estimations. The authors evaluate differences between 2D- and 3D-data in the population estimation, and show that 3D-building information is an appropriate data source for regional- and country-wide population estimations (ibid. p. 2). 3D-building information proves to be especially useful when the aim is to refine the population to a finer scale (ibid. p. 14). The prediction accuracy of such building-based population estimations, whether 2D or 3D, highly depends on the quality of the classified semantic building information. While in the past, classification has been performed manually, with the advancement in remote sensing technology the production of precise geo-referenced ancillary semantic building information has been automated and accessibility to public datasets has rapidly increased (see section 4.1.4). It should be noted, that the application of 3D-building information limits the methodology to those cases, in which such datasets are publicly available. With the increasing coverage of OpenStreetMap building information, LIDAR aerial scans and the digitisation of governmental cadastre information, global coverage of 3D-information might be available in the near future. Countries, such as the Netherlands take a vanguard position in this development and already provide country-wide 3D-datasets.

4.2.4 GEOSPATIAL METHODOLOGIES FOR POPULATION ESTIMATION

POPULATION ESTIMATION FOR BUILDING- AND STREET-LEVEL. Following particularly the recommendations of Biljecki et al. (2016), I will, in an initial step, derive building-precise population estimations from the combination of 3D-building information (LoD1) and classifications enriched by semantic information (ALKIS cadastre, OS AddressBase Premium and OSM points of interest). I will then evaluate the predictive power of the residential buildings, or to be more precise of the habitable volume, as well as of the volume of all buildings. This evaluation is done to verify the accuracy improvement through the use of semantic information. Once a predictive improvement has been tested, I will create a database of building-precise population estimates by methods of disaggregation. I will make use of disaggregation methods to calculate the

average person per habitable building volume (people per m³) for each building that falls into a GEOSTAT 1 x 1 km population grid. This volume based population estimate, is then spatially joined to the level of the street segment using an aggregation method. The aggregation is done by identifying all buildings in close adjacency to a street segment and link the total number of people present at a building address (people per m³) to infer a street-level population estimate (people per street segment).

Figure 32 shows a visualisation of the model components. The first step involves the extraction of building geometries from the region-wide 3D-model based on the semantic information forms (1); second, the disaggregation of population data to buildings, based on their habitable volumes (2); and finally, the spatial join of population data in the unit of people per habitable volume on spatial segments using methods of aggregation, based on the principle of adjacency (3). The result of this process is a reliable estimation of the number of people living in the proximity of a street segment. The population per street segments allows for large-scale correlations with regional spatial networks.
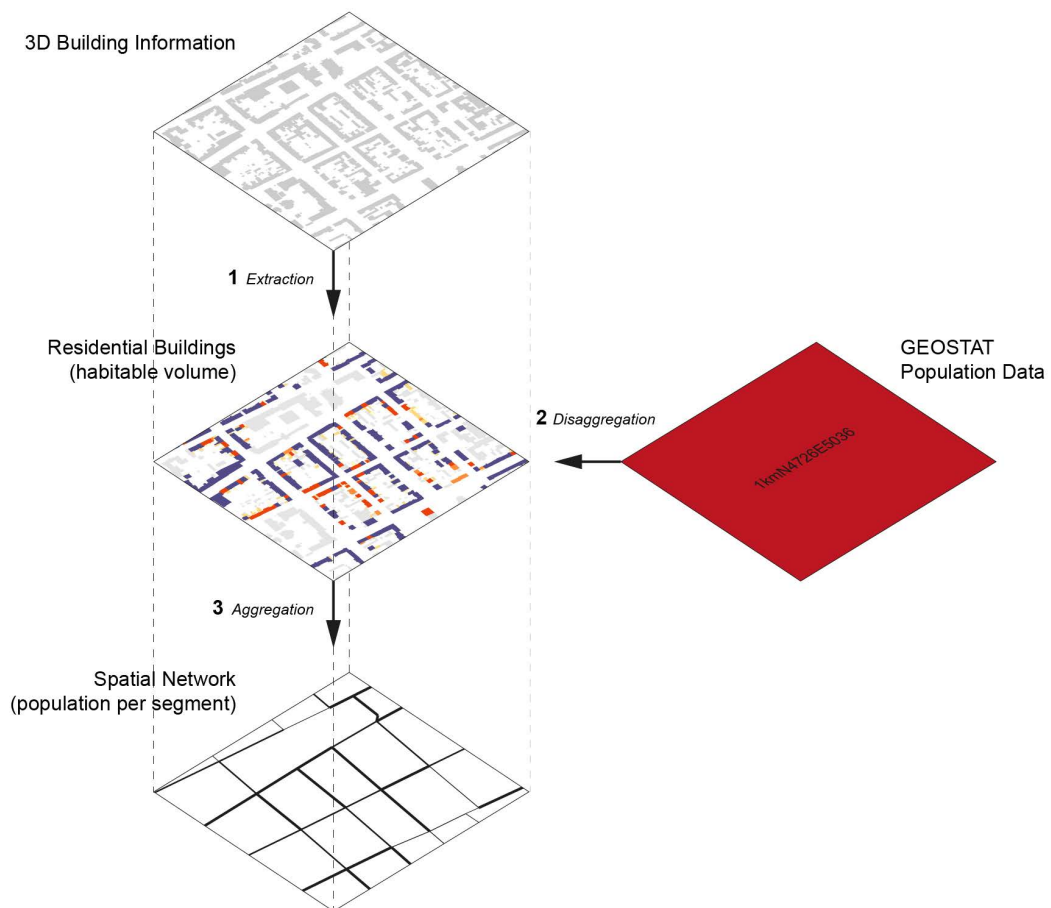
Figure 32: Model of population density estimation per building and street segment. Highlighting three core procedures: 1. extraction, 2. disaggregation and 3. aggregation.

HIERARCHICAL MODELLING AND THE PREDICTION OF POPULATION DENSITIES THROUGH SPATIAL METRICS. In order to validate the predictive potential of spatial metrics for population estimates beyond the level of the street segments, I conduct an area sampling approach. Using a three-level hierarchical model (Figure 33), I sample and compare the different datasets. Built environment data is aggregated at three different scale levels (1000, 500 and 250 metres) and the aggregation effect of the predictive potential compared.
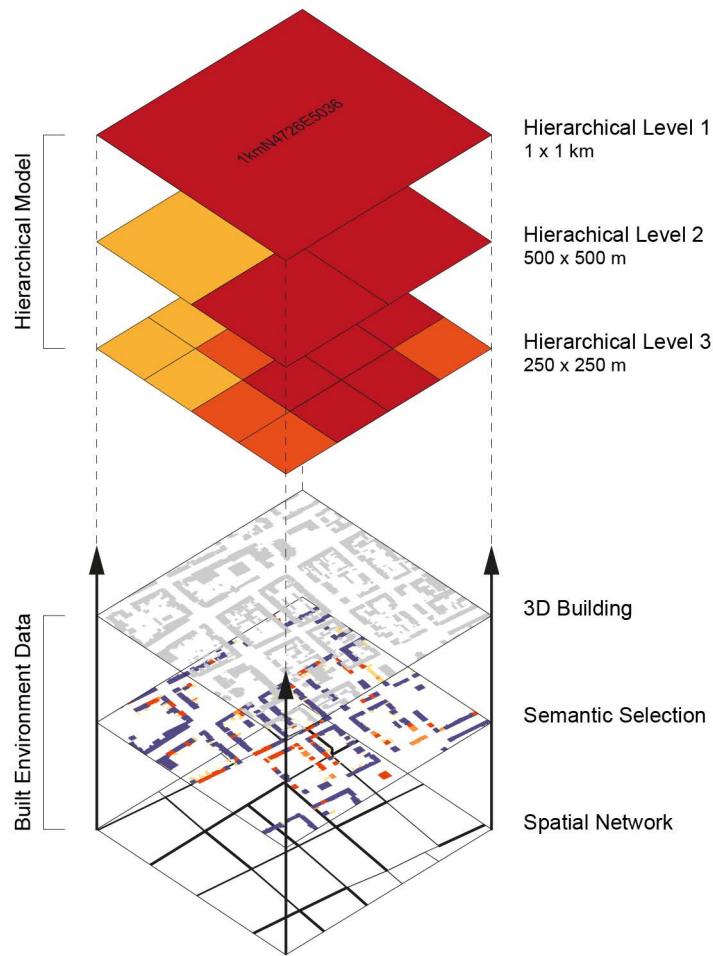
Figure 33: Hierarchical model and built environmental data.

PREDICTION OF HUMAN SPATIAL OCCUPATION BASED ON 3D-BUILDING INFORMATION. The next step is the exploration of the predictive power of spatial metrics and latent centrality structures for different spatial occupations. I will scrutinise whether the different theories can be verified by their physical realisations. In order to do this, spatial network metrics are correlated with a series of different land-use selections. I select four different building volumes based on the semantic information (see 4.1.4).

The first selection consists of all buildings present in the sampled areas, the second selection excludes all buildings with an industrial function, the third selection includes all buildings of residential usage, and the fourth and final selection includes all buildings with service and trade functions. The reasoning for these four different selections is to scrutinise the regional form-function relationship. If the central place theory is not able to account for the spatial organisation of post-industrial regions no relationship between service and trade functions and the centrality measurements should be observable. I spatially join the selections to their closest street segments and aggregate their number of occurrences, size of the area and building volume.
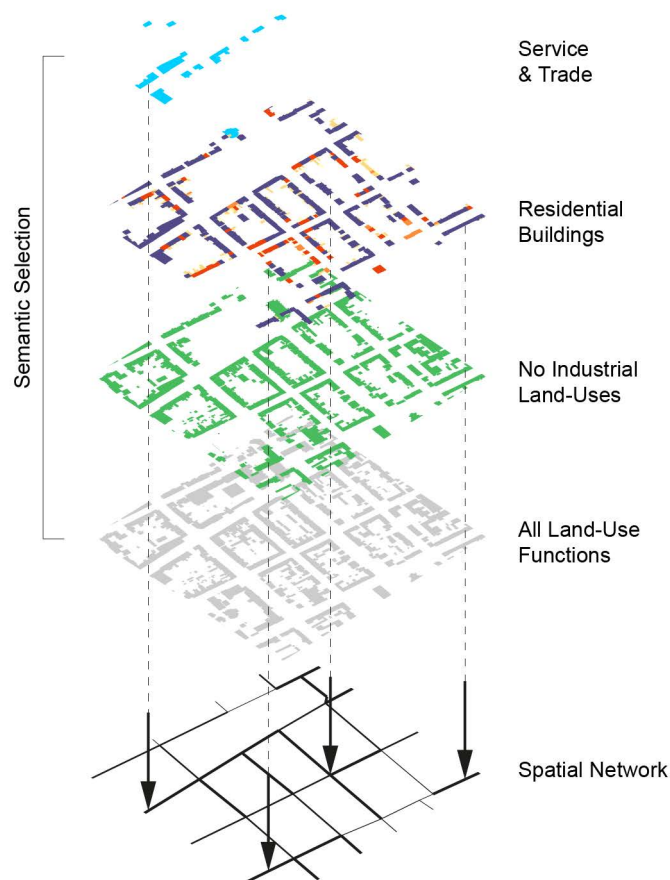
Figure 34: Model for the prediction of human spatial occupation. Four semantic selections of 3D-building geometries (service and trade, residential, all but industrial functions and all land-uses) are spatially joined and aggregated onto the spatial network segments.

4.2.5 CENTRES HIERARCHY CLASSIFICATION THROUGH SPATIAL NETWORK METRICS

In order to compare centres of commercial activity with different spatial metrics and their respective scales, I select all buildings with land-uses of service and trade and convert them into simple point representation. The selection classification is based on the previously introduced semantic information and includes governmental, as well as OpenStreetMap information. All geometric 3D-building information (volume, height, floorplan area and building stories) is kept at the attribute level of the generated point features. In an initial step, I identify those areas that can be classified as centres based on the number and density of commercial activity. I do this using a kernel density analysis (see Chapter 6). I select the kernel bandwidth with a distance of 200 metres in order to capture small-scale clusters of land-uses arranged in linear and circular shapes. The resulting kernel density estimates can be used to generate contour polygons. A minimum of 0.0003 densities per square meter resulted in initial tests in boundary areas that capture a minimum of 10 land-use functions per cluster. As a form of verification of this method, I compare all selected areas against the British 'Town Centre Boundaries' dataset (TCB) and all TBC areas that are captured in the model area are also captured by the newly introduced service and trade boundaries. The TCB dataset was initiated by UCL in 2004 and has since been updated by REVO (formerly The British Council for Shopping Centres); it includes all primary shopping areas and areas predominantly occupied by main town centre uses.

UK Service and Trade Centres Exemplary Selection

UK Centres Kernel Density for Service and Trade Functions [bandwidth 200 metres]
Natural Breaks [Jenks]

0 - .00002
.00003 - .00011
.00012 - .00025
.00025 - .00044
.00045 - .00108
.00109 - .00159
.00160 - .00275

Buildings
Service & Trade
Centres
English Town Centres
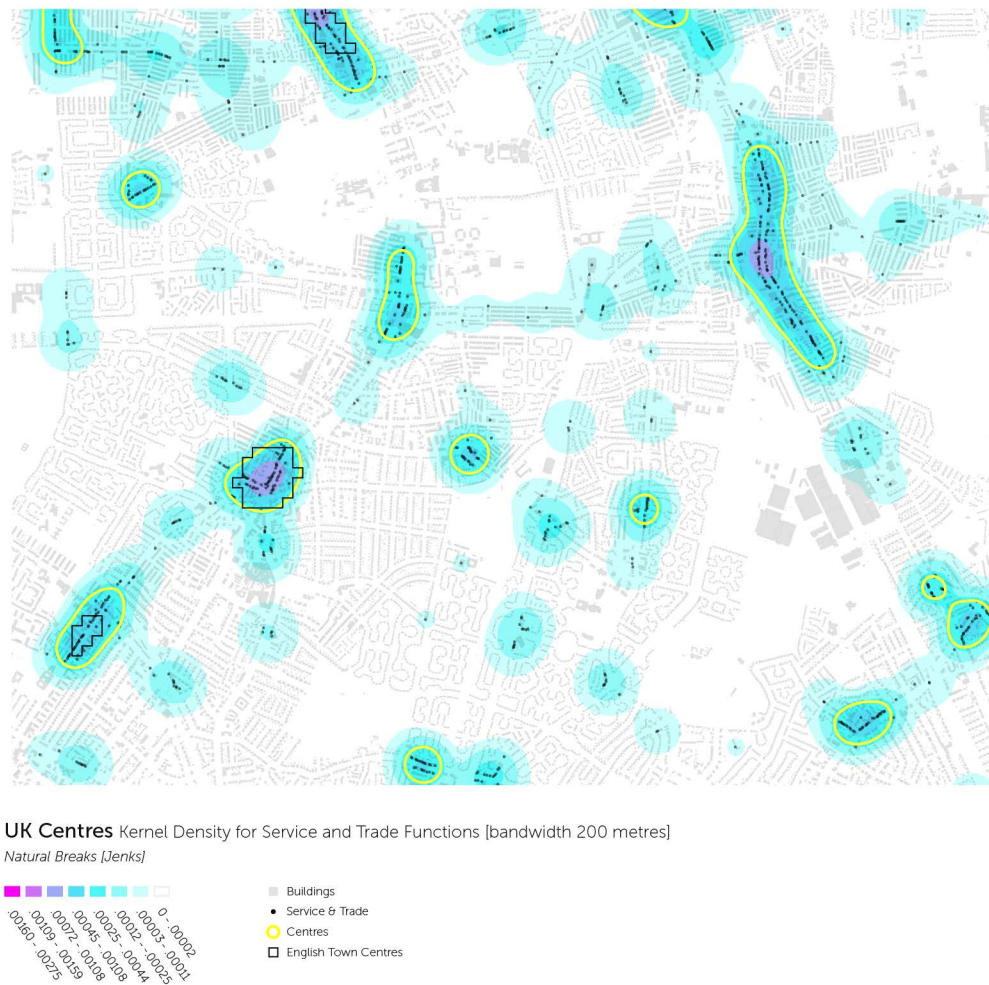
Figure 35: Visualisation of the centre selection process of urban centres of service and trade for an exemplary section of the UK model. Selected centres highlighted in yellow, TCB boundaries highlighted in black.

These generated centre boundaries are used to agglomerate spatial metrics, building geometries and land-use information. The boundaries are solely used to identify which buildings and land-uses form a cluster but not to establish geographic relationships, as this relationship is established at the level of the building. The spatial network metrics are selected by computing the nearest network segments to each respective building (maximum distance 25 metres), rather than to select those streets that intersect with the centre boundary. All of the segments with at least one match are agglomerated based on the maximum value obtained and spatially joined to the centre boundary. The table below gives an initial account on the number of identified service and trade agglomerations, the total of the unique land-uses taken into consideration, the mean land-uses per cluster as well as the mean cluster size. The German model features twice as many clusters with a much higher mean land-use density.

Table 16: Identified number of agglomerations, land-uses and mean cluster area.

Model	Agglomerations	Land-uses in units	Mean land-uses per cluster	Mean area in sqm
UK	345	21,918	63.53	84,974
GE	717	66,257	92.40	116,776

These 1,062 agglomerations are compared with their maximum value for each of the previously identified latent centrality structures in order to test the relevance of each of the scales for the respective cluster. The comparison is done by performing a hierarchical cluster analysis, a multivariate statistical technique that groups together those observations that share similar values across a number of variables. The aim is to identify service and trade agglomerations that share similar patterns of scale relationships.

CHAPTER 5

VOLUNTEERED GEOGRAPHIC INFORMATION AND
NETWORK ANALYSIS

5 CHAPTER

The aim of this chapter is to present a workflow and methodology that allows the use of OpenStreetMap (OSM) data in space syntax angular segment analysis (ASA). The reasoning behind employing such datasets is the increasing scale of analytical investigations in the context of space syntax and in this study in particular. This augmentation of scale has become particularly necessary due to the extensive global growth of cities and the development of their urban hinterland into large and complex urban regions. Such urban structures are simply too vast to be mapped manually or generated by automated algorithms. This magnitude of structures has created a situation in which the time and economic feasibility of traditional as well as algorithmically derived axial line maps needs to be revisited. Previous research has proposed to make use of governmental so-called road-centre line data as an alternative for a segmented axial line, more commonly referred to as segment maps (SM). However, very little has been said about the disadvantages of such approaches particularly when global comparability is needed, something in which space syntax is believed to be particular strong. OSM road-centre line data, on the other hand, I will argue, forms not only an appropriate alternative basis for models in these situations, but it also allows global comparability and is freely accessible on a large scale. Nevertheless, OSM data also come with some disadvantages. Particular caution should always be taken when dealing with datasets of such an abundance of information, which requires a simplification prior to any ASA application.

This chapter consists of three parts; the first revisits the foundation of space syntax axial line models and the sequentially developed analytical method of ASA and its segment map (SM) model. Here, I place particular emphasis on the model, which underlies the analysis and the difficulties generally arising in the model generation and particularly in large-scale applications. In this light, I review volunteered geographic information and governmental road-centre line data, such as the British Integrated Transport Network (ITN) as alternatives for SM models. Finally, I discuss the advantages as well as the disadvantages of OSM data and their effects on ASA outcomes. The second part examines the structure and particularities of the previously introduced OSM data, as well as the difficulties researchers are facing when employing such data in ASA. I discuss the three main difficulties, which are topological inconsistency, traffic management components and excessive or redundant nodal information. I propose a series of different GIS strategies to simplify and remove this redundant information and explain the theoretical reasoning behind them. The result is a newly developed simplified OSM network model, termed 'SIMP'. The third part evaluates the new SIMP

model against OSM, ITN and SM models in ASA. I do this, using descriptive statistics, visual comparisons, as well as a Pearson and Spearman correlation analysis. The results show an overall high correlation between the four models, confirming previous findings. The new SIMP model exhibits higher correlations with the segmented axial line model than both OSM and ITN network models, indicating that a simplified OSM network does not only constitute an appropriate alternative but one that presumably incorporates fundamental network characteristics of SM models.

5.1 EMPLOYING OSM DATA IN SPACE SYNTAX ANALYSIS

5.1.1 AXIAL MODELS AND ANGULAR SEGMENT ANALYSIS

Axial analysis forms one of the fundamental techniques of space syntax. At the heart of the axial analysis methodology lies the axial line map, a representation of the continuous structure of open spaces in urban settings. Hillier and Hanson (1984 p. 17) introduced the first axial line model during the early 1980's and defined it as a system of fewest and longest intersecting lines covering all open spaces. These lines are the result of a two-step process where the spatial system under investigation is first represented through a two-dimensional organisation of convex spaces. Convex spaces are polygonal representations of continuous open spaces, in which each part of a space must be visible from every other part. The underlying rule for drawing a convex space is that each polygon must feature the best 'area-perimeter ratio', starting with the 'fattest'. In a second step, this system of convex spaces is covered by a one-dimensional set of axial lines. Axial lines are linear representations of longest lines of sight and/or movement. Each convex space must be covered by at least one axial line, while each line needs to be the 'longest straight' line possible (ibid. p. 17).

Although Hiller and Hanson describe this process as reproducible and objective, there is some discussion and ambiguity about the comparability and making of axial maps. Problems arise for instance with differences in the level of detail or resolution in which convex spaces are produced, as this impacts the number and distribution of the resulting axial line map. Problems also arise with the difficulty of arriving at a comparable and reproducible solution for the same given urban context. In this regard, Peponis et al. acknowledge 'SpaceBox'[7], a software that automates the generative process of convex spaces. Yet, the authors criticise the lack of mathematical rigour of its

---

[7] SpaceBox is a software developed by Sheep Dalton (1988) and includes several *space syntax* related functionalities, one of which being the generation of an all convex space map. The software's partitioning algorithm extends a walls surface area collinear until the produced line reaches another wall surface. See Carranza and Koch for more recent work on convex spaces (2013).

computational algorithms to generate convex spaces (1997, 1998). According to Peponis et al. neither the initial principle of generating convex spaces based only on an economic partitioning, nor the principle of extending surfaces to the next opposite wall are sufficient methods. Both lead to multiple conflicting solutions, implying that a more sophisticated set of rules is necessary. Interestingly, although the methodology of convex spaces is conceptualised in an urban context most of the discussions are set in the context of buildings. This might be due to the time-consuming process of producing convex spaces for entire cities with the sole purpose of deriving an axial line map. The scale of the area under investigation and respectively the time necessary to produce such convex representations is certainly one of the most important factors.

Moreover, Desyllas and Duxbury argue that not only the production of convex spaces, in general, is difficult but that it constitutes a 'mathematically impossible problem' to link all maximal convex spaces with axial lines in an identically repeatable manner (2001 p. 27.6). The core problem here is that there are several solutions to axial lines that fulfil the criteria of being the longest as well as covering all convex spaces (Batty and Rana 2004; Ratti 2004). As a solution to this technical and theoretical problem Turner et al. (2005) – building on an initial but not ideal solution by Peponis et al. (1998) – propose an automated methodology that produces a fewest line axial map. The starting point of their method is vector information of open space boundary polygons. Based on this, a so-called 'all-line map' is generated (Penn et al. 1997). The 'all-line map' is a map that features all lines that connect each vertex of boundaries and buildings with all other visible vertices, i.e. all possible lines of movement. In a following step Turner et al. employ an algorithm to reduce this 'all-line map' to a fewest line axial map. Their results are reproducible and strikingly similar to the original Hillier and Hanson axial map (2005).

However, the method of Penn *et al.* (1997) for the generation of the fewest line axial map does not constitute a feasible solution to produce models for large cities and regions. There are two primary factors, which prevent the application in a city-wide and regional context. The first deals with the source of data and its definition of open space, a problem that the very initial convex space methodology already inherited. What to include and what to leave out in a graphical representation of the real world is left to the individual cartographer or researcher and forms core challenges in comparative cartography and map-making in general. This challenge is of particular importance when investigating suburban or rural areas. Suburban and rural areas often lack a continuous urban form and hence a clear limitation for movement and visibility. Consequently, the definition of what can be considered an 'accessible open space' becomes vague. A problem that researchers are also facing in the context of developing countries is that roads are often not solidified and boundaries between public and

private spaces are less clear. In these cases, an alternative could be to rely on other sources of geographic data of open spaces. Such sources are, for example, governmental agencies for cartography, geodesy and planning or volunteered geographic information, both of which follow precise definitions of what and how open spaces are mapped.

Computational time constitutes the second difficulty. With a rising number of mapped open space polygons and their vertices the necessary computational time to generate the fewest line axial map increases. Turner et al. report the computational time needed for their algorithm to compute fewest line axial maps as follows: a model of the small town of Gassin takes 119 seconds to compute and features 5217 lines in its initially generated all-line map and 38 axial lines in the final result (Turner et al. 2005). Thus, the computational process for an entire city or even a region with far more than one million street segments will take significantly longer[8] time. While, in theory, the algorithm could run for whichever time needed, in practice, this is limited by the software design dealing with large datasets. Currently the most commonly used software for this is depthmapX. Initial tests using the software on large urban systems generating fewest line axial maps have consistently caused the application to crash. Varoudis et al. state the maximum number of segments that can be computed by depthmapX as <1.500.000 (2013), resulting in an axial line map of approximately 15000 lines. At the moment, this renders an automated generation of axial lines for a metropolitan or regional system impossible.

5.1.2 ROAD-CENTRE LINES AS ALTERNATIVE FOR SEGMENT MAPS

Initially, the focus of axial line maps was to have a tool that would help to understand complex urban systems in a simplified comparable manner. Over the time, the primary use of this morphologically descriptive tool changed to help investigations into the deep relation between human behaviour and space. Since the development of the methodology, throughout the last 30 years, researcher consistently found a correspondence of the topological relationships of spatial systems and pedestrian movement (Desyllas and Duxbury 2001; Hillier et al. 1993; Hillier and Iida 2005; Penn et al. 1998) as well as vehicular movement activities (Hillier and Iida 2005; Law and Versluis 2015; Serra et al. 2015; Turner 2005) and even global transportation networks (Hanna et al. 2013). This has particularly been the case since the introduction of ASA in space syntax as an extension of axial analysis (Turner 2001). Overall emphasis, thus, shifted from a theory and tool to analyse spatial configurations, to a tool to predicting the potential of human behaviour in the form of movement and flows. There are four

---

[8] The total number of axial lines in cities with a population of 300,000 can range between 10,000 and 15,000.

studies that focus on alternatives that constitute possible models for an analysis of movement and flows in the built environment: the pioneering work by Thomson (2003), a study by Dalton et al. (2003), two studies by Turner (2005, 2007) and as a follow-up to these studies, most recently, the work by Dhanani et al. (2012). All authors investigate the possible application of different types of so-called road-centre line data. In all of the above studies, the reasoning is that their approach relies on replacing a segment map, which is used in angular segment analysis, rather than the traditional axial line model that is based on SM. This study will follow the path taken by the four studies above and base the comparison on a segmented axial line model, rather than emulating an axial line model, which would later inevitably be segmented in order to perform ASA.

Road-centre lines ideally represent the geographic centre of the public rights of way network, a transportation network of all paths on which the public has a legally protected right to pass and re-pass. These transportation networks are based on vector line information and can be generated through a variety of GIS methods such as automated processes of on the ground collected GPS data, generative processes based on cadaster boundary data or manual tracing of roads on aerial photographs. In a subsequent step, additional information can then be attributed to this line information such as road names, road type, travel direction, road geometry information as well as a large variety of other possible attributes.

This makes road-centre line maps a powerful tool for a variety of GIS-based applications. The ones applied the most are transportation modelling and navigation routing. Road-centre line data was first provided by local governments, such as the TIGER[9] dataset by the United States Census Bureau or the ITN[10] by the British Ordnance Survey, as well as commercial companies, such as the Dutch Company TeleAtlas[11] or the American-based Company Navteq.[12] The latter mainly provides line-based data for navigational systems. With the rise of the Internet and Web2.0, [13] publicly accessible road centre-line information became widely available through different sources. The most dominant

---

[9] TIGER is an acronym for Topographically Integrated Geographic Encoding and Referencing and an American based format used by the United States Census Bureau to describe land attributes such as roads, buildings, rivers, and lakes, as well as areas such as census tracts. The TIGER format forms a base for the US part of the OpenStreetMap project.

[10] The Integrated Transport Network, is part of the OS MasterMap and a format provided by the United Kingdom governmental Ordnance Survey.

[11] TeleAtlas is since 2008 wholly owned by navigation system company TomTom.

[12] Navteq is since 2011 fully merged into NOKIA.

[13] Web 2.0, is a term describing the state of the Internet as a collaboration focused information platform, where the user produces content. The term is set against Web 1.0, where content was provided as ‚ready-to-use’ and no interaction with the user was aimed (O’Reilly 2005).

sources are Google maps and Bing maps, both available under restricted license for non-commercial usage. The opposite of this governmental and proprietary-based information with restricted license is volunteered geographic information (VGI). VGI describes all geographic data, which is created, assembled and disseminated voluntarily by individuals (Goodchild 2007). Open source VGI projects such as OpenStreetMap (OSM) and MapQuest are available under a GUP license and therefore freely accessible to anybody. Due to the increasing number of online participants all over the world these projects are on the rise and constitute a commercially as well as academically meaningful alternative.

In the context of space syntax analysis 2003, Thomson (2003) pioneered by proposing to make use of street networks. His study focuses on theoretical and technical problems based on the model construction rather than an investigation on how different models affect the analysis. In the study, Thomson highlights the possibilities of generalizing road networks. Simultaneously Dalton et al. propose to make use of TIGER data and present their initial results of their analytical work (2003). TIGER is a data format used only in the United States, providing road-centre line information among other geo-referenced spatial data. Dalton conducts a fractal analysis and compares a TIGER dataset with a traditional hand-drawn axial map of Downtown Atlanta, US. He highlights differences in the results of both models and concludes that these differences are caused by the very different representations of space. While a long linear avenue with adjacent side streets is represented by one long axial line in a traditional axial line map in the TIGER dataset, road centre-lines are segmented by nature and have a node at each intersection, which is the case for any road centre-line map. Any topological investigation would thus lead to a highly-skewed outcome. Moreover, Dalton raises the theoretical problem of radii, emphasising the need for a 'relativisation' due to the differences within each system (ibid. p. 9). While Dalton does not propose a solution to the problem, his argumentation led to a series of investigations by Alasdair Turner.

In his study from 2005, Turner develops a methodology that overcomes this problem of segmentation and lack of 'relativisation' by drawing on advantages of space syntax and applying ASA to road centre-line maps in combination with a segment length weighted algorithm. The results of his 2005 and 2007 study indicate that metric radii in combination with weighted choice measures present not only a suitable alternative to SM models but, in fact, generate better correlations with flow data in the tested case studies. Turner emphasises that his measure holds configurational information while incorporating plausible cognitive and physical constraints (2007 p. 553). Turner's findings are reasonable since road centre-line maps are fundamental representations of the accessible – rights of way – movement network and incorporate detailed angular information.

Dhanani et al. (2012) follow Turner's findings and conduct a comparative study of an axial line model and two different types of road centre-line based models. As mentioned previously, there are different sources for road centre-line maps. Dhanani et al. studies' focus on two very particular networks: the governmental ITN dataset and the OSM VGI data. Their studies aim to understand whether a VGI-based dataset constitutes a reliable alternative compared to governmental datasets in the light of space syntax analysis. Beside of Dalton's (2003) and Turner's (2005, 2007, 2009) work, there are no other comprehensive studies where space syntax measures are applied to governmental road centre-line datasets correlating results with empirical data. This is surprising as both of the studies rely either on the American TIGER data or the British Ordnance Survey datasets. The difficulty here is that governmental road centre-line maps are presented as a reliable and coherent source of data, yet, this is only true for information within one dataset [14] and very little is being said about their comparability in an international context.

Differences occur between governmental datasets not only on an international level but also within countries. The British Ordnance Survey, for example, provides three different road centre-line data products: the OS MasterMap layer Integrated Transport Network (ITN) layer, the OS Open Roads layer and the Merdian 2 layer. All these datasets provide comprehensive road network information and are designed for routing and road network analysis, yet, their level of precision and coverage differs. [15] This means that the total number of nodes and coverage of real-world details such as roundabouts are not the same throughout the three datasets. More importantly, such datasets are not available in every country. Germany, Italy and France – to name only some – do not provide freely accessible datasets. This is why the question of comparability needs to be answered and investigated for each country individually, and alternative sources need to be found. The lack of comparable data makes it difficult for international comparative approaches making use of such datasets, particularly in the context of space syntax.

5.1.3 ADVANTAGES AND DISADVANTAGES OF OSM DATA

In the light of this lack of comparable data, OSM data becomes more interesting as an appropriate alternative to a segment map representation, which, in theory, provides a comparable representation of space all over the world. OSM data is produced according

---

[14] It shall be noted that governmental datasets can also feature errors, but usually undergo rigorous quality checks prior to their release.

[15] See: http://digimap.edina.ac.uk/webhelp/os/osdigimaphelp.htm#data_information/os_produ cts/os_open_map_local.htm for further information on the datasets and examples of their application.

to a guideline indicating the level of precision and the handling of particular situations such as divided highways, roundabouts, intersections or bridges (OpenStreetMap Wiki contributors 2016). This makes the data, in theory, globally comparable. However, differences in terms of the data quality arise due to the nature of its production and its contributors' heterogeneous understanding of street networks.

Understanding such differences in quality is a non-trivial task in the realm of OSM data. There is a set of ISO standardized quality measures to assess the quality of map-based VGI (OSM) data. These measures are of particular interest for routing and navigation application, namely *positional accuracy* and *topological consistency* (Senaratne et al. 2016 p. 6) and thus for a space syntax application. *Positional accuracy* is a quantifiable value reflecting the difference between a mapped location and its real-world location while *topological consistency* measures how well as topological relations ('disjoin', 'meet', 'overlap' or 'equal') are mapped. A simple example for low positional accuracy would be a mapped intersection, of which the GIS location is 20 metres further North than in reality. An example for bad topological consistency of an intersection would be the case, in which two streets, which in reality are connected and should share a common node, would not be doing so in GIS. To evaluate the two mentioned quality measures it is necessary to compare the dataset under investigation with the real world. This is usually done by comparing the VGI data with ground-truth data. The ground-truth means data that represents the respective exact location in reality. This is a theoretical value, rather than a goal that is truly achievable for most GIS datasets. GPS systems feature on average a positional accuracy of 6-10 metres to ground-truth. The ordnance survey MasterMap ITN data states its *positional accuracy* with 1 metre in urban and 6 metres in rural areas against ground-truth.

Throughout the past decade, several authors have conducted comparisons of volunteered geographic information with governmental as well as commercially produced geographic information (Flanagin and Metzger 2008; Ludwig et al. 2011; Neis et al. 2010; Zielstra and Zipf 2010)[16] to measure their quality. In the context of road centre-line information, the work by Mordechai Haklay was one of the first to evaluate the quality of OSM data (2010). Haklay used the British OS Merdian 2 road network as the control measure to test OSM data quality. His findings indicate highest mapping quality in urban and affluent areas and the lowest coverage in rural and poorer areas, while positional accuracy ranges from over 70% to occasionally drop down to 20% (ibid. p. 700). Overall OSM data covered 29% of England based on a network from March 2008.

---

[16] See Sehra et al. (2013) and Senaratne et al. (2016) for a comprehensive review of studies dealing with quality assessment of VGI data.

In a subsequent study, conducted in October 2009, this percentage was already corrected to 65% of coverage (Haklay 2009). This indicates a growth of the network coverage by 36% within one year. Another study by Neis et al. (2011), dealing with the case of Germany, compared the OSM network against the proprietary dataset of TomTom (formerly TeleAtlas) and estimated a complete coverage of the German OSM data by the year of 2012. Moreover, already in 2011, the OSM data exceeded the topological consistency and completeness of the TomTom network by 27% including pedestrian pathways (ibid.). The continuous growth and the pace at which the OSM dataset grows, does not only make a coverage and quality assessment difficult but indicates that it is only a matter of time until a full topological consistency will be reached. The number of total users in the OSM community as well as their nodal contribution to the network shows a growth of the total user number to 2,9 million since the start of the project 2004 and gives insights in the pace of this process.
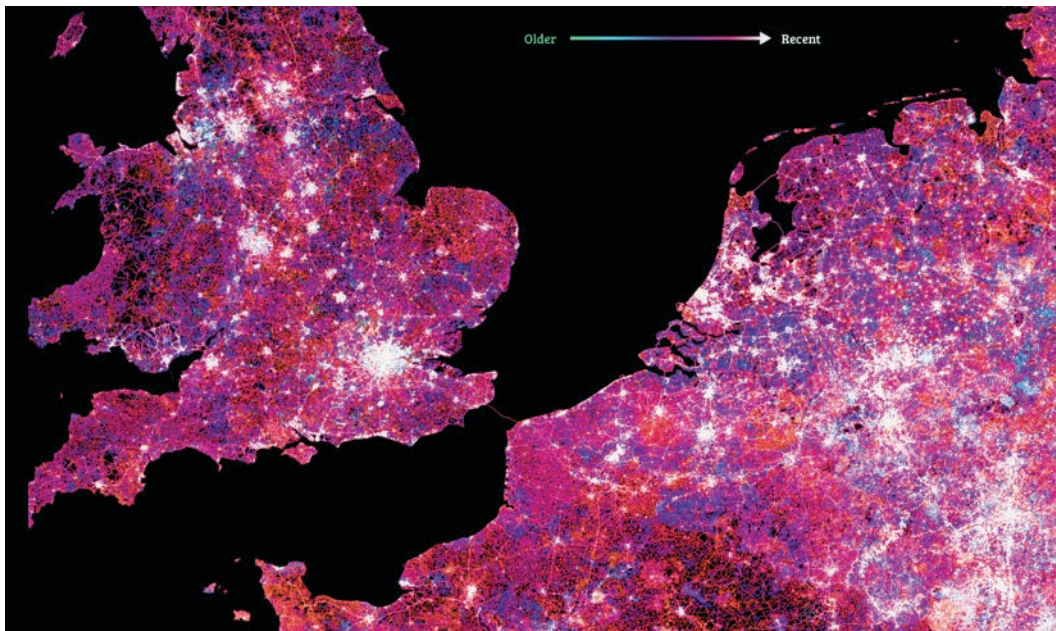
Figure 36: Visualizing road updates. All roads shaded by how recently they have been updated by users. Older imports are in green and blue, while cities with strong and active communities and the effect of recent automated editing makes areas glow red. (2013) Source: https://www.mapbox.com/osm-data-report/ (retrieved on 1 August 2016)

Hakley et al. (2010 p. 11) investigate how many volunteers are needed to map an area thoroughly concluding that areas mapped by more than 15 contributors per square kilometre feature a very good positional accuracy of below 6 metres for resulting VGI data. In regard to the growing numbers of contributors, this leaves us to expect an equal rise in topographic consistency and positional accuracy. An additional positive effect on the coverage of areas, beside the growing number of contributors, is the fact that governmental agencies increasingly provide their data for public usage. Likewise, are the American TIGER network as well as the AND Dutch road network fully implemented

in the OSM network, contributing not only to the coverage but positional accuracy of the OSM dataset. A visualised snapshot of the data and its topicality reveals updating intervals, as well as showing that Great Britain and Germany are part of the best-mapped countries of the OSM project (Figure 36). All of the above studies use ground-truth data for the evaluation of VGI quality. Still, such data is not available in every country, and more difficulties for the assessment of VGI data arise due to the lack of ground-truth data for comparison (Senaratne et al. 2016 p. 6). To overcome this lack of ground-truth data, Keßler and de Groot (2013) propose a method to indicate the quality of VGI via trust assessment models. Their approach is based on a trust assessment model of the independent contributions in an OSM dataset. Albeit presenting promising results, the methodology is at an early stage of development and does not propose an applicable method for the field. At the present stage, this leaves the research with an as-good-as complete network for some countries with reasonably accurate precision, but a manual control of the entire dataset by the researcher remains a necessity. With regard to future research, the OSM will very likely constitute the most coherent freely available dataset.

Dhanani et al. (2012 p. 30), assess the usage of OSM in space syntax to be problematic and describe the data as lacking 'of consistency [,…] accuracy and coverage'. Their study calls on the researcher to rely on governmental data such as the British OS MasterMap ITN. As mentioned earlier, because data is not accessible in every country and level of detail but also differs throughout different datasets, this approach remains unsatisfactory: the OS MasterMap ITN network covers only the vehicular network disregarding any path or street that is accessible only to pedestrians. The resulting vehicular centred spatial representation can therefore only be used to evaluate vehicular structures. Space syntax segment map representation, on the other hand, sees space through the eye of an individual moving in space and constitutes a sharp contrast to a vehicular only street network. There are also other difficulties within the ITN dataset that render an ad hoc use impossible. Dhanani et al. note that the ITN network comprises all traffic management features including traffic islands, artificial cul-de-sacs or roundabouts (ibid. p. 6). According to the authors, using such data creates a 'disjoint and fragmented network', particularly if a researcher is interested in other modes than a purely vehicular estimation. The usage of such data is not recommendable without any prior processing. Prior processing is also necessary for OSM data making it indispensable to develop a strategy to overcome said inconsistency and arrive at a comparable network for any given case.

5.2 OSM DATA STRUCTURES AND GIS SIMPLIFICATION PROCESSES

The following section gives an overview of the necessary components to create a road network based on OpenStreetMap data and the necessary steps of post processes to allow an application in space syntax ASA.

5.2.1 OSM DATA STRUCTURE

At present, OSM datasets are divided into four different elements: nodes, lines, surfaces and relations. For an ASA only line information is necessary, but not all of the available line information and categories are useful. The OSM wiki provides extensive accounts on all different key categories and their morphology (OpenStreetMap Wiki contributors 2017), it is important for each researcher working with OSM data to make herself familiar with all categories and morphologies. Decisions about which category to exclude might differ, for example, in cities in developing countries. The following steps should be considered as a general guidance: for the purpose of network analysis, only components with the key *highway=** shall be used. This key defines any kind of road, street or path and their respective importance in the network hierarchy (from the most important 'motorway' to the least 'service') and, thus, gives a good account of the rights of way network. The following list assesses which of the following elementd are recommendable to be included in a network for an application in ASA: *highway=motorway; trunk; primary; secondary; tertiary; unclassified; residential; motorway_link; trunk_link; primary_link; secondary_link; tertiary_link; living_street; pedestrian* (ibid.). Particular care needs to be taken with the key *pedestrian* as it includes pseudo polyline information of squares and these needs to be cleaned and afterwards broken down into individual segments. Other sub keys such as highway=*service; path* or *bridleways* can be included but are not recommended, as they are of a very small scale and might otherwise be eradicated in a subsequent simplification process.

With a view to this selection of data, there are three main difficulties that occur when the data is applied in a space syntax context.

1. Topological inconsistency occurs if street segments are supposed to share a connecting node but due to positional inaccuracy fail to do so. This is often the case at intersections of different contributors. Even a small gap between two nodal ends of 1 cm can create a network fragmentation. It is, therefore, necessary to process and clean the data from these inconsistencies.

2. Traffic management components are network details that are necessary for vehicular traffic management but have no immediate impact on cognitive route decision-making. Such details are for example roundabouts, small traffic islands or motorway trunks. Ideally, roundabouts are simplified into simple intersections whereas meandering trunk links are represented by single links. Moreover, this is also the case of dual line representations. Space syntax analysis is a non-directional approach in the sense that the possible travel directions are not taken into consideration, and each space is treated as equally accessible. A dual line representation constitutes only a reasonable option if directions are taken into consideration. Hence, the model needs to be cleaned from said dual line representations.

3. Redundant or excessive nodal information is often problematic when using OSM data. Although the OSM guide notes that nodes should be used economically, contributors often have different interpretations of what 'economic' means. This is particularly the case for curved roads but also occurs on straight lines. Ideally, each street is simplified to its fundamental segment.

5.2.2 GIS ROAD-NETWORK SIMPLIFICATION PROCESS

To overcome these difficulties, I developed a series of GIS algorithms. The following solutions that I propose are employing the GIS software ArcGIS Desktop 10.2 from Esri. I employ ArcGIS because it is the only software that provides solutions for all three said difficulties. At present, only a few of the solutions presented here can be achieved with open source GIS software packages. Figure 37 shows a workflow diagram for the proposed solutions, while Figure 38 provides an illustration of each obstacle and its favoured solution after the application of the simplification method presented here.

1 Topological
Inconsistency
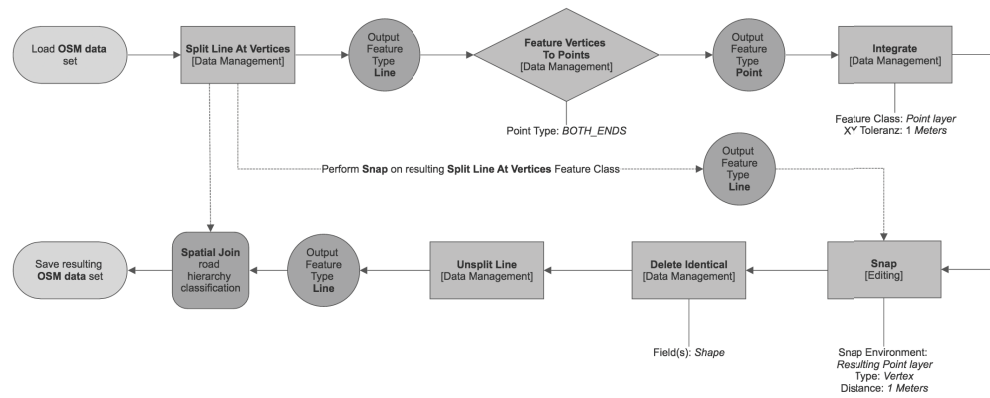
2a Dual Line
Removal
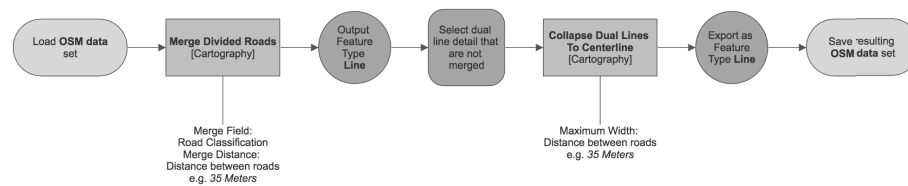
2b Road
Detail
Removal
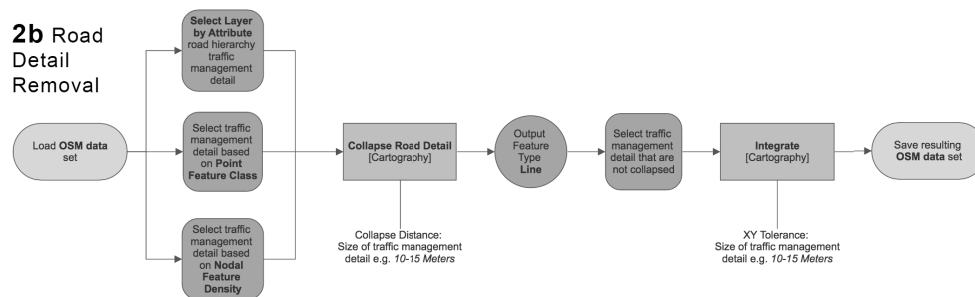
3 Line
Simplification
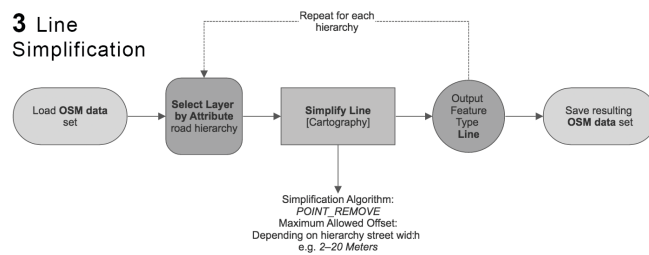
Figure 37: Workflow of ArcGIS tools and algorithms to solve: 1. topological inconsistency; 2a. dual
line removal; 2b. road detail removal and 3. line simplification

1 Topological Inconsistency
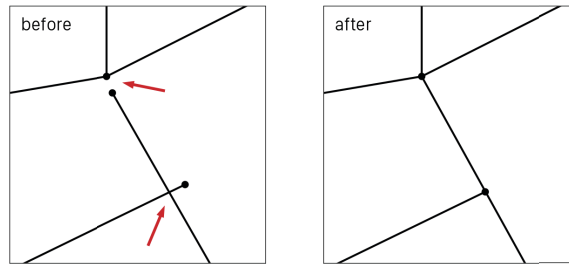
2a Dual Line Removal
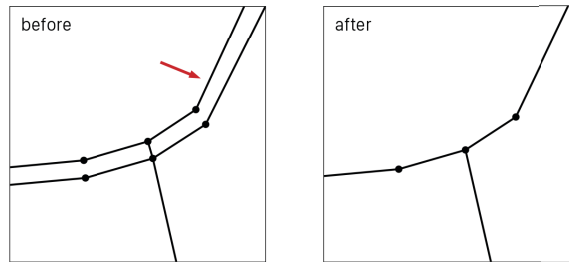
2b Road Detail Removal
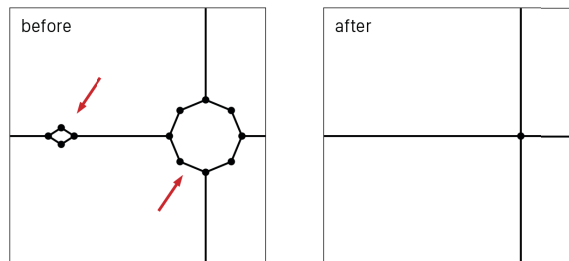
3 Line Simplification
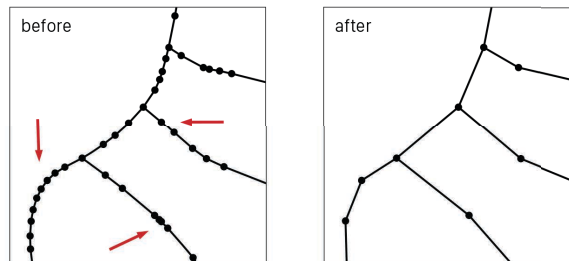
Figure 38: Illustration of each difficulty found in OSM data: 1. topological inconsistency; 2a. dual line removal; 2b. road detail removal and 3. line simplification as well as the condition after application of the simplification method.

1. Starting with the approach of solving topological inconsistency (Figure 37:1), it must be noted that a lack of network information, such as entirely missing streets cannot be solved through automated processing and that the OSM data needs to be carefully checked by the research prior to any post-production. Moreover, this is a strategy to overcome small inconsistencies that are difficult to identify manually. It will reconnect topological inconsistencies by a given tolerance distance and in a following step merge segments that can be considered as independent streets, from intersection to intersection. This will provide the researcher with a street network of real segments and

consistent topological information. The two core ArcGIS functionalities the workflow is based on are '*integrate*' and '*unsplit*'.

The *integrate* tool is applied to extracted nodal information, rather than the actual line information, to overcome misalignment at intersections. *Integrate* maintains the integrity of shared nodal feature information by making features coincident if they fall within the specified x,y tolerance. Features that are considered identical or coincident are merged. In a next step, the newly generated nodal point information is used as a basis for a *snap* command of the initial street network. This will connect lines that feature topological inconsistency at a new point based on the location of their nodal line ends.

The *unsplit* tool is then applied to the now topologically consistent line network. The aim is to aggregate single-part line features into multi-part features, to arrive at continuous street segments. *Unsplit* merges lines that have coincident endpoints. This can be done by relying on any given attribute information, or as in this case solely by geometric relationships. Merged lines are of particular importance with regards to further simplification processes.

2. The next difficulty is the existence of traffic management details and dual line representations in the datasets (Figure 37:2a & 2b). Not only do such details (roundabouts, traffic islands, etc.) create differences in angular movement, while the general journey direction stays the same, but more importantly they increase the total number of journeys (dual line highways) and skew analytical results towards an emphasis of such details. Especially in the light of none directed centrality analysis, dual lines make little sense. This could be negligible if traffic management details are normally distributed throughout the street network. This is, however, not the case in most examples and particularly not on inter-city and regional scales. There are four main ArcGIS components, '*merge divided roads*', '*collapse dual lines*', '*collapse road details*' and '*integrate*' that help to remove such dual lines and reduce low-level street network complexity.

The *merge divided roads* is an algorithm that merges road segments that are parallel along a significant distance into a single centre line. The merging process is based on common attributes that can be computed on the base of the initial highway keys. It is fundamental that the merge field parameters are established properly to avoid conflicts during the process. The divided roads algorithm can be applied to the entire datasets and maintains topological relations with adjacent streets.

The *collapse dual lines to road centerline* is an algorithm designed to derive centre lines from a base of street perimeters. It is hence a less sophisticated form of simplification, and it is not recommended to perform the algorithm on large datasets including

multiple-lane highways with interchanges, ramps, overpasses and underpasses. In individual cases, where the *merge divided roads* tool does not arrive at satisfactory results, the *collapse dual line to road centerline* tool can form a useful alternative.

The *collapse road detail*, on the other hand, is an algorithm that describes small road segment details and open configurations that interrupt the general trend of a road network and collapses or replaces them with a simplified feature. The collapse distance on which the tool performs is defined by the maximum size of the largest road detail and can differ for each model. If the *collapse road detail* tool does not solve or remove some of the details, the *integrate* tool explained earlier, becomes an appropriate alternative. Particular care needs to be taken when using *integrate* on road details as it can impact the topological consistency of the data and should hence not be performed on entire datasets, but single cases.

3. Line simplification is usually applied when segment records feature far more data than necessary for computer analysis or visual representations (Figure 37:3). In the case of space syntax and the use of VGI street networks, this is aside from excessive data also a conceptual question. While road-centre lines depict the centre of the road, an axial line (as the base for a segment map line) is based on the longest line of sight. A generic street usually features a much larger field of vision than that of a single line. While axial lines fundamentally connect convex spaces, these lines naturally pervade more than one space at once. Road-centre lines, on the other hand, simply represent the centre of the road and, hence, feature excessive angular information that does not impact the field of vision or accessibility and thus have no effect on the actual movement in space. A removal of such road details should be based on the field of vision of each street, namely the street width. Since road-centre lines give a precise account on the centre of each street segment, a simplification process should allow the newly generated feature to deviate from at least the extent of the field of vision. Such a process can be performed through the Douglas-Peuker Algorithm (DPA) (1973). The DPA is broadly recognised to deliver the best perceptual representations of the original segment and generates new segments based on a deviation tolerance. In ArcGIS this can be done by applying the *simplify line* tool.

The *simplify line* tool reduces and removes redundant nodes of line features. Among others when applied with the POINT_REMOVAL functionality it employs the DPA. The aim of the algorithm is to extract the essential segment form based on a previously selected off-set tolerance. The strength of the algorithm is its reproducibility and process speed, arriving at the same solution for the same problem.

If all said steps of the methodology are followed the simplified version of a road-centre line map (SIMP) will look visually as well as topologically much closer to an axial line representation.

5.3 SIMPLIFIED MODEL EVALUATION

To test if the theoretically laid out version of a simplified OSM network (SIMP) constitutes a comparable alternative to a segmented axial line map, which would make it suitable for an analysis of different scales and very large ones in particular, the model will be analysed and correlated with results from an ASA of a segment map, ITN and OSM model. The comparison builds on the methodologies of Eisenberg (2007), Turner (2007) and Dhanani et al. (2012) but also extends them.

Eisenberg (2007 p. 5) focused on the comparison of different axial line models for the same cities. The different models that Eisenberg compares are developed as a by-product of variations in analytical scales (pedestrian, bicycle and vehicular) and variations in the detail of the base information used for the production of the axial line maps. He highlighted that three indicators are most interesting for a comparison. First, the impact of base map scales, second, different levels of detail and third, different city morphologies (ibid. p. 5). All aspects are directly transferable to the different network models previously introduced. His findings suggest that in order to have a meaningful comparison, the analysis should focus on 'rank correlation measures' (ibid. p. 8). Eisenberg's 'rank correlation measures', apply to every kind of network representation. It simply compares values and their respective rank within the dataset. It therefore constitutes an appropriate method for the analysis, for which we are aiming, where numbers of lines differ significantly and the resulting values do not form a comparable unit.

In addition to 'rank correlation', this comparison will draw on the methodology of Turner (2007). Turner proposed an *angular*-based analysis in combination with *segment length-weighting* and the introduction of a *metric length* based *radius*. While an angular-based analysis incorporates the cognitive dimension of route choices, the reason behind a *segment length-weighting* is to overcome the large differences in segment numbers between the different representations (ibid. p. 541). Turner showed how his propositions advance space syntax analysis in general, but in particular in the context of road-centre line networks.

Finally the methods proposed above will be merged with a methodology by Dhanani et al. (2012). The authors conducted a comparison of road-centre line networks with axial line models, using a general description of the network characteristics, followed by a

topological and metric step depth analysis from the most central segment. Although the outcome of the topological step depth showed interesting results, the application of topology on a road-centre line network seems inappropriate, as road-centre lines topological information is highly skewed by its nodal information. The measure of topology in space syntax analysis is based on the cognitive and visual space; in that sense what is considered as one space in space syntax would result in several spaces in a road-centre line network. This analysis will hence draw only on the measure of metric step depth (MSD) for comparisons, as MSD is not affected by nodal information.

To summarise, the following comparison is based on four different road network models of the centre of the city of Leeds. The city of Leeds was selected because it features a variety of different network details, such as motorways, traffic management details as well as local paths. The road network models are the Ordnance Survey ITN network, the OSM network, a simplified version of the OSM (SIMP) and a segmented axial line model (SM). The ITN network and the OSM data are not simplified and used as they are provided by the organisations. Moreover, the ITN and OSM networks where controlled on topological consistency and no irregularities were found. Some network categories, such as were pointed to in the OSM data sections, have been removed from the OSM dataset, while traffic management details have been left as they were. The four models are compared on their network characteristics and analysed on 14 different radii, from *100* up to the entire system *n*,[17] using angular segment analysis with segment length weighting. The models are analysed on closeness and betweenness centrality. The resulting structures from three exemplary scales are compared visually and the resulting correlations are calculated using 'rank correlation measures'. To make a comparison possible mean values of coincident segments of the ITN, OSM and SIMP with the SM model are plotted on each respective SM segment.

---

[17] The applied scales are: *100, 150, 200, 300, 500, 800, 1300, 1800, 2500, 3200, 4100, 5000, 6100* and *n*.
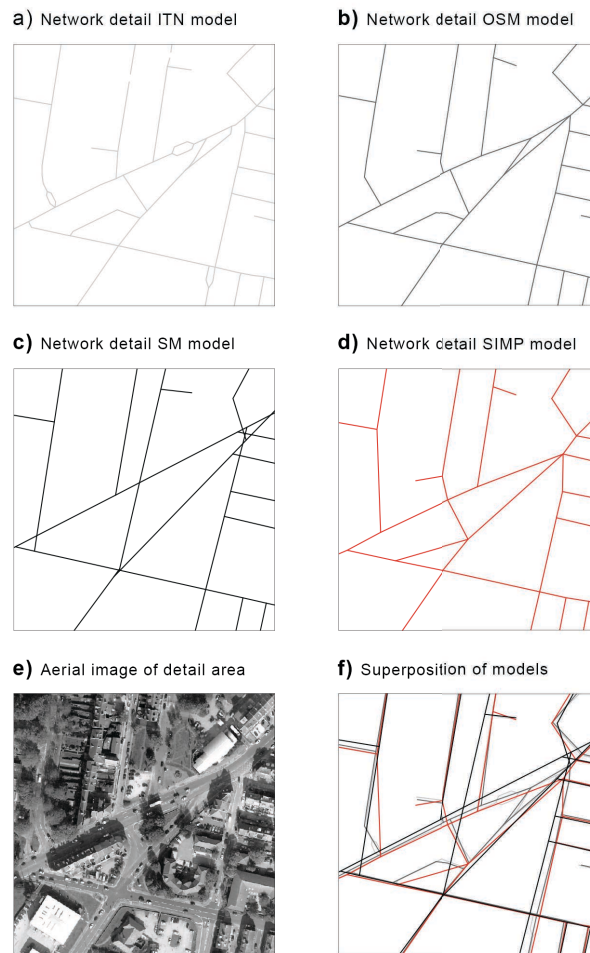
Figure 39: Detailed section of different network models of ITN, OSM, SIMP and SM model.

Figure 39 shows a small section of each of the modelled areas. The section of the ITN network shows traffic islands, as well as road interruptions. Moreover, some roads have significant angular turns just before their connection with the adjacent road. This is because for traffic management purposes rectangularity is preferred. In the light of angular segment analysis, Dhanani et al. (ibid. p. 10) have argued that this forms an important aspect and the most detailed and 'optimal' account of the street network. The aerial photo of the area (Figure 39:e) shows that at this point a straight connection is a more reasonable account of the real world situation. Also, visible at the lower right, the roads diverge into two separate lanes. A noteworthy detail is also that roads, which one might consider as intersecting in reality, do not share a common node in the road network, due to a 5-10 metre distance of their road-centre.

Table 17: Network characteristics for each model.

Statistics	ITN	OSM	Axial	SIMP
Segments	15049	9308	5072	3908
Total length (m)	283410	276388	240534	238848
Computation time (min)	14.31	4.49	1.21	0.44

A view on the network characteristics for the four models highlights their differences numerically (Table I). The ITN network features the longest total network length with 283410 metres. This is particularly caused by the several roundabouts and traffic management details within the model. One can get a rough account of the effect on the length of the network through traffic management details by comparing the length of the ITN and OSM networks, yet not to the fullest as the OSM network features streets and connections that are not represented in the ITN. The large difference of 40km of the ITN and OSM data in comparison to the segmented axial model is caused by the several multi-line motorway roads, which are represented by a single segment in a segmented axial line and SIMP model. Striking is the difference in the number of segments compared through all the networks. The ITN model has three times more segments than the segment map representation. This is because curved roads and roundabouts feature large numbers of segments to give precise accounts on the length of the lines. While this exemplifies the detailed account on angular changes in road centre-line networks, it also shows the inherent problem of this data when it comes to space syntax analysis. The computational time is $O(n^2)$ raised by the number of segments. Generally speaking, the ITN and OSM are similar in their measures, and the difference in the number of segments is as expected. With a view on the segmented axial line and SIMP model, an arising question is if the SIMP model stores less information as it features significantly fewer segments. This can be explained with the 'cleaning' of intersecting spaces. Whenever three segments intersect with each other, segmented axial line models tend to create a cluster of very short segments, also stubs that fall over 40% of the line length during the 'axial line to segmented axial line' transition, might contribute to this difference. The SIMP model features almost the same length as the segmented axial line model, hinting towards a similar degree of spatial representation.

Histogram of Segment Length Distribution for Different Models

a) Historgram for ITN model

Mean = 18.83
Std. Dev. = 22.42
N = 15,049

b) Historgram for OSM model

Mean = 29.68
Std. Dev. = 28.40
N = 9,308

c) Historgram for SM model

Mean = 47.41
Std. Dev. = 46.21
N = 5,072

d) Historgram for SIMP model

Mean = 61.05
Std. Dev. = 45.28
N = 3,908
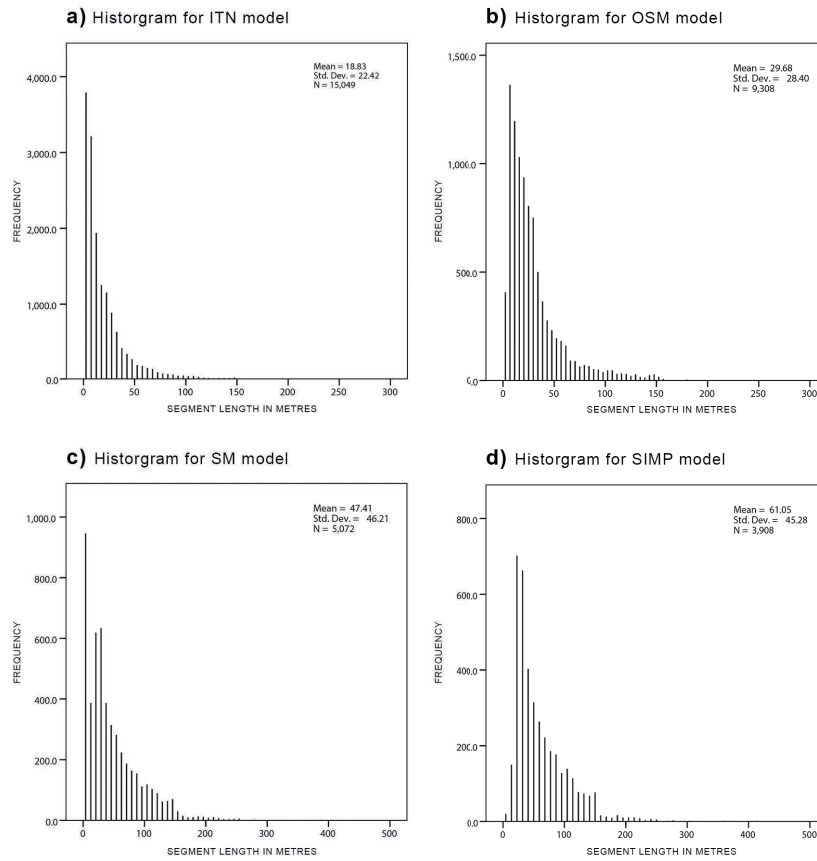
Figure 40: Histogram of segment length distribution of each of the four models.

These observations become pronounced with a view on histograms for segment length distribution for each model type (Figure 40). While the ITN network has an even increase of segment length with declining frequency, the OSM shows an initial increase indicating fewer stubs and curve segmentation than the ITN. Moreover, the short line cluster effect of the SM model becomes visible with almost thousand segments in the range of approximately 1-10 metres. Contrary to this, the SIMP model has a steep increase of frequency with a peak at a mid-range of approximately 30 metres, indicating less short line information. The simplification range used during the simplification process influences this peak.

Dhanani et al.'s (ibid. p. 25) study has shown that differences between road centre-line network and axial line models are consistent in their appearance and concluded that the different models do not form a fundamentally different structure of the spatial configuration. In the next step, I will compare the new SIMP model with this assumption. Figure 41 shows the number of segments for nine different radii; the maximum is 2,5km as this is the distance at which the entire system was captured (in other words n). For the four models, the total number of segments reached per metric distance increases in relation to the total number of segments. The semi-log plot

168

highlights these similarities and differences, especially at lower scales. The SM and SIMP model, exhibit a similar development, while the OSM and ITN – that were similar at the beginning – disperse towards the growing metric distance and increase of network details. Differently to the values for the central segment, the curve for the edge segment shows a slightly uneven development. This is better visible in the semi-log plot of the data. Here, particularly the development around the scale of 500 metres reveals that there are underlying differences in the complexity of the models, which might affect the analysis.
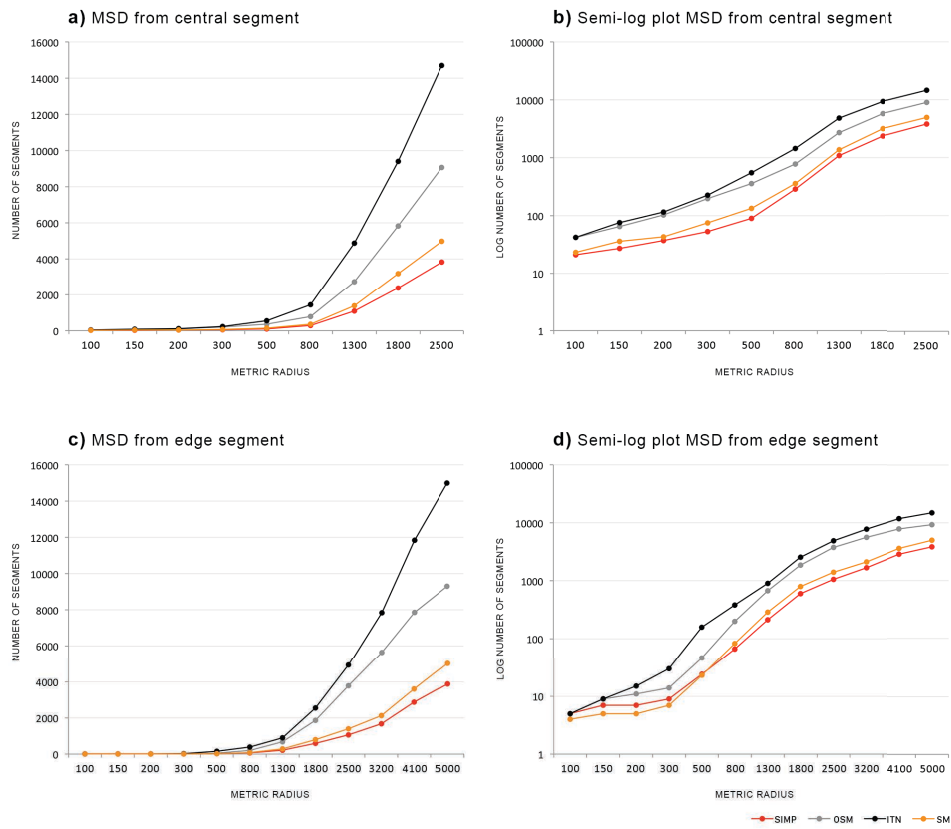
Figure 41: a) Number of segments for different metric step depth from the most central segment for ITN, OSM, axial and SIMP models. b) Semi-log plot of the same dataset. c) Number of segments for different metric step depth from an edge segment for ITN, OSM, axial and SIMP models. d) Semi-log plot of the same dataset.

To arrive at a better and more detailed account of the impact of differences in the network morphologies, I compare betweenness and closeness centralities using a segment angular analysis with segment length weighting. The models are analysed on 14 different radii. The applied scales are *100, 150, 200, 300, 500, 800, 1300, 1800, 2500, 3200, 4100, 5000, 6100* and *n*. Two of these scales, namely *800* and *n* are visualised in order to understand the geographic distribution of differences. Figure 42 shows the results for betweenness centrality, while Figure 43 shows the results for closeness

centrality. The values of each figure are broken down using a quantile colour break division. This is done to overcome significant outliers in the datasets that make a natural break highly skewed and the resulting maps illegible. These circumstances make it necessary to process the data in a GIS programme rather than applying the implemented symbology of depthmapX.

1 ASA SLW Betweenness Centrality
radius metric 800

1a **ITN**

1b **OSM**

ITN Integrated Transport Network
OSM OpenStreetMap
SIMP Simplified OSM
SM Segment Map

1c **SIMP**

1d **SM**

2 ASA SLW Betweenness Centrality
radius n

2a **ITN**

2b **OSM**

ITN Integrated Transport Network
OSM OpenStreetMap
SIMP Simplified OSM
SM Segment Map

2c **SIMP**

2d **SM**
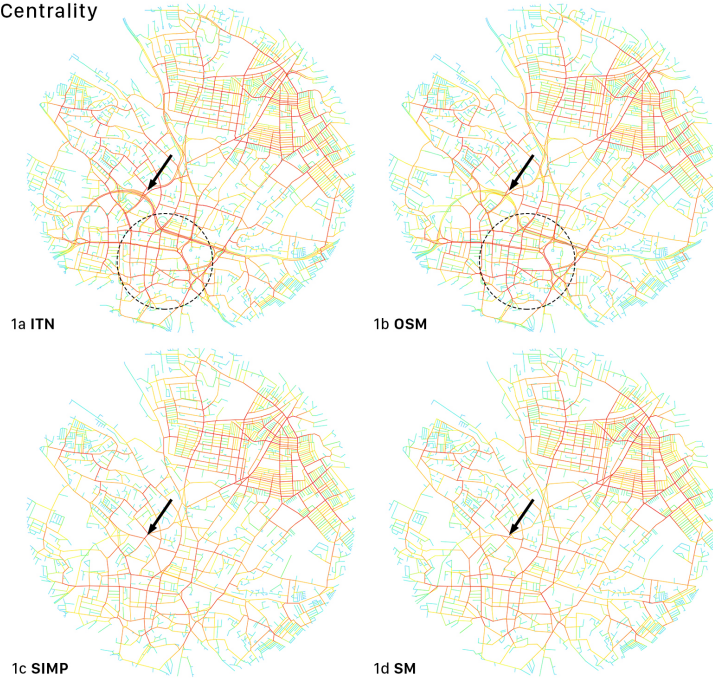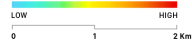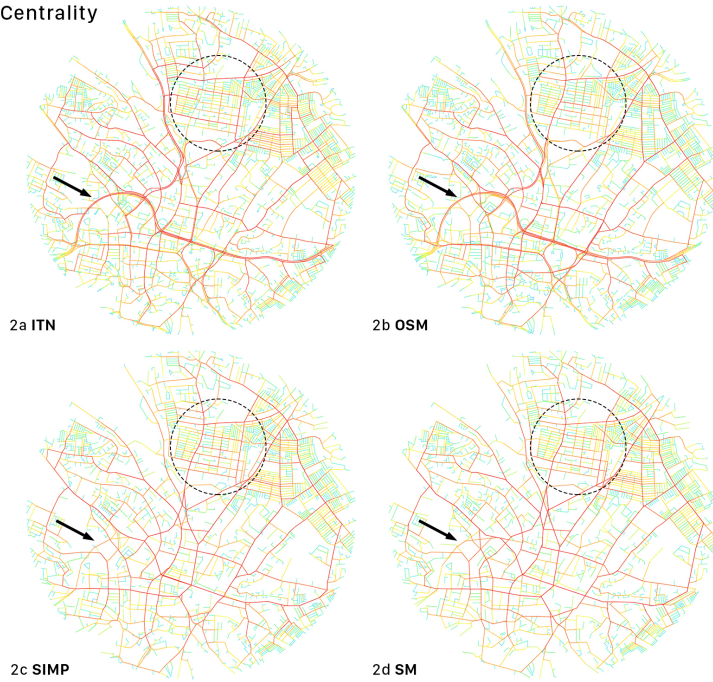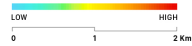
Figure 42: 1) ITN, OSM, SIMP, and SM models analysed on ASA betweenness centrality on radius metric 800. 2) ITN, OSM, SIMP, and SM models analysed on ASA betweenness centrality on radius metric n.

3 ASA Closeness Centrality
radius metric 800

LOW — HIGH

0 — 1 — 2 Km

3a **ITN**

3b **OSM**

3c **SIMP**

3d **SM**

ITN Integrated Transport Network
OSM OpenStreetMap
SIMP Simplified OSM
SM Segment Map

4 ASA Closeness Centrality
radius n

LOW — HIGH

0 — 1 — 2 Km

4a **ITN**

4b **OSM**

4c **SIMP**

4d **SM**

ITN Integrated Transport Network
OSM OpenStreetMap
SIMP Simplified OSM
SM Segment Map
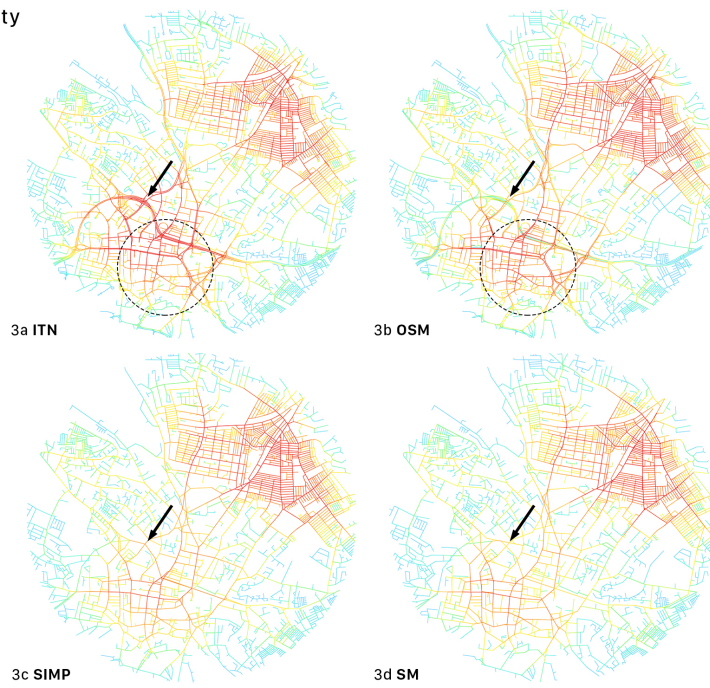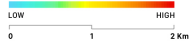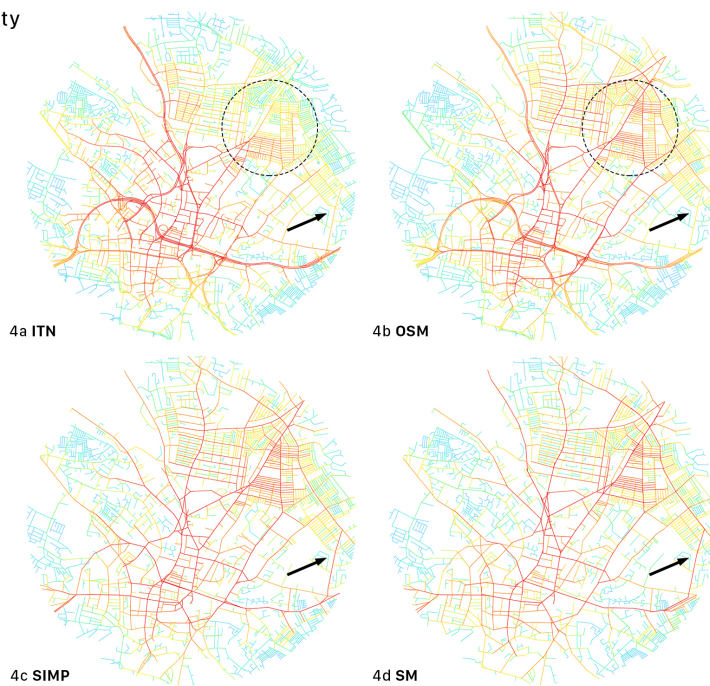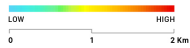
Figure 43: 3) ITN, OSM, SIMP, and SM models analysed on ASA closeness centrality on radius metric 800. 4) ITN, OSM, SIMP, and SM models analysed on ASA closeness centrality on radius metric n.

The results show that all models exhibit comparable patterns on all of the two visualised scales and both measures of betweenness and closeness centrality. This corroborates the initial findings of Dhanani et al. (2012). However, similarities in the results are much stronger between the OSM network and the SM than they are between ITN and SM.

172

Nominal segment differences appear to have a higher impact on betweenness centrality than on closeness centrality. Models with large numbers of short segments and a high degree of precision, such as the ITN network, are, thus, more likely to be affected by outliers and unexpected clusters, than models with fewer short segments. Moreover, the ITN network shows high values on all scales in the motorway network. The SIMP model showed patterns that were visually more strongly related to the SM model than to the ITN or OSM and more similar to the OSM compared with the ITN. This is rather unexpected, as SM models are thought to be intrinsically different.

After gaining an insight into differences and similarities in the geographical distribution of the data between the different models and the SIMP model, in particular, I conduct a final analysis of the statistical extent of these observations. This analysis will give an account of how each model behaves in comparison to each other across all scales. As elaborated earlier, the analysis draws on Eisenberg's proposed 'rank correlation measure'. Differently, to Eisenberg, this analysis will compare all segments that are intersecting, giving a more detailed account rather than only 10% of highest values proposed by Eisenberg (2007). This is done, by plotting mean values of the ITN, OSM and SIMP on the SM model. The SM model is used as a base and comparisons are only conducted at streets whose middle point falls within a 10-metre distance of an SM segment. These middle points are then snapped to the closest segment and plotted on the SM model. If more than one street segment of an ITN, OSM or SIMP model falls into this category, their mean is calculated and plotted on the SM model instead.

Eisenberg's rank correlation is based on Spearman's Rank correlation (ρ). Usually, Spearman's Rank correlation coefficient is used to identify and test the strength of a relationship between two datasets. The correlation method tests if the relationship of both variables can be described by a monotonic function. Ideally, the SIMP model could estimate the segmented axial line model through such a monotonic function. In addition to this, I will calculate a Pearson correlation. Rather than correlating the different ranks of each variable, a Pearson correlation works with the actual values of the variables and measures their linear correlation. Both correlations provide a coefficient of r^2, indicating how related the variables are with each other. A coefficient of 1 indicates that the two models are identical. Any value below 1 describes the degree of difference. One can hence compare the differences between all models statistically and provide a correlation coefficient to describe the fitness of the SIMP model for space syntax ASA. The analysis is based on 14 different scales for both space syntax measures of betweenness and closeness centrality. Figure 44 and Figure 45, show Pearson and Spearman correlations of ITN, OSM and SIMP compared with the segmented axial model and subsequently the same for all models correlated with the SIMP model.

Starting with Figure 44 the findings from the initial visual description are also displayed statistically. A first observation is that the Spearman rank correlation provides more consistent results across scales and measures, with weaker differences and higher scores. Pearson correlation, on the other hand, show much stronger differences of the four datasets, but features a significant outlier on the scale of 100 metres for closeness centrality. Regarding the single models, the ITN model shows lower correlations across both Pearson and Spearman measure and on both betweenness and closeness centrality. Particularly interesting is the significant drop towards higher radii, with the lowest correlation of 0,56 on Pearson for betweenness and closeness. This increases at Spearman's rank, however, the general tendency towards lower correlations at higher radii persists. In terms of the visual observations made earlier, this is caused by traffic details and the strong representation of motorway features. The OSM and SIMP model on the other hand show very comparable correlation developments, an exception constitutes the Pearson correlation for betweenness centrality of the OSM model, where similar to the ITN a sudden drop is visible at higher radii. The SIMP model correlates higher across all measures, with the highest scores of 0,983 for Spearman correlations of closeness centrality metric 1300 and 0,919 for betweenness centrality. Contrary to OSM and ITN, the correlations for SIMP are very consistent.

Pearson correlation of OSM, ITN & SIMP against SM for
ASA Segment Length Weighted Betweenness & Closeness Centrality

a) ASA SLW Betweenness Centrality

b) ASA SLW Closeness Centrality

Spearman correlation of OSM, ITN & SIMP against SM for
ASA Segment Length Weighted Betweenness & Closeness Centrality

c) ASA SLW Betweenness Centrality

d) ASA SLW Closeness Centrality
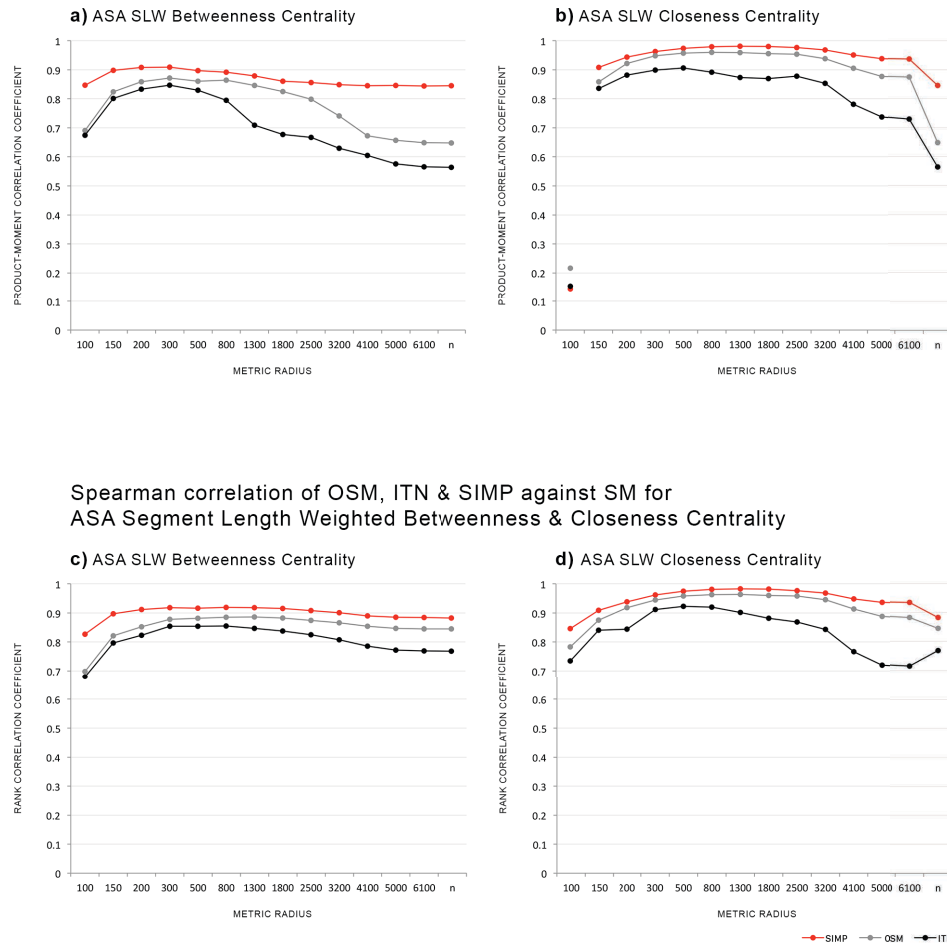
Figure 44: R2 of a Pearson correlation for ASA segment length weighted betweenness centralities (1a) and closeness centrality (1b) for 14 different metric radii (from 100 metres to n) for the three different network models SIMP, OSM and ITN against the SM model. 2: R2 of a Spearman correlation for ASA segment length weighted betweenness centralities (2a) and closeness centrality (2b) for 14 different metric radii (from 100 metres to n) for the three different network models SIMP, OSM and ITN against the SM model (left). Correlation is significant at the 0.01 level (2-tailed), N=3172.

Figure 45 shows Pearson and Spearman correlations for 14 different scales and closeness and betweenness centralities, this time, however, ITN, OSM and SM models are compared with SIMP. The general correlation developments are very similar to the ones we have observed previously, with a gradual decrease of values towards higher radii. Interesting is at this point how ITN and OSM behave compared to the SIMP model. While the ITN networks show a slightly weaker correlation, the OSM correlates much stronger. This could be expected, on the one hand, as the SIMP model is entirely based on the OSM, on the other hand, in the light of the overall comparison, it seems as if the simplification process were to bring the simplified OSM model much closer to the segmented axial line representation than expected.

Figure 45: R2 of a Pearson correlation for segment length weighted betweenness centralities (3a) and closeness centrality (3b) for 14 different metric radii (from 100 metres to n) for the three different network models axial, OSM and ITN against the SIMP model (left). 4: R2 of a Spearman correlation for segment length weighted betweenness centralities (4a) and closeness centrality (4b) for 14 different metric radii (from 100 metres to n) for the three different network models axial, OSM and ITN against the SIMP model (left). Correlation is significant at the 0.01 level (2-tailed), N=3172.

These differences become more striking in the log-log scatterplot of betweenness and closeness centrality of the global scale n (Figure 46). The diagram shows a log-log scatterplot of each of the measures, allowing a visual comparison of outlier distribution within each dataset. The more dispersed the values are, the less they are correlating, while linear consolidation implies stronger correlations. This is clearly visible for the log-log plot of axial and SIMP, while both other models show a stronger dispersion. The ITN model shows outliers across the values from low to high. This is particularly the case for closeness centrality.

176

Log-log plots of OSM, ITN & SIMP against SM for
ASA Segment Length Weighted Betweenness & ASA Closeness Centrality
radius metric n
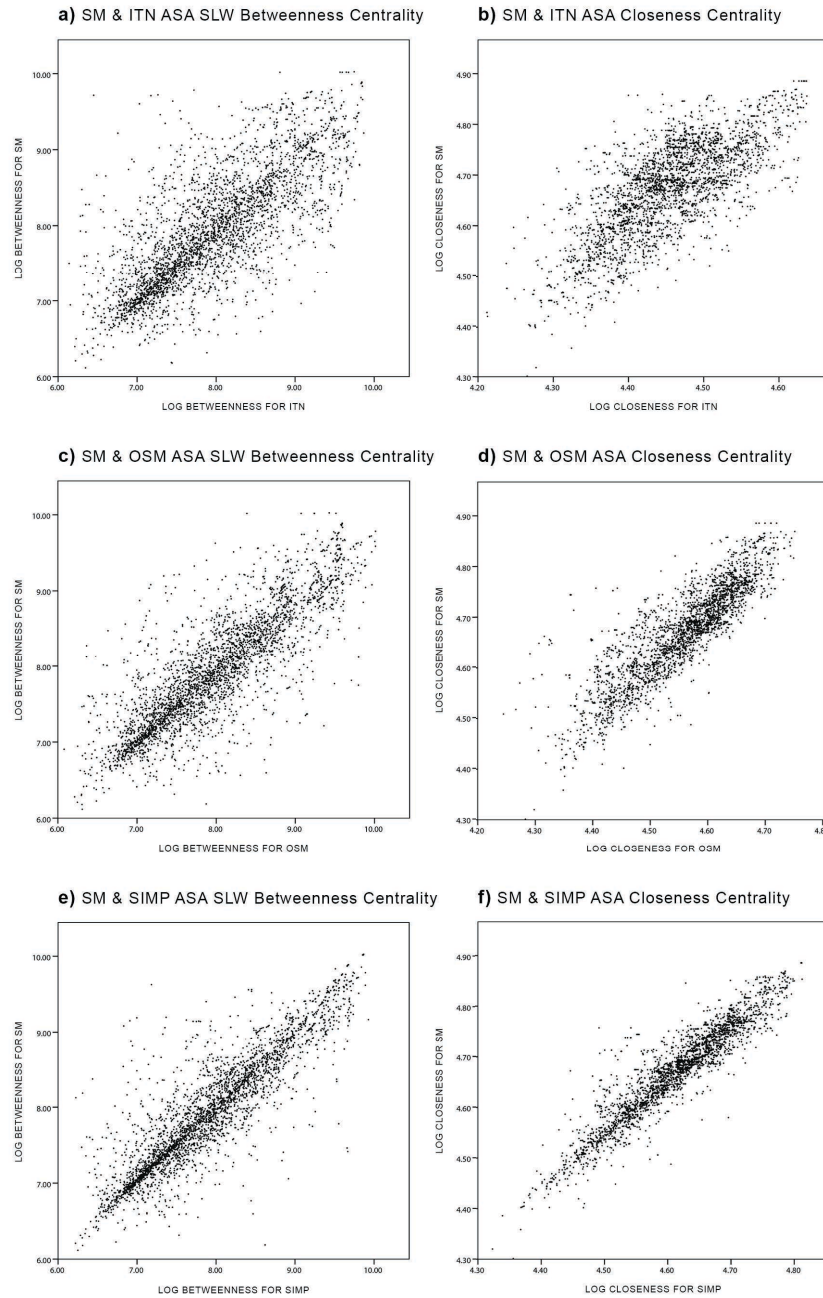
Figure 46: Log-Log plots for the SM model compared to ITN, OSM and SIMP respectively for ASA SLW betweenness and closeness centralities on radius n.

To summarise, the results show that the four models differ particularly in terms of the number of short length segments. This difference can be described by an exponential relation and has a significant impact on the computational time needed for the analysis. The results of the metric step depth analysis confirm the findings of Dhanani et al. (2012) and show that all models share a similar complexity in terms of their nodal distribution. However, the analytical space syntax analysis showed that despite a similar

distribution in the data in general, the geographic location of these differences has an impact on the results. The ITN network is strongly influenced by its emphasis on vehicular movement and traffic management details. This makes it less comparable to the segmented axial line model than the OSM model or the SIMP.

5.4 SIMP STREET NETWORK MODELS OF THE TWO CASE STUDIES

Following the presented methodology, I created two spatial network models by employing OSM data and the proposed simplification method. Figure 47 shows the resulting model for the case study region in Germany, and Figure 48 shows the model for the British region. Both figures feature a zoomed-in detail section on 5 different scales, highlighting the level of detail and providing an insight into the extent of both models. A detailed description of the model morphologies can be found in Chapter 6 and Chapter 7.
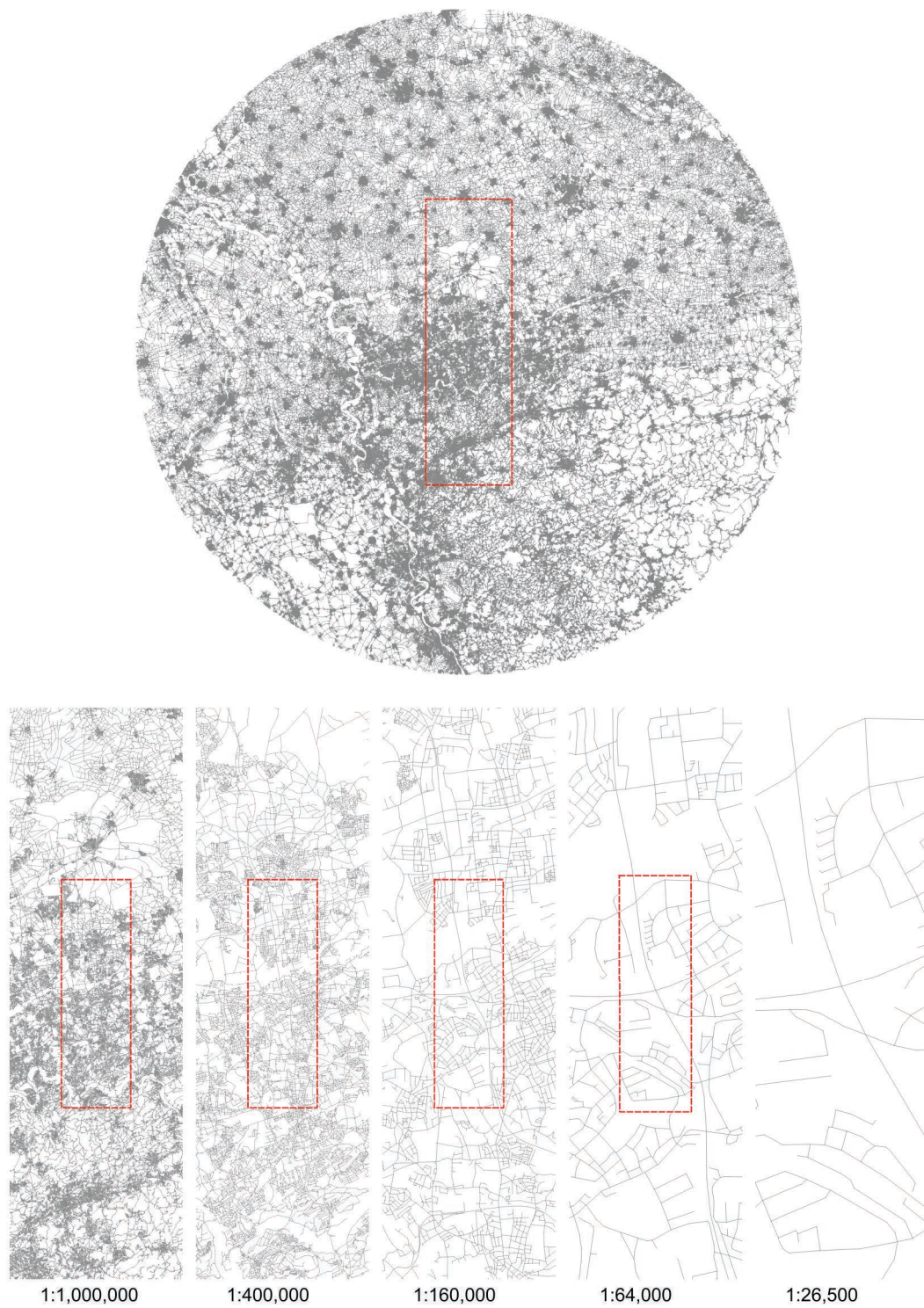
Figure 47: GE simplified network model with selected detail areas on five different scales of 1:1,000,000, 1:400,000. 1:160,000, 1:64,000 and 1:26,500.
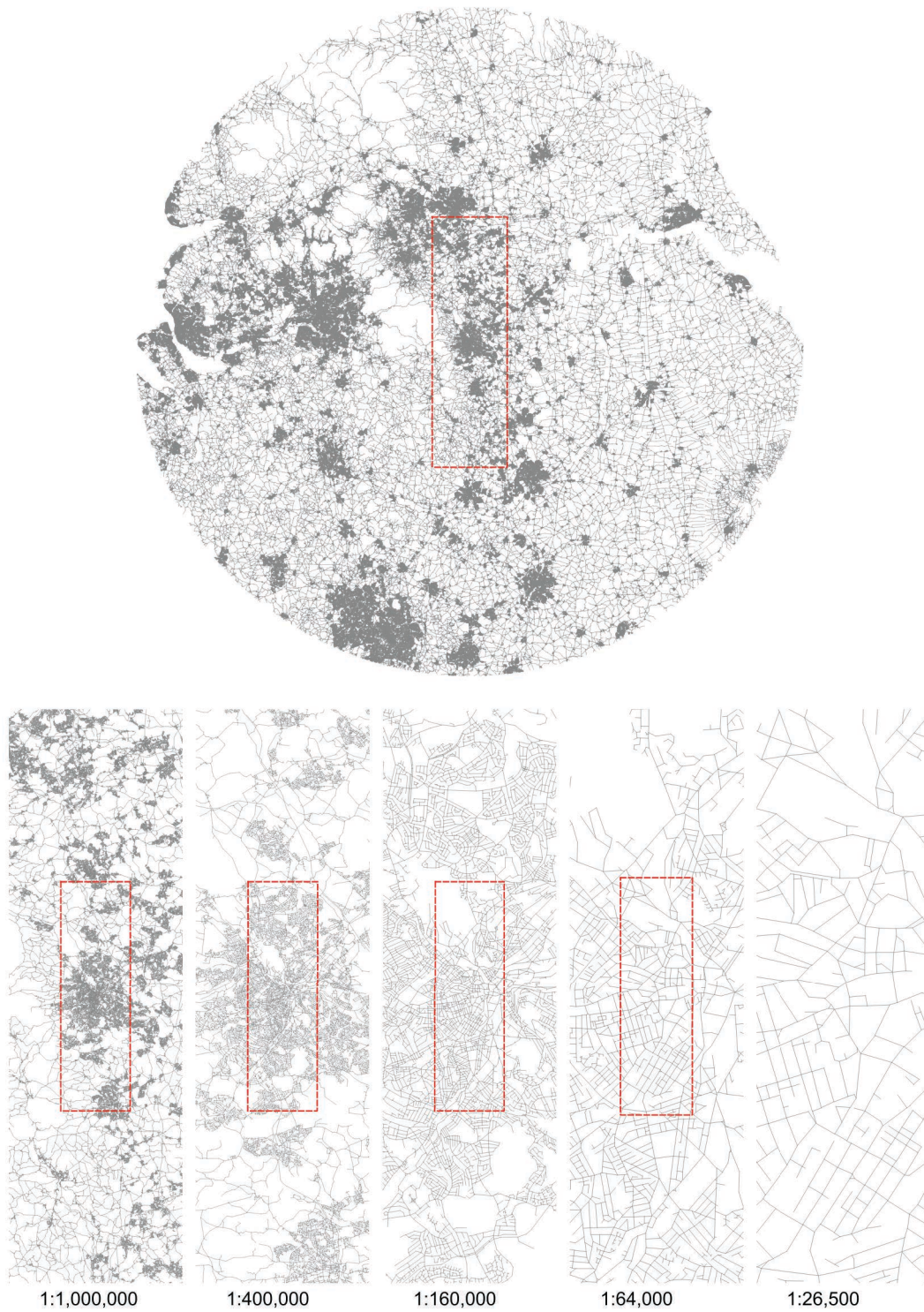
Figure 48: UK simplified network model with selected detail areas on five different scales of 1:1,000,000, 1:400,000. 1:160,000, 1:64,000 and 1:26,500.

1:1,000,000 1:400,000 1:160,000 1:64,000 1:26,500

5.5 SUMMARY

Concluding, this chapter elaborated the fitness of OSM data in space syntax analysis, it proposed an ArcGIS simplification workflow and presented the theoretic reasoning behind the method. The final fitness tests showed that the simplified OSM network (SIMP) exhibits very strong similarities with the traditional segmented axial line model across all investigations. It features the topological and angular information of the OSM network with the simplistic representation of a segmented axial line model. This is rather surprising because the alterations in the model are mainly based on segment nodal reductions and minor topological alterations. In fact, the Pearson and Spearman correlation analysis showed that the SIMP model is more strongly related to the segmented axial model than to the OSM model. The strong similarity between SIMP and segmented axial also poses the question of weather axial line models are such intrinsically different representations.

Overall, the findings suggest that a SIMP model constitutes an appropriate model for space syntax analysis, particularly in the light of regional investigations where the production of an axial line model is not a feasible option. Following the proposed method, two street network models have been generated and are presented.

CHAPTER 6

GENERATION OF RANDOM REGIONAL STREET
NETWORKS

6 CHAPTER

Due to the absence of comparable research on regional morphology, this research uses randomised street networks as a method of comparison. The aim is to overcome the previously in Chapter 2 discussed lack of appropriate concepts, as to what kind of scales can be expected in real-world networks, and what scales might be attributed to fundamental properties of spatial networks. This is done by comparing the results of real-world spatial networks to those networks that are the result of a randomised process. The aim of this exploration is to understand whether regional networks exhibit scale structures that are different to those of randomly generated street networks. For a generated network to be considered completely spatially random, the network must be independently and uniformly distributed over the region. However, I will demonstrate that such networks are in any regard barely comparable to real networks and instead propose a different approach for random street network generation. The underlying hypothesis is that human-shaped regional networks feature scale structures different from those of randomly generated street networks. If randomly generated street networks do not feature similarities in their scale structures then the occurrence of such difference might be attributed to human organisation in space. If, however, randomly generated street networks feature similar patterns, then it will be of value to gain insights into similarities and differences of such scales, as if spatial scales occur in completely random networks as well as in human-shaped networks, then any spatial network might feature such scale structures. One can then start to compare these structures and learn about the role of human-driven processes in their emergence from these insights.

Automated street network generation is not a trivial task. While there are, in general, several approaches to deal with random street network generation, approaches that produce entirely random networks are scarce. This is because knowledge of randomly generated street networks is still in an early stage. It will be argued that one model, the Erdös-Renyi random planar graph with radius restriction (henceforth ERPGr), is appropriate for the purpose of this study. This chapter will start to compare the network characteristics of existing regional street networks and formulate a series of criteria for the generation of the ERPGr. This is followed by a review of existing random generated network approaches and the reasoning behind the selection of an ERPGr for the purpose of comparison. Following this, is an introduction of point pattern analysis as a method to simulate a regional nodal pattern that can be used in random street network generations closer to real networks but which are nevertheless random in their generation.

6.1 GENERATION OF RANDOM REGIONAL STREET NETWORKS

6.1.1 CHARACTERISTICS OF REAL WORLD REGIONAL SPATIAL NETWORKS

Before I elaborate on the differences between network generative algorithms, it is important to establish some prior knowledge on the similarities of the characteristics of the street networks of both case studies. This is necessary to guarantee comparability between the random networks that I will produce and the case studies.

I will demonstrate that some of these characteristics are similar across both cases and should therefore be ideally met in the randomly generated planar graph. The effects measured on scale structures could then be attributed to the spatial configuration only and it could be determined that they are not influenced by differences in network characteristics caused by the network generation process. The descriptive exploration will compare the following fundamental network properties:

1. Number of segments n_s

   Segments are –as defined earlier– individual sections of streets that form an independent visual space. The more visually separated areas a street has, the larger is the number of segments per street. This often relates to the number of curves or turns without intersecting other streets. Segments are described by their Euclidean distance from start to end point.

2. Number of nodes n_n

   Nodes are said start and end points of segments and describe either the end of a street, an intermediate point of two continuously connected segments or an intersection of several segments. Nodes can provide information on the connectivity, or degree by counting the number of intersecting segments with the respective node.

3. Number of links n_l

   Links are defined by points of route decision making. A link can consist of one to n segments. The core difference between segments and links is, that links can provide information about the general network, but cannot account for local geometric differences within this network. They are a simplified version of the real spatial graph.

4. Number of link nodes n_{ln}

   Link nodes are the equivalent to standard nodes within a link-based network.

5. Frequency of segment length f_s

   Describes the value distribution of segment lengths within the model.

6. Frequency of segment connectivity f_{sk}

Describes the value distribution of segment connectivity within the model. Segment connectivity, provides a measure for the connectivity degree of each segment to all neighbouring segments.

7. Frequency of angular connectivity f_{ak}

    Angular connectivity measures the geometric relationship between each street segment and its immediate neighbours. This is done by counting the cumulative angular turn, whereas a straight intersection counts as 0, a 90° turn 1 and 180° the value 2 (Turner 2000).

8. Frequency of node connectivity f_{nk}

    Node connectivity[18] (degree) counts the number of segments intersections to the respective node. The frequency f_{nk} equals the probability distribution

9. Beta index β

    Beta is a measure that approximates the connectivity for the entire network. It relates the total number of edges to the total number of nodes ($\beta = \frac{s}{n}$).

Studies investigating distributions of street networks have been focused on models of urban areas and cities. While some authors argued that street segment length follows a power law distribution (Huang et al. 2016; Jiang 2009; Mohajeri et al. 2013), the work of others indicated that street segment length rather follows a log-normal distribution (Hillier 2002; Masucci et al. 2009, 2013). Differences in the classification of the nature of observed distributions as log-normal or power law have often been the source of disagreements in different disciplines in the past, which is due to the very close nature of both types (Mitzenmacher 2003).

Power law distributions, commonly known as Pareto distribution, heavy-tailed distribution or Zipfian distribution describe distributions where a relative change in one quantity is proportional to a fixed power of another quantity. Such power law probability distributions can often be found in data scrutinied in physical and social sciences, where the respective data is characterised by a substantially larger amount of small values than larger ones. In the urban context Felix Auerbach (1913) discovered in his work *The law of population concentration* that the population size stays in relation to a city's rank, also referred to as *rank size rule*. Simultaneously, Georg Kingsley Zipf (1932) came across such rank to frequency relationships for the first time in his linguistic work on word distributions of different languages. He found that the probability of encountering a word stays inversely proportional to its rank in the

---

[18] When the term connectivity is used, it refers to the concept of degree. This should not be confused with the term 'connectivity' in graph theory, where connectivity describes the concept of the minimum number of necessary elements that need to be removed in order to fragmentise the graph.

frequency table. This means that the most frequent word occurs twice as often as the second most frequent word and so on. Later Georg Kingsley Zipf (1949) published his first paper on the rank-size distribution of human settlements. Such power law relations point to intrinsic scaling laws and Auerbach's and Zipf's findings can hence be seen as early explorations into complexities between urban form and societies.

Log-normal distributed data on the other hand describes a variable whose logarithm is normally distributed. Such a distribution is sometimes also referred to as Galton distribution and can be found in many natural phenomena across sciences. The main difference to a Zipfian distribution is that a log normal distribution usually has a very stark increase in its lower tail, while both share similarities in their long upper tail. This is also the source of the aforementioned confusion between log-normal and power law distributions. Moreover, data can also without any contradictions follow both a log-normal distribution in the entire data and a power law distribution only in the upper tail. In fact, log-normal distributions are mathematically intrinsically linked to power laws (Mitzenmacher 2003 p. 3). Particularly double Pareto distributions can be technically indistinguishable from a log-normal distribution (Reed and Jorgensen 2004). However, power law distributions in the studies referred to above are reported as Zipfian or single Pareto.

In this context, Bin Jiang's study on street hierarchies argued that street length distributions in street networks follow a Zipfian power law relationship, where 'smaller streets are far more common than larger ones' (2009 p. 1033). A similar argument has been made by Huang et al. (2016). Both authors build their investigation on Zipf's law and support their argument by presenting a frequency to segment length plot. Moreover, also Mohajeri et al. present in their study on the city of Kerman, Iran a power law relationship for segment length distribution (2013 p. 3). In order to meet a Zipfian distribution the shortest street length needs to be the most frequent in the entire network, while the second shortest needs to be half as frequent and so on. The segment length is a continuous variable and in that sense the frequency can only be estimated through a binning process, because there are no streets with the exact same length. Different to this approach, Masucci et al. (2009, 2013) argue that street length follows a log-normal distribution. Their analysis is based on nine different time periods (from 1768 to 2010) of the city of London and shows throughout all cases comparable log-normal distributions. Moreover, the work of Hillier (2002) has shown that a theoretical construct of street networks, based on a set of simple rules leads to a log-normal distribution of segment length. None of these studies discussed, has considered the implication of larger systems beyond the size of a city. I propose that comparisons of large-sized systems, beyond the scope of the city, will give further insights into the

complex laws underlying their form, different to those findings established within a city context.

BASIC NETWORK PROPERTIES OF THE TWO CASE STUDIES. I will hence begin to investigate these characteristics in the two cases. Table 18 highlights that the street network model of the United Kingdom shows significantly smaller amounts of segments and nodes than the German model. This becomes also clear when comparing the total number of links to the total number of link nodes. With regard to their *beta* index, however, the German region although much denser and with respectively larger amounts of nodes, has a slightly lower value. This indicates that the British model has an overall slightly greater connectivity. Besides providing us with an indication on the connectivity of the graph, the beta index can also be interpreted as a measure of efficiency within the network. The greater the beta index value is, the more connected and hence more efficient is a system. Here, the beta score will be used as an indicator of the efficiency of regional spatial organisation. Despite their differences in segment and node quantities, both models show comparable beta index values. The last row of the table provides the average between both models. Particularly number for the segments (n_s) and the nodes (n_n) will be used as an approximation for the random graph generation process.

Table 18: Descriptive road network statistics of the German and United Kingdom case study within a 200 kilometre radius.

	n_s	n_n	β_{sn}	n_l	n_{ln}	β_{lln}
GE	1203173	1044762	1.152	703241	545948	1.288
UK	835145	711944	1.173	671597	549597	1.221
mean	*1019159*	*878353*	*1.161*	*687419*	*547772*	*1.254*

As this basic statistic only provides a very generalised description, I conduct a selective radial analysis to gain further insights into intra-regional differences of each model. The analysis starts with a centrally positioned 10km diameter circle. With each step the diameter expands by additional 10km. We count the number of n_s, n_n, n_l and n_{ln} and compare their frequency over 20 different radii. The results will highlight differences in the distribution and the degree of urbanisation. The stronger the incline of the line from one radius to the next is, the larger is the degree of urbanisation, while a lower incline indicates a decline in urbanisation and a stronger degree of rural character. An even or steady development of the line shows a homogeneous distribution of built and vacant areas, while sudden shifts highlight an inhomogeneous development.

a)

Segments per 10km Radius

Nodes per 10km Radius

b)

Links per 10km Radius

Link Nodes per 10km Radius
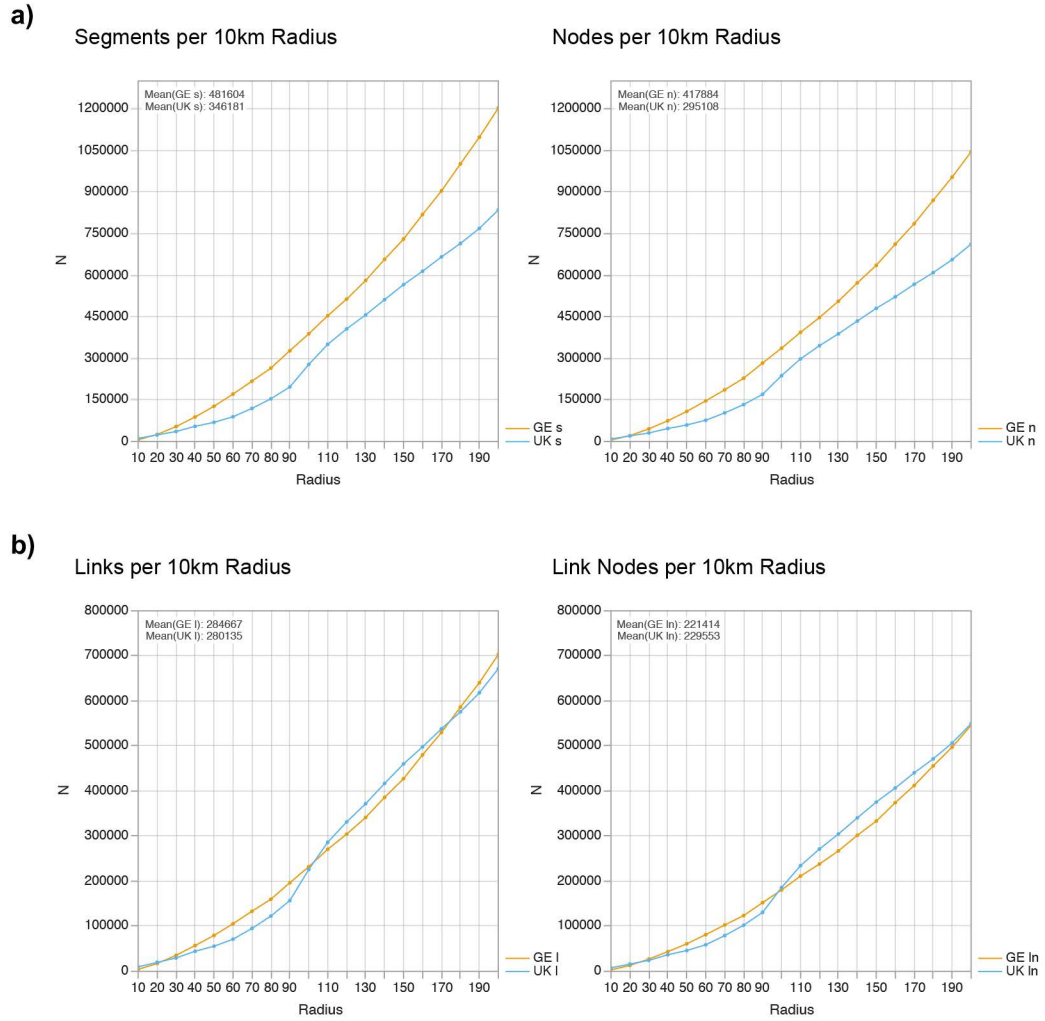
Figure 49: GE and UK: Frequency comparison of a) segments, nodes and b) links and link nodes on 20 different radii.

Figure 49:a and Figure 49:b illustrate the results of this analysis. Both cases exhibit comparable differences across the four measures. The German case is characterised by a steadily increasing development of the line and, hence, a more homogeneous development in each of the four analyses. The British case shows an uneven development with a lower increase on the first eight radii and a stark increase between 90 and 100 kilometres. These differences are particularly visible in the comparison of network links and link nodes, where the overall quantities are evenly distributed but the increase and the decrease of the curve highlight sudden changes of the level of urbanisation. Interestingly, the apparent quantitative differences between the two cases in Figure 49:a disappears when real network links are compared, as shown in Figure 49:b. This points to a fundamental difference in both networks. In the German case, network links are more often composed of several different segments, implying that more streets are sinuous and differ in their angular connectivity, while in the

British case visibility and network links are more closely related and segments coincide strongly with visual linearity.

Since network links and network link nodes showed a close relationship in their distribution over inter regional spaces and between both cases (Figure 49:b), I will make use of these two measures as an additional condition to be met for the following randomised model. This will be done by means of a bivariate analysis, which is a simple quantitative analysis to determine the relationship between two variables. A bivariate analysis provides tools for model fitting and data prediction. Initial tests have shown that the relationship between radius and number of links, as well as radius and number of link nodes is nonlinear and can be described by a polynomial of second degree. This allows a prediction of new data, or as in our case, a way to examine whether newly produced data meets the prediction. Both predictive models have a nearly perfect fit of an R-square of 0.994 for links and an R-square of 0.992 for link node prediction, with an adjusted R-square of 0.993 and 0.992 respectively. However, both plots (Figure 50) show that the UK model stays below the prediction on lower radii (40 – 90 km) and above the prediction on higher radii (110 – 170 km). This issue can be considered a limitation for the model's predictive generalizability, but it does not undermine its purpose as comparative control measure for the outcome of a randomised graph network. In general, randomised models are expected to feature homogeneous distributions and are, hence, expected to closer match the German model with a more even distribution of urban form.

Bivarial Analysis Links vs. Radius

$l = -21623 + 1395 \cdot Radius + 10.98 \cdot Radius^2$
R^2: 0.994

(y-axis: l, values 0 to 900000; x-axis: Radius 0 to 190)

Radius
- GE — l (Fit)
- UK — l (Prediction)
- l

Bivarial Analysis Links Nodes vs. Radius

$ln = -18277 + 1141 \cdot Radius + 8.639 \cdot Radius^2$
R^2: 0.993

(y-axis: ln, values 0 to 700000; x-axis: Radius 0 to 190)

Radius
- GE — ln (Fit)
- UK — ln (Prediction)
- ln

Polynomial Fit Degree=2

$l = -142663.9 + 3700.5784 \cdot Radius + 10.978766 \cdot (Radius-105)^2$

Summary of Fit

RSquare	0.994156
RSquare Adj	0.99384
Root Mean Square Error	17206.69
Mean of Response	282401.3
Observations (or Sum Wgts)	40

Lack Of Fit

Source	DF	Sum of Squares	Mean Square	F Ratio
Lack Of Fit	17	4765085313	280299136	0.9057
Pure Error	20	6189515422	309475771	**Prob > F**
Total Error	37	1.0955e+10		0.5778
				Max RSq
				0.9967

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Ratio
Model	2	1.8637e+12	9.318e+11	3147.329
Error	37	1.0955e+10	296070290	**Prob > F**
C. Total	39	1.8746e+12		<.0001*

Parameter Estimates

Term	Estimate	Std Error	t Ratio	Prob>|t|
Intercept	-142663.9	6423.917	-22.21	<.0001*
Radius	3700.5784	47.18148	78.43	<.0001*
(Radius-105)^2	10.978766	0.918268	11.96	<.0001*

Polynomial Fit Degree=2

$ln = -113522.3 + 2955.0586 \cdot Radius + 8.6390377 \cdot (Radius-105)^2$

Summary of Fit

RSquare	0.992791
RSquare Adj	0.992401
Root Mean Square Error	15267.15
Mean of Response	225483.6
Observations (or Sum Wgts)	40

Lack Of Fit

Source	DF	Sum of Squares	Mean Square	F Ratio
Lack Of Fit	17	3207157544	188656326	0.6965
Pure Error	20	5417025064	270851253	**Prob > F**
Total Error	37	8624182607		0.7723
				Max RSq
				0.9955

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Ratio
Model	2	1.1876e+12	5.938e+11	2547.580
Error	37	8624182607	233086016	**Prob > F**
C. Total	39	1.1962e+12		<.0001*

Parameter Estimates

Term	Estimate	Std Error	t Ratio	Prob>|t|
Intercept	-113522.3	5699.813	-19.92	<.0001*
Radius	2955.0586	41.86319	70.59	<.0001*
(Radius-105)^2	8.6390377	0.814761	10.60	<.0001*

Figure 50: GE and UK: Bivariate polynomial model fitting for link and link nodes and as a function of radii.

DISTRIBUTIONS OF GEOMETRIC NETWORK CHARACTERISTICS. In addition to these initial criteria, the following comparison sheds light on the distribution of aggregated attributes of both network models. I will compare the distribution of segment length, angular connectivity and segment connectivity. Whereas segment length and angular connectivity are geometric properties, segment connectivity (dual graph degree) and node connectivity (degree) describe distinctive properties of the network itself. Figure 51 and Figure 52 show the results of this comparison. The three connectivity measures share a common skewness to the right, with very few high values.

With respect to the street length distribution, both segment length histograms with fitted normal and log-normal distribution curves (Figure 51:b and Figure 52:b) show that the data is highly right skewed with few extreme high values. Particularly the log-

normal curve demonstrates that the data follows a logarithmic normal distribution rather than a power law distribution, as described by Jiang in the context of the city (2009). Cases of extremely right-skewed data are a common indicator for the application of a logarithmic transformation. Both models share very comparable parameter estimates with a mu and a sigma of 4.251 and 0.830 for the British and 4.271 and 0.825 for the German case for the fitted function. The second histograms of Figure 51:b and Figure 52:b approximate a normal distribution and support the hypothesis that the segment length distribution is of logarithmic nature in both cases. Here, the log segment length data of the German case follows a clear normal distribution. The British case now also approximates a normal distribution, although with some outliers in shorter length segments. Both normal quantile plots for log segment length exhibit a strong linearity. The closer the data points are to the central line in a quantile normal plot, the stronger there is a normal distribution in a dataset. Particularly the German case follows this principle strikingly closely, while the British case shows, as expected and observed previously, some skewness in low values but overall follows the principle of normality.

These findings could be expected, as the complexity of the network and its level of detail leads to a respective amount of small length segments that are lower than the length of highest frequency. One can imagine that the most frequent street length in a region is not 2 or 10 metres, but should instead measure a length of 30 to 50 metres. One can therefore expect that the data has a lower tail of declining segment length which is why it does not come as a surprise that a clear power law relation of a kind that Zipfian describes cannot be confirmed for the entire dataset. In the two tested models power law distributions, such as reported by Jiang, Huang et al. and Mohajeri et al., can only be verified, if almost 50% or more precise the lower (0 to 80 metres) and extreme upper tails (800 to 5000 metres) are excluded from the data (Figure 53). The high R-square of the fitted robust regression line of the British model (R^2 of 0.996) and German case (R^2 of 0.989), are a strong argument for the existence of a scaling mechanism for the mid part of the upper tail. One reason for the differences in these findings and the findings by Jinag, Hunag et al. and Mohajeri et al. could be that their analysis is based on incomplete or highly simplified datasets, that lack a sufficient level of detail to account for local differences.
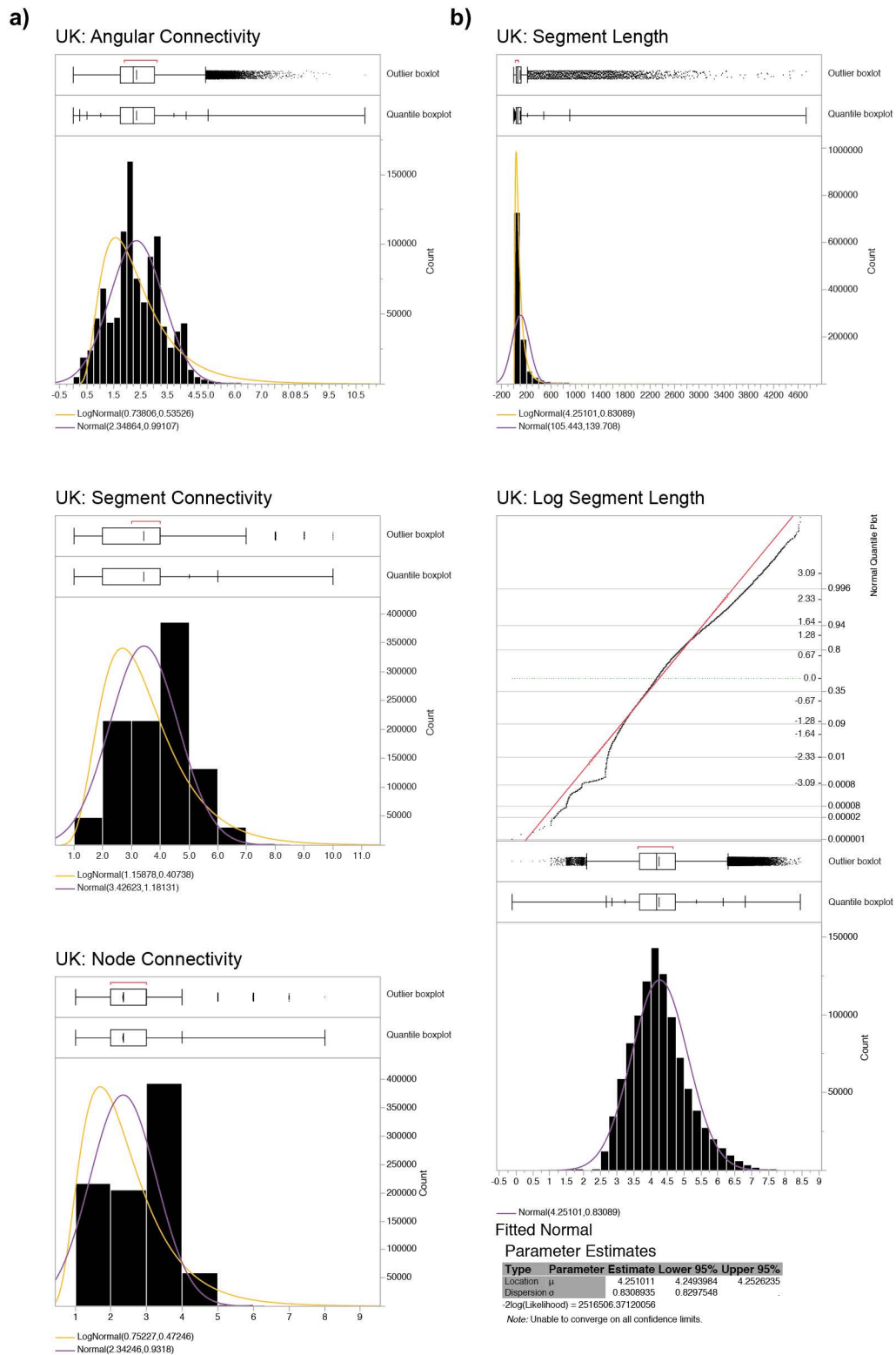
Figure 51: UK: Street Network Distribution Analysis. Outlier Box plot, box plot and histogram for a) angular connectivity (top), segment connectivity (middle) and node connectivity (bottom) and b) segment length (top) and logarithmic segment length (bottom).

a)

GE: Angular Connectivity

Outlier boxlot

Quantile boxplot

Count

— LogNormal(0.40246,0.79559)
— Normal(1.89304,1.06233)

GE: Segment Connectivity

Outlier boxplot

Quantile boxplot

Count

— LogNormal(1.07284,0.39084)
— Normal(3.13706,1.11911)

GE: Node Connectivity

Outlier boxplot

Quantile boxplot

Count

— LogNormal(0.76929,0.37764)
— Normal(2.30368,0.78163)

b)

GE: Segment Length

Outlier boxlot

Quantile boxplot

Count

— LogNormal(4.27153,0.82549)
— Normal(101.987,111.35)

GE: Log Segment Length

Normal Quantile Plot

Outlier boxlot

Quantile boxplot

Count

— Normal(4.27153,0.82549)

Fitted Normal

Parameter Estimates

Type		Parameter Estimate	Lower 95%	Upper 95%
Location	μ	4.271532	4.270057	4.2730071
Dispersion	σ	0.825495	0.8244533	

-2log(Likelihood) = 2952985.99257715

Note: Unable to converge on all confidence limits.

Figure 52: GE: Street Network Distribution Analysis. Outlier Box plot, box plot and histogram for a) angular connectivity (top), segment connectivity (middle) and node connectivity (bottom) and b) segment length (top) and logarithmic segment length (bottom).

194
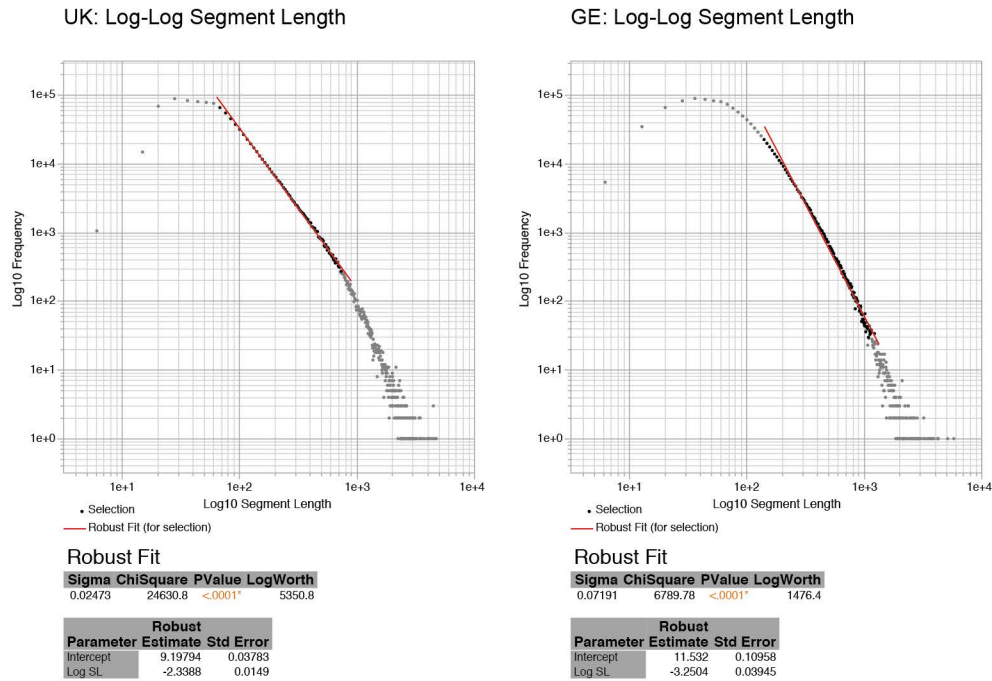
Figure 53: UK and GE Log-log plots for frequency vs. segment length with robust linear fit. Robust fit for UK of $Y=9.19794-2.3388*Segment Length$ ($R^2=0.996$) and for GE of $Y=11.532-3.2504*Segment Length$ ($R^2=0.989$).

The occurrence of power law or scaling, in general, is highly influenced by the set of observations. If insufficient samples are compared, it becomes impossible to reveal an underlying scaling relationship. At the same time a selective sample comparison can appear as a clear Zipfian distribution, while in fact a comprehensive dataset would exhibit a different distribution. A similar observation has been made in the context of distributions of settlement sizes. Jefferson (1939) described how the largest cities across the globe show significant outlier behaviour in relation to the next smaller city in their respective country. This means that the population of the largest cities is highly underestimated by the fitted power law function. In addition to this upper tail behaviour, several studies have shown that when settlements of smaller size are included in the analysis, a log-normal distribution provides a much better fit than a Zipfian model (Baker 1969; Parr and Suzuki 1973). This does not contradict the notion that street networks follow particular scaling laws. Instead, the point made here is that street networks are better estimated by a log-normal or double Pareto function. Particularly the findings of the log-log plot point to such a distribution in the two cases. The data distribution in both log-log plots shows a strong hyperbolic curve. Such a hyperbolic development is common for log-normal distributed data and can be well described through a double Pareto-log normal function (Reed and Jorgensen 2004).

Figure 54 shows a natural density curve for a double Pareto-log normal distribution as well as the same data in logarithmic scale, where β > 1. If we compare this with the log-density histogram and the fitted log-normal curves of the two cases in Figure 51:b and Figure 52:b as well as with the log-log plots in Figure 53, the similarities clearly show. Both density histogram curves are highly right-skewed with a monotonic increase in the lower tail and a power law behaviour in the upper tail.
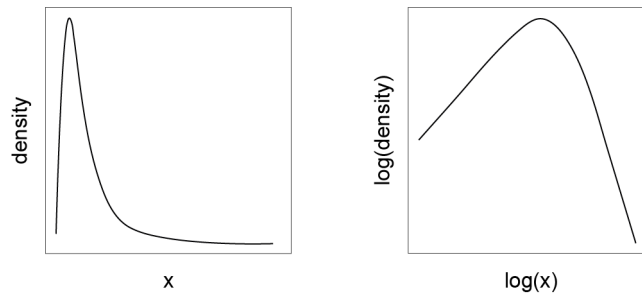
Figure 54: Double Pareto-log normal density in the natural scale (left) and logarithmic scale (right) for β > 1 (ibid.).

The findings presented above, seem to fall into this category and street length distributions follow power law regularity in parts of the upper tail, while the actual distribution of the entire dataset can be estimated better by a log-normal distribution. This coincides with the findings by Masucci et al. (2009, 2013) presented earlier, indicating a log-normal street length distribution for the city of London. Their findings can therefore be supported. This allows us to draw further conclusions about the scale of regions and due to the size of both models, it allows us to speculate on the generalizability of these properties. Regions can be described as a simple agglomeration of human settlements due to their sizes and nature. If regional models exhibit log-normal distributions, one conjecture is that individual settlements might be characterised by similar log-normal distributed street segments. Regions would in this case exhibit distributions that can be compared with added small-sized settlement distributions and each settlement should exhibit a similar distribution with respective similar mu and sigma values of their normal probability function. To test this assumption, I select the German model and divide the existing regional municipalities into 276 respective German administrative municipalities of the state of North-Rhine Westphalia. When doing this, it is important to acknowledge the potential arbitrariness of these boundaries, which can have an effect on the overall results. Generally, municipal boundaries contain at least one urban settlement, suburban as well as rural areas. Some exceptions occur in very densely populated areas, such as the central part of the Ruhr Valley, where cities have already merged into each other with no apparent

geographic separations. This will influence the log-normal distribution and it is expected that these areas show some form of skewness.

I hence analyse the logarithmic segment lengths of all 276 administrative boundaries by visually comparing their normal quantile plots and histograms. I fit a normal probability curve to the data, and test its goodness-of-fit. Where the number of segments was below 2,000 a Shapiro-Wilk test was performed and where the number was above 2,000, a Lillifors test (based on a Kolmogorov-Smirnov test) was employed (Lilliefors 1967; Shapiro and Wilk 1965). Both tests compare empirical data with a theoretical distribution. The Lillifors estimates the mean and variance of the data and looks for the most substantial difference between the empirical cumulative distribution function (CDF) and the empirical distribution function, while the Shapiro-Wilk test effectively compares the ratio of two different estimates of variance. The null-hypothesis for both tests is that the data is normally distributed.

Of all 276 areas, 96 are statistically significantly indistinguishable from a normal distribution (with $\alpha \geq 0,05$) whereas 62 have a p-value between 0,05 and 0,01 while 118 are less than 0,001. Figure 55:a shows a selection of the three cases with the highest test statistics (D) and the highest probability scores ($Prob>D$) of the Lillifors test (KSL), including an illustration of the area and its street network. The three histograms confirm the tested normality through their 'bell-shaped' distribution and symmetric distribution with a central peak. Figure 55:b shows the three cases with the respective lowest D and p-values. The null hypothesis of the goodness-of-fit test cannot be confirmed. In terms of the histograms, however, it becomes clear that the distributions of these cases do not differ drastically from the ones where normality was confirmed. The shape looks similar to a bell-shape with slight right-skewness for case 5974016 and slight left-skewness for case 551300. Moreover, all three cases have peaks that exceed the boundaries of a normal distribution probability. The normal quantile plot confirms this observation with data points departing from the ideal normal distribution and also leaving the 95% confidence limits in all three cases. Nevertheless, the general trend of the data can be approximated through a normal or skewed normal distribution. All six cases are very comparable in their overall data distribution.

What might contribute to the failed normality tests is the aforementioned arbitrariness of the administrative boundaries. All areas for cases where normality could statistically not be confirmed, feature boundary shapes that are rather linear than circular. A skewed boundary geometry might be an influencing factor. Additionally, the Ruhr Valley feature cities that have a continuous urban fabric throughout their administrative boundaries, such as in case 5566016 (Figure 55:b).

a) GE: Robust Normal Distribution for Log Segment Length

Figure 55: Normal quantile plots, Histograms with fitted normal probability density curve and goodness-of-fit KSL test for a) the three highest probability scores (*p* > 0.15) and for b) the lowest probability scores (*p* < 0.01*).

What might contribute to the central bars extending the estimated curve at their peaks in the histogram, is the fact that many more streets of urban character can be found for this case than in other areas. What all cases have in common is a uni-modal bell-shaped distribution or one that approximates the bell-shaped distribution. These findings are in accordance with the previous observations made in the regional models, which allow us to conclude that log-normal or double Pareto functions are well suited to estimate street segment length distributions. We can therefore use this to formulate an additional limitation for a randomised street network to guarantee comparability, namely that the street segment length distribution should approximate a log-normal distribution.

In summary, both existing regional models have a discrepancy of 20 and 30 per cent between their total number of segments and nodes. As these are fundamental network components, we will use these values as upper and lower limits (segments: 1203173 – 835145, nodes: 1044762 – 711944) for the newly generated model, while a value around the mean (segments: 1019159, nodes: 878353) is set as the ideal. The analysis of the segment length distribution has shown that regional as well as city street networks can be estimated by a log-normal function. The segment length distribution of the comparative model should approximate a log-normal function. Neither segment connectivity, nor angular connectivity, nor node connectivity show a comparable pattern. This is because they represent values of individual network characteristics and are beyond a generalizable trend. Hence, I will ignore these in the process of randomised model generation. Having established an understanding of fundamental similarities in regional street networks, I move on to details of network generative processes.

6.1.2 PARAMETRIC STREET NETWORK GENERATION

The available generative procedures often rely on parametric approaches to generate street networks. Such parametric approaches use either a set of generative rules in order to arrive at street networks (Marshall and Sutton 2013; Parish and Müller 2001), employ pattern-based approaches to generate networks (Sun et al. 2002), or a combinatory approach of the former (Chen et al. 2008). Parish and Müller (2001) introduce *CityEngine,* a procedural method that allows consideration of global goals and local constraints. Sun et al. (2002 p. 42) identify a series of existing frequent patterns in real-world networks and create a matching pattern template for each. Through the application of different pattern templates, they are able to generate new street networks that are combinatorial. Chen et al. (2008), on the other hand, combine Parish and Müller's (2001) procedural method with a tensor field to generate patterns. Most recently Marshall and Sutton (2013) presented the simulation tool *NetStoat* to model the growth of street networks. Their tool explores the potential of generative street

layouts. While these given examples should not be seen as a complete account on generative network tools, they provide a guide towards the general approaches taken in this field.

With regards to an application in the generation of random regional networks, none of these approaches is appropriate because they all emulate networks based on the ideal of cities. These are no suitable examples for comparisons with regions and polycentric regions in particular, as they are characterised by the occurrence of several cities of different sizes and the combination and gradual change of urban and non-urban form. It is, thus, questionable if any parametric approach, whose parameters and patterns are in their generation based on independent cities, could arrive at a comparable pattern or complexity of structures of regions. More importantly still, none of the parametric approaches can be considered to be completely random, although this is a necessity for the model we want to employ. More importantly, if the generated model does not feature a strong degree of randomness or ideally a complete spatial randomness, and is instead based on existing, observed street networks in cities, this will lead to one of two options: the results of the analysis to follow will exhibit either clusters and centrality patterns of the specific parameters used, or emulate human-shaped configurational environments. This would be contrary to the planned test, aiming to gain insights into fundamental network characteristics of regional-sized models that are not shaped by human interaction and are instead random by nature.

6.1.3 ERDŐS-RÉNYI AND RANDOM GEOMETRIC GRAPHS

At the core of any random street network generation stands the problem of creating a random graph that features spatial information. Random graphs are abstract mathematical models that consist of edges (lines, comparable to segments in street networks) and vertices (nodes, comparable to intersections, or start and end points of streets). Random graphs are described by either a probability distribution or a random process that generates them. These graphs have an infinite amount of possible arrangements because they do not incorporate any spatial information. From a mathematical point of view, the term random graph solely refers to the Erdős-Rényi random graph model (henceforth ER), introduced by Paul Erdős and Alfréd Rényi in the late 50's (1959). The ER model $G(n,p)$[19] is generated through a given number of vertices n, and the probability p of an edge being absent or present between two randomly

---

[19] Erdős and Rényi provide two definitions for random graphs; the alternative model is described by $G(n,l)$, where l is the total number of randomly placed links (1959). The difference is hence that the first model provides a probability p for two vertices to be connected, while l defines the total number of connections. For this study, only the former will be of interest, due to its wider applicability in network science.

selected vertices (*v,w*) with equal likeliness. In most cases, *p* will be constituted by a function of *n* with a given constant *d*, e.g. (*p* = *d/n*) (Blum et al. 2013 p. 229). Despite their random and independent selection process of vertix pairs (*v,w*), random graphs are characterised by global properties influenced by *p*. A small *p* (with *d* < 1) leads to small connected components in the graph, whereas if *p* is respectively large (with *d* > 1) the graph will feature a large connected component (ibid. p. 229). The threshold for this phase transition can be defined by *d* = 1. ER graphs can, in addition to their probability, be described by their degree (connectivity) distribution. Generally, ER graphs always feature binomial degree distributions. Erdös-Rényi random graphs are $O(n^2)$ problems (Gerke et al. 2008). These properties have led to numerous applications in comparisons to real-world examples.

$p = 0$ $p = 0.1$ $p = 0.2$
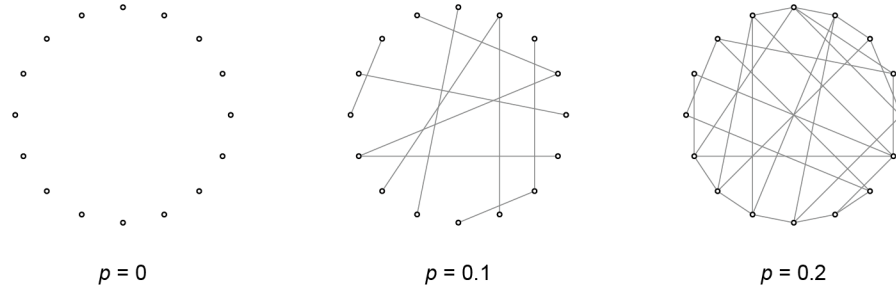
Figure 56: Erdős-Rényi Graph realisation for 16 vertices and three different probabilities from left to right: *p*=0, *p*=0.1 (not connected) and *p*=0.2 (connected).

Figure 56 illustrates three different probability realisations of the ER model. In the example (here with *n* equal 16) a probability of 0.2 leads to a connected graph. This is a graph where all vertices are connected to all others, either directly or indirectly. While a probability of 0.1 results in a fragmented graph with individual components, the arrangement of the graph vertices in a circle is chosen only for the purpose of visualising the graph. As mentioned earlier, vertices, as well as edges, do not feature any spatial or geometric information, instead only the information of the probability of edge occurrence is important for the characteristic and finalisation of the model.

If graphs, however, do incorporate spatial information their graph properties change significantly to none spatial graphs. This leads to a situation where established knowledge about random graph behaviour can only be applied limitedly. The presumably mathematically simplest form of spatial graphs is the Random Geometric Graph (henceforth RGG) (Barthélemy 2011). The process to generate a random geometric graph is comparable to the Erdős-Rényi random graph model. The process starts by placing *n* points uniformly at random in a given metric space. This is done by a homogeneous Poisson point process with a given intensity lambda *λ*. These points are then connected through edges if a point pair (*v,w*) lies within a given radius *r*. Edges in

RGGs are hence not independent because their occurrence depends on the distance between two vertices and r, rather than the probability p. Each point is connected to every possible point within r (Dall and Christensen 2002 p. 2). With the growing radius r also the occurrence of line intersections within the graph increases. Figure 57 exhibits the RGG process and shows a realisation of the model on a Euclidean plane [0,1] and a radius r of 0.75. RRGs are different to normal random graphs, which can be demonstrated particularly well in the occurrence of triangular cycles within the graph. This is regardless of the degree distribution of the random graph (ibid. p. 2).

Poisson Point Process   Radius Threshold   Random Geometric Graph
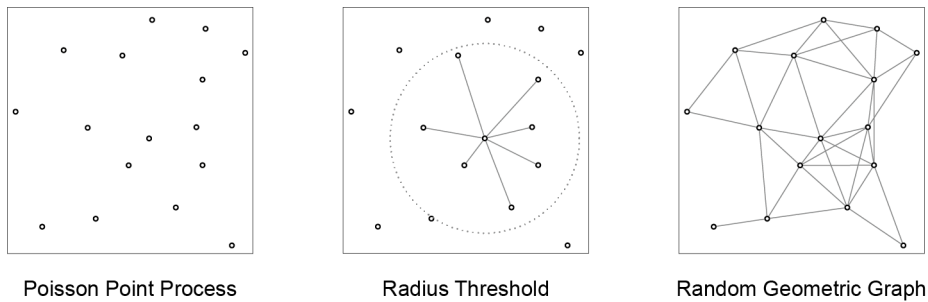
Figure 57: Computer simulation of a random geometric graph, from left to right: Homogeneous Poisson point process in a square area (x_1=0 to x_2=1 and y_1=0 to y_2=1) with intensity λ equal to 16 points per unit. Example of edge creation for used radius threshold of r=0.75. Final realisation of a random geometric graph.

The RGG constitutes a simple way to arrive at a randomised spatial network. The radius limitation distinguishes the RGG from the ER and has a series of implications for the final graph. RGGs are much more clustered compared to ERs; this is also related to the fact that long links are unlikely (Barthélemy 2011 p. 34). Instead of long tails in their segments length distribution such as observed in real street networks, random geometric graphs feature non-unimodal symmetric distributions, where segment lengths are more evenly distributed throughout the entire dataset and outliers do not exist. This is linked to the Poisson distribution used for the dispersion of points over the Euclidean plane space. This is why an application of this model in the context of street networks appears difficult, yet, this is also due to the large amount of intersecting lines occurring when larger radii are applied. Real street networks very rarely feature line intersections and this should, therefore, be seen as an outlier rather than an intrinsic feature of the network. Moreover, line segments of degree 1 only occur at the edge of the model and are very unlikely, because all edges that fulfil the radius restriction are connected to each other. This leads to a more evenly connected model and makes the RGG an inappropriate model for comparisons.

6.1.4 ERDŐS-RÉNYI RANDOM PLANAR GRAPH

A more complex approach is the Erdős-Rényi model realised in an Euclidean space and with a planar restriction. The spatial version of an Erdös-Rényi random graph is based on an initial Poisson point pattern of complete spatial randomness (henceforth CSR) similar to the RGG. The Erdős-Rényi random planar graph (henceforth ERPG) differs from a RGG in that the former rejects edges in its generative process if they do not fulfil the restriction of planarity, meaning that if generated edges intersect with existing edges, they are not added to the final model. Strictly speaking, ERPGs are not fully randomised graphs in the sense of CSR, because the planarity limitation affects the probability of any future edges. The ERPG approach leads to networks comparable with scale-free networks. Scale-free networks, such as the internet (router system) or the world wide web (linked pages) (Barabási 2009), are distinctively different from street networks in the way that a small amount of individual nodes exhibits degrees that are significantly higher than the average degree of the system. Real world street networks do not feature scale-free characteristics, because of their physical limitation of the number of streets meeting at an intersection. Figure 58 shows a realisation of such an ERPG. Due to the nature of the process one can observe that even a single long edge can divide the network into two parts. Such long edges restrict the probability of adjacent nodes and tend to accumulate further connections.

6.1.5 ERDŐS-RÉNYI RANDOM PLANAR GRAPH WITH RADIUS RESTRICTION

Since street networks are not scale-free, Masucci et al. (2009 p. 261) propose to make use of an ERPG with radius restriction (henceforth ERPGr) in this context similar to an RRG. The iterative approach by Masucci et al., starts with a Poisson point process in an Euclidean space. After this, a point pair is selected, based on the probability p if the pair falls within the previously defined radius r. Finally, a planarity test verifies if the newly generated edge validates the planarity of the network. An edge will, hence, only be added to the graph during the process if planarity is not violated, in other words, if no intersection with an edge that has already been added has been found. The radius restriction has a significant influence on the resulting model, as can be seen in Figure 58. Here, no edge is dividing the network into sub-regions, influencing the probability of a large number of adjacent edges. Instead, one arrives at a complete network and an edge length distribution that is mostly a combinatory effect of λ and r. This is because the intensity λ pre-defines the density curve of potential node pairs, which, based on the chosen radius r, becomes left or right-skewed. Therefore, an edge of length x will be more likely to occur if there are more point pairs within the distance x than for other distances.

Planar ER Network — Planar ER Network with Given *r*
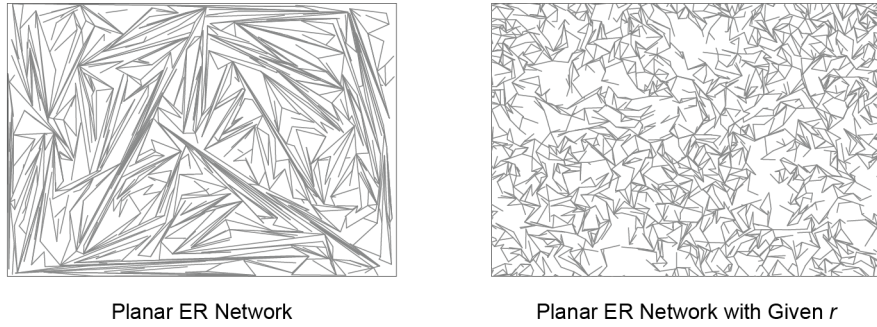
Figure 58: Erdős-Rényi random planar graph (left). Erdős-Rényi random planar graph with given radius influencing the probability of edge occurrence (right).

We can see this relationship by comparing the potential point pairs through a k-nearest neighbour analysis and plot their resulting kernel density estimation. The kernel density estimation is an estimation of the probability density function and will provide a precise estimation of the distance occurrence between all compared points. In order to do so, all nodes of both street networks were extracted and their nearest neighbour distance calculated. The nearest neighbour is calculated in R through the RANN fast nearest neighbour search L2 metric algorithm (Arya et al. 2017). This has been done for ten different *k* neighbours, where *k* is the maximum number of nearest neighbours to be compared. I then estimate the kernel density for each occurring pair distance of *k* neighbours and plot the respective density curve. Figure 59 and Figure 60 show the resulting graphs, where higher curve peaks indicate a higher density of point pairs of the respective distance.

One can see how the curves are developing from a rather log-normal like distribution with small sigma at small *k* (10–30) towards a log-normal like distribution with larger sigma (towards 1) at larger *k* (80–100). Particularly, the first three kernel density estimation curves (k ≤ 30) are comparable to the fitted log-normal curves which we observed earlier in segment length distributions (Figure 51:b and Figure 52:b). These density curves can be used to describe the probability of an edge occurrence for ERPG models, as they provide a precise estimation for the density of each distance. (Figure 61). Predictably, the distribution curve of a complete spatial random Poisson point process of any intensity λ will not be comparable to those we have just observed in existing regional street networks Here, the kernel density estimation curves exhibit a monotonic incline at the lower tail and a left-skewed peak. A resulting ERPGr will hence be constituted of many longer than shorter lines, potentially influencing the results.

Figure 59: GE: Segment Length Density per k Nearest Neighbour, for 10 different k groups (k = 10 – 100). N pairs range from 7,236,405 (*k*=10) to 103,431,438 (*k*=100), with a bandwidth range of 2.712 (*k*=10) to 6.31 (*k*=100).

Figure 60: UK: Segment Length Density per k Nearest Neighbour, for 10 different k groups (k = 10 − 100). N pairs range from 7,837,272 (k=10) to 86,209,992 (k=100), with a bandwidth range of 2.083 (k=10) to 4.978 (k=100).

ERPGr: Nearest Neighbour Distance Kernel Density Estimation per k Nearest Neighbour

N = 8107884 – 89186724
Bandwidth = 3.254 – 5.891
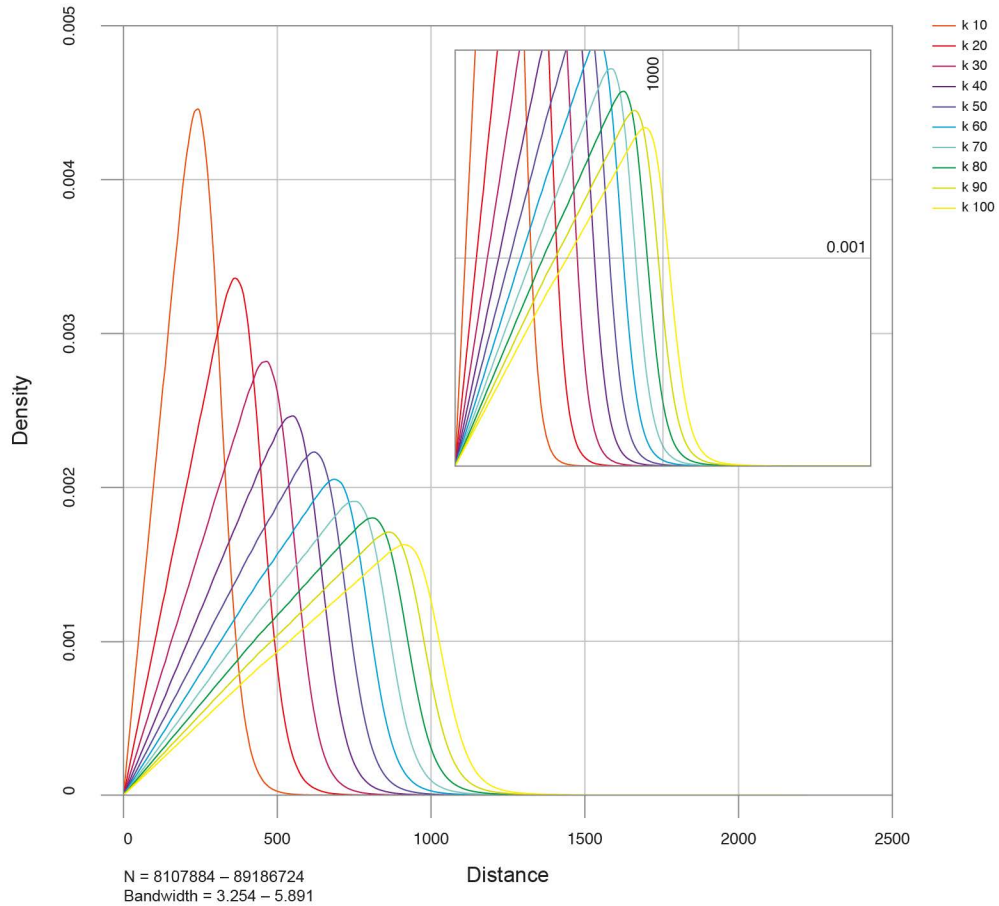
Figure 61: ERPGr: Segment Length Density per k Nearest Neighbour, for 10 different k groups (k = 10 – 100). N pairs range from 8,837,272 (k=10) to 86,209,992 (k=100), with a bandwidth range of 2.083 (k=10) to 4.978 (k=100).

At present the ERPGr forms the only method of spatial random network generation without disadvantages similar to the RGG or ERPG known to the author. This is why an ERPGr will be one of the models we will employ for comparisons. However, because the edge distribution of such ERPGr models is not comparable to those observed in existing street networks, I will propose a new random graph that – I will argue – is more suitable for a comparison with existing regional street networks and their fundamental component, their segments and the respective segment length distribution.

Since the influential factors are identified as the intensity of the point process lambda λ and the radius *r,* I will initiate a new model through an alteration of these two parameters. The idea is to use a point process with parameter estimates derived from the existing regions instead of a Poisson point distribution. I will use a generated randomised point pattern as the basis of a line production process that is based on the spatial point pattern observed in our real-world networks. This can be done by treating the regional node data as spatial point pattern data. Node information of cities and

regions must follow a complex array of dependent covariates, such as geographic location, topography, geology, economic activity, street linearity and so forth. It should therefore be noted that the aim is not to formulate a complex model that is capable of describing the evolutionary process of cities or regions. Instead, I aim at estimating a point process that is able to approximate the point pattern found in our nodal information dataset.

6.2 REGIONAL NODAL POINT PATTERN ANALYSIS

Point patterns are fundamental components of the study of spatial statistics. A point pattern describes the spatial arrangement of points in a two-dimensional space. Gatrell et al. (1996) define a point pattern as "a set of locations (s_1, s_2, etc.) in a defined 'study region', R, at which 'events' of interest have been recorded". The use of the terminology 'event' for such points has become a standard in point pattern analysis (Diggle 1983). Such events are then described by a vector of the form s_j, where s_{j1} refers to the 'x' coordinate and s_{j2} to the 'y' coordinate of an event (Gatrell et al. 1996 p. 258). These point patterns and the study of their underlying rule sets or spatial point processes plays a pivotal and long-standing role in many scientific fields, including geography (Jensen-Butler 1972), ecology (Wiegand et al. 2009), zoology (Andersen 1992), epidemiology (Gatrell et al. 1996), astronomy (Babu and Feigelson 1996) among others. The field of point pattern analysis has seen an increasing interest in the last decade, due to the works of Peter Diggle (1983, 2014), Jesper Møller and Rasmus Waagepetersen (2004), Illian et al. (2008) and most recently Baddeley et al. (2016) as well as the advancement in computational power and applications. The main object in such enquiries is to understand the spatial arrangement of events and their underlying spatial process that generated them.

COMPLETE SPATIAL RANDOMNESS. The most basic form of a spatial point process is the aforementioned Poisson process, also known as complete spatial randomness (henceforth CSR). CSR implies that events are conforming to the principles of independence and equal probability (Bivand et al. 2008 p. 160). Independence means that the position of any given point in a pattern is independent of the position of any other point, while an equal probability specifies that any point in a pattern has an equal probability of being at a location. This also applies reciprocally to any given location. Both principles relate to what is described as first- and second-order effects in a point pattern. First-order effects are trends and variations in point pattern distributions that are operating across an entire region. Because first-order effects operate at a general level affecting all points, they can be described by properties such as intensity and spatial density (ibid. p. 163). When variations in point patterns are caused by point-to-point interactions, we speak of second-order effects. Second-order properties give

insight into point pattern trends towards being clustered or dispersed (ibid. p. 163). Diggle provided a mathematical definition for such a CSR point pattern as one that asserts a) "the number of events in any planar region A with area $|A|$ follows a Poisson distribution with mean $\lambda |A|$" and b) "given n events x_i in a region A, the x_i are an independent random sample from the uniform distribution on A" (2014 p. 10). Hence, all regions within a study area have the same likelihood of events to appear or not to appear and events do not aggregate or disperse dependent on the occurrence of other events. CSR usually constitutes the H_o hypotheses with which point patterns are compared.

If a point pattern does not conform to this hypothesis, it usually falls into either of two categories, namely a clustered pattern or a dispersed pattern. Clustered patterns emerge through some form of attraction mechanism, while dispersed patterns emerge when there is some form of inhibition mechanism at work (Bivand et al. 2008 p. 160). To identify what kind of underlying pattern an observed dataset features, one can start with a visual observation of the data when plotted in a two-dimensional space. Figure 62 shows such a representation for the sections of the two datasets. In these plots (a, b), we can see that the observed pattern might be of a clustered kind, rather than a dispersed one and rather unlikely of uniform distribution. With regards to plot c), one can see that these clusters themselves might form a pattern, which here appears to conform to a dispersed formation.

Detailed section of the nodal point pattern for both cases

a)

b)

c)
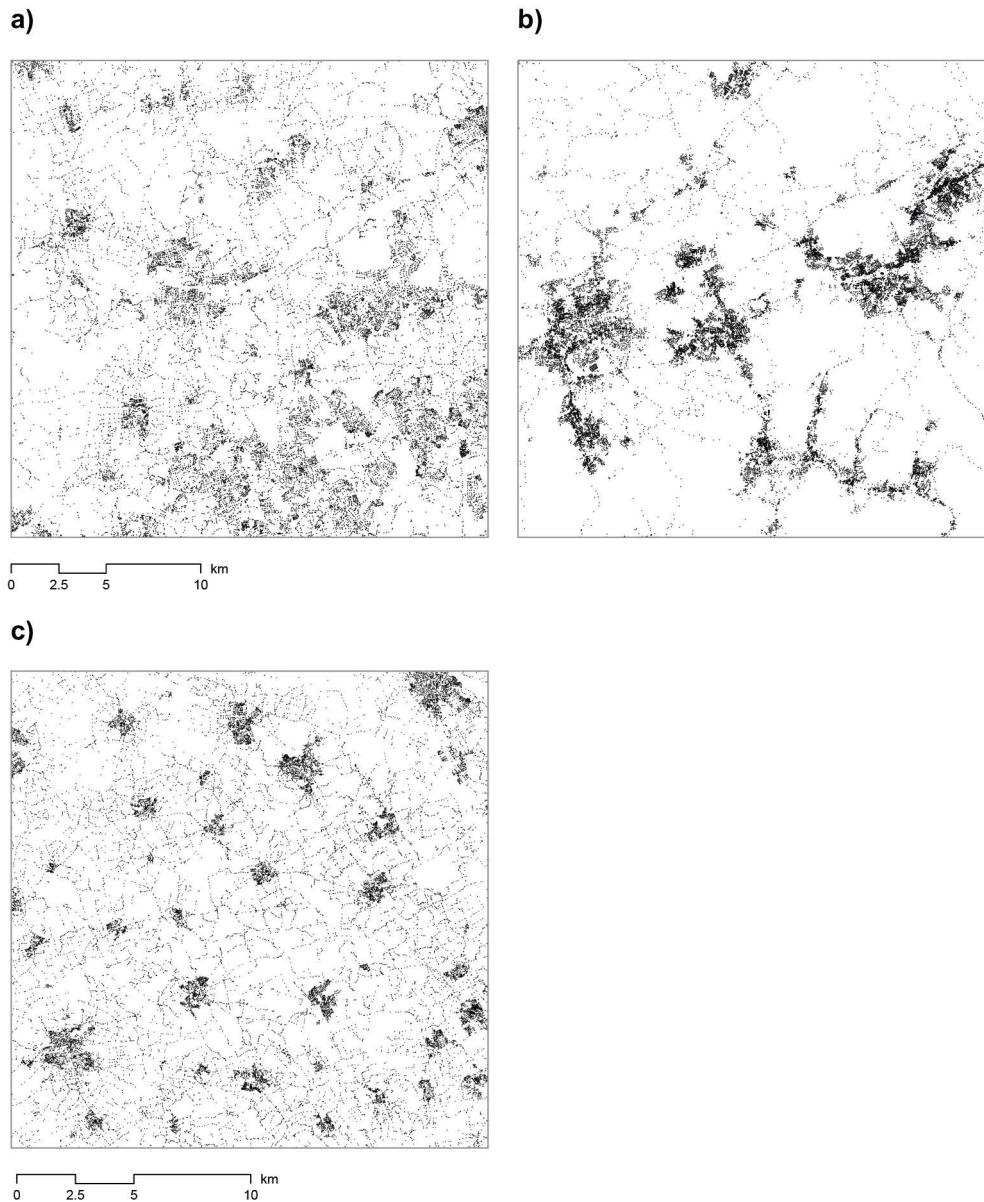
Figure 62: Detailed 30 by 30km (a, b) and 20 by 20km (c) section of the nodal point pattern for both cases GE (a, c) and UK (b).

However, visual observations can often be misleading or fail to identify underlying mechanisms. Exploring the point pattern or falsifying the observations made can be done by a series of statistical functions. These statistical functions can be divided into non-parametric summary statistics (intensity λ, empty function F or spherical contact distribution H_s), nearest neighbour distance distributions (function G and D), as well as second-order characteristics (g, K and L) (Baddeley et al. 2006 p. 9). These functions are then applied, while assuming that the observed point pattern is a stationary process, meaning that it is statistically invariant under translation (Baddeley et al. 2016 p. 146). A stationary point process has a homogeneous intensity because a shift of the

observation window will not affect the number of points per unit area. However, in reality, such characteristic is hardly found and only very few real-world processes follow actual CSR. Most of the observed point patterns are influenced by some form of underlying covariates. This is an important limitation for the statistical tests, which compare observed patterns against this assumption. I will nevertheless maintain the assumption that on sufficiently large regions settlement distributions may be homogeneous.

Generally, in point pattern analysis, the first step is to identify if the observed pattern follows CSR, which if confirmed means that a Poisson process can be used to reproduce a similar pattern, as the observed phenomenon has been the product of complete randomness and the properties of CSR are known. If, however, CSR cannot be confirmed, the first step is to identify the point pattern properties through an exploratory data analysis (EDA) to decide on an appropriate model. We will employ the function stated below in an EDA, which will be used to inform the subsequent model selection.

INTENSITY. Estimating the intensity function of a point pattern usually forms the first step in exploratory point pattern analysis (Diggle 2014 p. 57). This can be done by means of a density plot, which is an estimate of the intensity function of the point patterns' underlying point process. Intensity refers not to the density but to the expected number of random events per region. The units of intensity are stated as events per unit area. This is also known as local intensity $\lambda(x)$. Where $E[X]$ denotes the expectation of a random variable X; $Y(A)$ denotes the number of events in the planar region A; dx is an infinitesimal region that contains the point x; The integral of the intensity function over a spatial region gives the expected number of points falling in this region.

$$\lambda(x) = \lim_{|dx| \to 0} \left(\frac{E\{Y(dx)\}}{|dx|} \right) \qquad (6.1)$$

The Kernel density plots are produced as a method to overcome an inherent sampling problem of the so-called quadrat analysis, in which the study area is divided into small quadrats and the number of events falling into each quadrat is counted. Quadrat analysis is highly influenced by the size of the partitioning quadrats. To overcome this problem, kernel density plots employ a Gaussian kernel to smooth the counts making the interpretation of the analysis independent of quadrants. An important factor for the results of this method is the *sigma* value, which needs to be chosen by the researcher and determines the bandwidth of the kernel. Depending on the bandwidth the resulting plot will exhibit a smoothened pattern, where lower bandwidth produces a higher level of detail and higher values lead to higher smoothening effects. The bandwidth can also be

defined algorithmically through methods such as *likelihood cross-validation* (Loader 1999) or *mean-square error minimisation* (Berman and Diggle 1989). Yet, these approaches can differ substantially in their results, due to their underlying assumptions about the point relationship (Baddeley et al. 2016 p. 171). Hence, bandwidth selection needs to be well founded on a known relationship of the point pattern or its underlying process.

CLARK-EVANS TEST. A simplistic test for analysing spatial point patterns on aggregation is the Clark-Evans test (Clark and Evans 1954). This test is based on the Clark-Evans aggregation index, which calculates the average nearest neighbour distance for m random sample. The average is then divided by the expected value of a point process under CSR. A value lower than 1 suggests clustering, while a value higher than 1 indicates ordering. The significance at the 0.001 level can be tested by means of 999 Monte Carlo simulations[20].

F-FUNCTION. The empty space function F, sometimes also referred to as *spherical contact distribution*, or *point to nearest event distance* (Diggle 2014 p. 26) is used to statistically describe the size of gaps or average space between events in a point pattern. It provides the distribution function from a random point to the nearest random point of a point pattern. This is done by generating a set of random sample points m in the planar region A. The event distance function then calculates the proportional distance of m points to all points of the point pattern within a given radius r (where $r_i \leq r$).

$$\hat{F}(r) = m^{-1} \tag{6.2}$$

The estimate of $\hat{F}$ can then be compared against a Poisson distribution for inferential purposes. The true value of F for CSR is:

$$F(r) = 1 - \exp(-\lambda \pi r^2) \tag{6.3}$$

$\hat{F}(r)$ for clustered point patterns is expected to have a curve that is below $F(r)$. If the point pattern features an inhibition then $\hat{F}(r)$ exhibits a curve that is above $F(r)$.

G-FUNCTION. Similar to the *F*-function is the *nearest neighbour distance* function G the cumulative distribution function of the distance from a point of a point pattern to its nearest other point of the same pattern. Instead of arbitrary points like the *F*-function,

---

[20] Monte Carlo procedures are such procedures where distributions are simulated through random sampling.

G compares n points of the point pattern to the nearest neighbour within the same pattern (ibid. p. 24). $\hat{G}$ for point patterns can be calculated as (where $r_i \leq r$):

$$\hat{G}(r) = n^{-1}$$ (6.4)

The estimate of $\hat{G}$ can then be compared with the theoretical Poisson distribution of which the true value is:

$$G(r) = 1 - (1 - \pi r^2 |A|^{-1})^{n-1}$$ (6.5)

G is, however, only an approximation, because of the inevitable edge effects of point patterns. $G(r)$ ignores this by noting $|A|$ as the area of A where $\pi r^2 |A|^{-1}$ is the probability under CSR of a point being in the distance or r of another point (ibid. p. 24). $\hat{G}(r)$ needs to be inversely interpreted to $\hat{F}(r)$. A $\hat{G}(r)$ that is above $G(r)$ indicates a clustered point pattern. If the point pattern features inhibitions then $\hat{G}(r)$ exhibits a curve that is below $G(r)$.

J-FUNCTION. The J-function was first proposed by Lieshout and Baddeley (1996). The purpose is to measure the strength and range of inter point interaction in a point pattern. J makes use of the *nearest event distance* function F and the *nearest neighbour distance* function G, which were introduced above. Inter point interaction can be quantified through the function J that is defined as follows:

$$J(r) = \frac{1 - G(r)}{1 - F(r)}$$ (6.6)

The theoretical value for CSR will lead to a $J(r)$ value of exactly 1. Deviations from 1 can then be interpreted as indicators for clustering or dispersion. Where $J(r) < 1$ indicates clustering and $J(r) > 1$ indicates dispersion.

K-FUNCTION. The K-function also called *Ripley's K-function* or *reduced second-moment function* (Diggle 2014 p. 57) is similar to the J-function, a method to analyse inter point relationships. K counts the numbers of events within a defined distance of another event. One of the benefits of K is its ability to describe point pattern characteristics at many different distances. This makes it possible to identify clustering or dispersion effects operating at different scales at the same time. A pattern could exhibit clustering at smaller radii and simultaneously feature a dispersed pattern at larger radii (see Figure 62:c), which will be legible through a plot of the K-function estimate:

$$\lambda K(r) = E[N_0(r)]$$ (6.7)

$N_o(r)$ is the number of additional events within distance r of a random event (ibid. p. 57). This can again be compared against the theoretical K-function for CSR, which can be calculated as follows:

$$K(r) = \pi r^2 \qquad (6.8)$$

Deviations from the theoretical K curve reveals spatial clustering or spatial dispersion within the point pattern analysed. Given the deviation of $K(r)$ is larger than πr^2 the observed point pattern features clustering, while a deviation smaller than πr^2 indicates a regular pattern.

6.2.1 POINT PATTERN ANALYSIS FOR REGIONAL CASES

The following part will present the results of the point pattern analysis methods introduced above to the two regional cases. It will be shown, that the regions under investigation are both characterised by clustering mechanisms of different degrees measurable even at small radii ranging from 100 to 500 metres. The two regional nodal point patterns are statistically described through their kernel intensity estimation at a 1x1km square region, a Clark Evans test as well as the G, F, J and K function estimators. All calculations in this section are done with the use of the *Spatstat* software package (Version 1.53-2) (Baddeley et al. 2016) for the statistical software R (R Development Core Team 2016).

Figure 64, shows said kernel intensity estimator for a regional section of 140 by 140km and two smaller detail sections of 40 by 40km each. The detail sections are selected based on their distinct pattern, which we divide into 'metropolitan' and 'rural'. This separation should not be seen as classification, but rather as an attempt to differentiate two observable patterns that are characterised through a) a clear delimitation from few point-like intensity clusters to their surrounding low intensity areas and b) a continuous area of alternating patterns of medium to high intensity with few low intensity areas in between (Figure 63 and Figure 64). These patterns are present in both cases and can – to a certain extent – be interpreted inversely to each other. Due to their characteristics, we will compare all sections and test their differences through a Clark-Evans test. This is also in an attempt to better meet the underlying assumption of the pattern under observation as a stationary point process. With regards to the kernel intensity estimator, all plots are using a kernel σ of 500, which is equivalent to 500 metres. A value of 500 for *sigma* has been chosen because a G, F, J and K function estimation has indicated that 500 metres serve as an appropriate distance for clustering across all estimators (we will elaborate on this later). Moreover, a point distance of 500 metres can also be linked to network distances of 500 to 1000 metres, which can be related to perceptions of

neighbourhood sizes. The result is hence expected to give a good indicator for urban clusters without oversimplification.

Figure 63: UK: Node kernel intensity estimation for a regional section of 140 by 140km, and two detailed sections of 40by 40km for a metropolitan and rural area. Intensity values range from 0 to 4e-04 with sigma = 500.

The British regional section depicts the location of each of the cities and their relative geographic centres very well (Figure 63). While Manchester stands out as the largest intensity agglomeration, its delimitations from Bolton, Bury and Rochdale become blurred. Depicted similarly is the intensity cluster around Leeds and Bradford (also visible on a larger scale in the metropolitan section). All cities can be identified through their nodal intensity bandwidth. The majority exhibits what can be described by the

concept of distance decay, where the number of nodes is the highest in the relatively central location of each city and then declines radially with the distance to the central location. This becomes more evident in the northern part of the German region section (Figure 64).
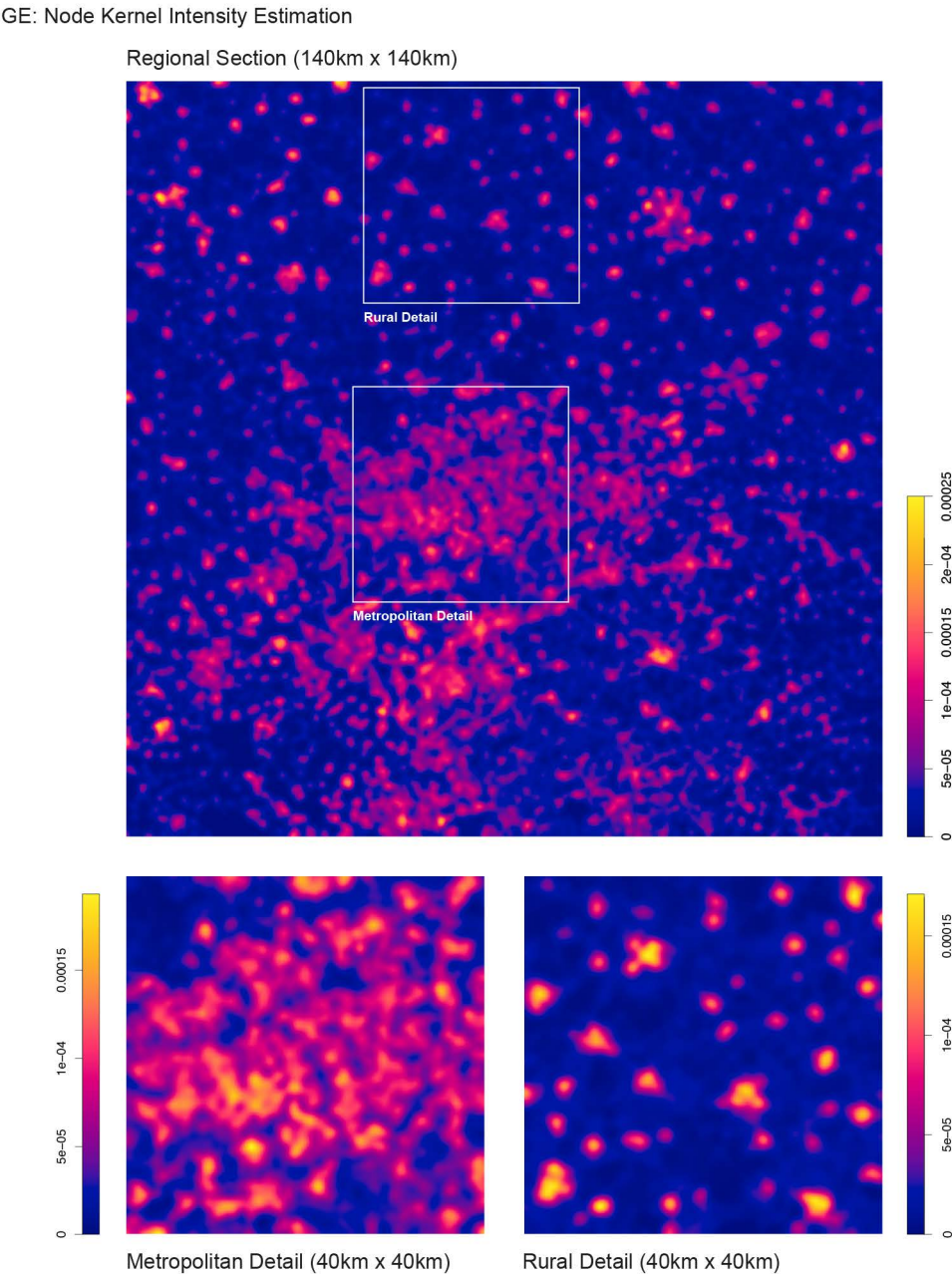
Figure 64: GE: Node kernel intensity estimation for a regional section of 140 by 140km, and two detailed sections of 40by 40km for a metropolitan and rural area. Intensity values range from 0 to 25e-04 with sigma = 500.

Different to the British cases, the German regional section exhibits a higher number of evenly spread small to medium-sized cities with relatively high intensity and clear demarcations to its surrounding areas (see rural detail Figure 64). Cities that are larger in size exhibit, once more, an effect similar to the one we observed in the British regional

and metropolitan section, where boundaries become fuzzy, and fringes blur into each other. The metropolitan core of the Ruhr Valley, with its continuously built urban form, can be identified very clearly and resembles the patterns observed in Greater Manchester, Leeds and Bradbury in the British regional section. When inspected more closely, the metropolitan details of both cases (Figure 63 and Figure 64) show a larger number of smaller cores of high intensity at what first seemed to be single core agglomerations. This becomes particularly clear with regards to cities such as Essen in the German case, where three high-intensity cores are present.

What we can draw from these initial observations is that nodal point patterns of regional street networks appear to cluster with higher intensities around geographic centres of cities. Moreover, the pattern of rural areas appears to feature a form of an initially dispersed pattern on larger radii, followed by high-intensity clustered centres around these dispersed cores. This is more so the case for the German model, as it is for the British, but the pattern is observable in both. The metropolitan pattern seems to feature a similar pattern, however, through extension and growth this pattern becomes indistinguishable from its initial seed pattern. Instead a continuous area of diverging intensity with above average intensity is present.

Table 19: Clark-Evans Test with Donnelly edge correction. Monte Carlo test based on 999 simulations of CSR with fixed n. Alternative hypothesis: two-sided for a) the German case and b) the British case.

a)

Point pattern	R	p-value
Region	0.66103	0.002
Metropolitan	0.78261	0.002
Rural	0.69098	0.002

b)

Point pattern	R	p-value
Region	0.57963	0.002
Metropolitan	0.69821	0.002
Rural	0.54424	0.002

A Clark-Evans tests, confirms this initial observation, in the sense that all six sections (regional, metropolitan and rural of both regions) score a value below 1, on the two-sided test indicating the presence of clustering patterns. The test was performed using a Donnelly edge correction, where the value of R represents the ratio between observed mean nearest neighbour distance and the edge-adjusted theoretical mean. All findings have a significance level of 0.002, based on 999 Monte Carlo simulations with fixed n. Notably, both metropolitan sections have a higher R, meaning lower cluster behaviour than the regional and rural sections. The British region exhibits stronger clustering patterns with $R = 0.544$ for the rural and 0.579 for the region, which are 0.1 index points higher than the German region. We can conclude that statistically there is indeed a

clustering mechanism at work, which is at a higher force in rural than metropolitan areas, as observed in the kernel intensity plots.

The following four pages (Figure 65, Figure 66, Figure 67 and Figure 68) show the results of F, G, J and K function, as well as the result of 99 simulations of a uniform Poisson process, for the British and German regional model. All analysis results indicate that a cluster mechanism is at work in both regions. This mechanism, is of stronger magnitude in the British case, compared to the German region.
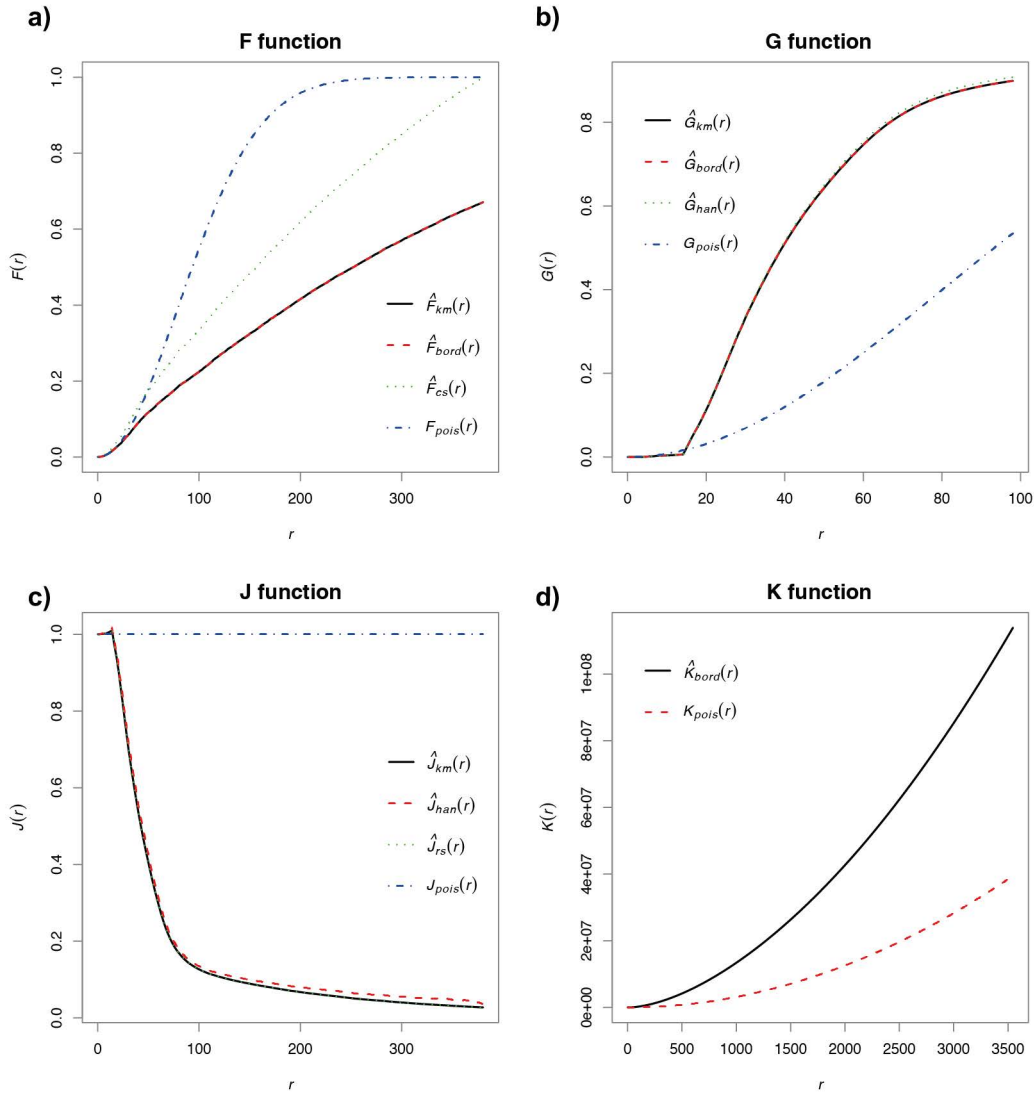
Figure 65: a) F, b) G, c) J and d) K-function for the UK regional nodal point pattern and distances up to 3km. Where r is the distance argument, $F_{km}(r)$, $G_{km}(r)$, $J_{km}(r)$ and $K_{km}(r)$ refers to the spatial Kaplan-Meier estimator, $F_{bord}(r)$, $G_{bord}(r)$ and $K_{bord}(r)$ are the border correction estimator and $J_{rs}(r)$ the reduced sample estimator, $F_{cs}(r)$ and $G_{cs}(r)$ are the Chiu-Stoyan estimator and $J_{han}(r)$ the Hanisch-style estimator, $F_{pois}(r)$, $G_{pois}(r)$, $J_{pois}(r)$ and $K_{pois}(r)$ are the theoretical Poisson (CSR).
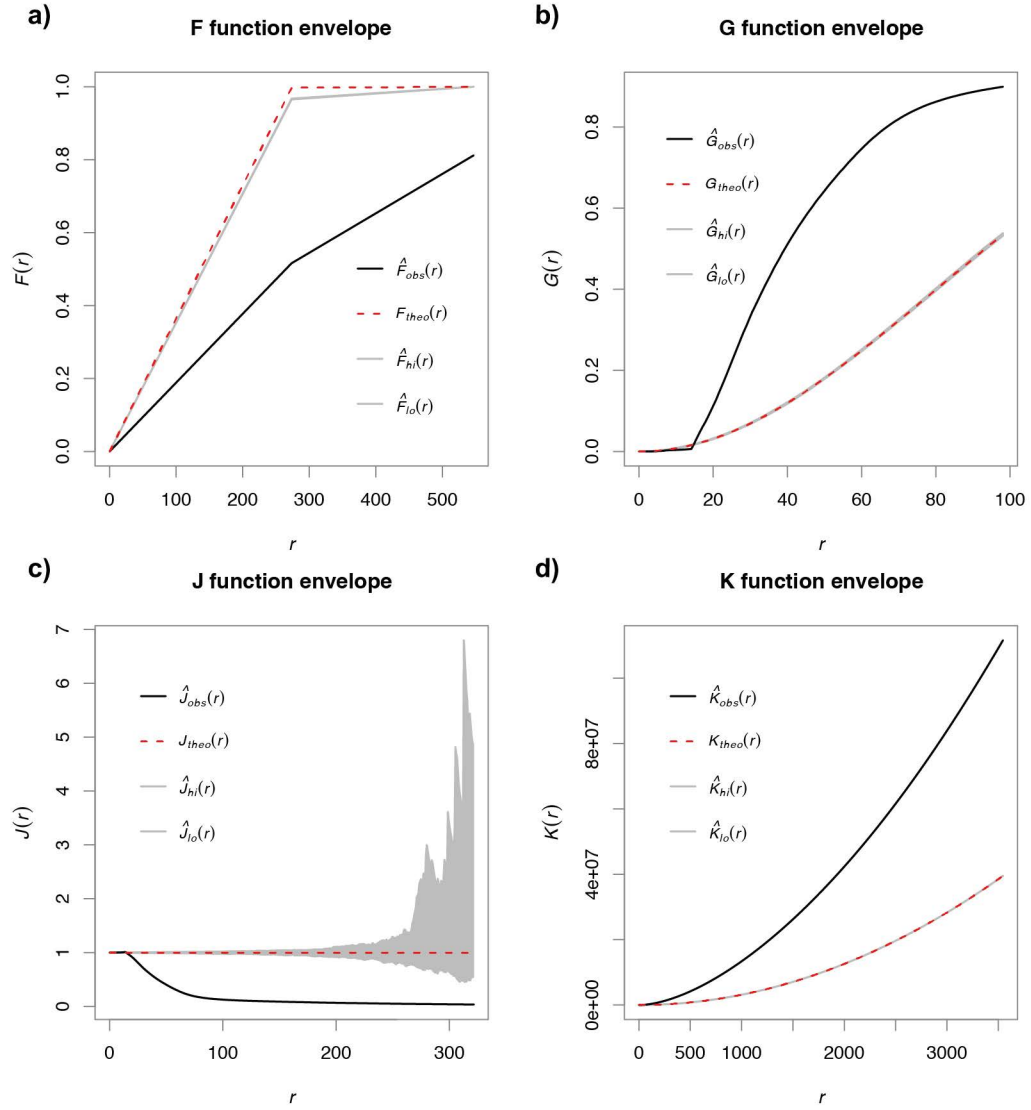
Figure 66: Upper ($F_{hi}(r)$, $G_{hi}(r)$, $J_{hi}(r)$, $K_{hi}(r)$) and lower ($F_{lo}(r)$, $G_{lo}(r)$, $J_{lo}(r)$, $K_{lo}(r)$) envelopes for 99 simulations of a uniform Poisson process with the same intensity as the British regional pattern. Observed values ($F_{obs}(r)$, $G_{obs}(r)$, $J_{obs}(r)$, $K_{obs}(r)$) that are outside the simulated envelopes are significant at the 0,01 level. The dashed lines ($F_{theo}(r)$, $G_{theo}(r)$, $J_{theo}(r)$, $K_{theo}(r)$) show the expected theoretical curve for each estimator.
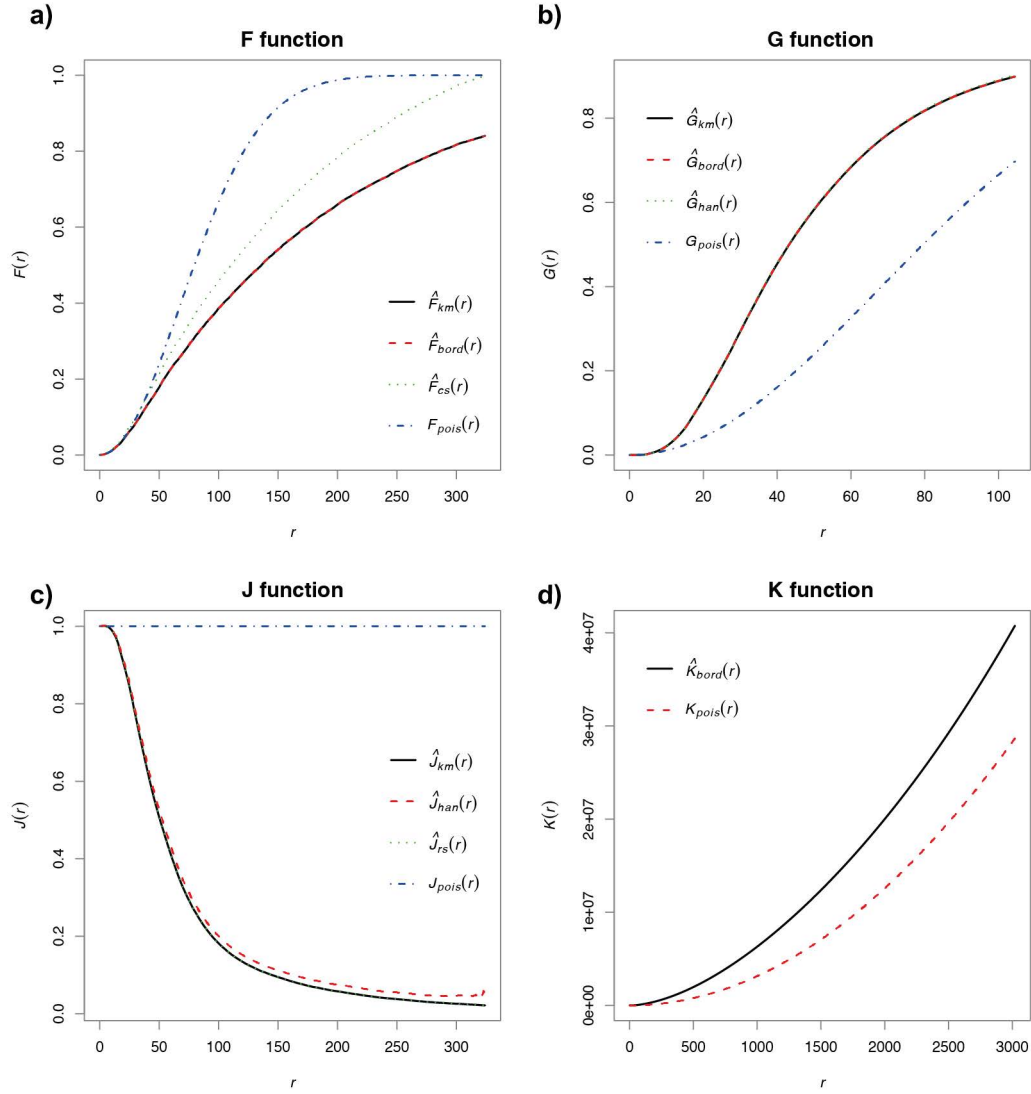
Figure 67: a) *F*, b) *G*, c) *J* and d) *K*-function for the GE regional nodal point pattern and distances up to 3km. Where *r* is the distance argument, $F_{km}(r)$, $G_{km}(r)$, $J_{km}(r)$ and $K_{km}(r)$ refers to the spatial Kaplan-Meier estimator, $F_{bord}(r)$, $G_{bord}(r)$ and $K_{bord}(r)$ are the border correction estimator and $J_{rs}(r)$ the reduced sample estimator, $F_{cs}(r)$ and $G_{cs}(r)$ are the Chiu-Stoyan estimator and $J_{han}(r)$ the Hanisch-style estimator, $F_{pois}(r)$, $G_{pois}(r)$, $J_{pois}(r)$ and $K_{pois}(r)$ are the theoretical Poisson (CSR).

a) **F function envelope**

b) **G function envelope**

c) **J function envelope**

d) **K function envelope**
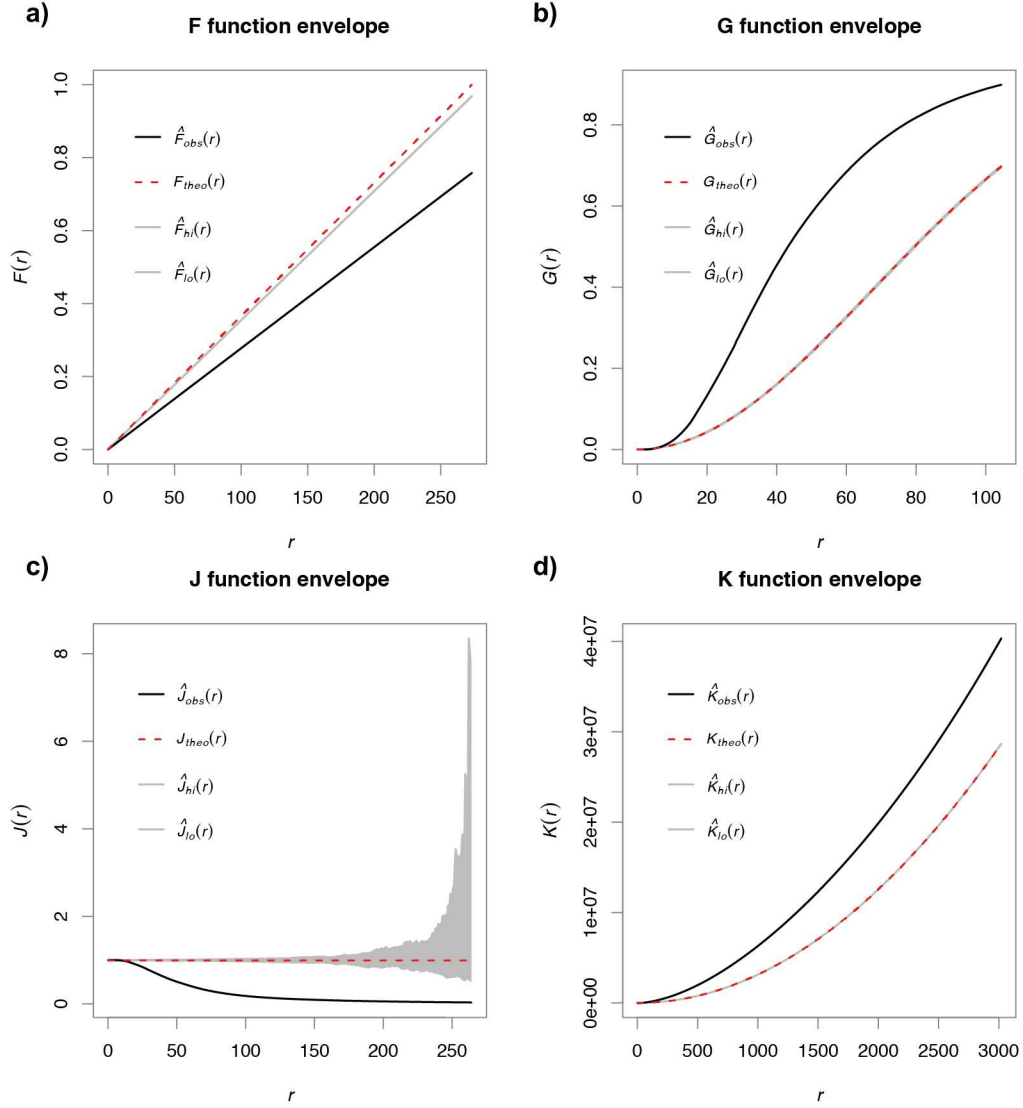
Figure 68: Upper ($F_{hi}(r)$, $G_{hi}(r)$, $J_{hi}(r)$, $K_{hi}(r)$) and lower ($F_{lo}(r)$, $G_{lo}(r)$, $J_{lo}(r)$, $K_{lo}(r)$) envelopes for 99 simulations of a uniform Poisson process with the same intensity as the British regional pattern. Observed values ($F_{obs}(r)$, $G_{obs}(r)$, $J_{obs}(r)$, $K_{obs}(r)$) that are outside the simulated envelopes are significant at the 0,01 level. The dashed lines ($F_{theo}(r)$, $G_{theo}(r)$, $J_{theo}(r)$, $K_{theo}(r)$) show the expected theoretical curve for each estimator.

6.2.2 SIMULATING COMPLEX SPATIAL POINT PATTERN

The previous part demonstrated that regional node patterns feature predominantly point clusters. The cluster mechanisms underlying this pattern are particularly observable at point relationships of smaller radii of 20 to 2000 metres. Such clustered point patterns can be simulated through fitted cluster point processes, given the assumption that the observed pattern is a realisation of a stationary stochastic process (Baddeley et al. 2016 p. 459). The basic concept for clustered point processes was first proposed by Neyman and Scott (1958, in Baddeley et al. 2016) in the context of cosmology in an attempt to model patterns of galaxies. Cluster point processes can be described by a simple two-step operation. First, a parent point pattern of a homogenous

Poisson process is generated. In the following step, a random number of independent and identically distributed offspring points is generated from the initial parent pattern (Figure 69). The final cluster point process then constitutes only the offspring points. This is also called a single-generation cluster, in opposition to a multi-generation process, which proceeds generating further offspring generations of each previous offspring generation, because the generative process stops after the first generation. Such a cluster process is usually referred to as Neyman-Scott process.

| Initial Parent Pattern | Offspring Cluster | Final Cluster Point Process |
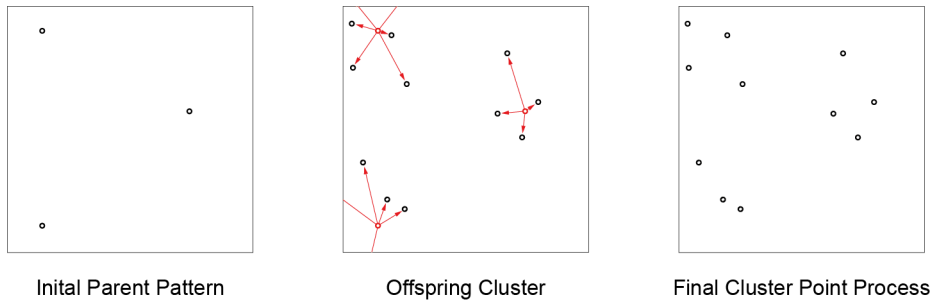
Figure 69: Two-step cluster process. Initial Poisson parent point pattern (left). Subsequent offspring cluster point generation of every parent point (middle). Final cluster point process, constituted by the offspring cluster (right).

Baddeley et al. (2016 p. 460) list a more rigorous definition of four main model assumptions in order to classify a clustering process as Neyman-Scott processes. These are a) the parent point process is a CSR Poisson process, b) clusters are independent of each other, c) clusters have the same distribution and d) the offspring are independent and identically distributed (ibid. p. 460). There are also alternations of the initial Neyman-Scott process. Such subcategories can be divided into processes that also fulfil the criteria e) the number of offspring depends on a Poisson random variable for each parent and f) the probability density of an offspring depends on its distance to the parent (ibid. p. 460). In the following simulations, I will focus on two particular cases of a subcategory of the Neyman-Scott cluster process that – it will be argued – are suitable to emulate the pattern previously observed at the kernel density plots. These two processes are the Cauchy cluster process (Ghorbani 2013) and the Variance Gamma cluster process (Waagepetersen 2007). There is a large array of other different cluster processes available that form variations of the above or use additional parameters in the process. I will not review them at this point and instead refer to the comprehensive work of Baddeley, Rubak and Turner (2016 p. 459) as well as Diggle (2014 p. 101).

What differentiates a Cauchy and Variance Gamma cluster process from a simple Neyman-Scott cluster process is a modification of the probability density of offspring. A simple Neyman-Scott generates offspring that are identically distributed, meaning that the probability of a point to be generated is not linked to its distance to the parent.

As shown in Figure 69, this leads to a uniform pattern around parent points. What we have observed in the kernel intensity plots earlier, is an increased intensity that decays with distance to its core and intensity cores (parent pattern) that are denser distributed in metropolitan areas and more dispersed distributed in rural areas. This pattern is better visible in Figure 70 where at both bandwidths of sigma 500 and 250, agglomerated intensity cores are visible, particularly in the rural detail sections. This implies that point intensity in regions decays with the distance from the intensity cores, or in other words the probability density of a node declines with its distance to the centre of the pattern.

Figure 70: Node kernel intensity estimation with intensity contours for a metropolitan and rural small detail of 10 by 10km. Intensity values range from 0 to 1e-04 with sigma = 500 (top) and 0 to 2e-04 with sigma = 250 (bottom).

Figure 71 illustrates how such an underlying mechanism would need to be incorporated into a cluster point process. First, a Poisson parent point pattern is produced followed

by a modified probability density depending on the distance to the parent. The probability density needs to be of a heavy-tailed kind in order to allow point occurrence at far distances to the core, This also coincides with the observed log-normal or Pareto distributions function of existing regional street length distribution and the general geography law of distance decay.
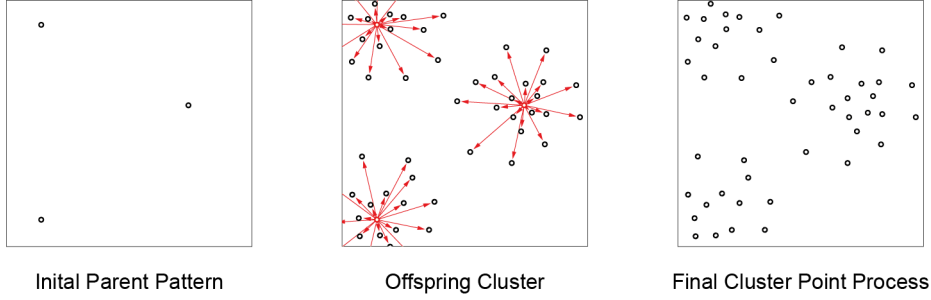
Figure 71: Two-step cluster process with modified probability density. Initial Poisson parent point pattern (left). Subsequent offspring cluster point generation of every parent point where the probability density declines with distance to the parent (middle). Final cluster point process, constituted by the offspring cluster (right).

Both, Cauchy and Variance Gamma cluster processes are characterised by such a modification of the probability density of the offspring. The probability density in both processes is extremely heavy-tailed, allowing offspring at a far distance from its parent (Baddeley et al. 2016 p. 463).

CAUCHY CLUSTER PROCESS. The Neyman-Scott cluster point process with Cauchy kernels, also known as Cauchy cluster process, modifies the probability density of offspring through a bivariate Cauchy distribution,

$$h(u) = \frac{1}{2\pi\omega^2}\left[1 + \frac{\|u\|^2}{\omega^2}\right]^{-3/2}$$

(6.9)

where ω is the scale parameter that modifies the tail behaviour of the distribution (ibid. p. 463). Jalilian et al. (2013) have proposed a fitting process that estimates the necessary parameters from existing data. A fitted model can then be used in the following step to simulate realisations of the fitted model. The fitting process estimates the parameters through the method of minimum contrast (Diggle and Gratton 1984), which compares the theoretical K function to the observed K function and computes the minimum distance between both theoretical and empirical curves. The estimated parameters for the simulation process are *kappa*, the intensity of the Poisson process for parent points, *mu* the number of offspring points per cluster drawn from a Poisson distribution and *scale* the parameter for the Cauchy kernel. The model returns a realisation of the Neyman-Scott process with a modified distribution in the given region. It should be noted that the process also generates parents and offspring outside the given region to

overcome potential edge effects. A theoretical density distribution for cluster offspring of the Cauchy point process in a three-dimensional space is shown in Figure 72.
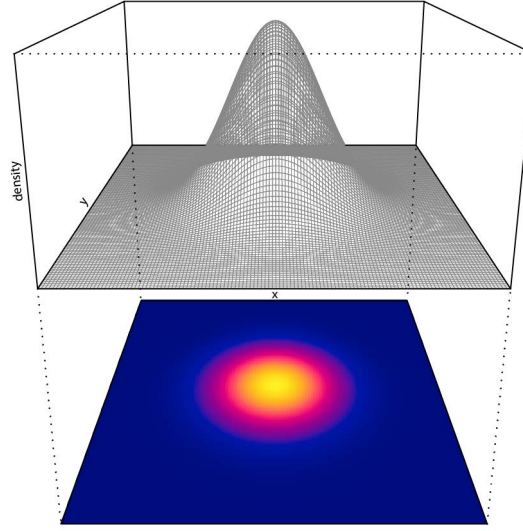
Figure 72: Bivariate Cauchy density distribution (grey mesh) and the corresponding intensity kernel (bottom).

VARIANCE-GAMMA CLUSTER PROCESS. Similar to the Cauchy cluster process, the Neyman-Scott cluster point process with Variance Gamma kernel, also known as Bessel, modifies the probability density of offspring. This is done by means of a Variance Gamma distribution, as follows

$$h(u) = \frac{1}{2^{v+1}\pi\eta^2\Gamma(v+1)} \frac{\|u\|^v}{\eta^v} K_v\left(\frac{\|u\|}{\eta}\right)$$

(6.10)

where η is the scale parameter (Baddeley et al. 2016 p. 464). Additionally, v controls the shape of the density, while the gamma function is Γ and K_v is the modified Bessel function of order v (ibid. p. 464). The model fitting process is similar to Cauchy, and the estimate parameters are comparable to the Cauchy cluster process, with an additional shape parameter index nu, which determines the shape of the kernel (Jalilian et al. 2013). In general, the model fitting procedure will initially fit the intensity of the empirical data and then the cluster parameter (Baddeley et al. 2016 p. 474). As mentioned earlier this is done under the assumption that the observed pattern is stationary.

6.2.3 POINT PROCESS SIMULATIONS

Both algorithmic cluster processes, Cauchy and Variance Gamma, have been fitted to the regional, the metropolitan and the rural section of the British and German region. This is to compare the parameters of each fitting process, which can be found in Table

20 and Table 21. The table allows a direct comparison of the effects of both approaches to potential cluster simulations of these models.

Table 20: Parameter estimates for stationary cluster point process models for the entire British regional point pattern and two selected point pattern details (metropolitan and rural). Models fitted by minimum contrast and using the K-function summary statistic.

Fitted point pattern	Cluster model	Uniform intensity	Kappa	Scale	Mean cluster size (in points)
Region	Variance Gamma process (nu=-0.25)	2.500E-05	6.940E-09	2.280E+03	3646.094
Region	Cauchy process	2.53E-05	6.238E-09	1.328E+03	4059.115
Metropolitan	Variance Gamma process (nu=-0.25)	5.620E-05	5.395E-08	1.205E+03	4059.115
Metropolitan	Cauchy process	5.990E-05	4.594E-08	7.245E+02	1110.720
Rural	Variance Gamma process (nu=-0.25)	9.140E-06	1.032E-08	1.602E+03	1304.429
Rural	Cauchy process	9.890E-06	8.031E-09	1.021E+03	958.040

Table 21: Parameter estimates for stationary cluster point process models for the entire German regional point pattern and two selected point pattern details (metropolitan and rural). Models fitted by minimum contrast and using the K-function summary statistic.

Fitted point pattern	Cluster model	Uniform intensity	Kappa	Scale	Mean cluster size (in points)
Region	Variance Gamma process (nu=-0.25)	3.480E-05	7.940E-08	9.950E+02	439.059
Region	Cauchy process	3.490E-05	6.292E-08	6.267E+02	554.403
Metropolitan	Variance Gamma process (nu=-0.25)	6.010E-05	3.814E-07	6.650E+02	158.567
Metropolitan	Cauchy process	6.050E-05	3.025E-07	4.180E+02	199.904
Rural	Variance Gamma process (nu=-0.25)	2.250E-05	7.764E-08	7.489E+02	292.984
Rural	Cauchy process	2.270E-05	5.929E-08	4.865E+02	383.677

The *uniform intensity* is the estimated intensity of the observed pattern, whereas *kappa*, *scale* and the *mean cluster size* are the fitted parameters. The British fitted models exhibit expected values for the estimated uniform intensity. Highest values are to be found in the metropolitan section with the densest urban area, and the lowest values

appear in the rural section where streets are rather scarce, while the regional uniform intensity is somewhat in-between. All parameter estimates of Cauchy and Variance Gamma appear to be comparable with the only outlier between the *scale* parameter for the regional section and the *mean cluster size* for the metropolitan section. The fitted pattern of the British region features a smaller *kappa* value, leading to fewer clusters and a larger *scale* value with a larger *mean cluster size*. The pattern will therefore have fewer clusters with more points and a larger radius. If we compare this with the observed pattern of the British case, and in light of the model assumptions made earlier, the model parameters appears plausible. The British region features fewer cities with a larger population size and larger urban areas. Hence a simulated model should represent a smaller *kappa* value with larger *scale* and larger *mean cluster size* accounting for the real-world pattern.

On the contrary, the German case features much smaller *mean cluster sizes* in the regional section with values of a maximum of 554.403. The clusters do not only have lower mean values but also comparably lower *scale* values. Different from this, all sections feature much higher *kappa* values than observed in the British sections. Simulated point patterns for the German case are hence characterised by a larger amount of smaller clusters with fewer points, which is in agreement with what the kernel intensity plots have indicated earlier. However, variations appear in the parameter estimates of the two rural sections, with a *kappa* value of 7.764E-08 for Variance Gamma and 5.929E-08 for Cauchy. A similar difference appears on the *scale*, while inversely the *mean cluster size* is higher for the fitted Cauchy model and lower for the Variance Gamma. Nevertheless, both fitted models result in the same *uniform intensity*.

Generated spatial realisations of each simulation provide a visual representation of the differences between each fitted model, which can be seen in Figure 73 to Figure 76. Each model simulation features three realisations.

UK: Cauchy Point Process Simulation
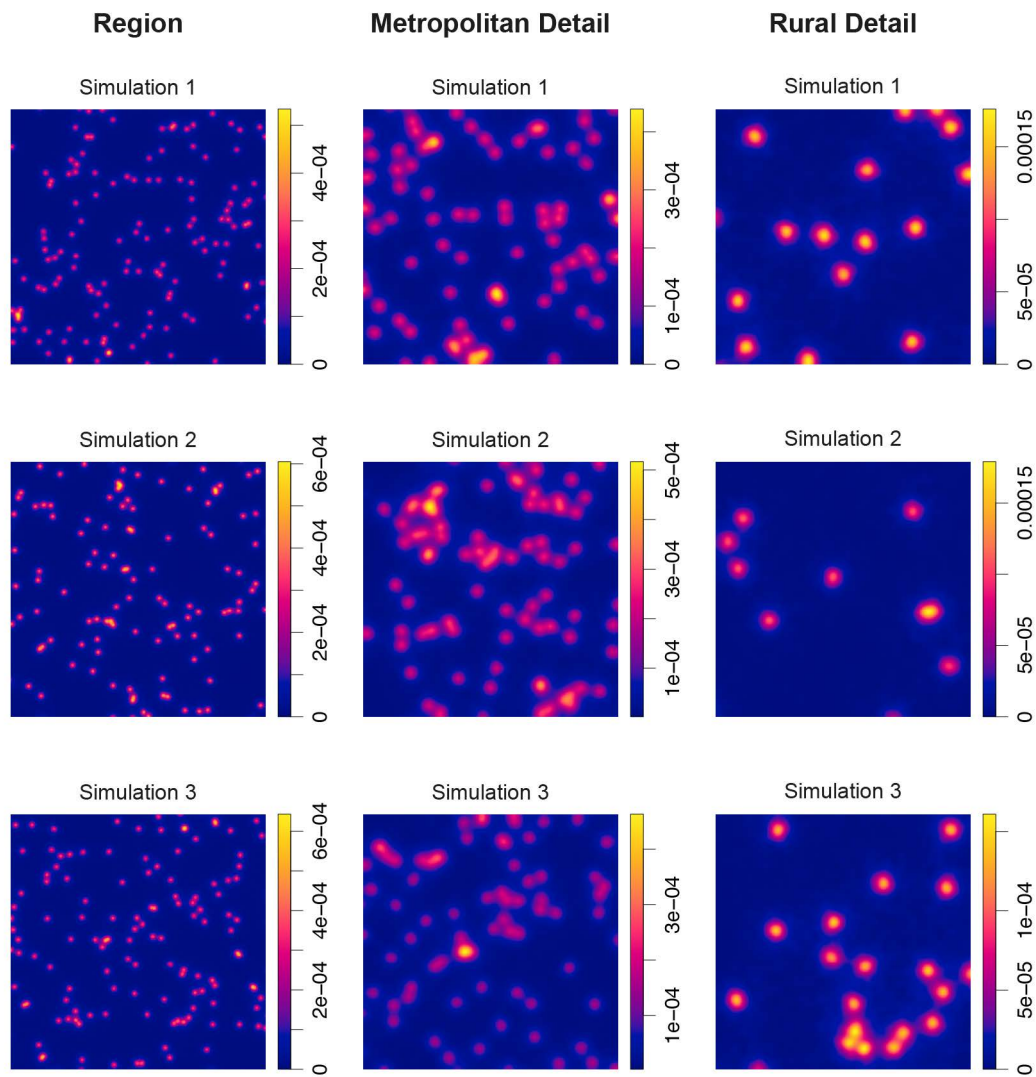
Figure 73: Simulations for the British regional, metropolitan and rural detail of a fitted Neyman-Scott cluster process with Cauchy kernel and the method of minimum contrast, using the K function.
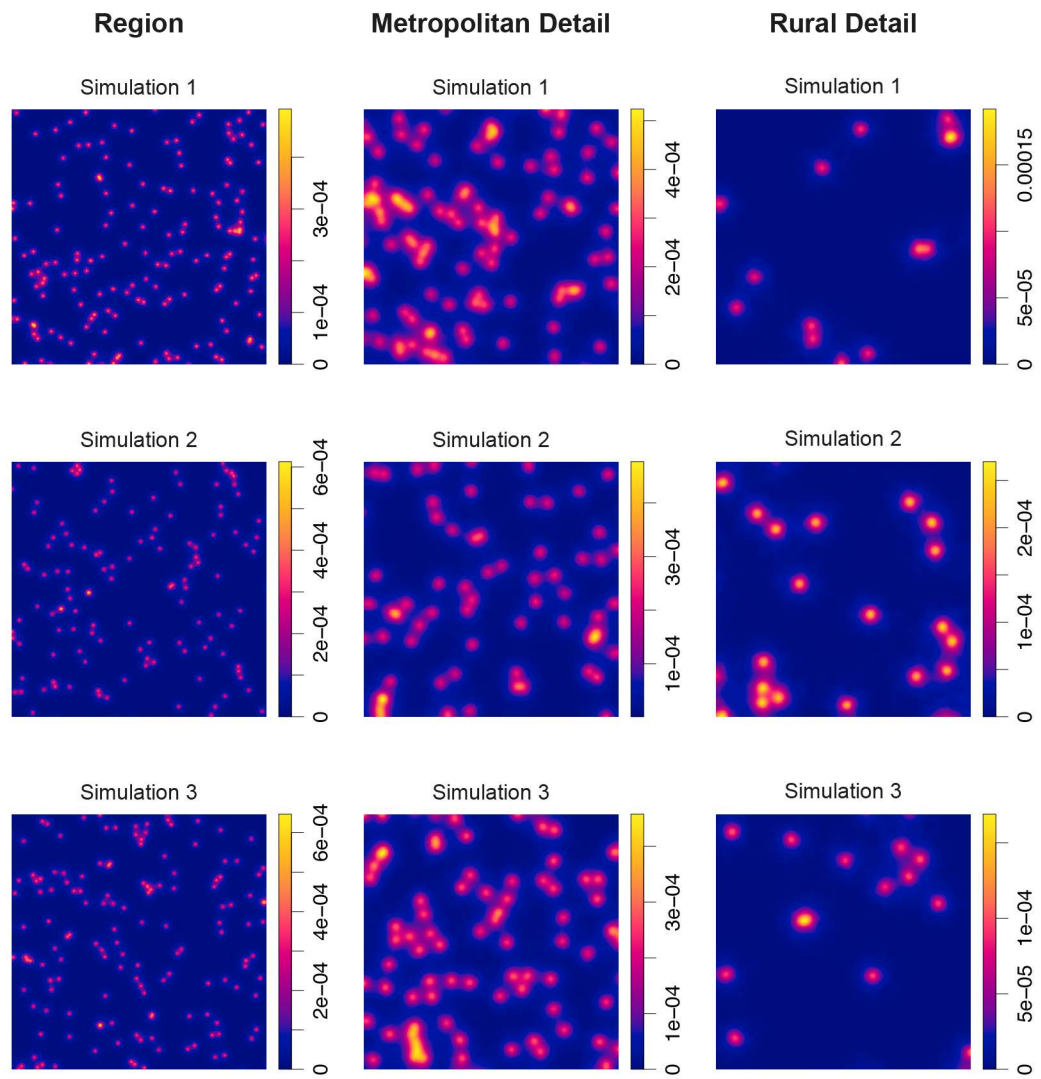
Figure 74: Simulations for the British regional, metropolitan and rural detail of a fitted Neyman-Scott cluster process with Variance Gamma kernel and the method of minimum contrast, using the K function.
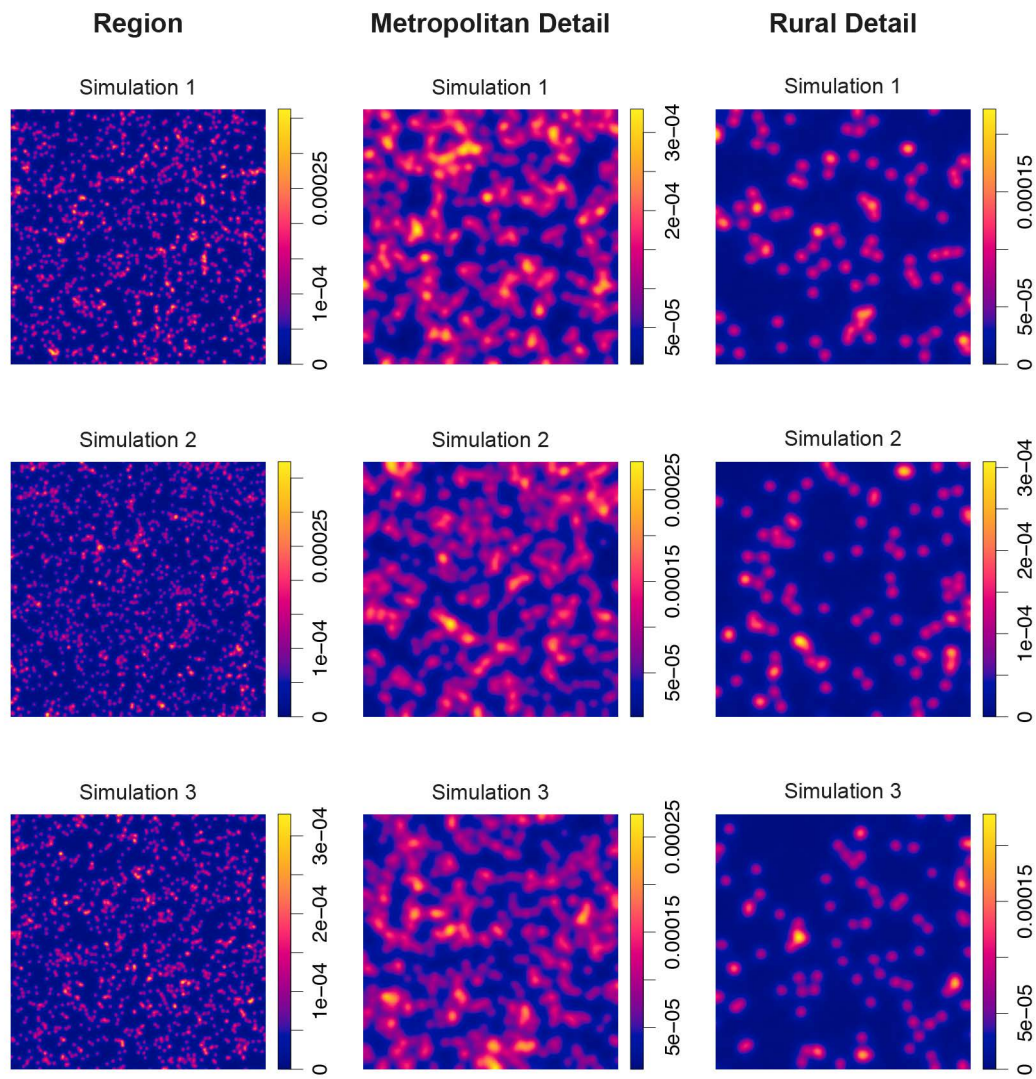
Figure 75: Simulations for the German regional, metropolitan and rural detail of a fitted Neyman-Scott cluster process with Cauchy kernel and the method of minimum contrast, using the K function.
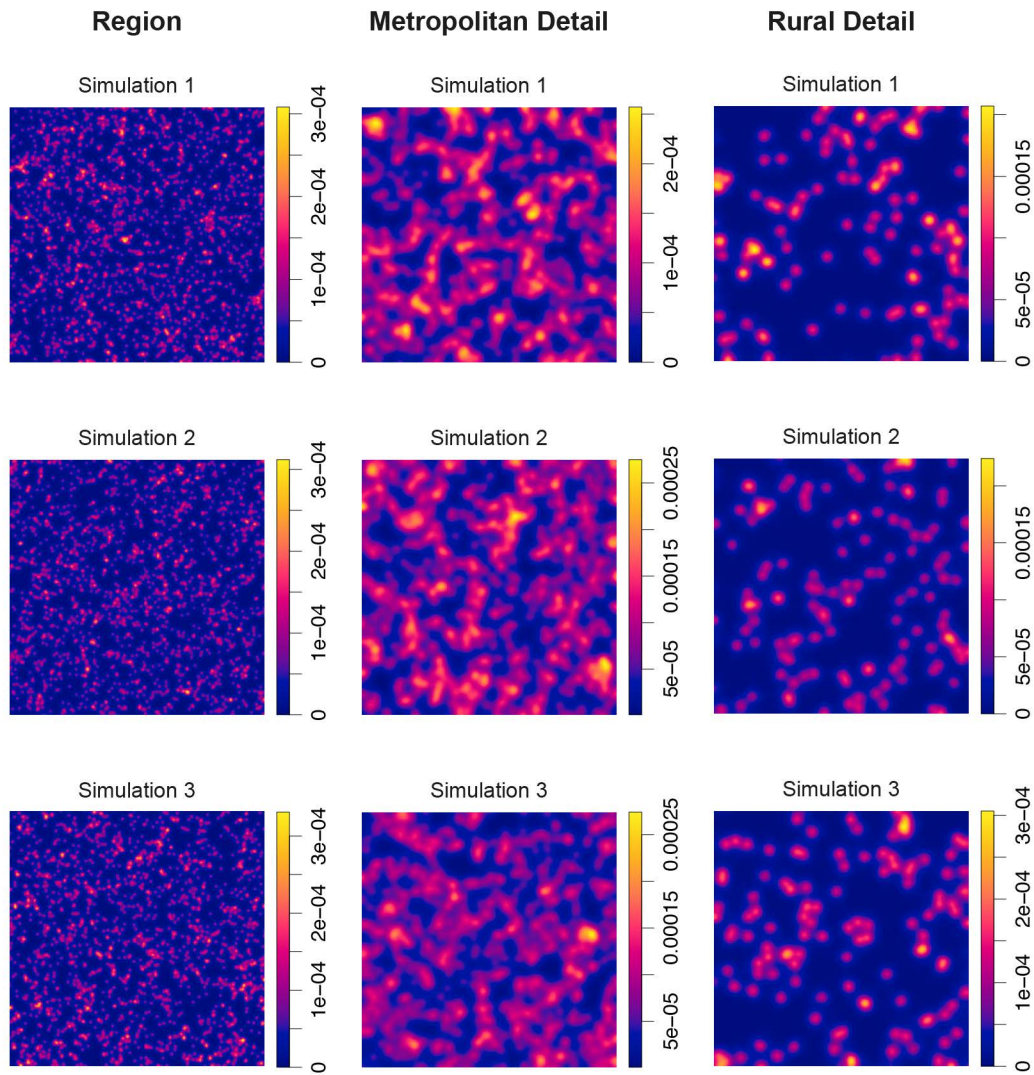
Figure 76: Simulations for the German regional, metropolitan and rural detail of a fitted Neyman-Scott cluster process with Variance Gamma kernel and the method of minimum contrast, using the *K* function.

All simulations constitute statistically feasible realisations of the theoretically set out simplified model of regional nodal density clusters. Due to its higher uniform intensity, I have at random chosen one of 99 generated simulations of the fitted Variance Gamma model of the German region to proceed with the street network model generation. In order to the test whether the set aim of emulating the segment length distribution observed in existing regions can be fulfilled in the newly generated Variance Gamma point cluster and hence in the final Variance Gamma planar graph with radius restriction model (henceforth VPGr), I have estimated the nearest neighbour distance kernel density per k nearest neighbours. Similar to the previous kernel density estimation (Figure 59), the generated VPGr model (Figure 77) exhibits a log-normal or Pareto-like distribution of potential segment lengths. Once again and similar to the

German case, on which the model is based, the density curves are developing from a rather log-normal like distribution with small sigmas gradually towards a log-normal like distribution with larger sigmas. This stands in contrast to the non-clustered ERPGr model (Figure 61), which featured a pyramid-like curve development without a long-tail behaviour.
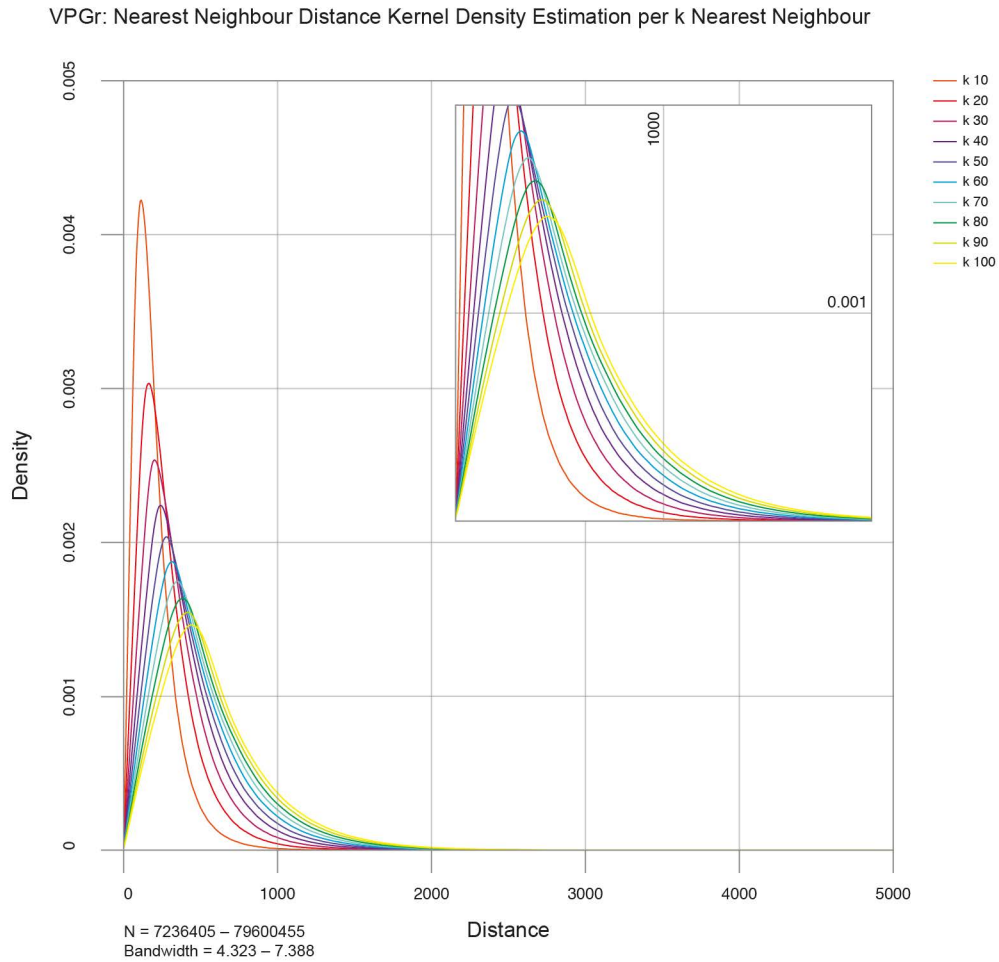
Figure 77: VPGr: Segment Length Density per k Nearest Neighbour, for 10 different k groups (k = 10 – 100). N pairs range from 8,837,272 (k=10) to 86,209,992 (k=100), with a bandwidth range of 2.083 (k=10) to 4.978 (k=100).

6.3 TWO RESULTING RANDOMLY GENERATED STREET NETWORK MODELS

Following the presented methodology, two spatial network models have been generated. These models were the Erdős-Rényi Random Planar Graph with radius restriction (ERPGr) and a Variance Gamma planar graph with radius (VPGr). Figure 79, shows the realisation of the ERPGr model, and Figure 79 shows the model realisation of the VPGr. Both figures present a zoomed-in detail section on 5 different scales, highlighting the

level of detail and providing an insight into the extent of both models. A detailed description of the model morphologies can be found in chapter 7, together with a comparison against the two real-world case study regions.
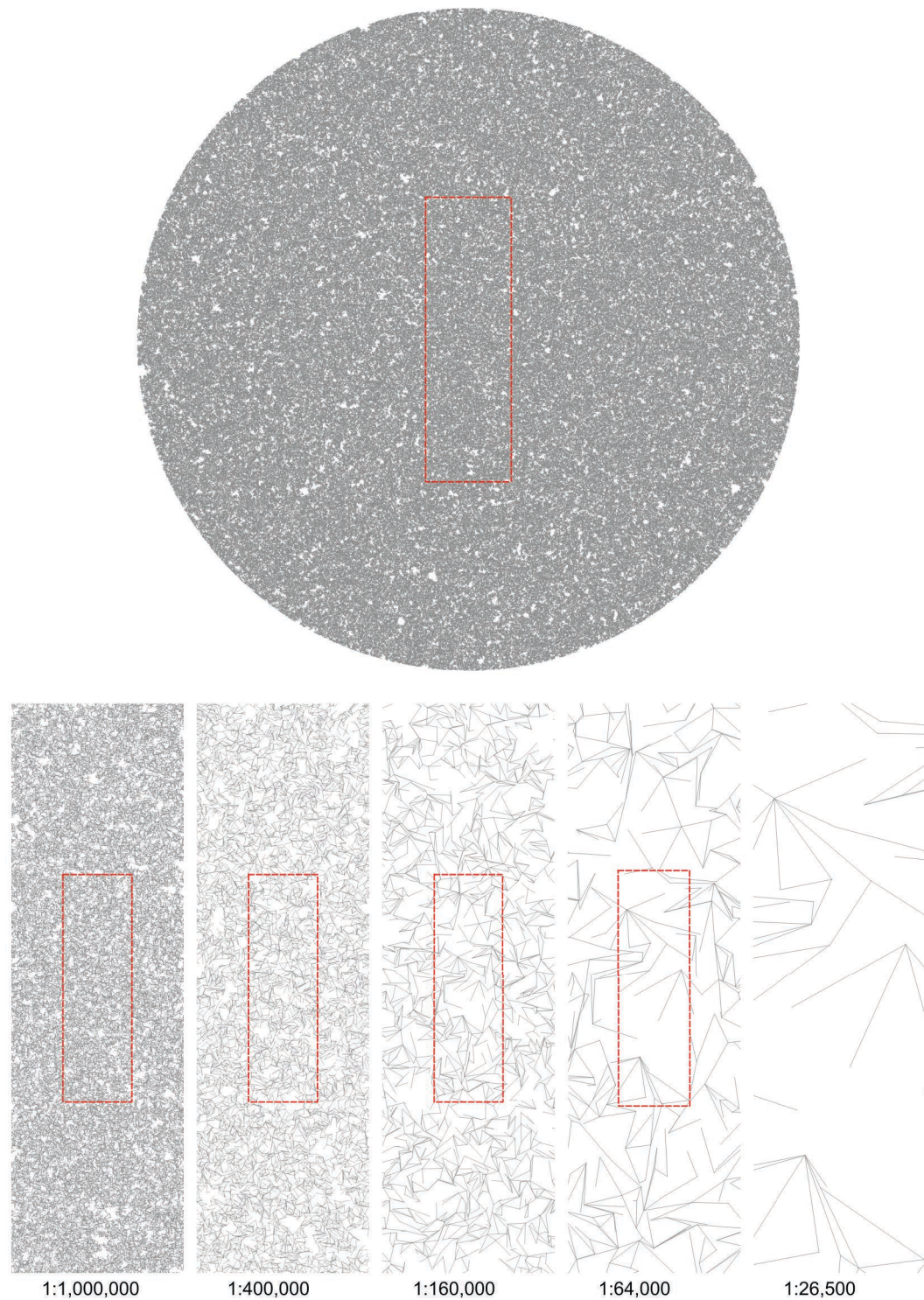
Figure 78: ERPGr network model with selected detail areas on five different scales of 1:1,000,000, 1:400,000. 1:160,000, 1:64,000 and 1:26,500.
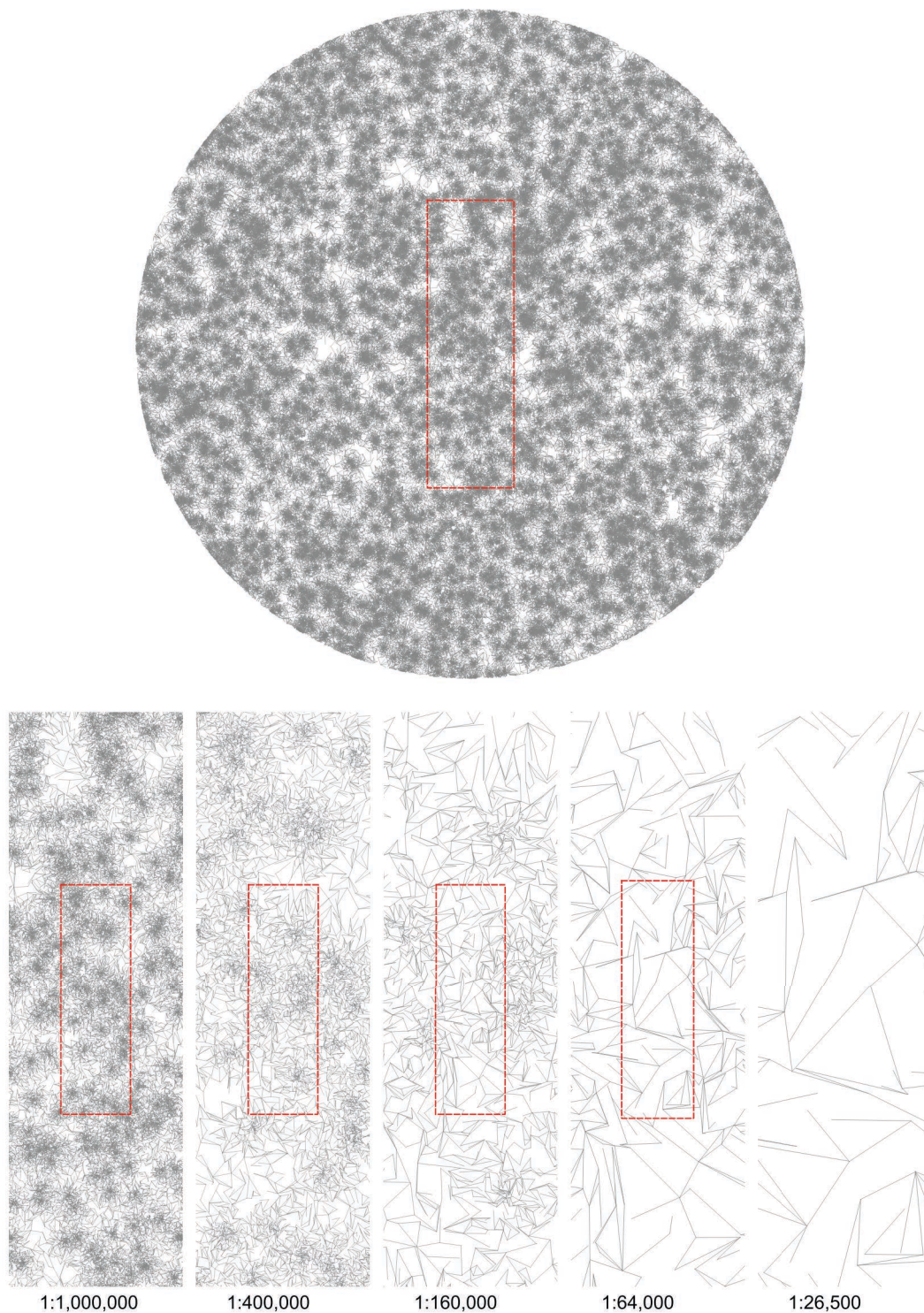
Figure 79: VPGr network model with selected detail areas on five different scales of 1:1,000,000, 1:400,000. 1:160,000, 1:64,000 and 1:26,500.

6.4 SUMMARY

This chapter discussed the difficulties in generating random regional street networks for the comparison of spatial metrics. It demonstrated that a traditional Erdős-Rényi Random Planar Graph with radius restriction differs significantly from real world street networks in their segment and segment length distribution. These characteristics, however, are important factors for the comparability of spatial networks. This is specifically the case when Euclidean distance plays a role in the analysis. I proposed the Variance-Gamma Planar Graph with radius restriction as a new method for the random regional street network generation. Kernel density estimations for segment length densities have shown that the VPGr is highly comparable with real-world street networks in their network characteristics, but are random in their spatial configuration. This allows comparisons of the effect of the spatial configuration of real-world networks with those that are a product of a random process. In order to verify these initial observations, I will compare an ERPGr, as well as the VPGr against both real-world regional street network models in the following analysis.

The following chapter will present the result of such a comparative analysis. It will propose an exploratory factor analysis as a method to reveal the fundamental structural differences in regional spatial networks and compare the two real-world regions (UK and GE) against the proposed randomised model VPGr and the ERPGr.

CHAPTER 7

LATENT CENTRALITY STRUCTUES IN POLYCENTRIC URBAN REGIONS

7 CHAPTER

This chapter presents exploratory factor analysis (hereafter EFA) as a method to analyse regional street networks to identify underlying structures of scale. It elaborates further on previously identified issues of scale and radius selection within space syntax in particular and spatial network centrality measures in general, as previously laid out in chapter 2. I will start by introducing factor analysis as a method to reveal latent centrality structures in regional street networks as well as presenting a justification for the method in the context of space syntax. This is followed by the presentation of results of comparing four models: the Erdős–Rényi Random Planar Graph with radius restriction (ERPGr), the Variance Gamma Random Planar Graph with radius (VPGr), the UK model as well as the GE model, which will be compared on two different network centrality measurements of betweenness and closeness centrality on a set of 49 different radii.

It will be argued that spatial networks feature inherent latent centrality structures that can be revealed through the proposed exploratory factor analysis method. I will highlight the fundamental differences between human-made street networks and randomly produced street networks and will discuss suggestions on the nature and cause of these differences. I then introduce a method of data representation in order to visualise the results of such an analysis, which enables morphological interpretations of the findings. Finally, I propose a combined multi-dimensional latent centrality model. This combined model allows simultaneous statistical comparisons across scales and provides a solution to arbitrary radius selection within space syntax and spatial network analysis.

7.1 LATENT CENTRALITY STRUCTURES IN POLYCENTRIC URBAN REGIONS

7.1.1 FOUR STREET NETWORK MODELS

This chapter's statistical exploration is based on the four previously introduced street network models (Chapter 5 and Chapter 6). A detailed review of these models can be found on the following pages (Figure 80–Figure 83). The figures provide an overview of the entire model, as well as five detailed sections of each respective model on different scales. This allows a superficial comparison of the four models and their morphological differences. The first two models are the randomly generated Erdős–Rényi Random Planar Graph with radius restriction (ERPGr) and the Variance Gamma Random Planar Graph with radius restriction (VPGr). The latter two models are the real-world

simplified regional model, based on the simplification method for OpenStreetMap data introduced in Chapter 5, one of the case study area in the United Kingdom (UK) and the other of a simplified regional model for the case study area in Germany (GE). While the general network characteristics, size, number of segments and number of nodes are comparable across all cases, the distribution of segment lengths is only comparable between the latter three models (VPGr, UK and GE). The core difference between the randomly generated networks (Figure 80 and Figure 81) and the real-world case studies (Figure 82 and Figure 83) is the random nature of the spatial configuration.

In their spatial arrangement, neither the ERPGr model nor the VPGr model exhibit any apparent top-level structure, i.e. a structure where linearity occurs continuously over long distances. The probability of segments is modulated only by the occurrence of other segments and a restricting radius. In addition to this, the VPGr model features a clustering mechanism influenced by the underlying Variance Gamma distribution. This becomes clearly visible in the 1:1,000,000 and 1:160,000 scale detail sections of both models (Figure 80 and Figure 81), where the VPGr exhibits segment clusters that are based on the Variance Gamma distribution. The ERPGr has a uniform segment distribution. Neither of the two model networks feature any symmetry or regularity in pattern in their network. Linearity, in the sense of several segments forming a linear path with little angular differences between them, however, do occur occasionally in both models, but they never form a continuous network. Blocks formed by street segments are either triangular or multi-polygonal, but very rarely quadratic and never consecutively regularly quadratic as can be seen in the 1:64,000 and 1:25,600 scale detail sections (Figure 80 and Figure 81).

The UK and the GE model on the other hand are characterised by contrasts between dispersed and clustered segment agglomerations (Figure 82 and Figure 83). Unlike the randomly generated models, real-world networks feature a large degree of ordered patterns, with occurrences of consecutively regular blocks and quadratic arrangements. This can be seen in the 1:64,000 and 1:25,600 scale detail sections (Figure 82 and Figure 83). Moreover, both regional models feature characteristic networks of long, linear and continuous motorway roads. These are a result of either historical reinforcements of existing path and way connections or newly introduced patterns through large-scale planning processes. These continuous networks play a significant role on large radii as they form efficient connections of shortest path journeys through the system, and as such are not present in the randomly generated models. Additionally, there are two main differences between the regional model of the central NDY region of the United Kingdom, with regards to the intensity of segment densities, namely that the UK model exhibits few segment clusters of large scale urban agglomerations. The British urban region can be described by a small number of highly centralised cities that grew in size

and presumably at the expense of the rural surrounding areas in time. The German urban region, on the other hand, features a much larger number of small to mid-size towns and cities, as well as large-scale conurbations; this leads to a much denser network of settlements.

| 1:1,000,000 | 1:400,000 | 1:160,000 | 1:64,000 | 1:26,500 |
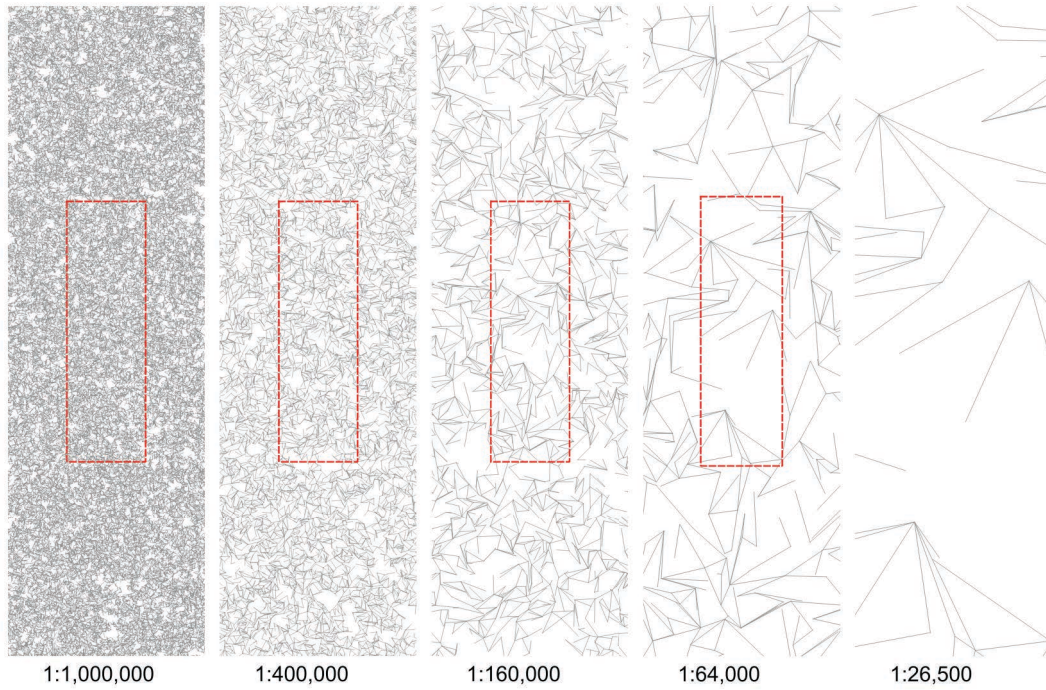
Figure 80: Selected detail areas of the ERPGr network model on five different scales of 1:1,000,000, 1:400,000. 1:160,000, 1:64,000 and 1:26,500.

| 1:1,000,000 | 1:400,000 | 1:160,000 | 1:64,000 | 1:26,500 |
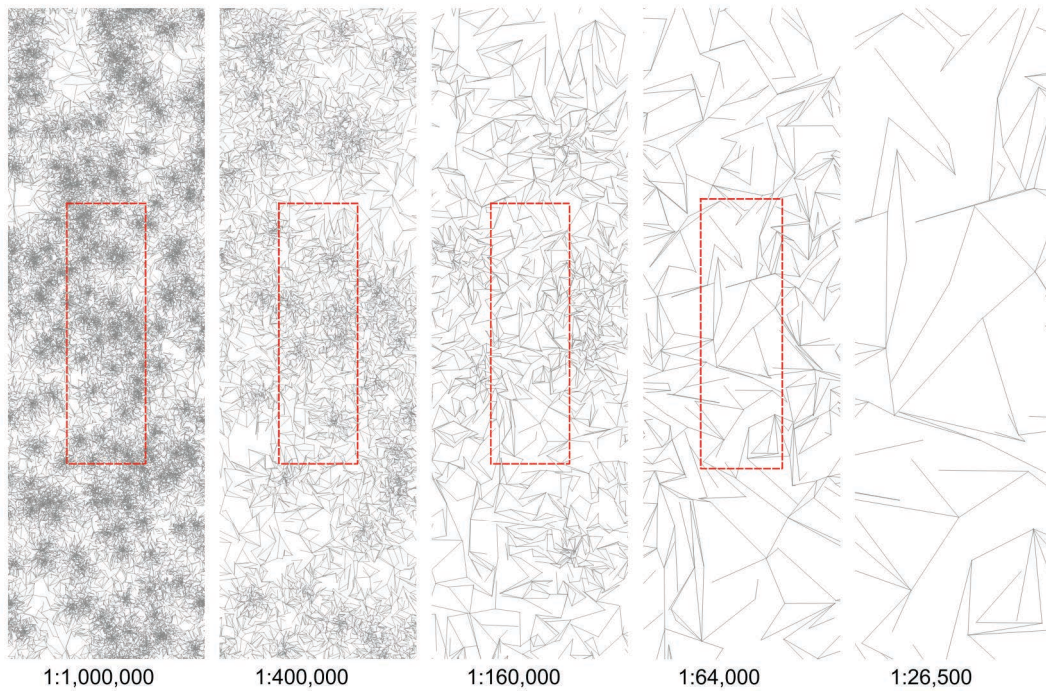
Figure 81: Selected detail areas of the VPGr network model on five different scales of 1:1,000,000, 1:400,000. 1:160,000, 1:64,000 and 1:26,500.
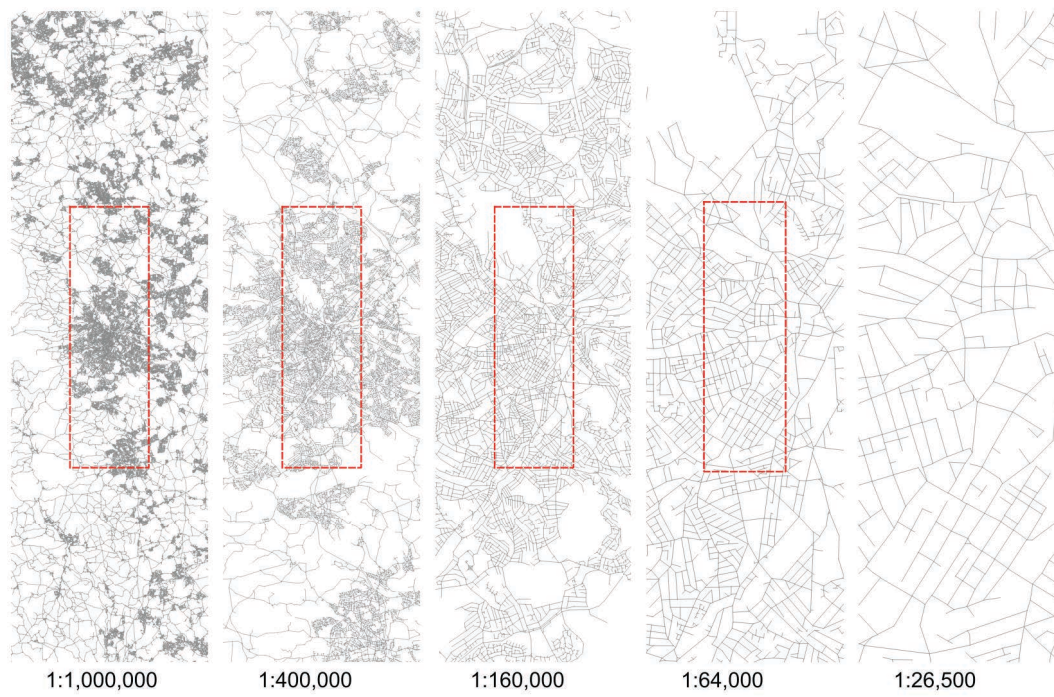
Figure 82: Selected detail areas of the UK simplified network model on five different scales of 1:1,000,000, 1:400,000. 1:160,000, 1:64,000 and 1:26,500.

Figure 83: Selected detail areas of the GE simplified network model on five different scales of 1:1,000,000, 1:400,000. 1:160,000, 1:64,000 and 1:26,500.

7.1.2 EXPLORATORY FACTOR ANALYSIS

This section introduces factor analysis as a statistical method that allows the extraction of a simpler mathematical description of centrality variables. It will be argued that this method can be employed to reveal the hidden and latent centrality structures of spatial networks. The factor analysis is a statistical procedure used to reveal a potentially existing lower degree of unobserved variables in an existing larger set of correlated variables. Looking back on a long history of applications in the social sciences, this approach has regularly been used in scenarios where researchers are facing large batteries (collections of variables) of correlating variables that might lead back to one underlying mechanism. The aim of this statistical procedure is to 'determine the number of distinct constructs needed to account for the pattern of correlations among a set of measures' (Fabrigar and Wegener 2011 p. 3). More specifically, factor analysis performs a series of correlations and tests whether the variation in a larger number of observed variables might be explained through a smaller number of unobserved variables. In the context of network analysis with a series of different centrality radii, this means that the analysis might yield new and unobserved variables that are able to explain the variance in the observed, measured radii. In practice, each network should have specific, structural scales that are significant for the respective analysis radii.

Imagine the following situation of a simplified network of two cities connected by a single street. Each city's edge-to-edge distance is shorter than the length of the connecting street between these two cities (Figure 84).
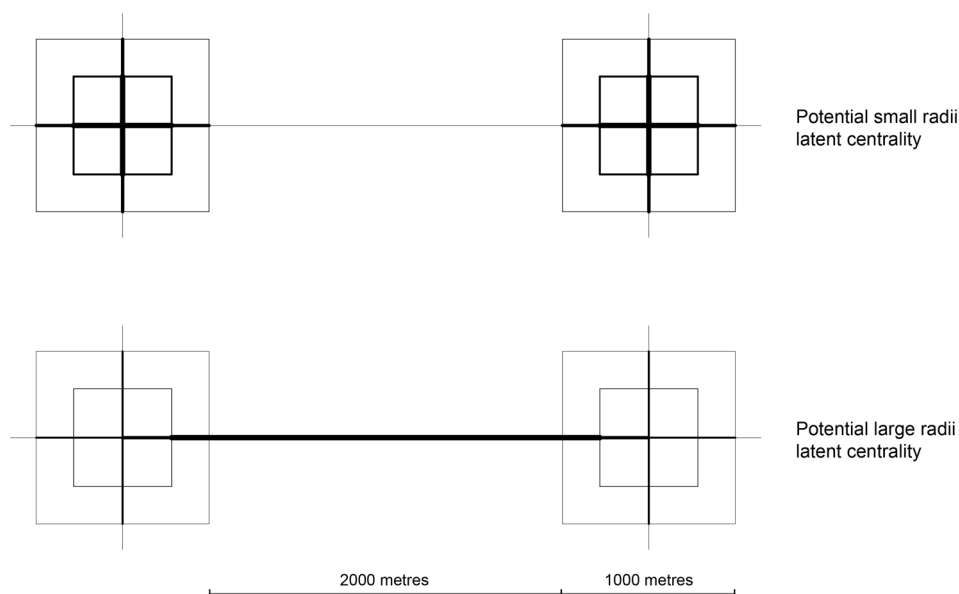
Figure 84: Potential latent centralities in an abstract network of two cities. Betweenness centrality values for two scenarios, small radii and large radii, thicker strokes indicate higher values.

If one performs a betweenness centrality analysis on different small radii (e.g. 100, 200, 300, 400) most likely each radius will correlate well with each other, but the correlation will show that one of these radii correlates the most with all the others. This most strongly correlating factor might be the closest to a potential, small-scale, latent centrality structure that governs the result for each radius. This specific pattern will not change as long as the selected radii do not reach a point where a different structure existent in the network can be captured, such as the connection between the two cities. This means that unless the analysis uses a radius that is long enough to reach from one city to the next city, all chosen radii will always describe a fraction of the small core radii's latent centrality structure. If, however, a set of larger radii is chosen (e.g. 3000, 3100, 3200, 3300) where journeys are also possible between both cities, the observable structure will change from an internal cross to that of a linearly connecting beam. All further radii will then look substantially the same, describing a potential, large-scale latent centrality structure, even if all segments are compared to each other.

A factor analysis, would in these cases, find a factor that is able to explain the different variance in the selected radii. In the example, an input dataset of 8 different radii (i.e. 100, 200, 300, 400, 3000, 3100, 3200, 3300) might, thus, arrive at two latent centrality factors that represent the fundamental difference in the spatial structure. With regards to network centralities and space syntax analysis, this would imply that instead of selecting a specific radius (or in some cases 'searching for the right radius'), the researcher could employ a large set of different radii and use an EFA to reveal such latent centrality structures. This might be of particular use, when the traditional boundary of the city dissolves into a regional network of cities and urban spaces and the need for an appropriate radius becomes more pronounced. As laid out in Chapter 2, based on Christaller's CPT and Hillier's movement economy, a set of distinctive latent centrality scales should emerge giving insights into the particular polycentrality of PURs.

Figure 85 shows an abstract regional context of the concept of latent centrality structures outlined above. In the example, the same distance segments connect four similarly sized cities with the same structure in a quadrangle arrangement. In such a case, two fundamental, latent centrality structures are expected to occur. In real world regions, as laid out previously and is visible in Figure 82 and Figure 83, spatial network feature a much higher complexity than can be intuitively grasped by a researcher, such as shown in the example below. It is not clear how many of such structures are there to be uncovered. Following the notion of Bill Hillier, at least two of such structures must be apparent in the context of cities, whereas Christaller's CPT hierarchy would predict at least seven of such structures. Together, all structures should form a hierarchy of latent centralities.

a) Regional Spatial Network

b) Latent Centrality Structure I

c) Latent Centrality Structure II
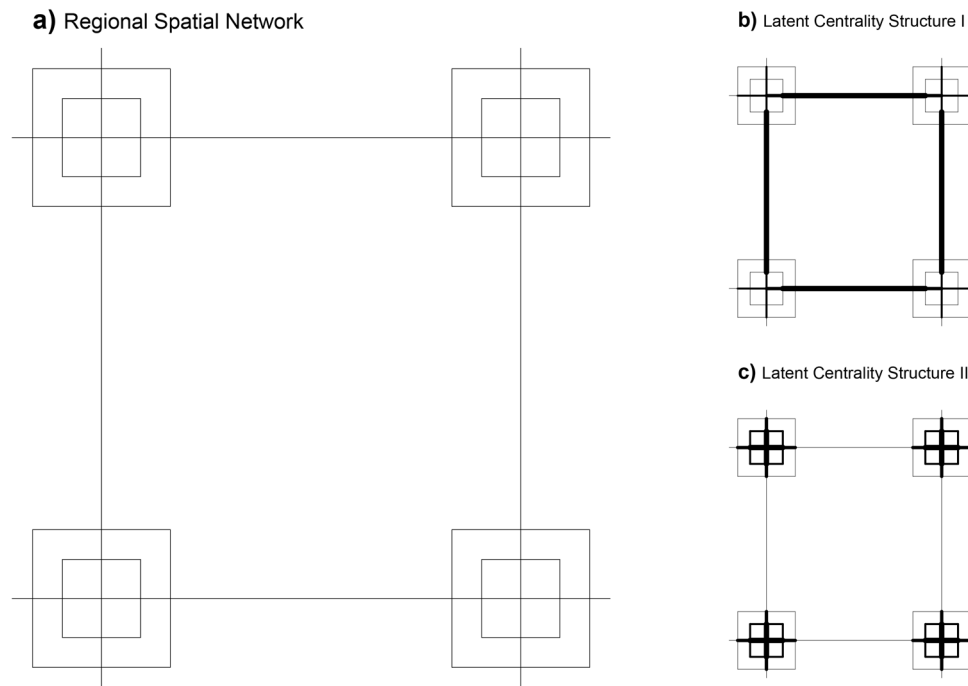
Figure 85: Abstract regional network (a), and two potential latent centrality structures expected to be present in the network (b, c).

JUSTIFICATION FOR EXPLORATORY FACTOR ANALYSIS. In their seminal study on the structure of the metropolitan city of Oporto, Serra and Pinho (2013) have dealt with a similar problem. They investigated closeness centrality structures on 15 different radii and proposed a principle component analysis (PCA) to arrive at a reduced dimensionality of these radii (ibid.). Their analysis yielded three components theorised as neighbourhood, city and regional scale. The authors described these components as 'natural centrality scales' and 'intrinsic hierarchical organisation of metropolitan centres' (ibid.). The reason for using a PCA analysis in their study was to arrive at 'variables that are contained, albeit not explicitly, in the original one' (ibid. p. 189). As outlined in the previous section, one of the aims of this study is to reveal the latent structures that cause the emergence of centrality patterns. For this purpose, I apply an exploratory factor analysis (EFA) to a series of radii. PCA and EFA are often confused and believed to be similar, or in the case of PCA, it is often believed to be a simpler form of EFA. In fact, however, the PCA functions on a different mathematical model than the EFA and can be distinguished from the EFA in several other aspects (Fabrigar and Wegener 2011; Widaman 2007).

PCA 'was not originally designed to account for the structure of correlations among measured variables, but rather to reduce scores in a battery of measured variables to a smaller set of scores (i.e., principle components)' (Fabrigar and Wegener 2011 p. 31). The main purpose of components derived by a PCA is to explain as much variance as possible from the original variables, rather than to explain the correlations among them

244

(ibid. p. 31). In that sense, the PCA is an efficient method to represent information in measured variables, while the EFA on the other hand produces common factors. These factors are unobservable latent constructs that are conjecturally causing the measured variables (Costello and Osborne 2005; Fabrigar and Wegener 2011 p. 31). In contrast to the PCA, which constructs components directly from the measured variables, the EFA's common factor model divides the variance in measured variables into common variance and unique variance (Figure 86).

Exploratory Factor Analysis

Graphical Representation of the Common Factor Model
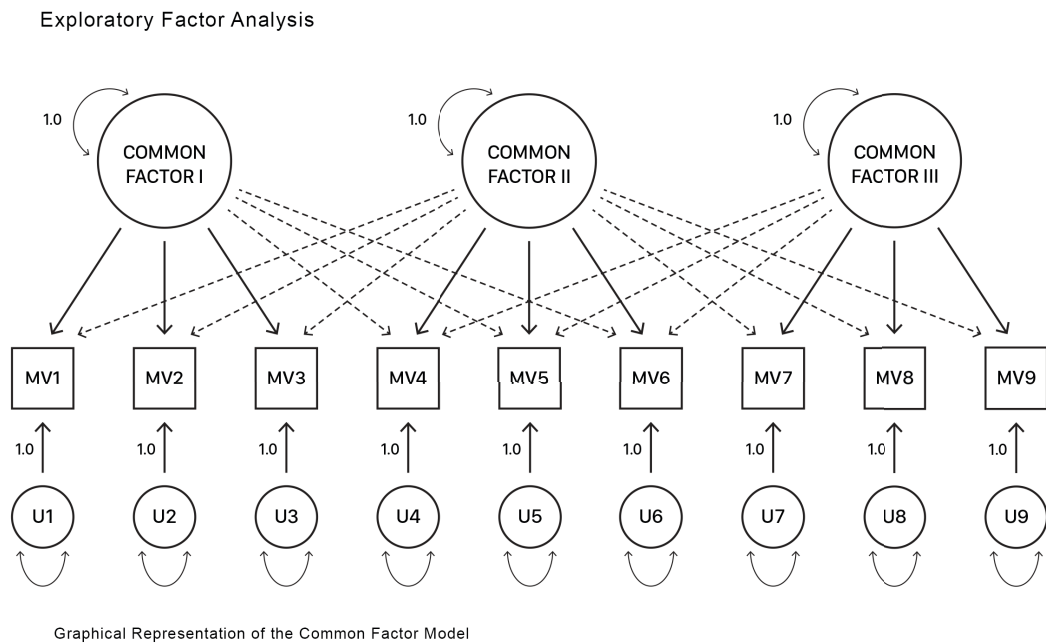
Figure 86: Illustration of the Common Factor Model for an example involving three common factors and nine measured variables. Where, MV equals measured variance, U equals unique variance.

The reasons why the EFA has been chosen over the proposed PCA in this study is because i) the general aim of this research is to identify latent constructs (spatial scales) that are thought to cause the measured variables (centrality pattern), to ultimately inform a broader theory. Is has been argued that EFA is the appropriate method for this purpose (ibid. p. 32),

ii) EFA is designed for cases 'in which the researcher has no clear expectations or relatively incomplete expectations about the underlying structure of correlations' (ibid. p. 4) as it is the case for patterns of centralities in spatial networks, and

iii) differently to PCA, EFA generates parameter estimates that allow a generalisation beyond the measured variable collection on which they are based (Widaman 2007). This means that the components and component loadings resulting from a PCA change with every time an additional variable is changed or removed. In the case of EFA, however, adding more measured variables (or radii) does not alter the parameter estimates, such

as the respective factor loadings for original measured variables, unless they rely on a new common factor that was not present in the original measured variable collection (Fabrigar and Wegener 2011 p. 33). If one includes a sufficient number of radii in the analysis, making sure that the differences between each radius are small enough, one can assume to capture all existing latent centrality factors. These advantages make EFA more robust in the context of variable/radii selection and investigations of scale structures. This study will therefore employ an EFA to extract latent centrality structures, conceptualised as spatial scales that are presumed to cause centrality patterns of different radii.

Still, this method does not come without any disadvantages. The EFA is a complex multi-step process with multiple options that need to be carefully defined in order to arrive at meaningful outcomes. Particularly difficult is the selection of extractions, the selection of the rotation type and the definition of the number of factors that should be retained (Costello and Osborne 2005), as I will elaborate below. For an application in network analysis all extraction methods have been tested, but the results are only reported on Principle Axis Factoring, as these have produced reliable results across all measures and across all four models.

FACTOR ROTATION. Rotation describes the method with which factors are rotated in order to achieve what has been termed a 'simple structure'. Simple structure is a condition at which it is clear which variable is related to a specific factor. In general, one can divide rotation methods into orthogonal and oblique methods. Orthogonal rotation produces factors that are uncorrelated, whereas oblique rotations allows factors to be correlated with each other (ibid.). Since the main purpose of an EFA in a network centrality context is to arrive at a latent centrality structure that can account for distinctively different patterns in the network, an orthogonal rotation is chosen. This means the results will yield factors that are correlated with each other as little as possible. The orthogonal rotation method offers three different rotation types, namely Varimax, Quartimax and Equamax. Varimax maximises the sum of the variances of squared correlations between variables and factors. This minimizes the number of variables with high loadings on more than one factor. Quartimax, on the other hand, minimizes the number of factors needed to explain each variable by generating a general factor on which most variables load to a high or medium degree. This will account for the most dominant factor in the data set. Lastly Equamax is a combination of Varimax and Quartimax. Equamax arrives at solutions that minimize the number of variables with high loadings on a factor and the overall number of factors needed to explain a variable. For this reason, the analysis is conducted with Equamax as the method of rotation.

DETERMINING THE NUMBER OF FACTORS. Determining the number of factors is a major concern in exploratory factor analysis, as it can introduce substantial bias into the analysis. Which method is suited best to determine, which of the factors is of actual relevance for the interpretation of results, is still contested. The default method in most of the statistical software, determines the number of factors based on eigenvalues that are greater than or equal 1. The reasoning behind this approach is that factors that can account for as much variance as a single input variable are worth keeping. Another widely-established method looks at the graph development of a so-called scree plot (ibid.). In a scree plot the researcher can observe the fraction of total variance in the data as explained or represented by each factor. This method advises the researcher to select all factors that fall on the line development before it slope drops markedly, a feature that is often referred to as the 'elbow' of the line. This visually operating method has been proven to be particularly reliable where a clear change in slope is visible. With data where this jump in the line development is subtler the identification of the right number of factors is more difficult. An alternative to this is the parallel analysis (PA) that formally tests the probability at which a factor occurs by chance (Horn 1965). Parallel analysis is the only approach that relies on a probability analysis to determine the number of factors and it is argued that PA is therefore superior to the eigenvector or scree plot approach. PA usually generates +1,000 Monte-Carlo simulations on randomly generated data that is comparable to the observed variables. Based on the difference between the 95th percentile and the raw data eigenvalues one can define which factors are beyond chance. This is why, I will compare all three results and test the degree of interpretability of the resulting factors against each outcome.

7.1.3 RADIUS SELECTION

Until today, there is no established method to define or justify radii selection within the field of spatial network analysis. Due to this limitation, this research bases its analysis on a large set of 49 different radii[21]. Radius selection is fundamentally arbitrary, with no established method to solve this issue. As an approach to reduce potential bias, this study will use a large set of radii. The difference between each radius is smaller on lower radii and increases between larger radii. This is to account for the growing computational time needed to compute the model, as well as the fact that variance in the resulting data decreases with large radii. The smallest radius is selected based on the mean segment length found in the two regions (GE: 101.99m, UK: 105.44m), while the distance differences between each radius are smaller than the longest segment in each system (GE: 5777.72m, UK: 4732.79m). The reason behind this is to analyse a large

---

[21] An initial selection included radius n (all-to-all), and the total number resulted in 50. However, radius n proved to be not computable with any of the released depthmapX version.

collection of radii with only small differences between them, in order to be able to capture any potentially occurring scale embedded in the system and to reduce the selection bias. If each of the radii correlates strongly with the next one, one can make the assumption that there are no hidden scales between the two that are not covered by the analysis. The selected radii are: *100, 150, 200, 300, 500, 800, 1300, 1800, 2500, 3200, 4100, 5000, 6100, 7200, 8500, 9800, 11300, 12800, 14500, 16200, 18100, 20000, 22100, 24200, 26500, 28800, 31300, 33800, 36500, 39200, 42100, 45000, 48100, 51200, 54500, 57800, 61300, 64800, 68500, 72200, 76100, 80000, 84100, 88200, 92500, 96800, 101300, 105800* and *110500* metres. The resulting data for each of the centrality measures ranges above 50,000,000 values per model, yielding a very large data set from which we can make observations. The radii selection can be further extended by the power law equation $y = 50x^2$, which results in radii that increase in a similar fashion as the computational time. Resulting values should be rounded to the nearest hundredths decimal place. It should be noted that this equation should regarded as a general guidance, rather than a solution to the fundamental problem of radius selection.

7.1.4 EXPLORATORY FACTOR ANALYSIS FOR BETWEENNESS CENTRALITY

Beginning with the exploratory factor analysis for betweenness centrality of the German region, I first determine the total number of factors by means of a parallel analysis, the eigenvalues that are larger or equal 1 and finally a scree plot analysis. For the parallel analysis, a total of 1000 Monte-Carlo simulations of the data using permutations of the raw datasets have been computed. The simulated datasets are based on 49 variables and 1,203,173 features for the German region and 1,019,915 features for the British region to match the initial raw dataset. The difference between the 99th percentile and the raw data eigenvalue is displayed in Table 22. Those factors whose 99th percentile is higher than that of the raw data are considered to occur not only due to chance. In the case of the German region, the parallel analysis results in no more than eight factors. For the British region, however, the simulated data generates 15 factors.

Table 22: EFA parallel analysis, eigenvalue results for ASA SLW BC of the German region.

Root	Raw Data	Mean	99th Percentile
1	27.981851	0.011954	0.012521
2	11.003571	0.010944	0.011574
3	5.017728	0.010266	0.010907
4	2.234959	0.009451	0.009597
5	1.019676	0.009079	0.009409
6	0.47149	0.008504	0.008716
7	0.244395	0.007695	0.008181
8	0.134522	0.007291	0.007622
9	0.073648	0.006781	0.007125

Table 23: EFA parallel analysis, eigenvalue results for ASA SLW BC of the British region.

Root	Raw Data	Mean	99th Percentile
1	28.685539	0.012869	0.012869
2	10.434712	0.01211	0.01211
3	4.872896	0.011228	0.011228
4	2.392501	0.010208	0.010208
5	1.017032	0.009883	0.009883
6	0.521716	0.008561	0.008561
7	0.269733	0.008053	0.008053
8	0.149446	0.007741	0.007741
9	0.07966	0.006655	0.006655
10	0.049061	0.006433	0.006433
11	0.029977	0.006309	0.006309
12	0.018138	0.005387	0.005387
13	0.011699	0.005182	0.005182
14	0.006657	0.004635	0.004635
15	0.004677	0.00425	0.00425
16	0.002954	0.003868	0.003868

The results of the parallel analysis indicate that a PA in the context of network metrics appears to overestimate the number of factors. It has been reported that PA can in some cases tend to indicate a higher number of factors than are actually relevant (Buja and Eyuboglu 1992). In such cases, the eigenvalue for trivial or negligible factors surpass the corresponding random data eigenvalues of the same roots. This makes it necessary to employ additional analysis types to trim trivial factors. This is particularly the case for

the data of betweenness centrality, as the eigenvalues surpass the corresponding random data eigenvalues in the ERPGr model as well as the VPGr model. The results of the parallel analysis show that there are statistically significant, latent centrality factors that can explain the variance of the 49 observed radii by means of a much smaller set. In contrast to that, the two alternative methods, namely an eigenvalue greater than or equal to 1 and the slope identification in the scree plot indicate much fewer factors. For both regions, the number of factors with an eigenvalue greater than 1 is 5 (compare raw data in Table 22, Table 23), with the fifth factor just above 1 for both regions (GE: 1.019 and UK: 1.017). The eigenvalues are highly comparable between both cases. This can also be observed in the slope change in the scree plot (Figure 87:a and b). The scree plot exhibits a drastic decline in the eigenvalue from factor 1 to factor 5 with an almost flat line after the slope at the 6th and 7th factor.
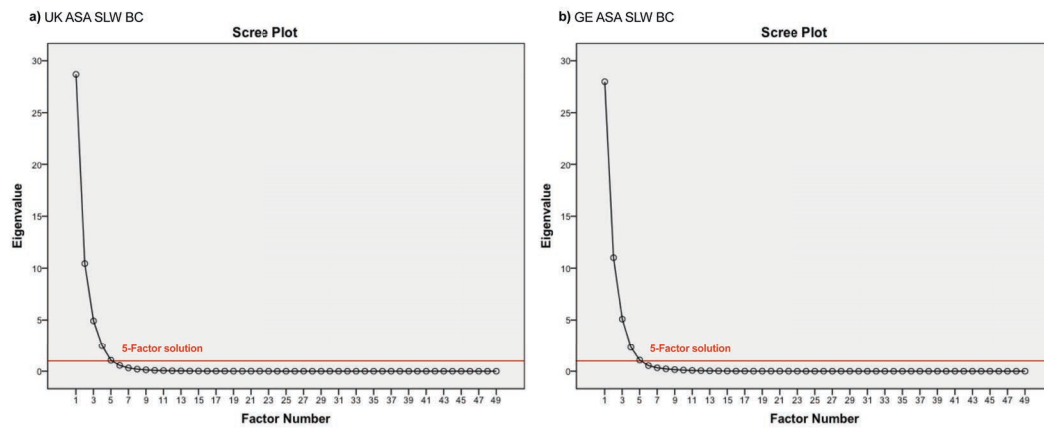
Figure 87: Scree plot for ASA SLW BC for GE and UK. A drastic change in the slope is market with a red line, indicating the number of factors i.e. 5.

Similar observations can be made for the ERPGr and VPGr models. However, here the eigenvalue of the fifth factor falls below the greater than or equal to 1 eigenvalue threshold. The ERPGr model's fifth root has an eigenvalue of 0.850, whereas the fifth root of the VPGr model is 0.895. The scree plot (Figure 88) for both randomised models are more difficult to interpret than those of the real-world networks. Particularly the ERPGr has a dominant first root of an eigenvalue of 36.092. It remains questionable whether the data of the randomised model is best explained by 4 or 5 factors. In these cases, the rotated factor matrix can help. The purpose of the rotation is to arrive at a simple structure that exhibits a clear difference between the variable loadings for each factor.
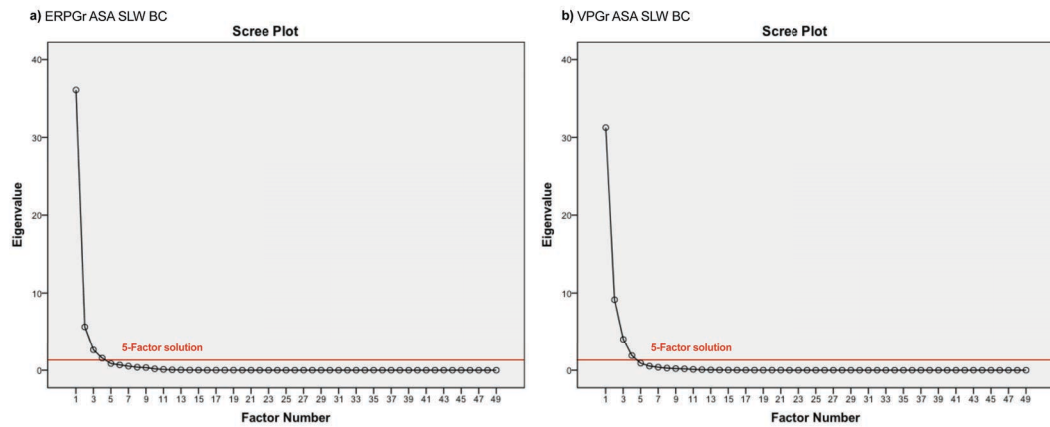
Figure 88: Scree plot for ASA SLW BC for ERPGr and VPGr. A drastic change in the slope is market with a red line, indicating the number of factors i.e. 5.

This can be achieved by plotting the value of the rotated factor matrix for each of the radii. The rotated factor matrix displays both the loadings and the correlations between the variables and factors. In a rotated factor matrix plot the course of the curve reveals whether a simpler structure could be achieved, or whether a different number of factors is better suited to describe the observed data. Figure 89 shows the rotated factor matrix plot. The y-axis shows the factor loading, or correlation with each of the radii (x-axis) and the respective factor. The more the line in the graph increases, the stronger is the correlation. Each line graph represents one factor and the factor loading of the radius, which it is influencing. Based on the factor loading, one can observe associations of different radii and each factor. This allows interpretations for each of the factors and a collection of measured variables.

Figure 89 shows that the fourth and third factors for the ERPGr do not meet the criteria for what is described as a simple structure. For the VPGr model, however, the number of factors is suitable, as all four factors have a clearly distinctive peak and are separated correlating with a differing number of radii. If instead of four, five factors are extracted, the rotated factor matrix displays a much clearer picture: both models exhibit a clear differentiation between the factors, and very comparable curve developments. The five factors have their peaks at different points (approximately at radius 200, 1300, 7200, 24200 and 110500 metres) and correlate with different radii. The largest difference is between factor three of both models; the VPGr model peaks at 8500, whereas the ERPGr model peaks at 4100. In the case of betweenness centrality for the ERPG we can observe that almost half of all radii (33,800 – 110,500m) are influenced by factor I. Larger radii report the strongest correlations, which means that factor I can estimate parameters more precisely than the remaining factors II-IV. Radii between 6,100 and 33,800m are associated with factor II, radii between 500 and 6,100m with factor III, and radii between 100 and 500m are influenced by factor IV.

Exploratory Factor Analysis: Rotated Factor Loadings
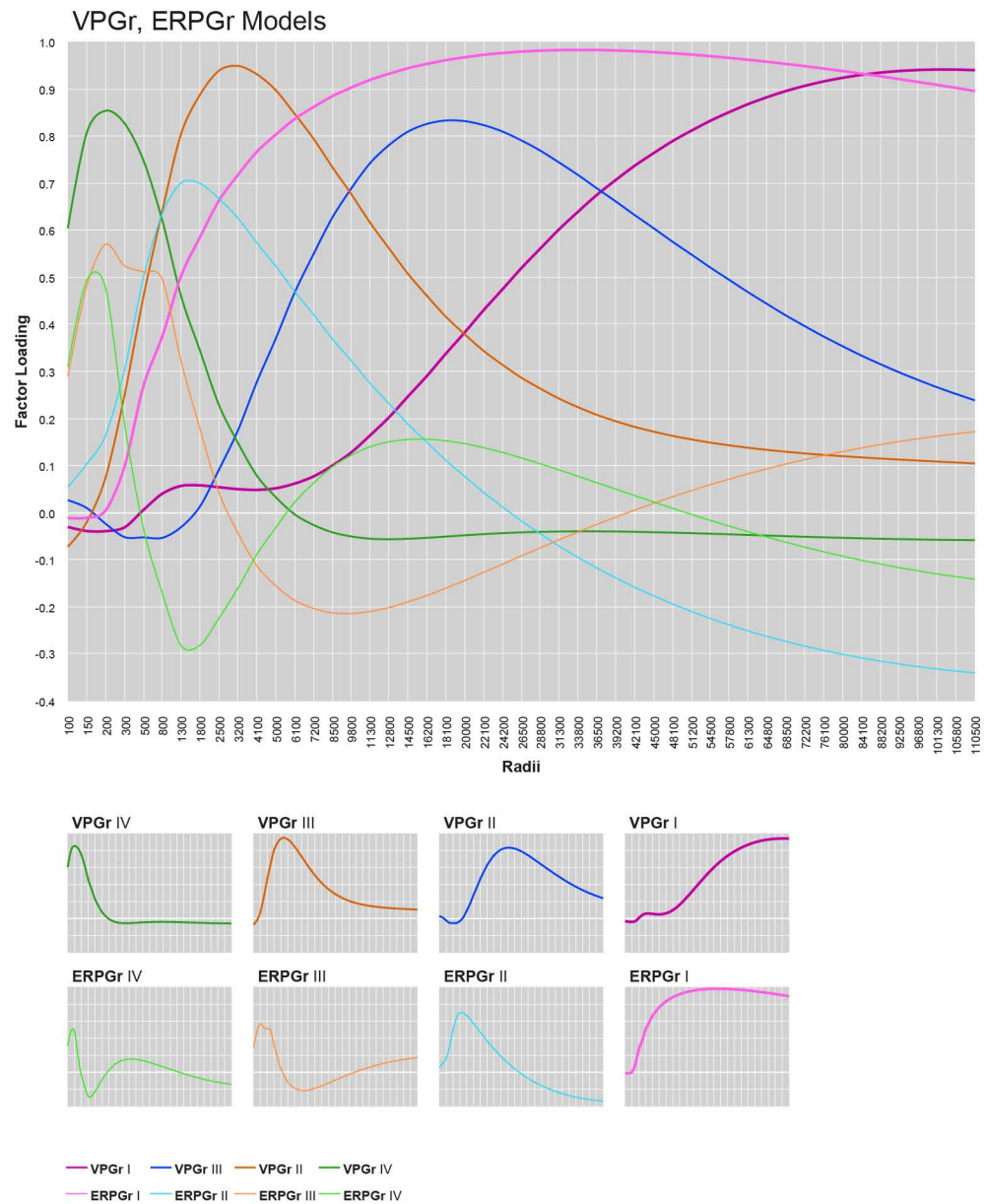
VPGr, ERPGr Models

Figure 89: Explorative Factor Analysis rotated factor loadings for 49 metric distances of ASA SLW between centrality for VPGr and ERPGr Models. Extraction method: Principle Axis Factoring. Rotation method: Equamax with Kaiser Normalisation. Rotation converged in 34 iterations for VGPr and 44 iterations for ERPGr respectively.
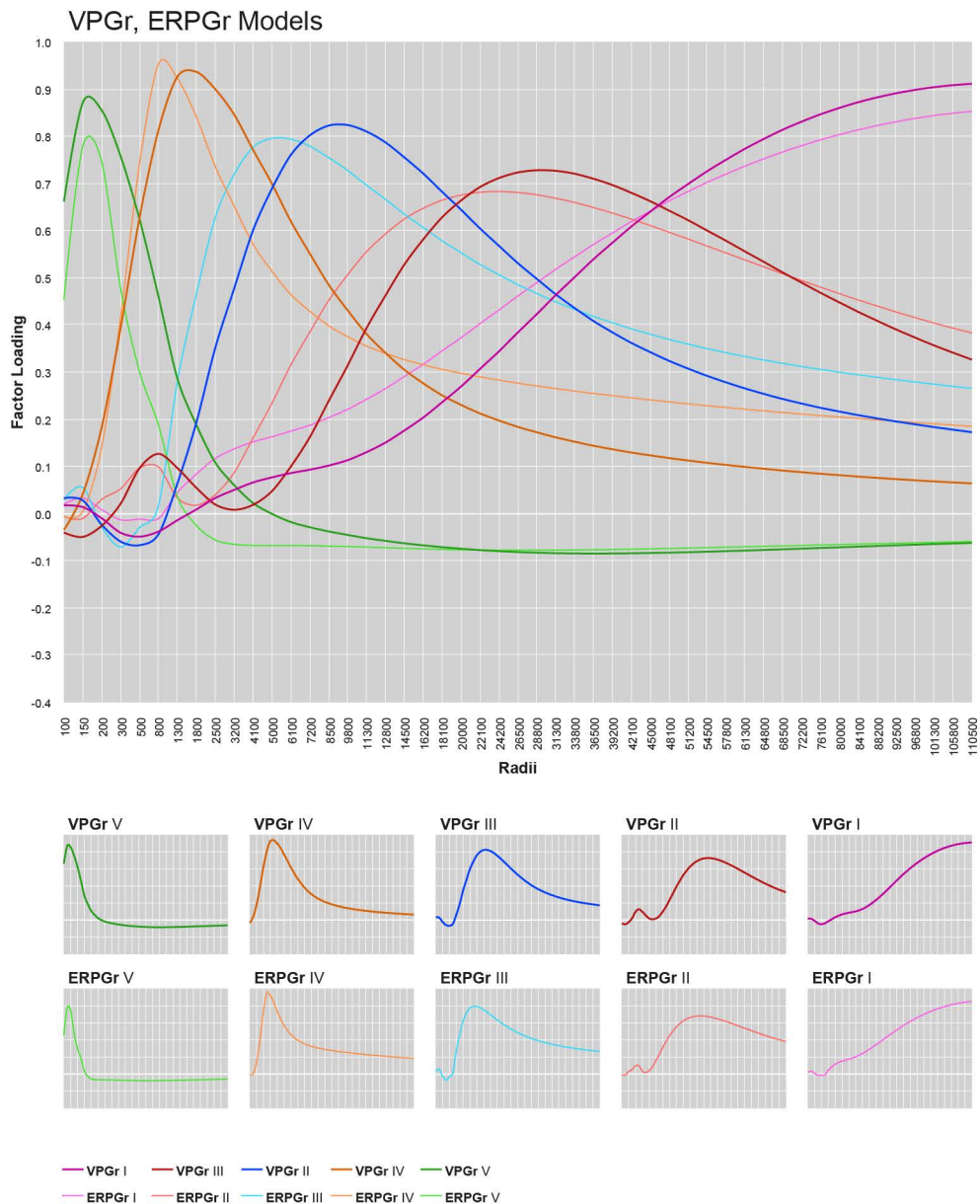
Figure 90: Explorative Factor Analysis rotated factor loadings for 49 metric distances of ASA SLW between centrality for VPGr and ERPGr Models. Forced factor extraction for 5 factors. Extraction method: Principle Axis Factoring. Rotation method: Equamax with Kaiser Normalisation. Rotation converged in 34 iterations for VGPr and 44 iterations for ERPGr respectively.

The fact that EFA produces these factors and that they form a clear pattern in their rotated factor loadings provides insights into the general behaviour of centrality patterns in planar graphs. Independently of how the spatial configuration is structured, there are always shortest paths and locations in the system that have an advantaged or disadvantaged accessibility. Yet, presumably these shortest paths do not exhibit a large variation throughout radii of comparable distance, but a large variation between radii of significantly larger distance. This is why we can assume that a certain phase transition

takes place from lower radii to larger radii, meaning that if a journey takes place between two points on a radius of 1,300 metres and another one on a radius of 1,800 metres these two journeys are more likely to select the same path within the network. Yet, if journeys of 1,800 metres are compared to a journey taking place between two points on a distance of 41,800m then a fundamentally different spatial structure is used. Such switching of structures between two spatial scales is not sudden but transitions smoothly. The gradual difference of rotated factor loadings provides us with evidence for the assumption that spatial graphs inherently feature best-fit latent structures or scales for certain distance modes. If this is the case, we should find similar structures in human-shaped configurations. These similar structures might exhibit a level of optimisation in relation to each distance mode. This is because the movement of human beings as well as other natural processes have evolved through mechanism of optimisation. Accordingly, Barthélemy (2011 p. 59) points to the existence of such spatial network characteristics as indicators of 'evolutionary processes'. Batty (2007), as well as Allen and Sanglier (1981a, 1981b) made similar arguments in their work.

If the observations made for ERPGr and VPGr models are compared against the rotated factor matrix of the two real-world models (UK and GE), a much clearer picture emerges. Figure 91 shows not only the rotated factor matrix of both regions, but also the result of an EFA where the ASA SLW BC values for each street and region have been combined and computed as a single model (dotted line). The observable pattern of the factor correlation for both regions is strikingly similar, with a simple structure and peaks at exactly the same radii. The first factor peaks at 110,500 metres with a loading of 0.952 and then drops monotonously. The second factor peaks at between 33,800 and 36,500 metres. The third factor peaks at 11,300 metres and the fourth factor peaks at 1,800 metres. Finally, the fifth factor peaks at 200 metres.

Each of the factors correlates strongly with a set of radii with one to two radii defining the maximum correlation. From these peaks a steady decline of the loadings can be observed with some radii correlating to an even degree with two factors; this is the case for radii between 500 and 800, 5,000 and 6,100, 22,100 and 24,200, as well as 48,100 and 51,200 metres. The small peaks of the third and fourth factor around radii of 500 and 800 metres might indicate that a larger set of smaller radii might enrich the model, as there is still variance of smaller radii explained by parts of large-scale factors. The general development of these latent factors of the two real-world street networks is also highly comparable to that of the randomly generated VPGr (and to a lesser degree to the ERPGr). Figure 92, shows a superposition of the rotated factor matrix plot of the combined model and the VPGr model, highlighting that a similar scalar pattern emerges independently of the spatial configuration.
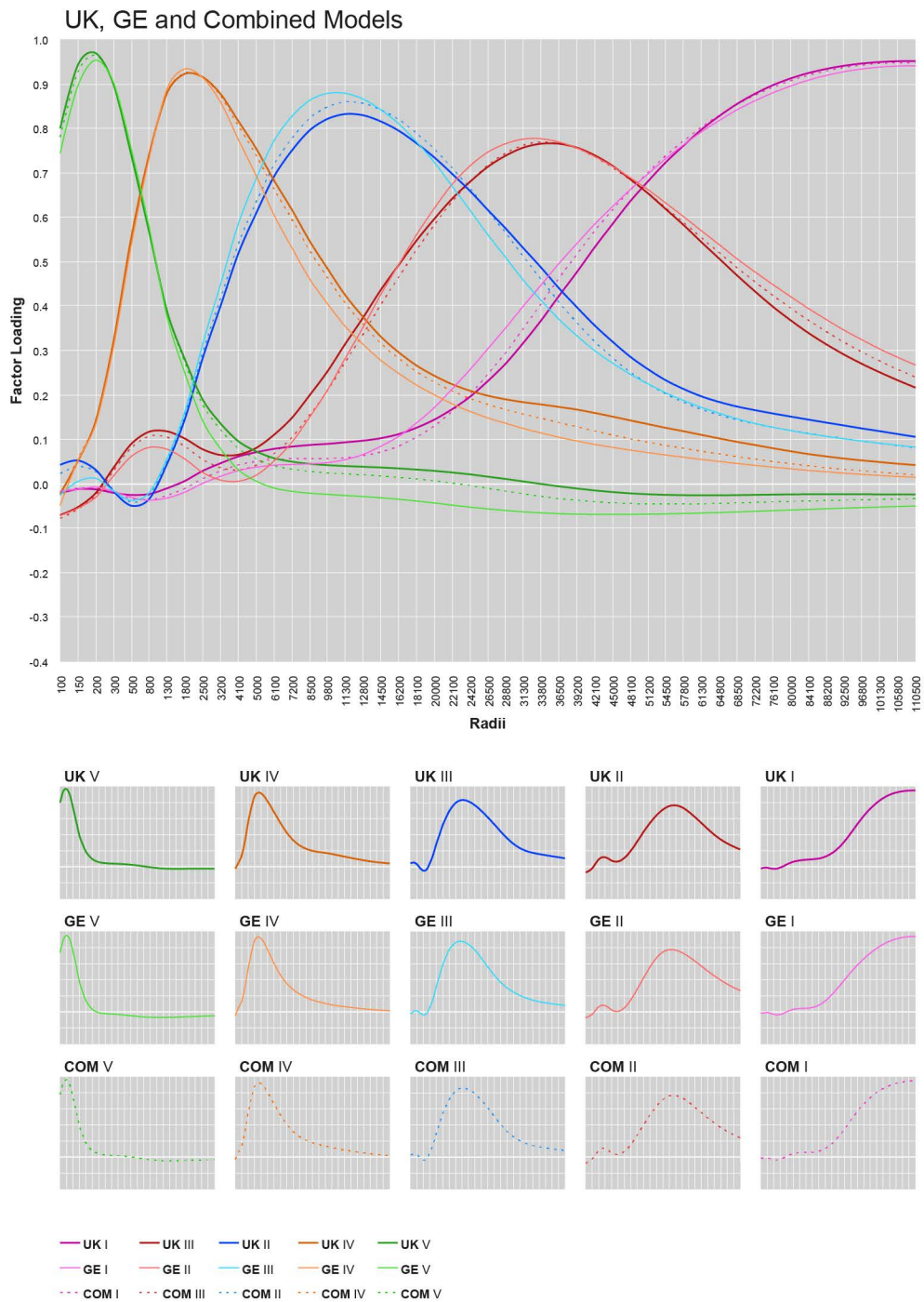
Figure 91: Explorative Factor Analysis rotated factor loadings for 49 metric distances of ASA SLW between centrality for UK, GE and a combined dataset. Extraction method: Principle Axis Factoring. Rotation method: Equamax with Kaiser Normalisation. Rotation converged in 26 iterations for UK, 23 iterations for GE and 24 iterations for the combined model respectively.
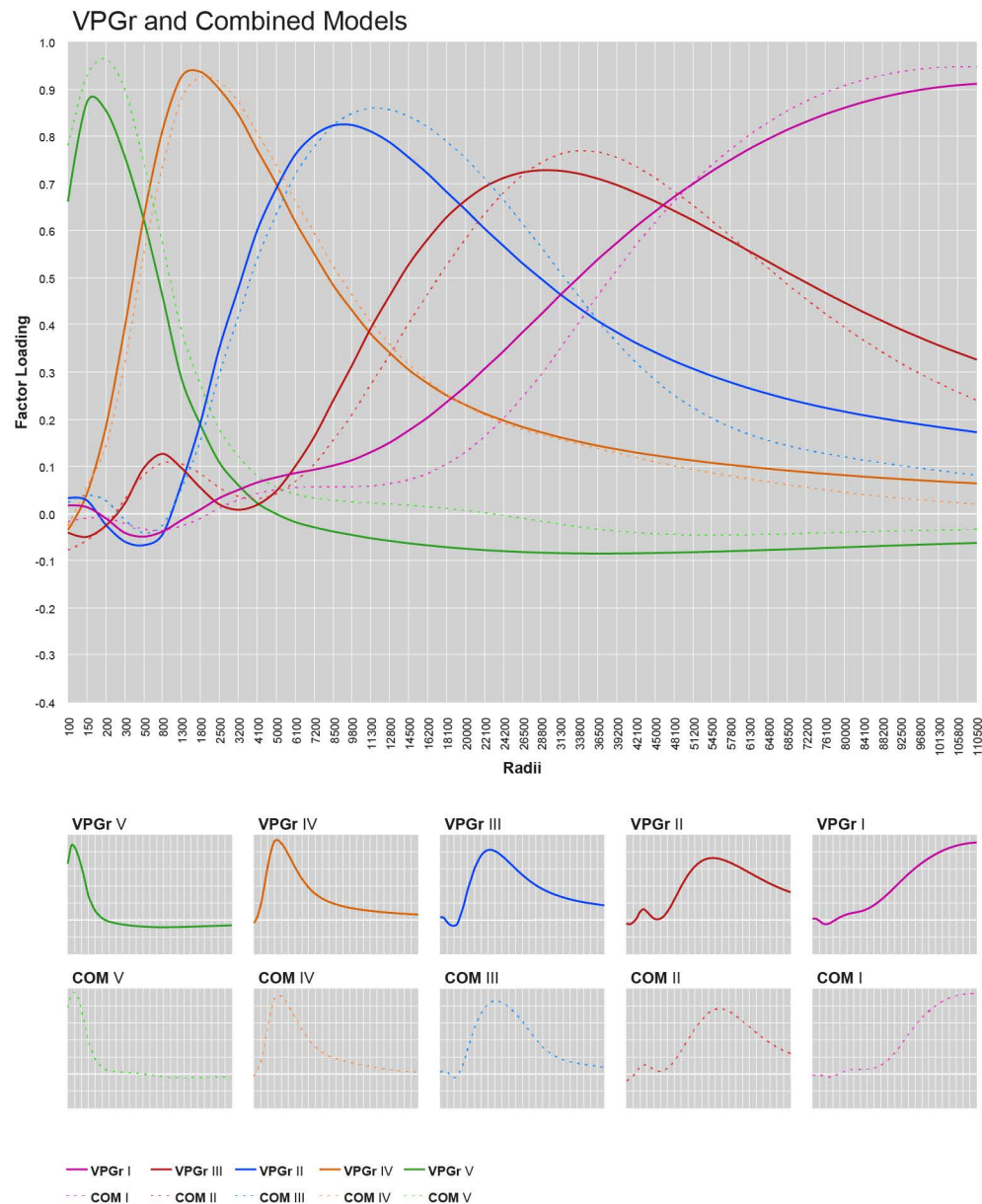
Figure 92: Explorative Factor Analysis rotated factor loadings for 49 metric distances of ASA SLW between centrality for VPGr and the combined dataset of GE and UK. Extraction method: Principle Axis Factoring. Rotation method: Equamax with Kaiser Normalisation. Rotation converged in 34 iterations for VPGr and 24 iterations for the combined model respectively.

Overall, the real-world networks show a more distinct pattern of correlations. This could be an indicator for a hierarchical organisation in human activity patterns that underlie the shaping process of the spatial configuration and defines spatial scales. Hillier argues that betweenness centrality provides insights into the location of economic activities (2009), which is the reason for this his conjecture on the emergence of these spatial scales. Walter Christaller's CPT (1933) points to a hierarchical relationship between differently sized urban areas and their respective market spaces. In relation to human

activity and movement this means that individuals are more likely to seek everyday goods in local markets and rare goods in higher hierarchies. In reality this implies that grocery shopping is more likely to take place in a local neighbourhood, while specialised services such as those of a lawyer would benefit from being situated at centres of a higher hierarchy, as they need to extend their market area in order to suffice the need of frequent customers. Further, one can assume that certain activities take place more often than others, such as the daily commute to the workplace for the majority of the population, while other activities will occur less often, such as buying electric goods. These constantly reoccurring patterns of human activity have an impact on the spatial organisation of societies in a way that an optimisation process shapes the spatial configuration. This entails that these repetitive everyday activities are more effectively distributed in the system, meaning that they should manifest themselves in the form of spatial scales. Moreover, the spatial product of this process will have an impact on the possibility of future activities and, subsequently, influence the former. This allows us to make assumptions about the nature of these extracted spatial scales.

Comparing those radii that the extracted factors (latent centrality structures i.e. scale) predict best, with those radii theorised by Christaller's CPT for different market areas (Table 24), a relationship between the two empirical and theoretical radii becomes apparent. The factors extracted are similar to the radii defined by Christaller's CPT. Three of these five extracted factors exhibit estimate parameters on exactly those radii that Christaller estimated for each of his central place hierarchies (Table 24). The exceptions to this are very local centrality patterns (factor IV and V), which might be caused by the fact that smaller centrality patterns are influenced more strongly by cultural process such as the development of particular types of coal-mining settlements, or social housing estates, rather than by economic activity. Whereas three of Christaller's seven central place types are represented by the factors extracted, the remaining four are not captured by the EFA. However, with regards to the randomised graph ERPGr and VPGr models, we have already gained insights into the inherent scales embedded in planar graphs and find that only four of such scales emerged. This might indicate that the remaining centres are not pronounced enough to constitute independent spatial scales. Based on this assumption, it is worth considering as points of interest the intersectional area of each of the latent centralities; these intersections are points that can load on either of the two factors. Table 24 includes additional sub-categories of potential spatial scales between the factors found. Again, one can observe similarities between the distances in Christaller's CPT and the factors extracted.

Table 24: Comparison of the betweenness latent centrality structure with Christaller's central place system and their respective closest scales.

Latent Centrality	UK Region	GE Region	Market Radius (m)	Christaller Type
Neighbourhood	**200**	**200**	-	-
City	**1,800**	**1,800**	-	
-	-	-	4,000	Marktort (M)
Between City/Metro	6,100	6,100	6,900	Amtsort (A)
Metropolitan	**11,300**	**11,300**	12,000	Kreisstadt (K)
Between Metro/ Intra-Regio	22,100	22,100	20,700	Bezirksstadt (B)
Intra-Regional	**33,800**	**33,800**	36,000	Gaustadt (G)
Between Intra/Inter-Regio	45,000	45,000	62,100	Provinzstadt (P)
Inter-Regional	**105,800**	**105,800**	108,000	Landstadt (L)

These findings indicate that regional morphologies might, indeed, be able to provide insights into economic processes and human activity patterns causing this formation. The findings also suggest that Christaller's theory can, to a certain extent, be used to explain the spatial organisation of the two European regions under scrutiny. Very little has been said so far about the actual spatial configurations that the different factors produce. Advancing a strategy proposed by Serra and Pinho (2013), I visualise each factor based on its respective loadings. Figure 93, Figure 94, and Figure 95 provide a loading plot for each factor for each of the three models, VPGr, UK and GE. The factors include loadings at different intensities for each segment and, therefore, provide a rich pattern. Here, only loadings above a standard deviation of 1.0 are highlighted in a colour that matches those in the rotated factor loading graphs. Furthermore, the line thickness of each plot increases with the factor loading. Hence, those segments, which are significant for the respective latent centrality, are displayed by thicker lines and are better visible in the plot. This way of representation increases the distinctive difference between each factor visually and facilitates the interpretation of their morphologic nature. The patterns observed in the rotated factor loadings are now mapped on the respective spatial network.
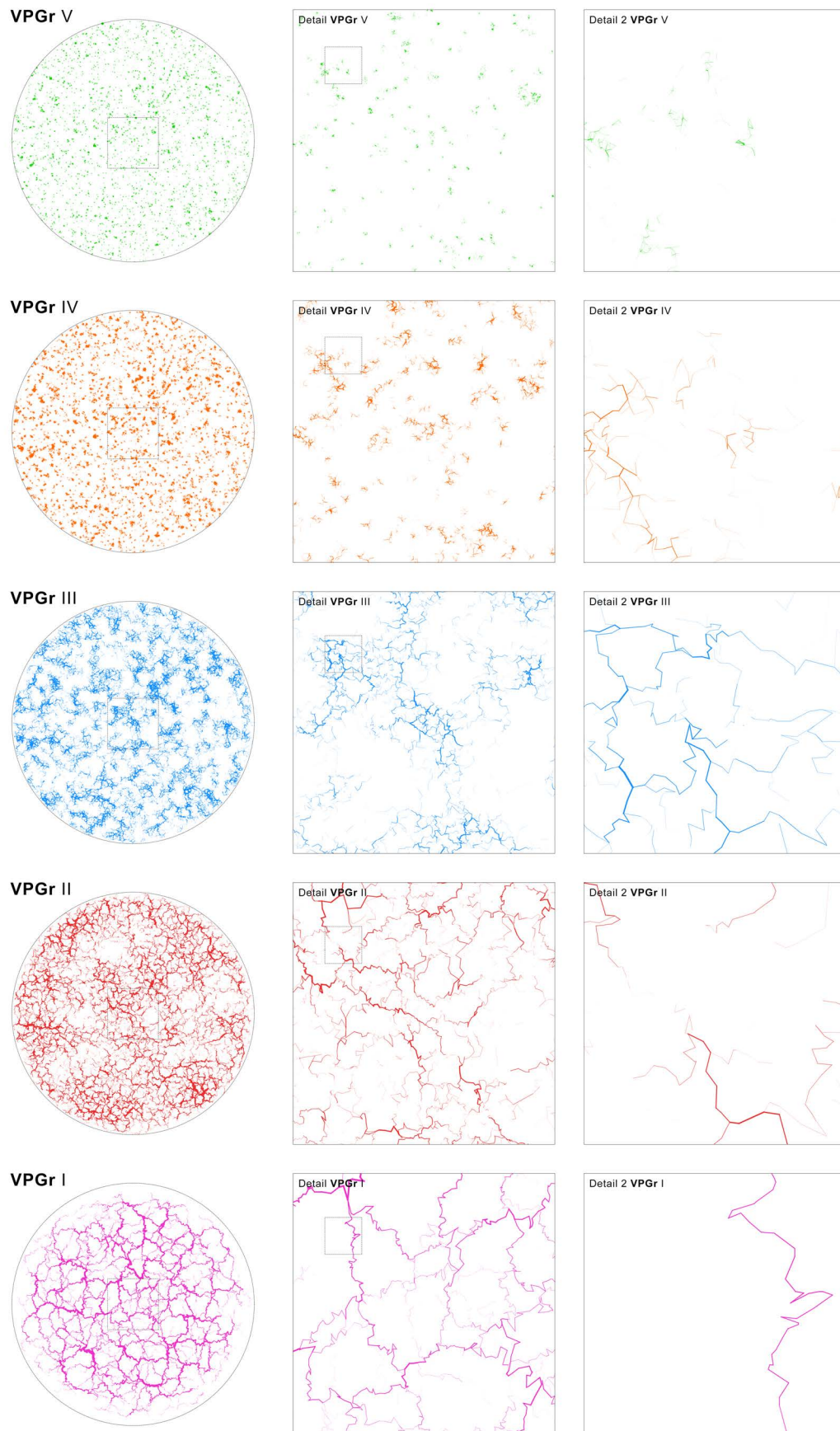
Figure 93: VPGr latent centrality structures for ASA SLW BC. EFA Factor Analysis Scores for each of the five factors and cases with score values above 1.0.
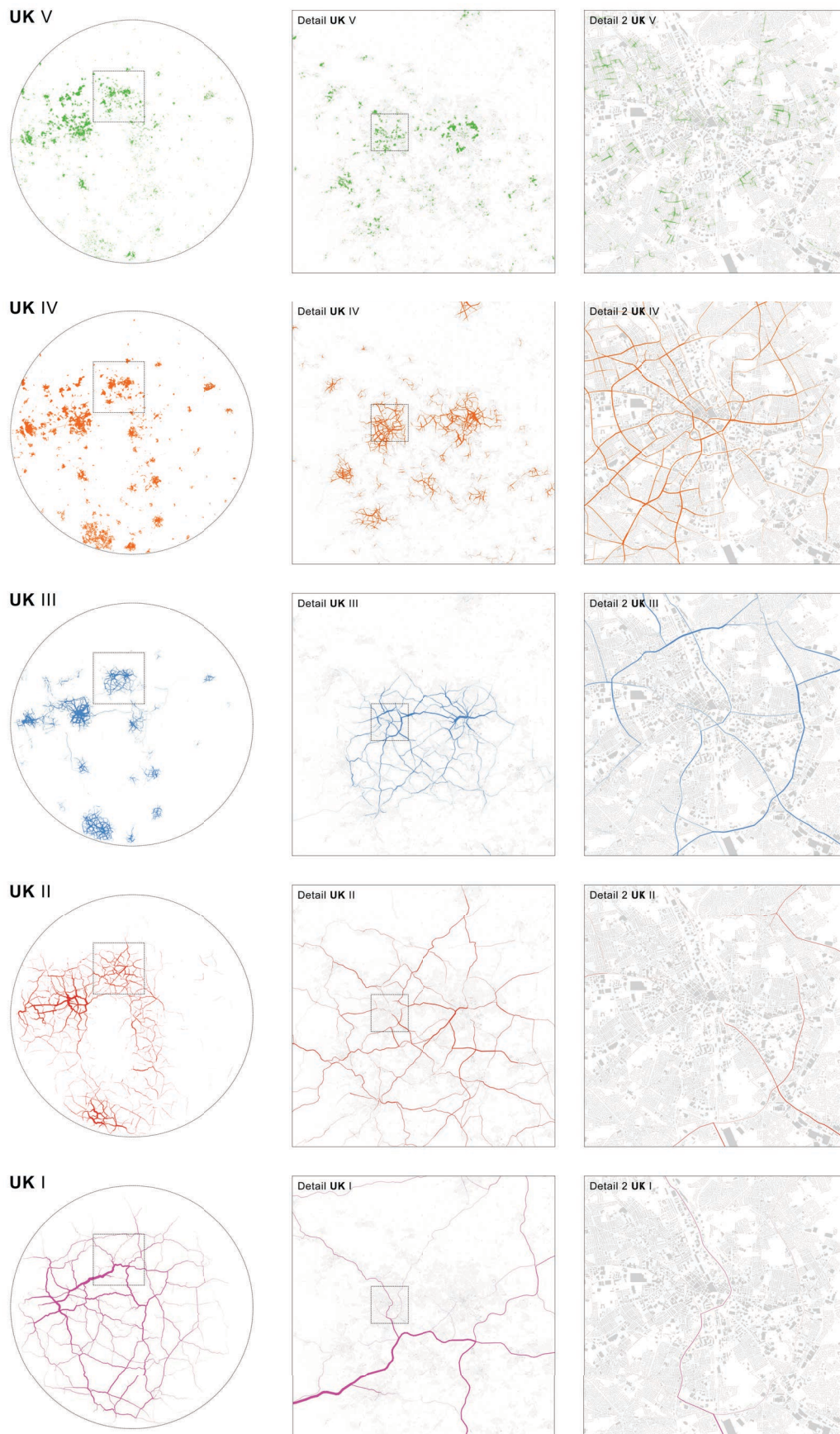
Figure 94: UK latent centrality structures for ASA SLW BC. EFA Factor Analysis Scores for each of the five factors and cases with score values above 1.0.
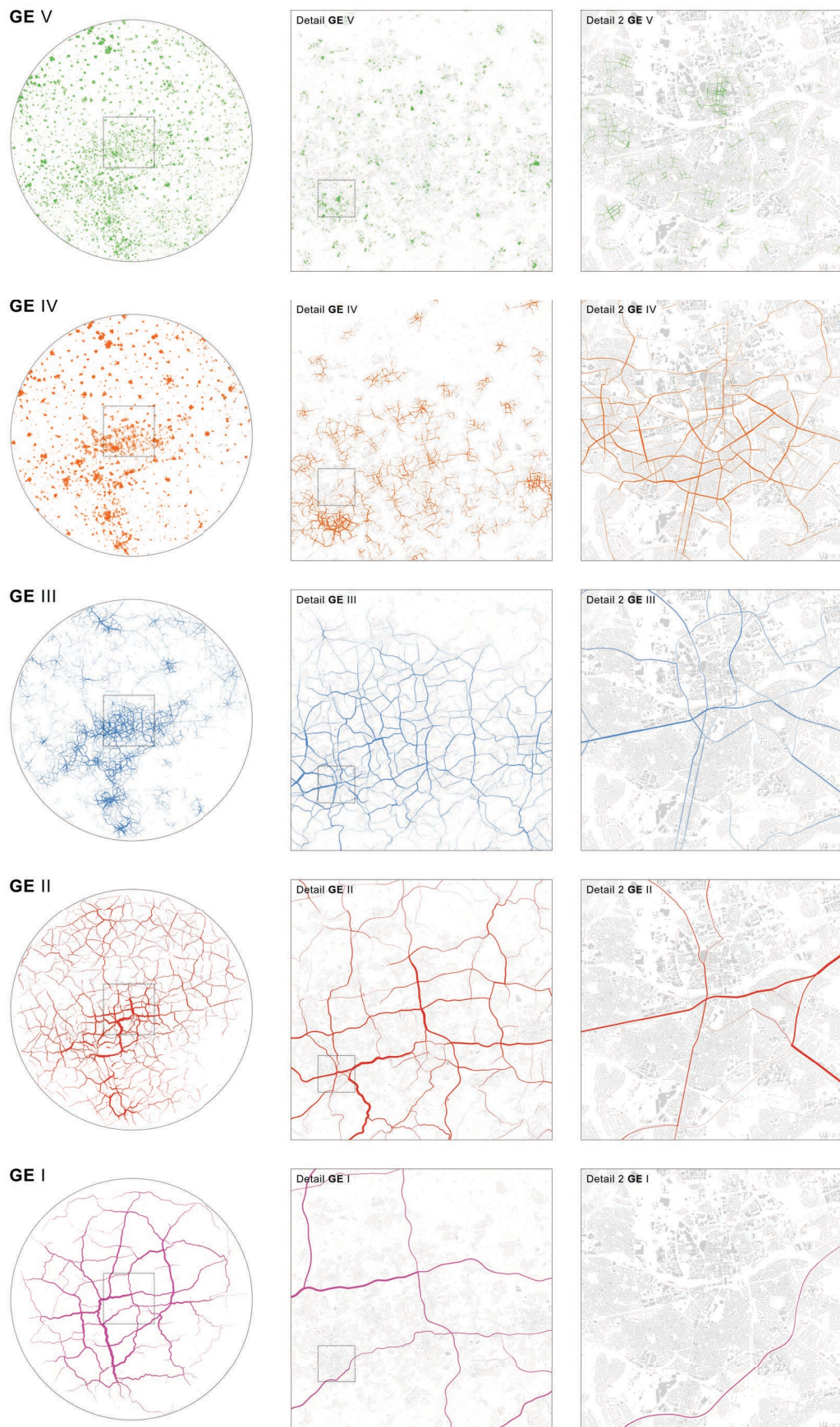
Figure 95: GE latent centrality structures for ASA SLW BC. EFA Factor Analysis Scores for each of the five factors and cases with score values above 1.0.

Some initial findings can be inferred from these visualisations: the visualised latent centrality structures of the two real-world models exhibit striking similarities in their spatial manifestation. The fifth factor of both real-world models highlights highly local or neighbourhood areas and small clusters in predominantly urban areas. The fourth factor for both real-world models reflects the location of distinctive cores of settlements and large cities. The third factor shows metropolitan agglomerations. That way, the Ruhr Valley becomes clearly visible as a distinctive region in the form of the third factor, as well as the metropolitan area around Leeds and Bradford. The second factor highlights intra-regional relationships. In the German model, the Rhein-Ruhr area (Ruhr Valley and the cities Düsseldorf, Cologne, Bonn) is now merged into one consecutive network, whereas in the UK the agglomeration of the NDY region is a continuous network too. Finally, the first factor strongly highlights the top-level motorway network and spatial inter-regional relationships. In the case of the UK, one can clearly trace how the network evolved around the Peak District National Park, with the highest values between Leeds and Manchester. The German cases exhibit an almost city-like circular wheel structure around the Ruhr Valley, with the highest values along the river Rhine and the cities of Bonn, Cologne and Düsseldorf towards the Ruhr Valley.

A detailed comparison between these real-world networks and the randomised model provides further insights into the morphology of such latent centrality structures. Figure 96 shows a comparison of the detailed sections of the fourth factor, i.e. fourth latent centrality structure, for the three models VPGr, UK and GE. All three models show clear clusters of segments with higher values. However, in both real-world cases these clusters are much more distinctive and larger in size. Each cluster can be clearly related to an existing city. Thus, the two biggest clusters in the case of the UK, the city of Leeds and Bradford, are very pronounced, as well as the city of Essen in the German model. A clear distance decay mechanism can be observed in both real-world models, where the strongest clusters are located within large cities and smaller clusters are situated around them. This pattern can be linked back to a form of centre distribution that would be in alignment with Christaller's CPT.
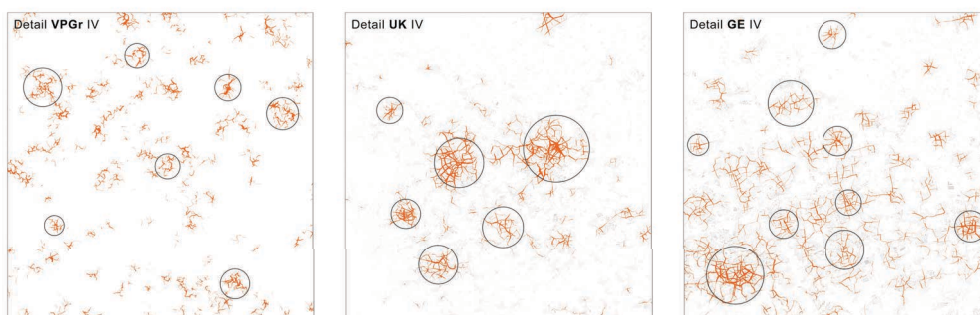
Figure 96: Visualisation of EFA BC IV, latent centrality structures for VPGr, UK and GE.

With regards to the third factor, i.e. the third latent centrality structure (Figure 97), the network morphology changes significantly. Now, instead of highlighting urban and inner city cores, the network represents much more strongly the spatial morphology of the relation between these centres. In both real-world cases, the structure follows linear connections between urban areas with a distinctive structure that resembles what Hillier called a 'deformed wheel' in the context of cities (Hillier 1999). The structure in both real-world models is characterised by a large number of linear connections. The randomised model (VPGr) exhibits only an agglomeration of segments arranged in a zigzag fashion. This highlights that in the evolutionary process of real-world networks a specific optimisation mechanism must have been at work. Otherwise the observed structure would not follow such clearly linear and comparable pattern across countries. The third factor can also be seen as a strong indicator of a centre distribution. Betweenness centrality values are the highest on those streets that efficiently connect a large number of segment clusters with each other. In this sense, betweenness centrality appears to be an appropriate indicator for i) the existence of centres at a particular distance and ii) the degree of the spatial relationship between these centres. The more diverse and complex the observed latent centrality structure is , the more complex and polycentric the existing urban agglomeration are.
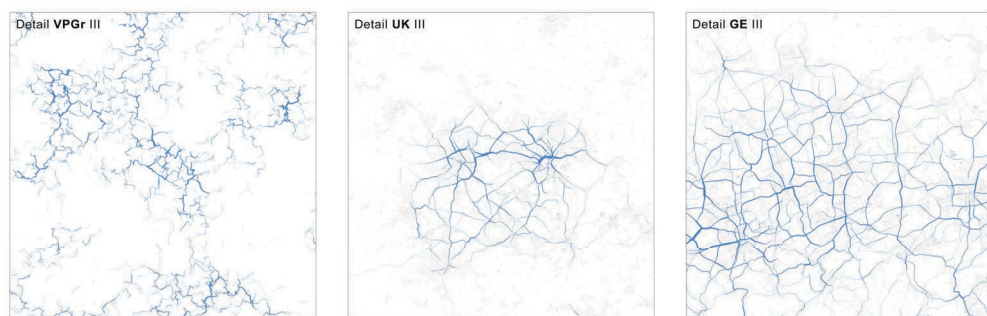
Figure 97: Visualisation of EFA BC III, latent centrality structures for VPGr, UK and GE.

Finally, after gaining insights into the morphology of these networks, we can observe that each factor or latent centrality structure represents a distinctive spatial pattern of the spatial network. Together all of these latent centrality structures form a hierarchical relationship that might be shaped by and might represent different modes of movement through the network. While the exploratory factor analysis helped to reveal these structures and hierarchy, it is important to highlight that only a combination of all factors can give a comprehensive account of the overall model. I will, thus, propose to combine the extracted factors into a combined model that is able to express the fundamental spatial structure of the spatial network. The result is a multi-scalar model that represents the hierarchy of the network. In order to do so, each factor can simply be combined by calculating the maximum value for each EFA factor and segment (7.1).

$$EFA\ BC\ COM = \max\left(EFA\ I, EFA\ II, EFA\ III, EFA\ IV, EFA\ V\right) \qquad (7.1)$$

The strength of such an approach is i) that the issue of radius selection can be solved by relying on a single model, and ii) that results from different models can be compared against each other independent of differences in model sizes, given that effects of model size (i.e. maximum possible distance) and edge effects (model size multiplied by maximum analysis radius) are considered. The following pages introduce a number of visual representations of this newly combined model for the British region (Figure 98, Figure 99 and Figure 100) and the German region (Figure 101, Figure 102 and Figure 103).
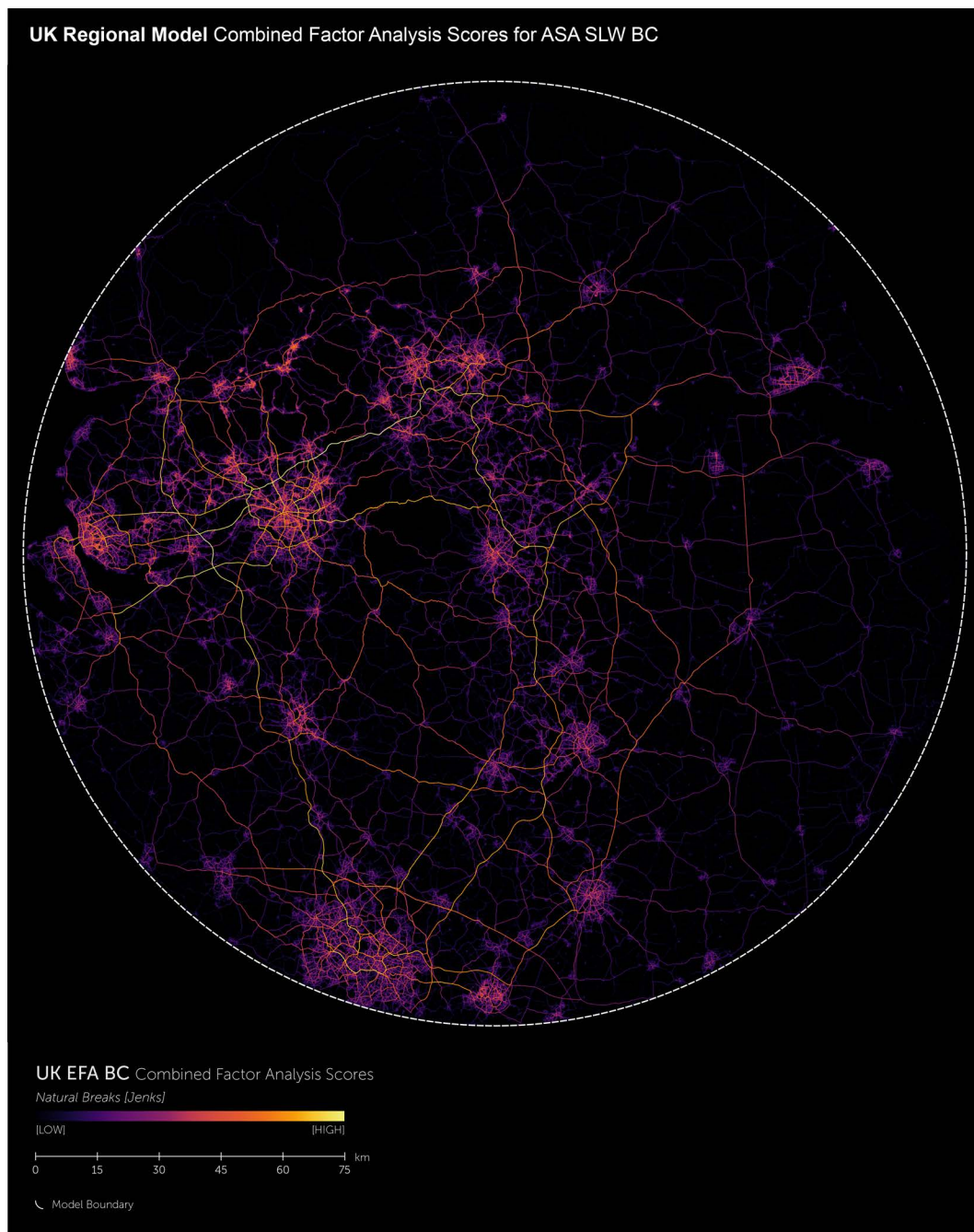
Figure 98: Combined model of 5 EFA factors for betweenness centrality of the UK region. Colour breaks according to the natural breaks algorithm starting from black (lowest value) to bright yellow (highest value).
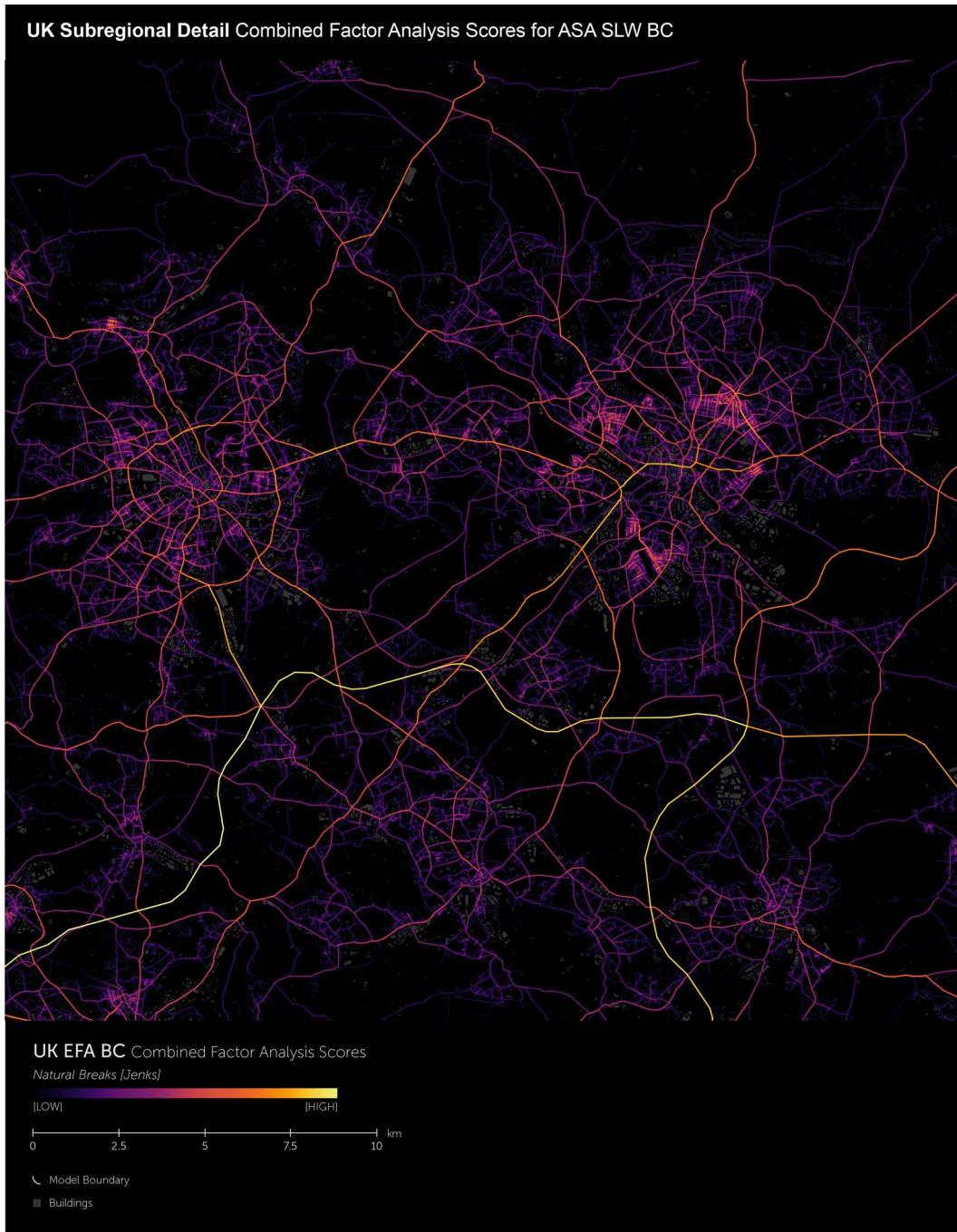
Figure 99: Zoom in of the combined model of 5 EFA factors for betweenness centrality of the UK region. Colour breaks according to the natural breaks algorithm starting from black (lowest value) to bright yellow (highest value). Buildings highlighted in dark grey.
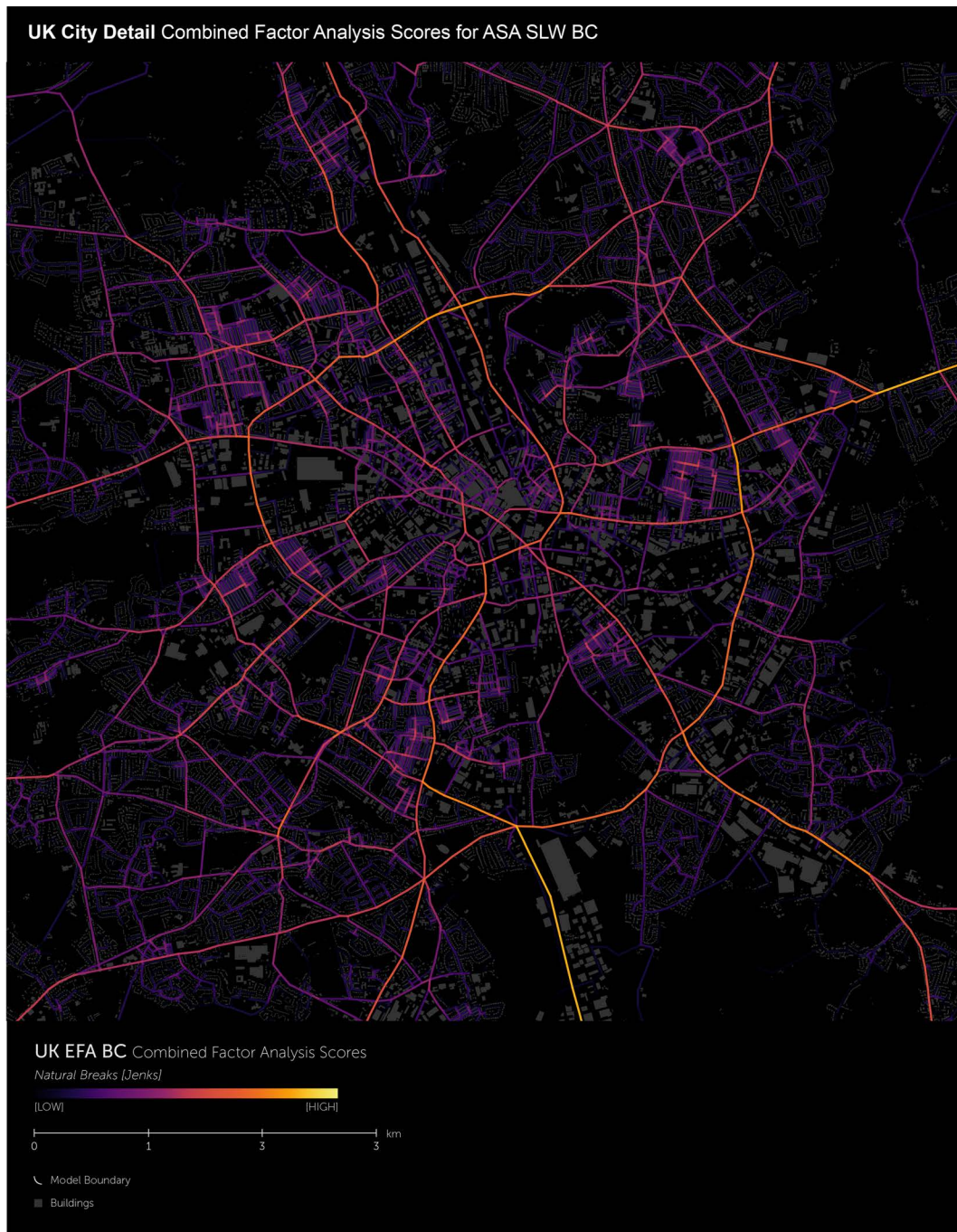
Figure 100: Detailed zoom in of the combined model of 5 EFA factors for betweenness centrality of the UK region. Colour breaks according to the natural breaks algorithm starting from black (lowest value) to bright yellow (highest value). Buildings highlighted in dark grey.

GE Regional Model Combined Factor Analysis Scores for ASA SLW BC

GE EFA BC Combined Factor Analysis Scores
Natural Breaks [Jenks]

[LOW] [HIGH]

0   15   30   45   60   75 km

Model Boundary
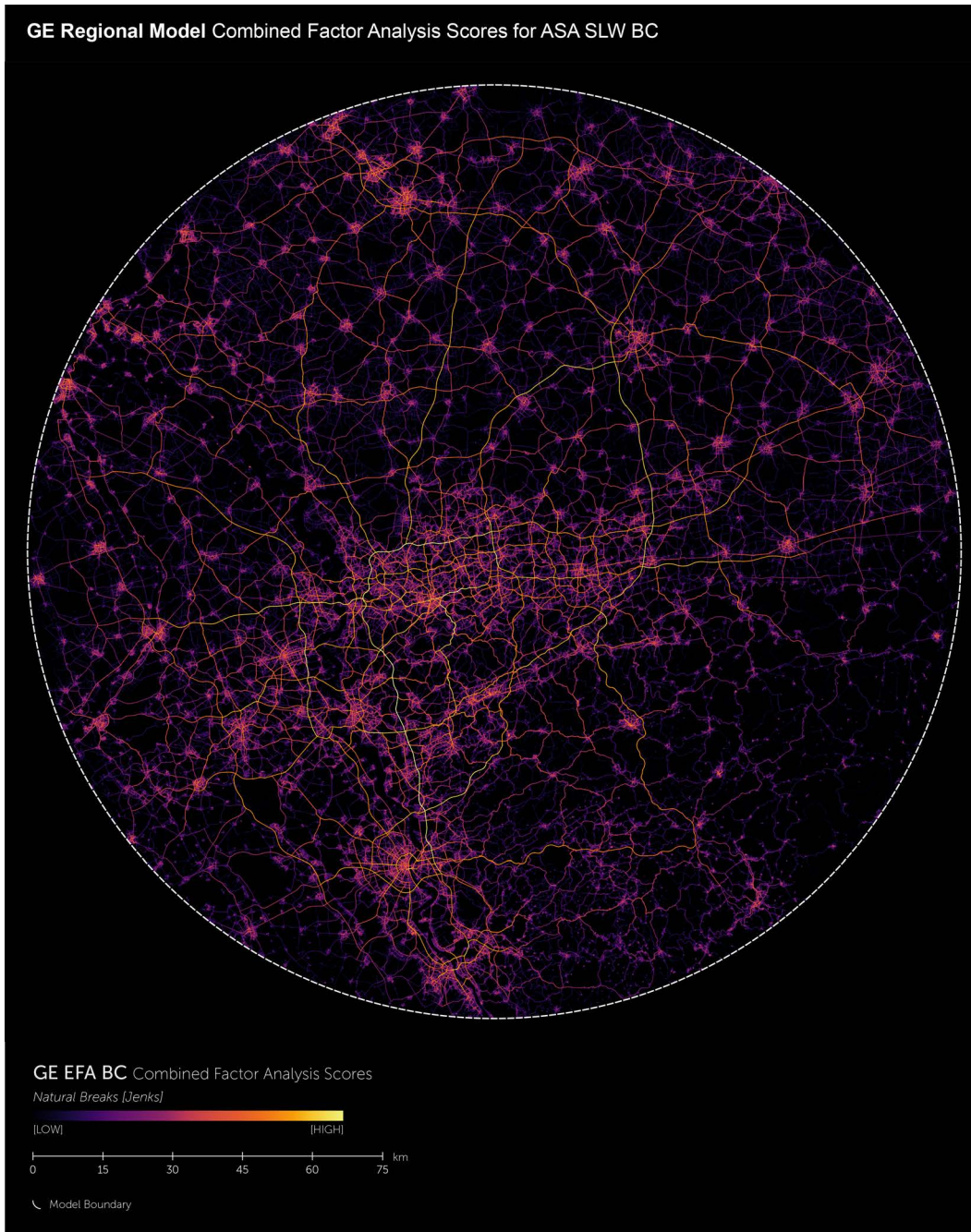
Figure 101: Combined model of 5 EFA factors for betweenness centrality of the GE region. Colour breaks according to the natural breaks algorithm starting from black (lowest value) to bright yellow (highest value).

GE Subregional Detail Combined Factor Analysis Scores for ASA SLW BC

GE EFA BC Combined Factor Analysis Scores
Natural Breaks [Jenks]

[LOW]   [HIGH]

| | | | |
0   2.5   5   7.5   10   km

↘ Model Boundary
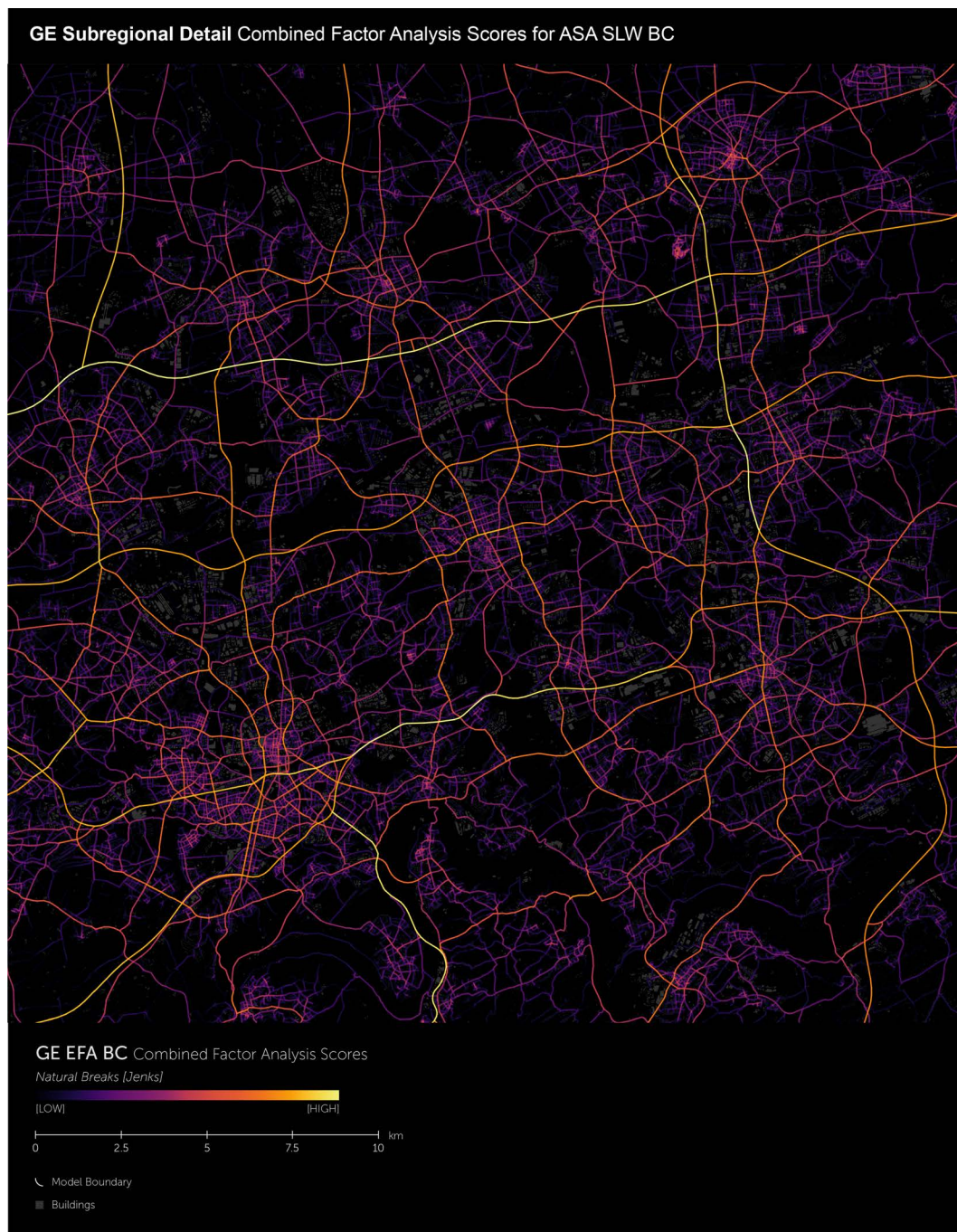
▪ Buildings

Figure 102: Zoom in of the combined model of 5 EFA factors for betweenness centrality of the GE region. Colour breaks according to the natural breaks algorithm starting from black (lowest value) to bright yellow (highest value). Buildings highlighted in dark grey.
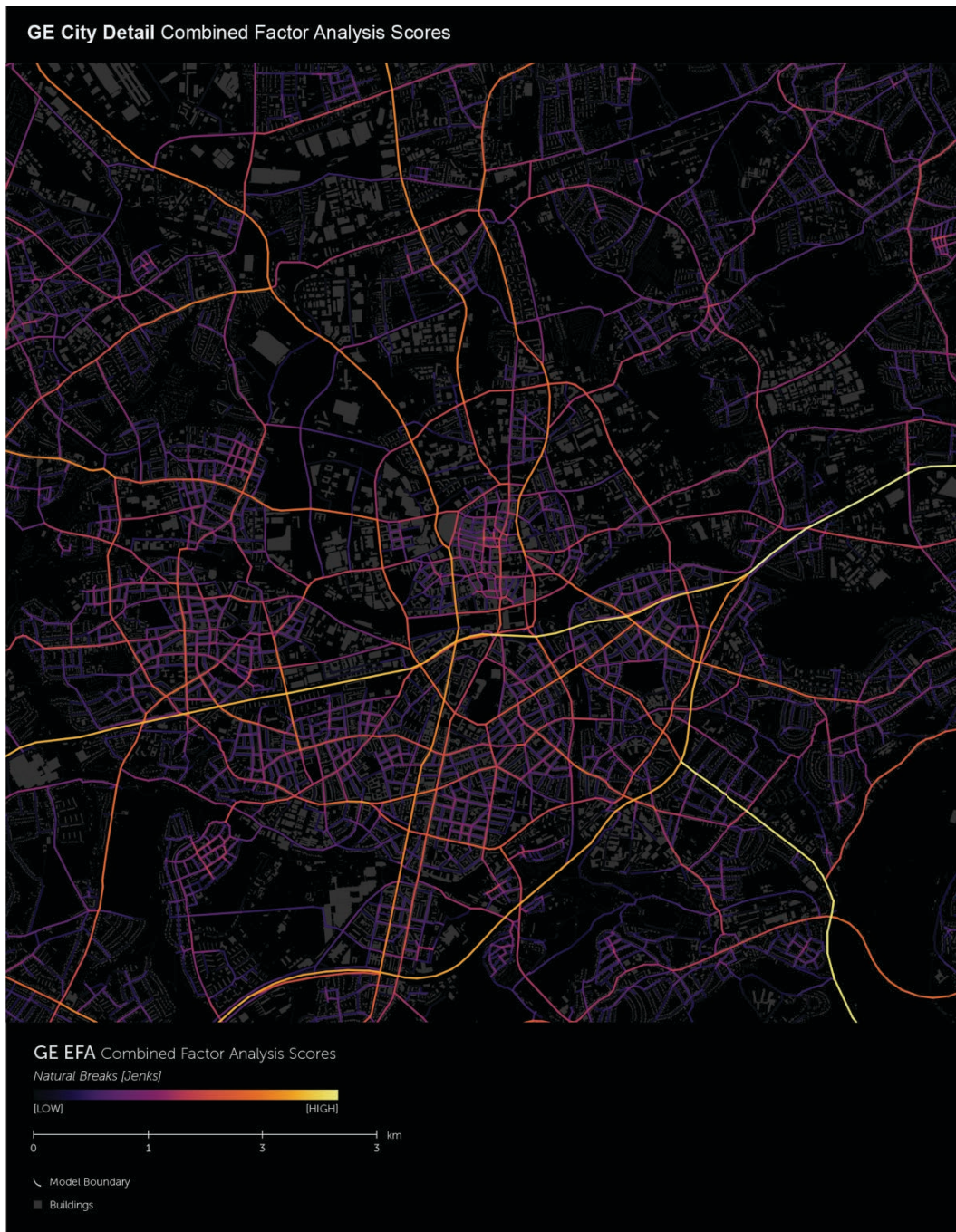
Figure 103: Detailed zoom in of the combined model of 5 EFA factors for betweenness centrality of the UK region. Colour breaks according to the natural breaks algorithm starting from black (lowest value) to bright yellow (highest value). Buildings highlighted in dark grey.

7.1.5 EXPLORATORY FACTOR ANALYSIS FOR CLOSENESS CENTRALITY

The following section builds on the initial findings on betweenness centrality reported above. It also presents the results of the exploratory factor analysis for closeness centrality of the two randomised models and the two real-world regions. Again, I will first determine the total number of factors and compare visualised spatialisations of the resulting latent centrality structures. Before, this is done I will critically discuss the application of angular segment analysis on closeness centrality in a regional context.

UNEXPECTED OUTLIER CASES FOR ASA CC. An occurring problem from such an application is that there are a few cases of high value clusters in areas of low urbanisation where one would have expected lower values. This is particularly the case on small radii (100 – 2500m) but measureable up to a radius of 5km for the cases investigated. One can compare this effect to a similar problem that Hillier et al. (2012, p. 191) faced when introducing normalised least angle choice. The problem seems to be related to long linear segment structures or cul-de-sacs (Figure 104:b). Such linear structures and dead-ends are very common for rural or less-urbanised areas but also highway systems exhibit this effect. The reason for these outliers is a mixture of discretisation from the tulip analysis used in dephtmapX, the length of the segment and the fact that segments are leaf segments, i.e. only connected to one other segment in the analysis). From a mathematical point of view these outliers are to be expected due to the way the values are calculated, however, from a conceptual point of view these outliers are not reasonable and substantially skew the distribution of the data. As partially urbanised areas and rural structures are very common in large cities and also an intrinsic part of regions, it is necessary to detect such outliers. The majority of outlier cases can be identified by means of a visual comparison. Still, some cases of lower value ranges are difficult to identify visually. These exceptional cases can only be identified by comparing values of surrounding segments. This makes it difficult to manually clean the data. However, all segments with an unexpectedly high value share a common characteristic, namely very low total depth (TD) values in relation to the value of closeness centrality (CC). This allows for an objective, reproducible strategy to identify outliers, even in cases where a visual comparison of the data does not allow detection. The following equation can be used to detect outliers in angular segment analysis closeness centrality results. By adding the constant of 3 to CC and TD and dividing the logarithm of CC by the logarithm of TD of the respective radius, one arrives at a new set of values, as is shown in the formula below.

$$CCTD_r = \frac{Log(CC_r + 3)}{Log(TD_r + 3)}$$

(7.2)

In the event, in which the obtained value is equal or above 1 the respective segment can be considered an outlier. Figure 104:a shows a scatterplot that visualises this effect. All values on the right side of the red cut-off line can be considered to be outliers, whereas the distance to the left indicates the amplitude of the outlier effect. I will employ this method on each radius variable that exhibits such outlier behaviour so that all subsequent calculations are based on these 'cleaned' variables.
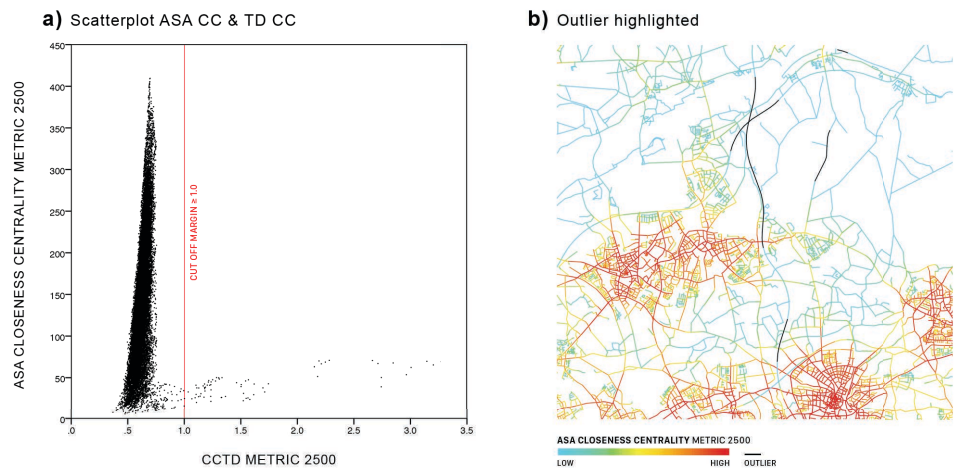
Figure 104: Scatterplot for ASA Closeness Centrality and TD CC radius metric 2500, highlighted in red the cut off margin ≥ 1.0 (a). Detail section of a ASA Closeness Centrality metric 2500 with outliers highlighted in black (b).

EXPLORATORY FACTOR ANALYSIS RESULTS FOR CLOSENESS CENTRALITY. Starting with the two real-world models, the results of the parallel analysis (PA) once more yield a substantial overestimation of the number of factors, which is why in the context of regional spatial network metrics PA appears to be an inappropriate method. Instead, I use the eigenvalue and scree plot method to define the number of factors, as it has been done for betweenness centrality. For the German and British model, the eigenvalue method leads to four factors in both cases. The eigenvalues for the fourth factor are 1.568 for the German case and 1.397 for the British case (Table 25). A potential fifth factor is with 0.914 (GE) and 0.835 (UK) clearly below the threshold of 1 (Table 25).

Table 25: EFA eigenvalue results for ASA CC of the German and British region.

Factor	Eigenvalues GE region	Eigenvalues UK region
1	35.05116	35.08080
2	6.69724	6.63659
3	2.71092	3.23102
4	1.56829	1.39711
5	0.91368	0.83537
6	0.60147	0.58431
7	0.41221	0.39102

This can also be observed in the slope change in the scree plot for both regions (Figure 105:a and b). The scree plot exhibits a drastic decline in the eigenvalue from factor 1 to factor 4 with an almost flat line after the slope at the 5th and 6th factor. A four-factors model appears to be a reasonable solution of closeness centrality in both real-world models.
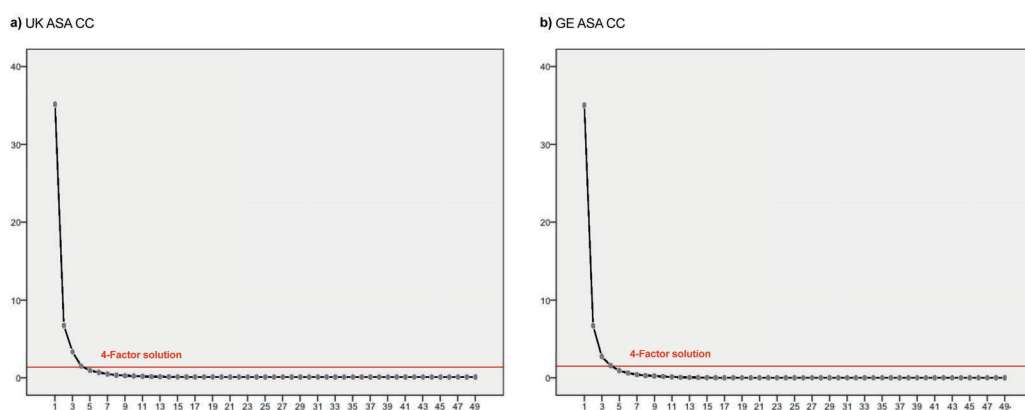
Figure 105: Scree plot for ASA SLW BC for GE and UK. A drastic change in the slope is market with a red line, indicating the number of factors i.e. 4.

This clear picture changes drastically with regards to the randomised models. The ERPGr model features six factors that have an eigenvalue of above 1, with the lowest value being 1.01 (Table 26). The EFA for closeness centrality of the VPGr model results in 5 factors, with the lowest value above 1 being 1.535 (Table 26).

Table 26: EFA eigenvalue results for ASA CC of the ERPGr and VPGr models.

Factor	Eigenvalues ERPGr	Eigenvalues VPGr
1	28.55161	25.42866
2	10.31541	10.91744
3	2.86992	5.14350
4	1.96045	2.37279
5	1.25583	1.53457
6	1.01343	0.90121
7	0.77437	0.66248

The scree plot (Figure 106) visualises these differences as well. The course of the curve of the ERPGr model exhibits a sudden decline and a jump in the slope after the third factor, as well as a second one after the sixth factor. Compared to the VPGr, the curve does not develop as smoothly and it becomes much more difficult to identify the number of factors. The results of VPGr model, on the other hand, are more comparable to the two real-world examples, with a much more distinguishable jump in the curve and a five-factor solution. The differences between randomised and real-world networks are stronger than for the spatial metric of betweenness centrality. This could be caused by the phenomenon, that closeness centrality values can change drastically when a model is characterised by a homogenous distribution of segments, as is the case in the ERPGr model, as well as when a model is characterised by a homogenous distribution of clusters as is the case in the VPGr model.
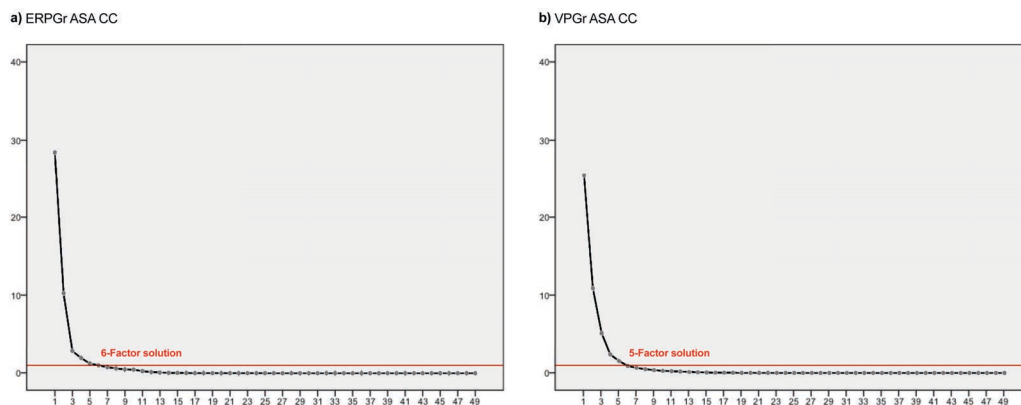
Figure 106: Scree plot for ASA SLW BC for ERPGr and VPGr. A drastic change in the slope is market with a red line, indicating the number of factors i.e. 6 and 5.

Plotting the value of the rotated factor matrix for each of the radii gives further insights into the differences between real-world and randomised models. Figure 107 shows the rotated factor matrix plot for the ERPGr and VPGr model, revealing that both factor

solutions can be considered as simple structure, as they show clear curve developments and no substantial overlaps between the variables explained by each factor. Yet, only the first (EFA ERPGr I and EFA VPGr I) and the last two factors (EFA ERPGr VI and EFA VPGr V) show similarities. All other factors seem to describe different underlying structures, which indicates the existence of fundamental differences in the spatial configuration of these two models.
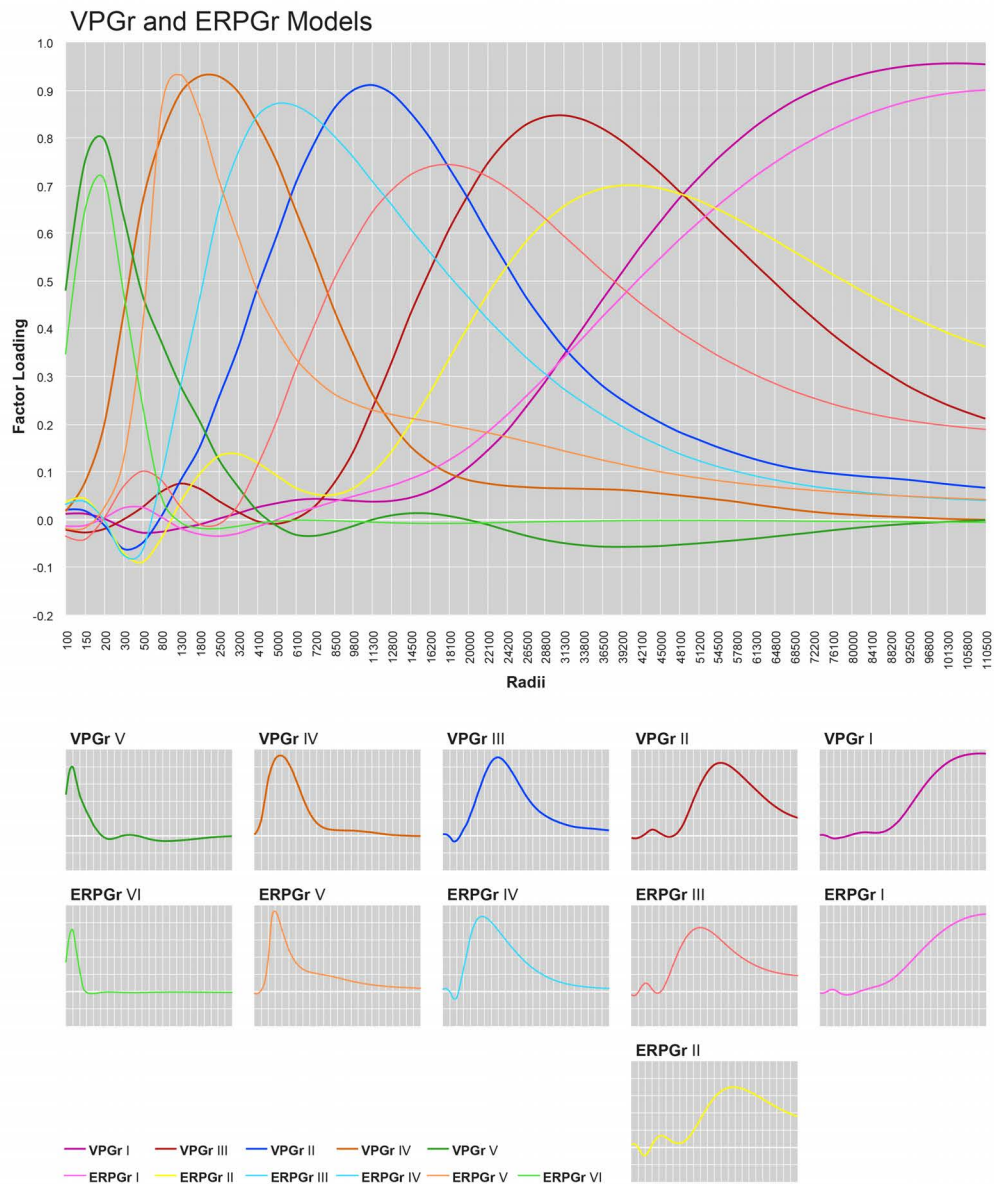
Figure 107: Explorative Factor Analysis rotated factor loadings for 49 metric distances of ASA SLW between centrality for VPGr and ERPGr Models. Forced factor extraction for 5 factors. Extraction method: Principle Axis Factoring. Rotation method: Equamax with Kaiser Normalisation. Rotation converged in 34 iterations for VGPr and 44 iterations for ERPGr respectively.

The ERPGr model features six distinctive latent centrality structures peaking at 200, 1300, 5000, 18100, 39200, and 110500 metres, whereas the VPGr model features five distinctive latent centrality structures peaking at 200, 1800, 11300, 31300 and 110500 metres. The factor loadings, i.e. correlation of the VPGr model, visible in the maximum peak, are substantially higher than the factor loadings of the ERPGr, indicating a more pronounced structure in the VPGr. This seems reasonable, as the VPGr model features segment clusters, which lead to higher and more persistent closeness centrality values. Closeness centrality value distribution behave in a way that large segment clusters naturally attract higher values, because the higher the number of potential journeys in proximity the 'nearer' is a place to other places. Stronger spatial clustering might as well indicate mechanisms of spatial optimisation, as the higher the cluster degree is, the higher the number of accessible spaces is.

Figure 108, shows the plot of the rotated factor loadings of the two real-world models (UK and GE) as well as a model where both regions are combined into a single model (dotted line). The observed factor distribution is again strikingly similar, as was the case with the spatial metric of betweenness centrality earlier. Both models feature four factors as a result of the EFA, with peaks at 200, 4100, 22100, 101300 metres for the UK model and 200, 1800, 14500, 92500 metres for the GE model. The factors of the German model explain in general a smaller radii range than the British model. This could be caused by the different degrees of clustering visible for both regions. As shown in Chapter 6, the British region features fewer, but larger clusters, whereas the German region featured many, but smaller clusters. This particular spatial organisation is reflected in the different latent centrality structures, i.e. scales, that emerge from the EFA analysis. The combined model compares better to the British model, than to the German, which again is an indicator for a stronger cluster mechanism in the British region.

Closeness Centrality
Exploratory Factor Analysis: Rotated Factor Loadings
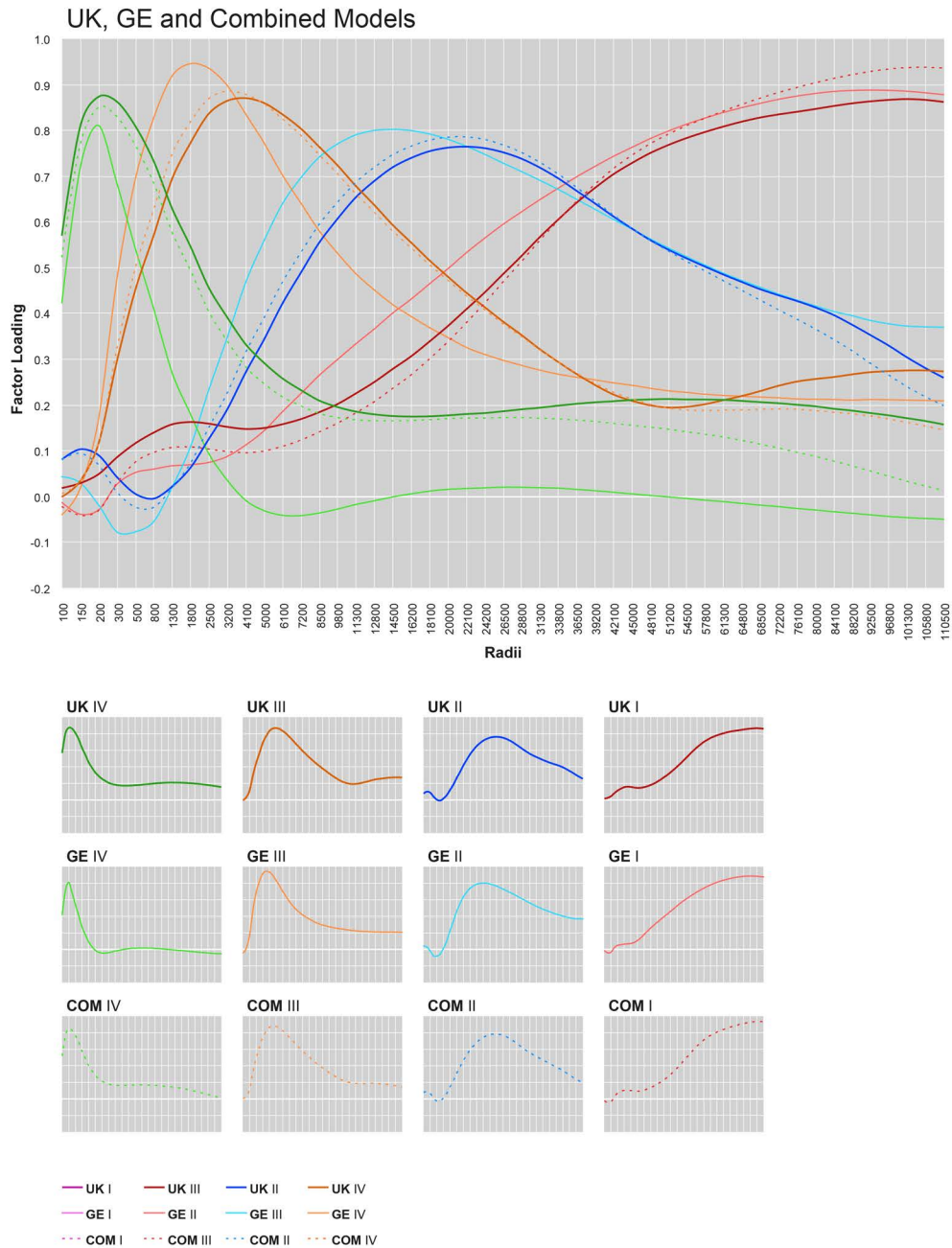
Figure 108: Explorative Factor Analysis rotated factor loadings for 49 metric distances of ASA C for UK, GE and a combined dataset. Extraction method: Principle Axis Factoring. Rotation method: Equamax with Kaiser Normalisation. Rotation converged in 26 iterations for UK, 23 iterations for GE and 24 iterations for the combined model respectively.

When the results of the EFA for closeness centrality of the combined model are compared against the VPGr model it becomes apparent that human-shaped regions feature a simpler latent centrality structure, as they exhibit fewer factors. This might point to a higher degree of organisation and a stricter hierarchy in real-world networks, compared to such randomised networks.

Figure 109: Explorative Factor Analysis rotated factor loadings for 49 metric distances of ASA closeness centrality for VPGr and the combined dataset of GE and UK. Extraction method: Principle Axis Factoring. Rotation method: Equamax with Kaiser Normalisation. Rotation converged in 34 iterations for VPGr and 24 iterations for the combined model respectively.

Further insights can be gained by moving on to a comparison of the observed latent centrality structures for angular segment analysis closeness centrality to Christaller's hierarchy of central places. First, similarities between the two regions are less pronounced than there were for the spatial metric of betweenness centrality. This points to the fundamental differences of cluster degrees in both regions, but also highlights that centre hierarchies are less strict than Christaller's CPT lead us to expect. Second, the third and fourth factors highlight centres that are either not described by CPT or fall into the lowest category (Marktort, M). Fundamentally, these centres occur in the scope of urban spaces, which is excluded in the operationalisation of CPT. As explained in

Chapter 2, the existence of such rather local centres in cities has been demonstrated by many researchers before (Griffiths et al. 2008), and highlights the strength of the approach and the lack in CPT to acknowledge the existence of centres within central places. Nevertheless, a relationship between the estimated market radius and those centres, which are highlighted by the EFA, can be reported. The British region features centres at scales of 200, 4,100, 22,100, and 101,500 metres, of which three are corresponding to the market radius proposed by Christaller. The radii are 4,000, 12,000 and 108,000. The German model features latent centrality structures highlighting closeness centrality clusters at 200, 1,800, 14,500 and 92,500, of which at least one corresponds with Christaller's Kreisstadt (K) with a market radius of 12,000.

Table 27: Comparison of the closeness latent centrality structure with Christaller's central place system and their respective closest scales.

Latent Centrality	UK Region	GE Region	Market Radius (m)	Christaller Type
Neighbourhood	**200**	**200**	-	-
City	1,800	**1,800**	-	-
-	**4,100**	-	4,000	Marktort (M)
Between City/Metro	-	6,100	6,900	Amtsort (A)
Metropolitan	11,300	**14,500**	12,000	Kreisstadt (K)
Between Metro/ Intra-Regio	**22,100**	-	20,700	Bezirksstadt (B)
Intra-Regional	36,500	33,800	36,000	Gaustadt (G)
Between Intra/Inter-Regio	-	-	62,100	Provinzstadt (P)
Inter-Regional	**101,500**	**92,500**	108,000	Landstadt (L)

Whereas betweenness centrality provided insights into the structure that evolves from the relative distance between urban agglomerations, closeness centrality provides insights into the relative size of such agglomerations. Comparable to the previous analysis for betweenness centrality, a series of visualisations have been produced following the proposed method. Each factor is visualised based on its respective loadings and can be seen in Figure 111, Figure 112, and Figure 113, which provide a loading plot for each factor for each of the three models VPGr, UK and GE.

If one compares the randomised model results with the results of the real-world examples, two observations can be made; first, the order in which VPGr clusters are located in each of the factors appears to be rather unorganised. By unorganised, I refer to the fact that clusters of factor V do not necessarily appear as parts of clusters in factor

IV or factor III and so on. In that sense, randomised models do not show a spatial relationship of clusters of one factor to clusters of other factors. This is especially visible in the detail sections of Figure 111, where each factor seems to reflect a somewhat inverse pattern of the other factors. Second, both real-world models show a clear pattern, where clusters of smaller radii factors appear to simply increase in size. This is especially the case for the IV, III and II factor. This pattern shifts in factor I for both regions, where the cluster shifts also to areas that have not been part of a cluster of another factor. With regards to the interpretative value of these latent centrality structures, one can relate factor IV to very local agglomerations of either historic cores or particular settlement patterns. Factor III shows what is semantically referred to as city centres. Factor II highlights metropolitan areas as well as the PUR agglomeration, and factor I seems to be an indicator for a future potential of further regional and metropolitan developments.
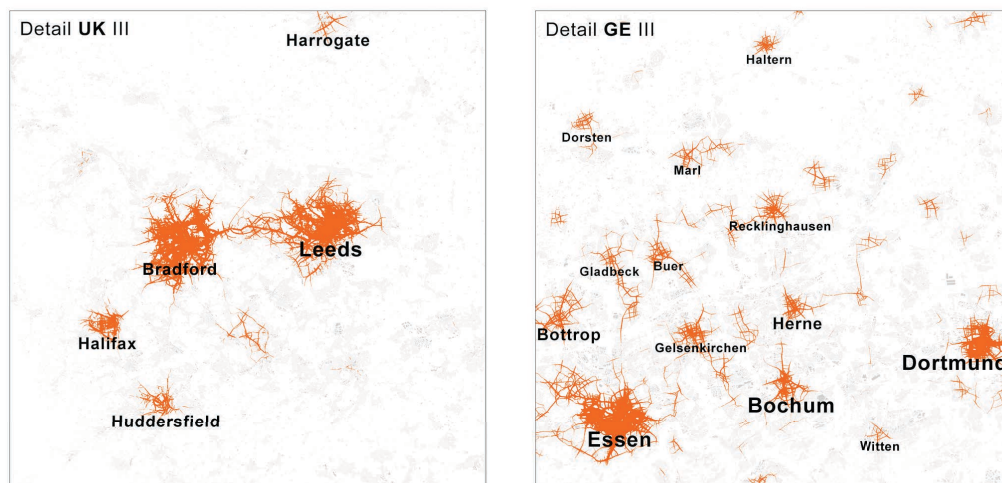
Figure 110: Visualisation of EFA CC III, latent centrality structures for UK and GE.

These interpretative relationships become more apparent when compared next to each other (Figure 110). Factor III allows a clear identification of each historic core and what a resident in the region would refer to as the centre of each city. The difference between the two regions in terms of size and number of the latent centrality clusters is clearly visible. The German region features a respectively denser number of similarly sized clusters. The spatial arrangement of these clusters, however, does not follow any of the spatial organisations described by Christaller. The cluster size can be related to the CPT theory, as well as to a certain degree the spatial distribution of these clusters. The distance distribution of clusters across the region provides us with a picture that reflects such an ordering. However, what clearly does not follow a CPT logic are the extensive urbanised areas between these centres, highlighted by factor II.
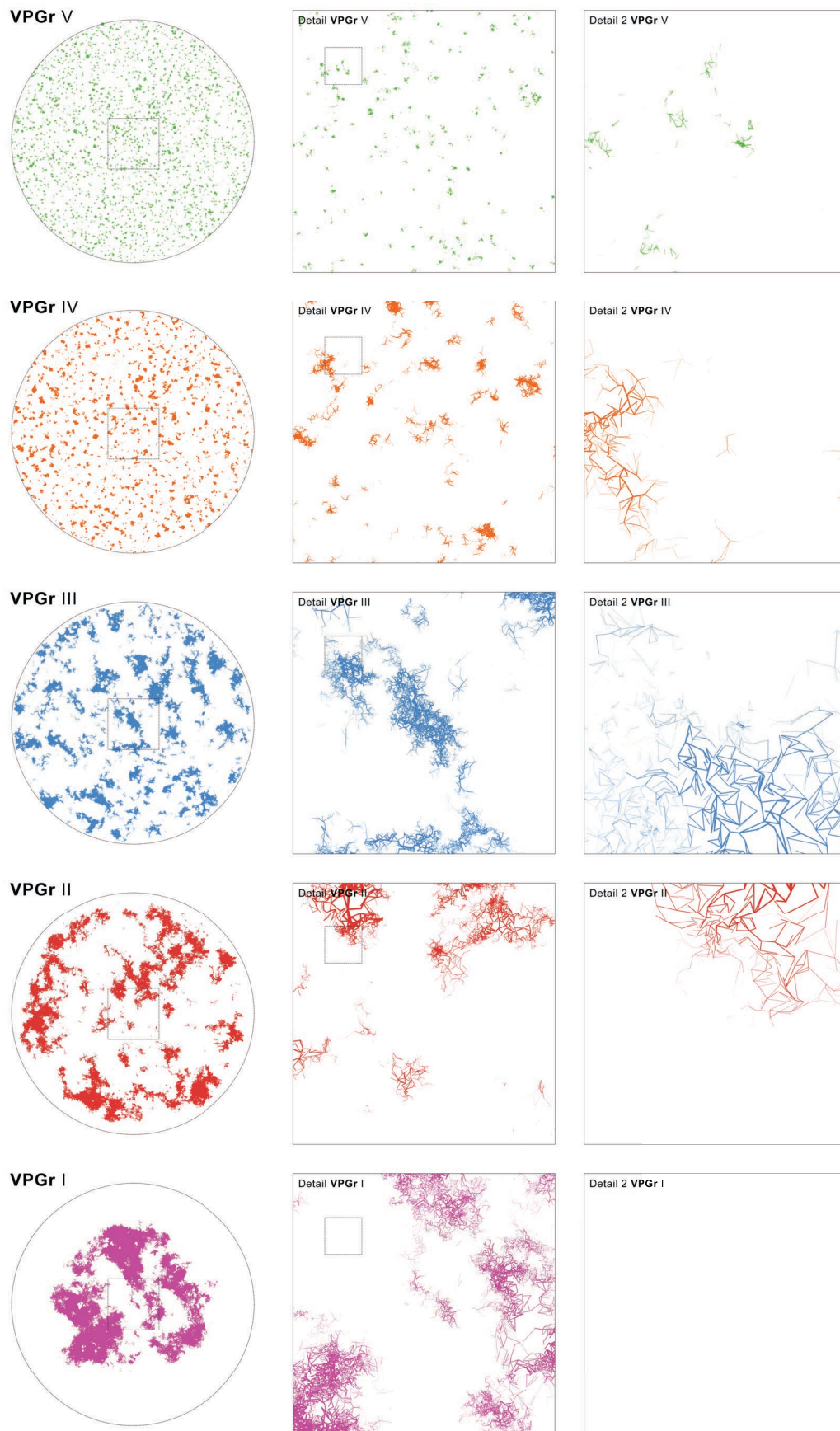
Figure 111: VPGr latent centrality structures for ASA CC. EFA Factor Analysis Scores for each of the five factors and cases with score values above 1.0.
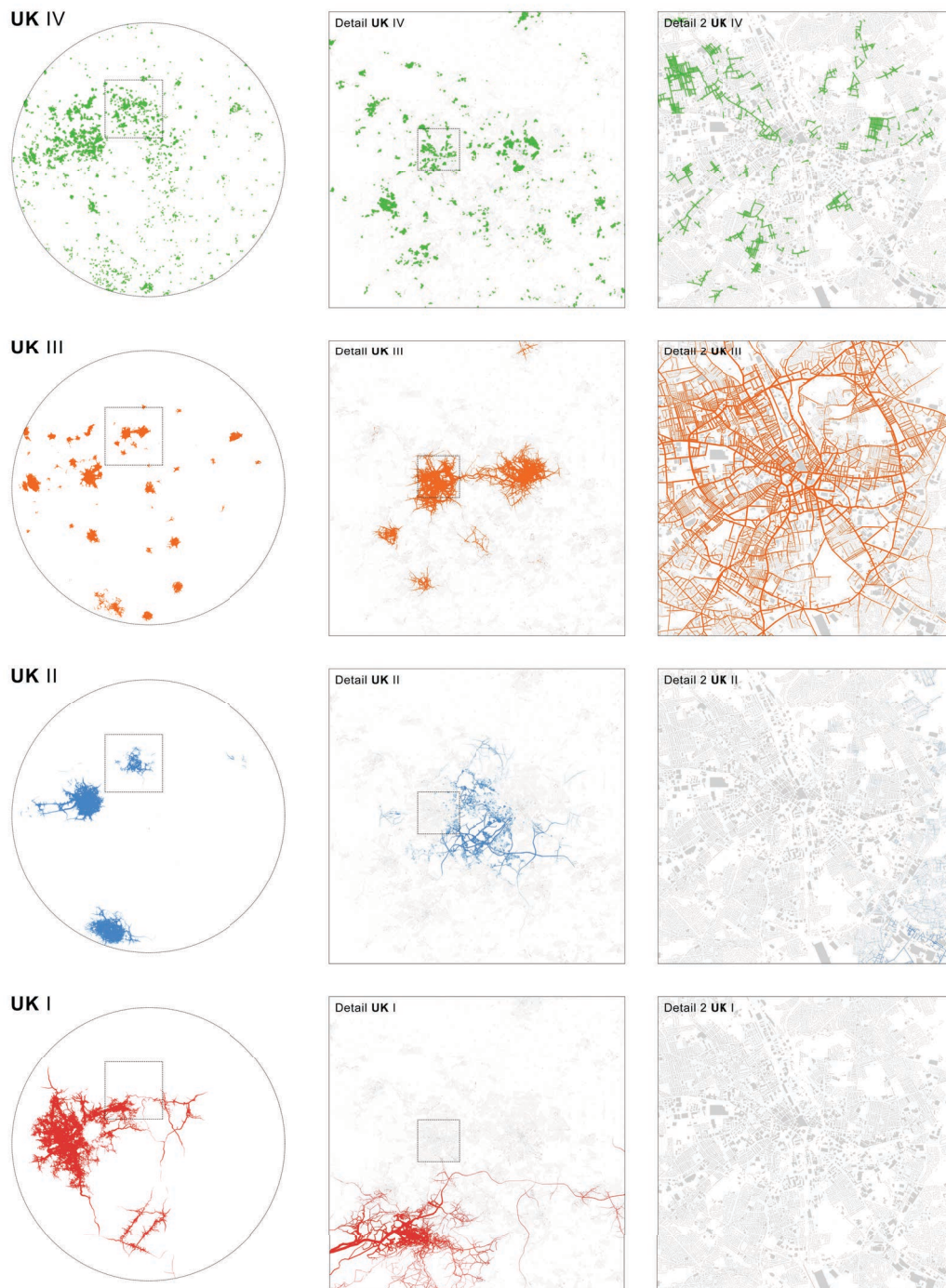
Figure 112: UK latent centrality structures for ASA CC. EFA Factor Analysis Scores for each of the four factors and cases with score values above 1.0.
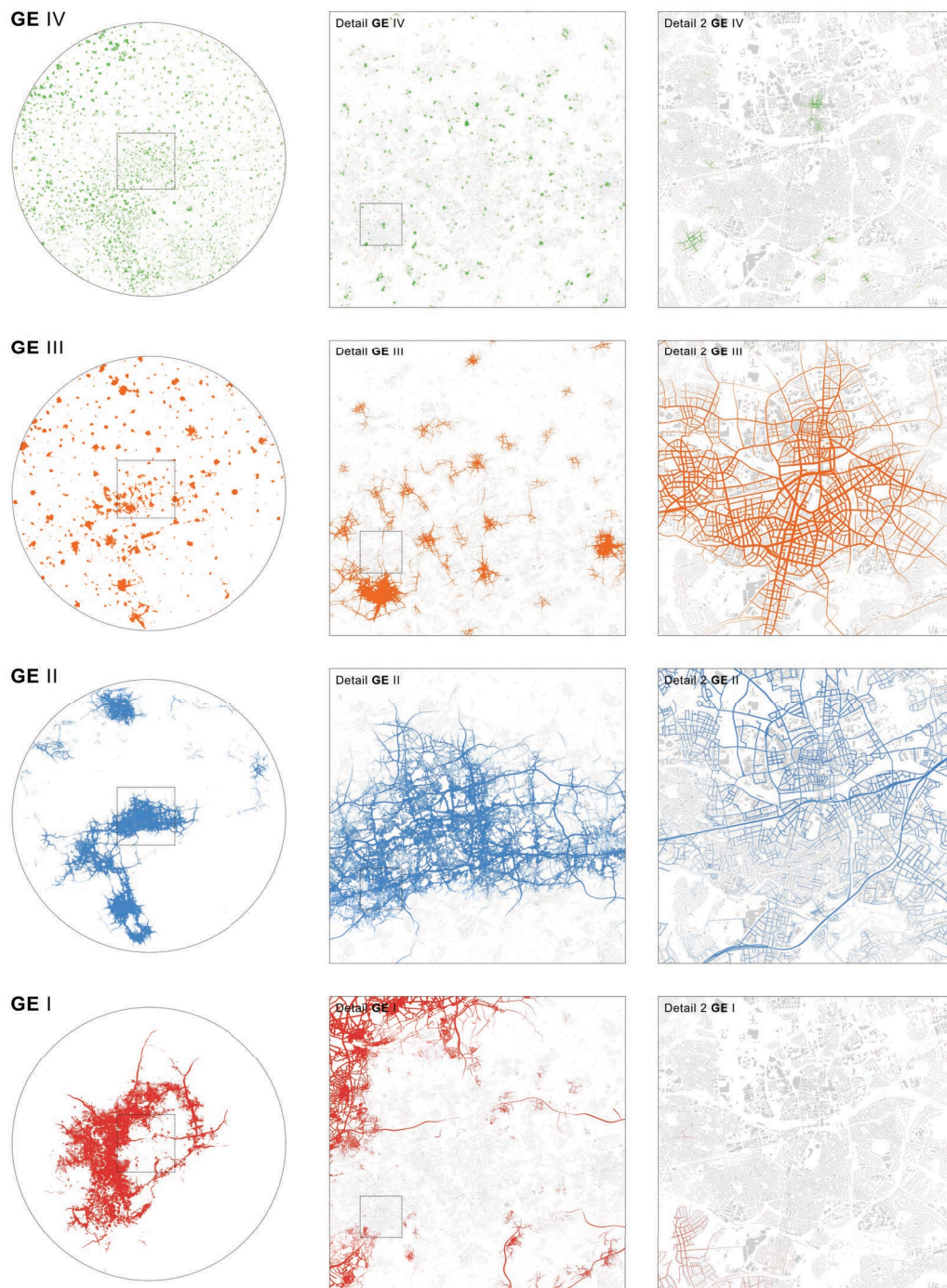
Figure 113: GE latent centrality structures for ASA CC. EFA Factor Analysis Scores for each of the four factors and cases with score values above 1.0.

Finally, after having gained insights into the morphology of latent centrality structures of randomised and real-world street networks for closeness centrality, I will turn to the proposition of another combined model, namely EFA CC COM. The latent centrality factors for closeness centrality are a reflection of the fundamental centrality of a place in terms of its closeness to other surrounding places, rather than representations of potential different modes of travel such as observable in EFA BC factors. The visual interpretation of the closeness centrality factors has shown, that particularly factor III and factor IV can be related to a city distance. Factor I and II had shown a pattern of metropolitan and regional nearness. It appears reasonable to define spatial hierarchy for closeness centrality rather through the relative size of a cluster within each factor, than through the interrelation of different factors. Once again, it is important to highlight that only a combination of all factors can give a comprehensive account of the overall model. I will, hence, propose to combine the extracted factors into a combined model that is able to express the fundamental spatial structure of the spatial network in terms of their closeness centrality. The result is a multi-scalar model that represents the cluster hierarchies existing within the network and each factor. In order to do so, each factor can be combined by using the maximum value of each EFA factor segment (7.3).

$$EFA\ CC\ COM = \max (EFA\ I, EFA\ II, EFA\ III, EFA\ IV) \qquad (7.3)$$

It should be noted that in contrast to betweenness centrality, it remains questionable to which extent the observed structures can be filled with meaningful interpretations. Particularly, the first and second factors appear to be highly influenced by the overall regional network and leave the field of conceptual interpretation. Due to this reason, the following pages present visual representations of the IV and III factors of this new, combined model for the British region (Figure 114, Figure 115 and Figure 116) and the German region (Figure 117, Figure 118 and Figure 119). In both cases, the maps give new insights into the overall regional functioning and the fundamental distribution of centres across the regional space.
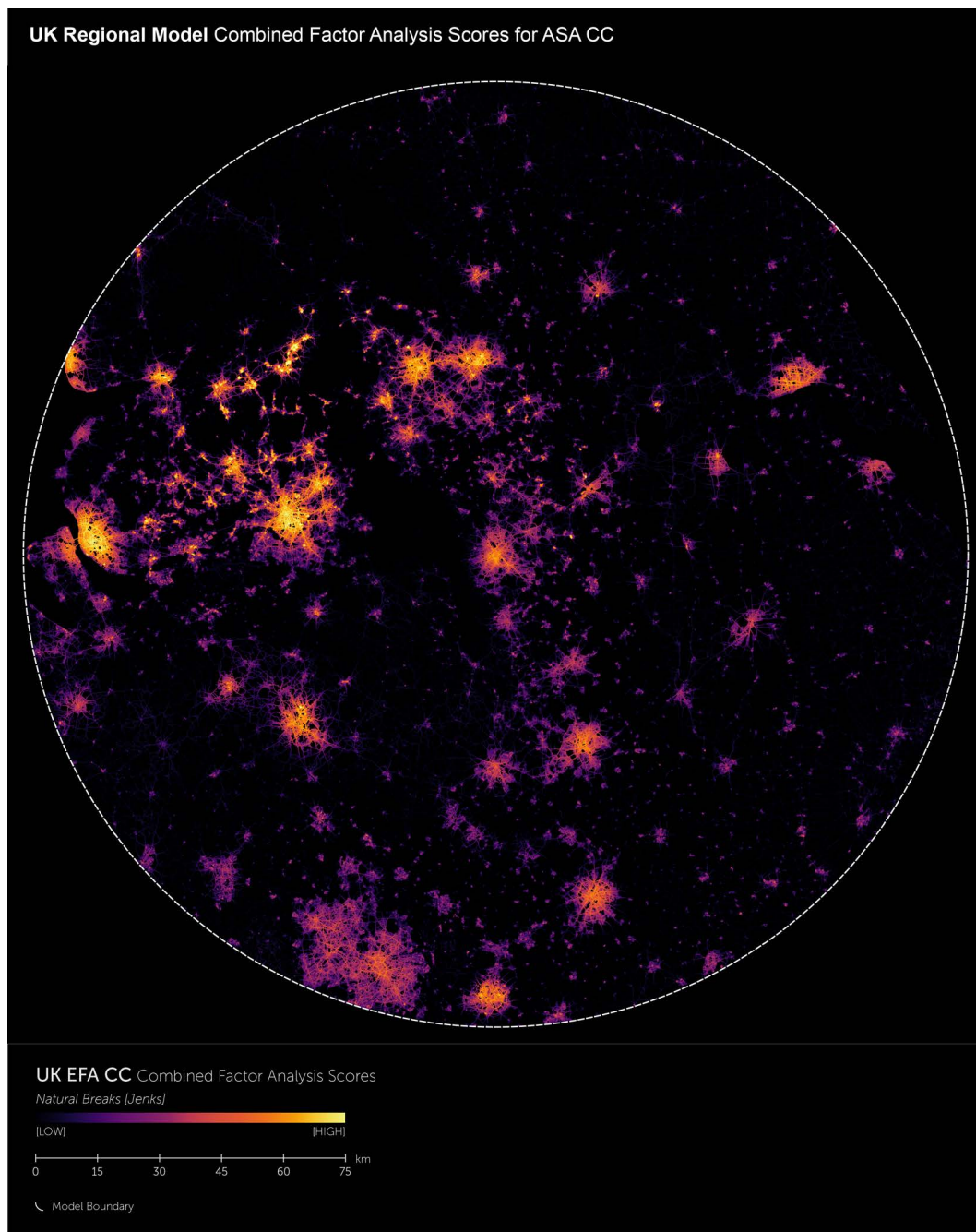
Figure 114: Combined model of 2 EFA factors for closeness centrality of the UK region. Colour breaks according to the natural breaks algorithm starting from black (lowest value) to bright yellow (highest value).

UK Subregional Detail Combined Factor Analysis Scores for ASA CC

UK EFA CC Combined Factor Analysis Scores
Natural Breaks [Jenks]

[LOW]   [HIGH]

0   2.5   5   7.5   10   km

⌐ Model Boundary
▪ Buildings

Figure 115: Zoom in of the combined model of 2 EFA factors for closeness centrality of the UK region. Colour breaks according to the natural breaks algorithm starting from black (lowest value) to bright yellow (highest value). Buildings highlighted in dark grey.

UK City Detail Combined Factor Analysis Scores for ASA CC

UK EFA CC Combined Factor Analysis Scores
Natural Breaks [Jenks]

[LOW]   [HIGH]

0   1   3   3   km

↘ Model Boundary
■ Buildings
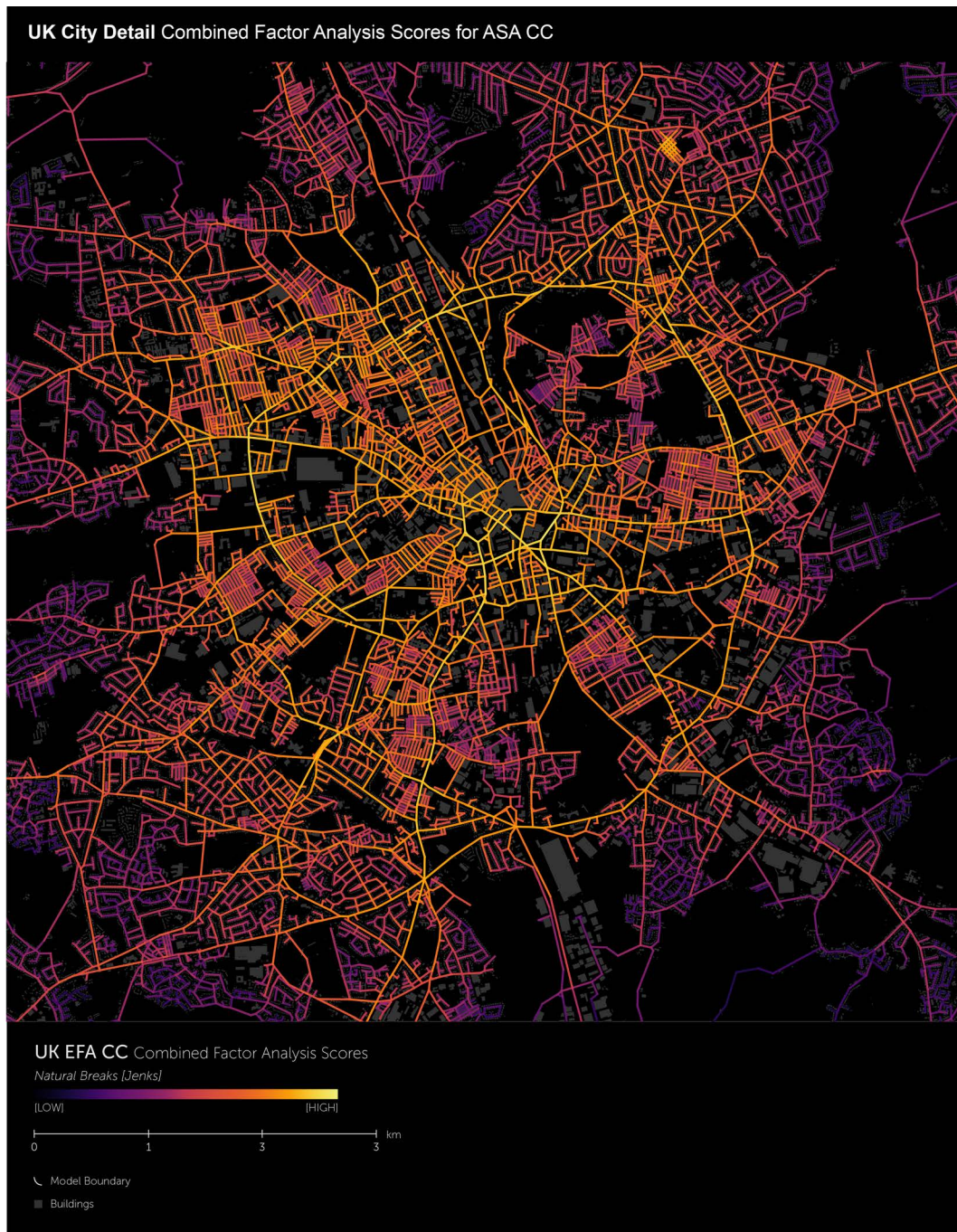
Figure 116: Detailed zoom in of the combined model of 2 EFA factors for closeness centrality of the UK region. Colour breaks according to the natural breaks algorithm starting from black (lowest value) to bright yellow (highest value). Buildings highlighted in dark grey.
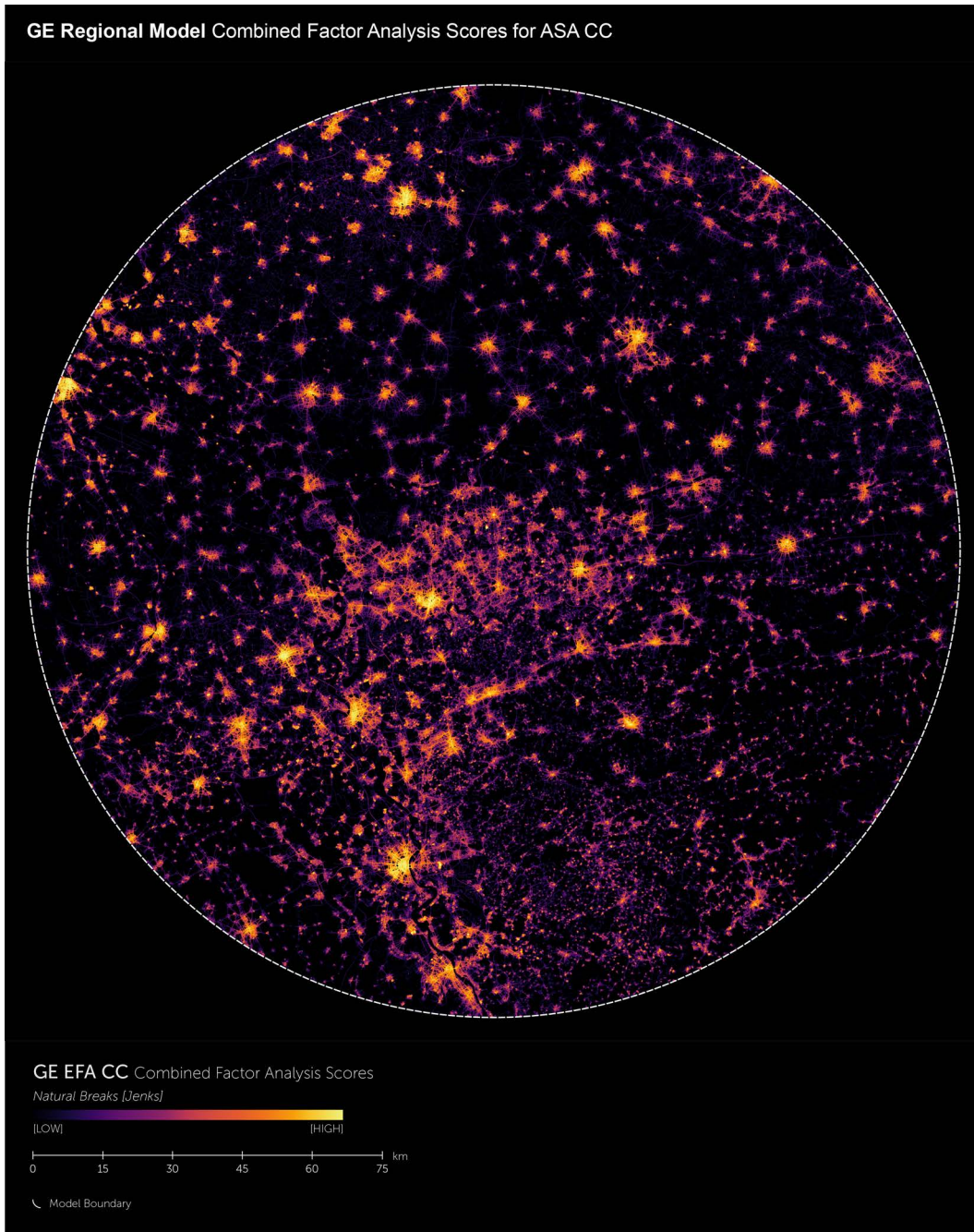
Figure 117: Combined model of 2 EFA factors for closeness centrality of the GE region. Colour breaks according to the natural breaks algorithm starting from black (lowest value) to bright yellow (highest value).
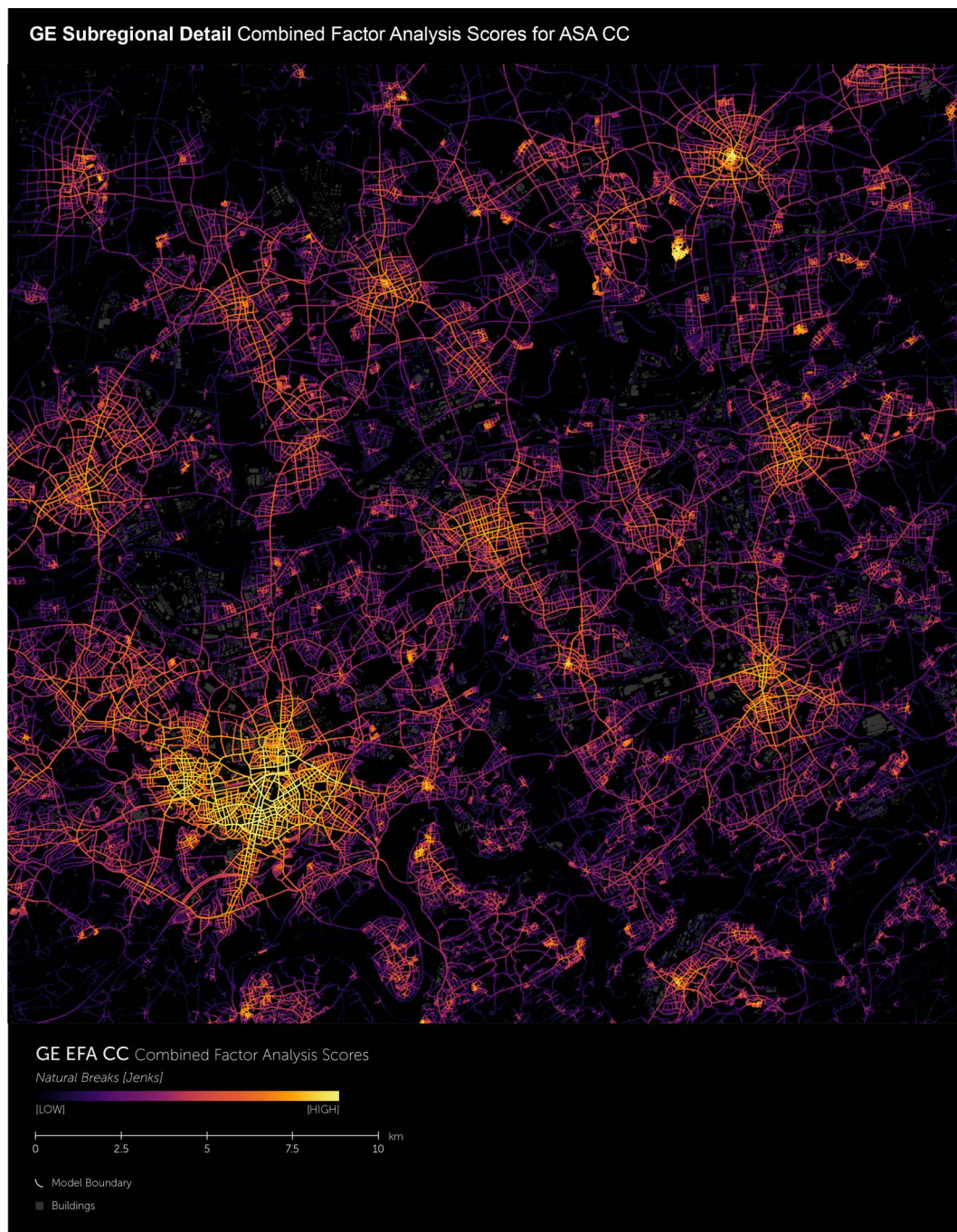
Figure 118: Zoom in of the combined model of 2 EFA factors for closeness centrality of the GE region. Colour breaks according to the natural breaks algorithm starting from black (lowest value) to bright yellow (highest value). Buildings highlighted in dark grey.

GE City Detail Combined Factor Analysis Scores for ASA CC

GE EFA CC Combined Factor Analysis Scores
Natural Breaks [Jenks]

[LOW]   [HIGH]

0   1   3   3   km

↳ Model Boundary
▪ Buildings

Figure 119: Detailed zoom in of the combined model of 2 EFA factors for closeness centrality of the GE region. Colour breaks according to the natural breaks algorithm starting from black (lowest value) to bright yellow (highest value). Buildings highlighted in dark grey.

7.2 SUMMARY

This chapter presented the results of an exploratory factor analysis (EFA) for betweenness (ASA SLW BC) and closeness centrality (ASA CC) for four different regional spatial network models. It introduced EFA as a method to reveal the formerly conceptually sketched latent centrality structures. The analysis could show that for both regional models (UK and GE), such fundamentally different spatial structures exist. An exploratory factor analysis for ASA SLW BC resulted in five distinctive spatial patterns in both real-world regions. For ASA CC, the exploratory factor analysis resulted in four distinctive spatial patterns. A comparison of two randomised spatial models provided insights into the fundamental difference between human made spatial networks and those that are generated by chance. With regards to betweenness centrality, these differences are particularly visible in the way the spatial network is organised. Both the Variance Gamma Random Planar Graph with radius restriction (VPGr) and the human made spatial networks (UK and GE) exhibited similar, latent centrality structures in terms of the radii that each of the factors explained. However, the spatial manifestation of these factors highlighted that human-shaped networks are characterised by a clear pattern of linearity connecting segments for all factors. With regards to closeness centrality, the differences are of a different kind: in human-shaped networks clusters of high values are spatially arranged in such a way that the size of the clusters grows from factor to factor, whereas the randomised model exhibits an inverse pattern between factors. A comparison with the proposed central place hierarchy by Christaller highlighted that all found factors can to a certain degree be related to the CPT.

These observations are all made from a structural or morphological point of view and with regards to the network itself. In the following chapter, I will investigate to which extent these latent centrality factors can be used to bring meaning to the actual, observed distribution of human spatial organisation. This is done by a series of spatial correlations of socio-economic variables with these latent centrality scales.

CHAPTER 8

NETWORK CENTRALITIES AND ESTIMATION OF SOCIO-ECONOMIC VARIABLES

8 CHAPTER

Following the discovery of comparable latent centrality structures in both real-world networks of the case studies, this chapter presents the results of the evaluation of these latent centrality structures for the estimation of socio-economic variables. This chapter tests if we can use spatial network metrics (i.e. betweenness and closeness centrality) as estimators for human affordance for regional planning purposes to aid transport, social and economic strategies. It explores if we can use network centralities as an appropriate method for the regional prediction of spatial occupation and movement patterns; or in other words whether latent centrality structures are an appropriate predictor for human affordances, such as where people are located in space and where people move in space. As a part of these inquiries, this study will be the first to employ 3D-building information for population estimation in a cross-country comparison. As a further result of these inquiries, I produce a novel dataset of street-level population estimates of unseen precision and extent.

I will begin by presenting the results of the prediction of regional movement by spatial metrics, followed by regional population predictions and finally by an identification of the relationship between service and trade centres and latent centrality structures, i.e. scales. The tests show that spatial network metrics and latent centrality structures hold substantial explanatory power for the prediction of regional movement and the location of service and trade centres on the level of the spatial network segment. With regards to places of residential occupation, such a relationship can only be reported on lower spatial resolutions. Simple geometric characteristics of the spatial network are proven to be better predictors for residential occupation than complex centrality factors.

8.1 NETWORK CENTRALITIES AND ESTIMATION OF SOCIO-ECONOMIC VARIABLES

8.1.1 MOVEMENT PREDICTIONS THROUGH SPATIAL NETWORK CENTRALITIES

I will begin with a series of correlations of traffic flow data with the identified latent centrality structures (see 4.1.1, page 102 for the used data and 4.2.1, page 132 for the employed methodology). The aim is to identify the role of each scale on the predictive power of the overall traffic flows. Values of EFA factors are relatively normally distributed around 0. This means, that streets that are well represented, e.g. factor 1, have a positive loading, while streets that are better correlated with factor 2 usually feature a negative loading in factor 1 in relation to the mathematical distance of both factors. A comparison with all traffic counts and EFA factors needs to take this

distribution of values into account. These comparisons should only be made with those streets that are of significance in the respective factor. As presented previously, whether a street is of significance for a latent centrality factor can be identified through a value equal to or higher than 1. Correlations with traffic counts and a latent centrality factor are based on all streets that are significant for the respective factor (i.e. value $\geq$ 1). A bivariate linear fit has been performed on each variable pair of average annual daily traffic flow (AADF) and latent centrality structures (EFA BC and EFA CC) and the resulting R-squared are compared with each other.

Table 28 and Table 29 show the results of this analysis. For both regions, correlations of betweenness centrality and the four AADFs values. The first three factors (BC5 to BC3) are not able to explain any variance in the data ($r2 \leq 0.050$). This means local, city-wide and metropolitan scale structures are not able to account for traffic flows within either of the two regions. Meaningful correlations can only be reported for factor BC2 (for AADF_All $r2=0.485$), the intra-regional scale, and BC1 (for AADF_HV $r2=0.507$), the inter-regional scale. BC2 appears to have better explanatory power for all vehicular traffic (AADF_All) and all passenger traffic (AADF_PA). Both results can be related to commuting patterns that seem to exist more on an intra- rather than an inter-regional scale. BC1 correlates strongest with all heavy goods vehicular traffic (AADF_HV) and all freight traffic (AADF_FR). The relationship of the explanatory power of BC2 and BC1 for passenger and freight traffic respectively is significant in both models. The strongest correlations for all AADFs can be reported for BC COM, the combined factor model with the highest R-squared of 0.634 (UK) and 0.733 (GE) for all vehicular traffic. Correlations between BC COM and freight and heavy goods traffic are markedly higher than scores of individual factors.

Table 28: R-squared of latent centrality structures of betweenness centrality and different annual daily traffic flows for both model areas.

	Variables	EFA BC COM	EFA BC5	EFA BC4	EFA BC3	EFA BC2	EFA BC1
UK	N	854	60	150	188	200	206
	AADF_All	0.634	0.025	0.010	0.033	0.485	0.312
	AADF_PA	0.606	0.024	0.015	0.053	0.479	0.238
	AADF_FR	0.650	0.026	0.000	0.000	0.432	0.488
	AADF_HV	0.592	0.023	0.001	0.011	0.295	0.507
GE	N	3,055	71	667	1,652	1,238	414
	AADF_All	0.733	0.008	0.001	0.045	0.414	0.539
	AADF_PA	0.718	0.007	0.000	0.059	0.431	0.450
	AADF_FR	0.687	0.014	0.007	0.001	0.269	0.667
	AADF_HV	0.656	0.013	0.005	0.000	0.220	0.655

Table 29: R-squared of latent centrality structures of closeness centrality and different annual daily traffic flows for both model areas.

	Variables	EFA CC COM	EFA CC4	EFA CC3	EFA CC2	EFA CC1
UK	N	854	45	196	196	224
	AADF_All	0.231	0.052	0.048	0.408	0.108
	AADF_PA	0.249	0.052	0.051	0.416	0.095
	AADF_FR	0.162	0.050	0.028	0.343	0.136
	AADF_HV	0.100	0.077	0.020	0.226	0.159
GE	N	3,055	56	566	879	1,250
	AADF_All	0.337	0.000	0.017	0.278	0.083
	AADF_PA	0.352	0.000	0.018	0.288	0.078
	AADF_FR	0.221	0.008	0.006	0.191	0.093
	AADF_HV	0.200	0.009	0.009	0.171	0.084

The explanatory power for all of the variance in the data by the variable closeness centrality is considerably lower than that by the variable betweenness centrality. The strongest relationship can be observed between CC2 and AADF_PA ($r2=0.416$). The two small-scale latent centrality factors (CC4 and CC3) show no linear relationship with traffic flows in both regions. In the British model, only the large-scale factor CC2 has noteworthy correlations of an R-squared of 0.416 (AADF_PA) and 0.408 (AADF_All). Similarly, in the German model CC2 scores an R-squared of 0.288 (AADF_PA) and 0.278 (AADF_All). The combined models have relationships of different strength between both regions, while overall in the German region more variance can be explained by a factor combination (an R-squared increase of 0.064 at worst). This is not the case for the British model (an R-squared decrease of 0.181 at best). Closeness centrality cannot account for a sufficient share of the variance of the data to employ the measurement for regional movement flow predictions.

The existence of a linear relationship between betweenness centrality and traffic flows and a weaker, non-linear relationship between closeness centrality and traffic flows becomes visible in the comparative scatterplots of AADF_All, EFA BC COM and EFA CC COM (Figure 120). Both regional models share a similar linear relationship between EFA BC COM and AADF_All, offering the possibility to employ either of the fitted models in the respective other region. The scatterplot for EFA CC COM and AADF_All shows no linear relationship, reassuring previous findings. Unlike previous research in the context of the individual city, no strong relationship can be reported between closeness centrality and movement for network models of regional extent. Human movement in space within regions, it appears, is following the principle of through-movement, rather than to-movement. Betweenness centrality might be a closer proxy

for human movement, because when humans move in space, they choose the same space to pass through more often than the same destination. Simultaneously there are more potential destinations (each segment forms a potential destination) in a regional spatial network than there are spaces that foster efficient journeys through the system (principle of hierarchy).
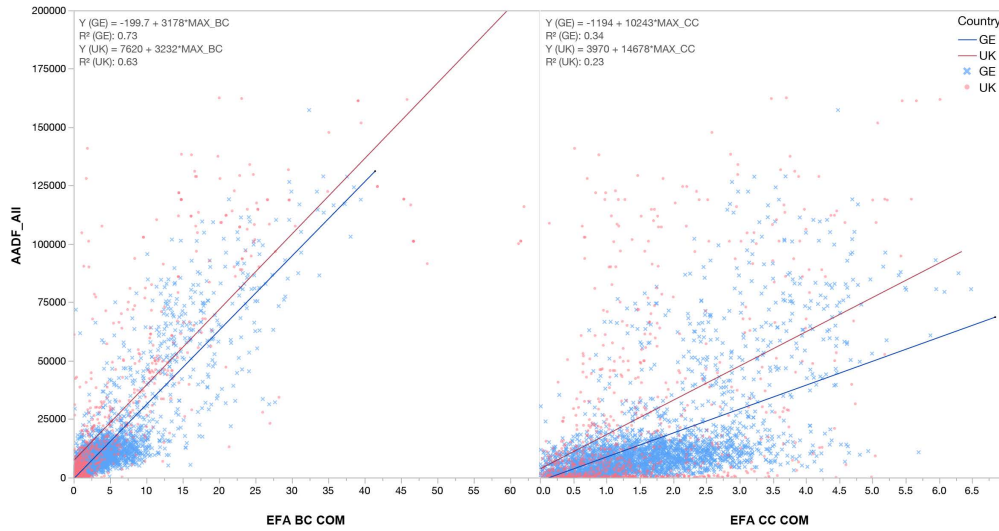
Figure 120: Scatterplot of EFA BC COM and EFA CC COM combined models correlated against the annual daily traffic flow (AADF) of all motored vehicles.

A further comparison of BC factors with each analysis radius and AADFs exhibits the strength of the combined model (EFA BC COM). Figure 121, shows the development of R-squared values of 196 individual bivariate analyses between EFA BC COM and EFA CC COM. The combined betweenness centrality factor model features two distinctive patterns between passenger and all vehicular movement and freight and heavy goods, with the former peaking at 48,1 kilometres and the latter at 88,2 and 110,5 kilometres. This is once again an indicator for a situation where two distinctively different scales are at work. Passenger movement can be related to intra-regional movement, whereas freight and heavy goods exhibit a pattern of intra-regional and potentially nation-wide scales. The analysis further demonstrates the usefulness and explanatory power of latent centrality factors. The very same pattern could be observed through the use of EFA BC factors, although with weaker correlations for the individual factors. With regards to the explanatory power of the combined model, however, this is marginally better than the highest correlation of an individual radius. Accordingly, the R-squared value is between all annual daily traffic flows (AADF_All) and a radius of 48,1 kilometres 0.674, and 0.677 for EFA BC COM. This shows that a combined factor model can account for a similar degree of variance, but without the involved bias of radii selection.
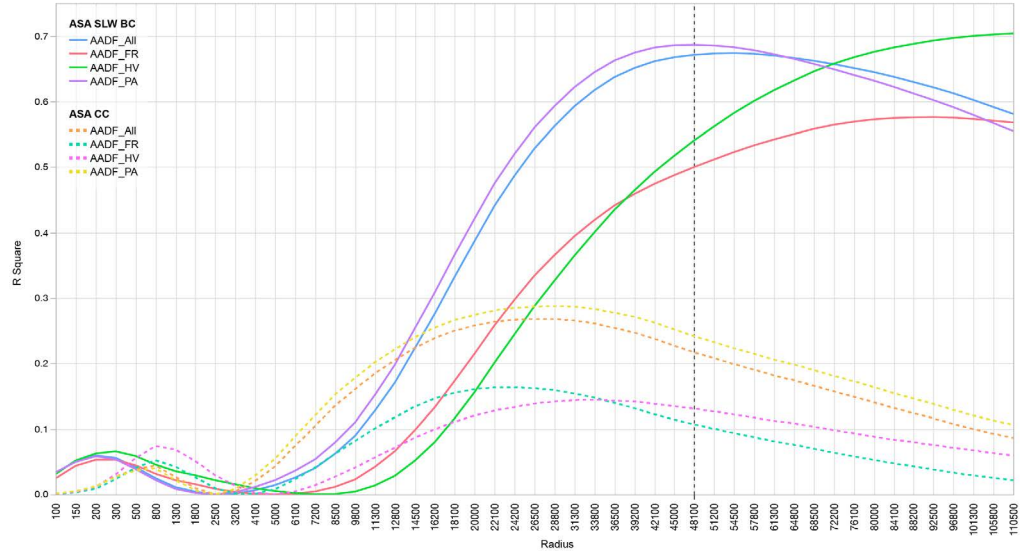
Figure 121: R-squared of all AADF counts against ASA SLW BC and ASA CC on 49 different radii.

The tests have shown, that particularly the combined latent centrality factor model can be employed in PURs and explain almost 70% of the observed variance. The combined latent centrality factor model hence forms an appropriate tool for early traffic estimations, in polycentric urban regions where traffic counts are not available. Moreover, the individual factors hold the explanatory power of identifying scale dependences.

8.1.2 REGIONAL POPULATION PREDICTION

Following the previously set out methodology for population predictions, this sections presents the exploration of the predictive power of 3D-building information for population estimates in a cross-country comparison (see 4.1.2, page 109 and 4.1.3. page 113 for the data used and 4.2.4, page 140 for the methodology employed). This exploration is the first study employing 3D-building information for population estimation in a cross-country comparison. It will show, that 3D-building geometries in combination with semantic building information is an appropriate method for the estimation of population data. The results highlight that single model approaches that bridge country datasets, are ineffective due to significant the differences in socio-cultural appropriation of space. Moreover, I will argue that data disaggregation of observed population data to the level of 3D-building geometries constitutes an adequate method for the resolution enrichment of population data. This is a pivotal step in order to enable robust comparisons with spatial network metrics. I will first present the suitability of both datasets and the improvement gained by the employment of semantic information enrichment. This is followed by a comparison of different aggregation methods for the aggregation of population data per spatial network segment.

I will begin with the exploration of the suitability of semantic information enrichment for the purpose of population estimation. With the use of semantic information of the ALKIS, AddressBase and OSM POI datasets, I create four different building selections. These four selections consist of volumetric information, building footprint area and number of units for a) all building geometries of the model area, b) all building geometries with an exclusion of buildings with an industrial function, c) all residential buildings and d) all residential buildings with a correction coefficient for volumes that contain retail or service functionalities. Each of these four selections have been spatially related to the GEOSTAT population grid. This process has been done by spatially joining all buildings whose geometric centre falls within the grid polygon. All four selections are compared by a simple correlation approach. Table 30 shows the results of this exploration. The aim is not to test the actual effect on the correlation, but to retrieve the extent of the relation for further analysis.

For both models, the strongest relationship exists between volumetric information of optimised residential buildings and population data (GE: 0.959 and UK: 0.937). The number of buildings (i.e. units) produces better correlations for the UK model, than for the GE model, yet both correlations increase through information enrichment. If volumetric information is employed, these relationships decrease significantly to only 0.665 for the German model and 0.424 for the British model. It is clear that semantic information enrichment improves the relationship of all three different measures with population data. However, the use of semantic information becomes an essential procedure when using 3D-building information on a regional scope. If volumetric information is employed without a subsequent selection of buildings of residential use, then the number of significant outliers is much higher. Furthermore, the semantic information becomes less effective when pure building units are compared.

Table 30: R2 values for correlations between different building information and population data per 1 x 1km grid (For all probability > |t| is < .0001*). GE n=10,928 and UK n=5148.

Model	Class	Units	Area	Volume
GE	All Buildings	0.776	0.716	0.666
GE	No Industries	0.753	0.882	0.874
GE	Residential	0.823	0.913	0.956
GE	Residential Optimised	0.820	0.914	**0.959**
UK	All Buildings	0.904	0.705	0.424
UK	No Industries	0.901	0.833	0.627
UK	Residential	0.922	0.916	0.936
UK	Residential Optimised	0.923	0.917	**0.937**

The effect of said outlier behaviour is observable in the following scatterplots (Figure 122, Figure 123). Both selection categories, 'all buildings' and 'no industries' exhibit an over-prediction with data points of high volume values and low population values. If volumetric information is used for all buildings, outliers are particularly visible in lower population ranges. The number of outliers decreases when I remove buildings of industrial use. This changes with regards to the relationship between the residential and optimised residential selections and population data. Both selection categories show strong linear relationships. Outliers are only present in the upper population ranges with comparable results in both models.

The residential and optimised residential selections are almost identical in their residual distribution, with an increasing dispersion in both positive and negative directions with a higher level of population. However, the British case exhibits improvements in high population estimations. Both UK and GE models exhibit only marginal differences between the residential volume and optimised residential volume, visible in the 0.001 differences in their R-squared scores. Volumetric adjustments for retail and service spaces seem to have only insignificant effects on the overall relationship.
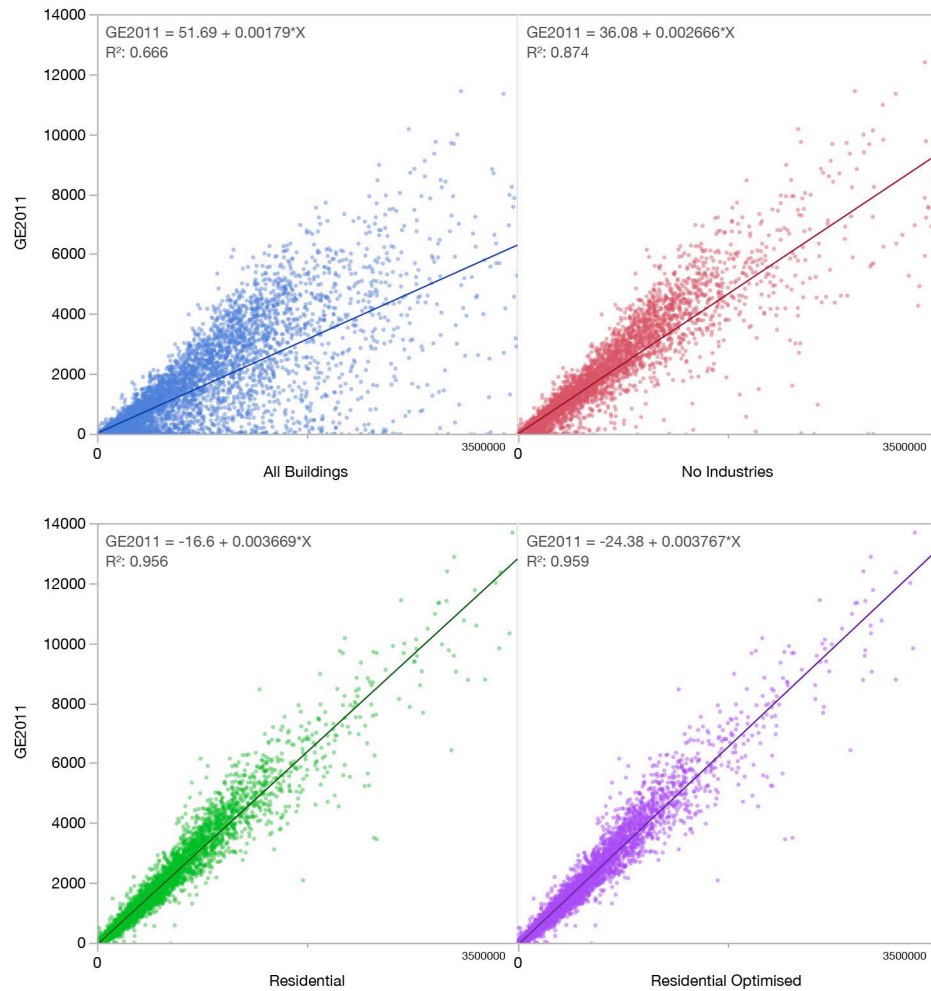
Figure 122: Scatterplot for correlations of German population data (total population) with four selections of volumetric building information (m3). The four selections are: all buildings, no buildings with industrial usage, all residential buildings and an optimised selection of residential buildings (n=10,928).
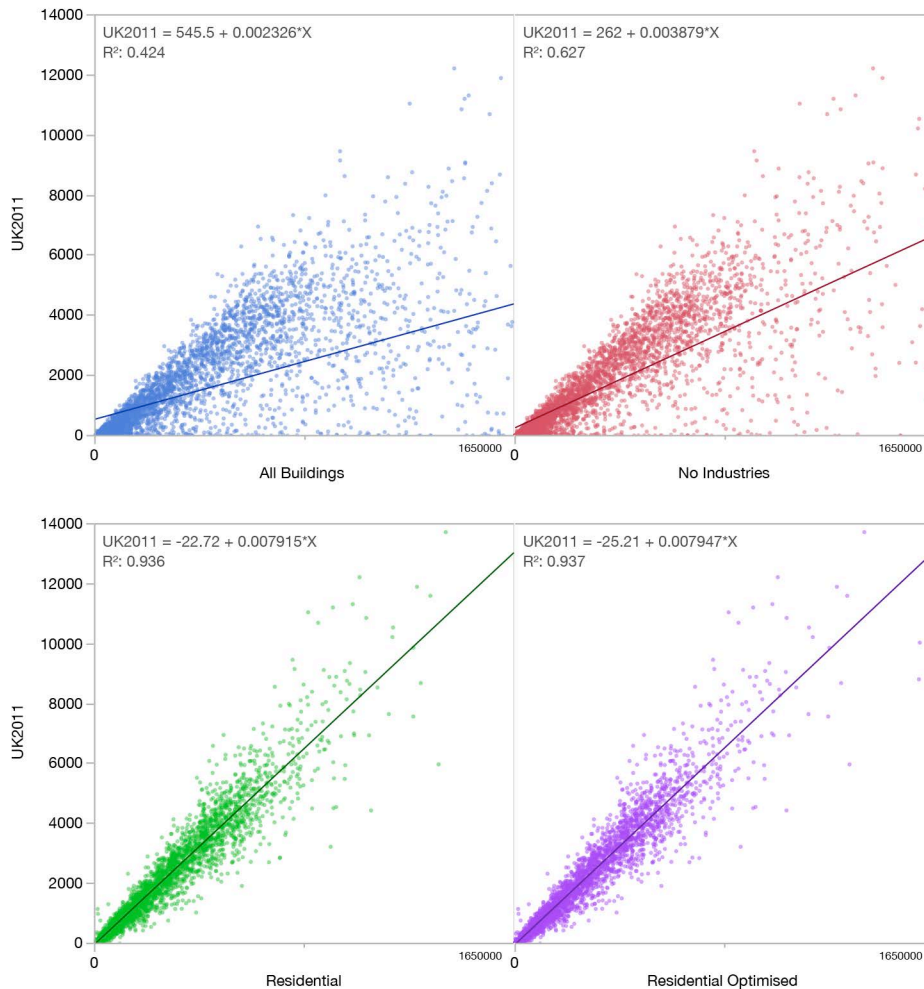
Figure 123: Scatterplot for correlations of UK population data (total population) with four selections of volumetric building information (m3). The four selections are: all buildings, no buildings with industrial usage, all residential buildings and an optimised selection of residential buildings (n=5148).

A comparison of the cases with the highest population misprediction (lower and upper band) provides further insights into the varying effect of building precise volumetric adjustments. Figure 124, shows the highest three over and under-predicted cases for both regions. The presented values (RVO and RV), are inverse residuals, i.e. the difference between observed population and estimated population. Negative values indicate estimates below the observed population, while positive values indicate an over-prediction. All twelve cases are found in urbanised areas; over-prediction, on the one hand, occurs specifically in densely urbanised areas, i.e. city centres, under-prediction, on the other hand, occurs more often in areas of medium to high density. One reason for the effect of under-prediction might be that buildings in highly urbanised areas, which are in adjacent to the city centre, have smaller flats and a lower per person volume due to the economic pressure in the rent market, which is potentially caused by higher accessibility. Over-predictions in city centres are most likely caused by wrongly

classified building volumes. The higher the degree of functional overlaps, the lower the precision of residential volumes.

	ID: 1kmN3160E4144	ID: 1kmN3150E4095	ID: 1kmN3136E4099
GE	RVO -3848.518	RVO -3733.740	RVO -3477.685
	RV -3621.166	RV -3909.245	RV -3543.595

	ID: 1kmN3131E4095	ID: 1kmN3128E4096	ID: 1kmN3128E4095
GE	RVO 5054.429	RVO 11483.761	RVO 4870.481
	RV 4913.055	RV 11560.077	RV 5040.528

	ID: 1kmN3438E3513	ID: 1kmN3470E3549	ID: 1kmN3472E3552
UK	RVO -4029.307	RVO -3790.288	RVO -4395.470
	RV -3988.917	RV -3809.400	RV -4418.988

	ID: 1kmN3373E3574	ID: 1kmN3443E3512	ID: 1kmN3443E3513
UK	RVO 4025.306	RVO 3802.848	RVO 4461.147
	RV 4152.232	RV 4105.543	RV 4717.792

RVO = Residential Volume Optimised   RV = Residential Volume

Figure 124: Mapping of over and under-predicted population. Six cases for each model are shown (GE at the top, UK at the bottom). Red negative values indicate lower prediction than observed values; blue positive values indicate higher prediction than observed values. Coloured buildings are part of the residential optimised volume selection, whereas grey buildings are all other existing buildings in the model.

A comparison of the residuals shows that for the selected outlier cases a volumetric adjustment through semantic building information improves over-prediction, but has a varying effect in cases where under-prediction occurred. However, overall volumetric adjustments with semantic building information improve the population prediction. Of 200 cases with over- or under-prediction, 79,5% cases showed improvements between predicted and actual population data over the non-volumetrically adjusted residential selection. This corroborates previous findings by Biljecki et al. (2016) and shows that the method is appropriate in cross-country comparisons. For this reason, I will employ the optimised residential volume model in all the following sections.

With regards to the overall regional distribution of residuals, the comparison of two regional models reveals the previously identified patterns. Figure 125 and Figure 126, show the predicted distribution of the total population (see Figure 20 and Figure 21 for the distribution of the observed population), the distribution of over- and under-prediction (based on residuals), as well as a map highlighting the location of cities, the major road network and the respective city name. Particularly, urbanised areas feature the largest disparities between observed and predicted values. This is the case for both the negative as well as the positive direction. In Figure 125:b and Figure 126:b, red areas indicate substantially lower predicted values, whereas over-predictions are highlighted in blue. In both cases, substantial over- and under-predictions are the case for urbanised areas and occur only in exceptional cases in rural parts of the region. At the same time, one can observe an additional pattern where the city core of large cities (Duisburg, Essen, Bochum, Dortmund (GE) and Leeds, Bradford, Sheffield (UK)) is characterised by over-predictions and the adjacent surroundings are under-predicted. This stark contrast does not appear in rural areas and might, hence, be related to an erroneous 3D-building classification or, to be precise, in the complex overlay of land-use functions in the volumetric geometries. Only building information that is precise in terms of floor plans allows us to account for such complex environments yet these details are usually not available for large-scale applications. Comparisons of residuals are particularly effective for the identification of high-value outliers. With regards to outliers that are low in value but represent a substantial percentage difference between observed and predicted value it is important to also observe the percentage difference between both.
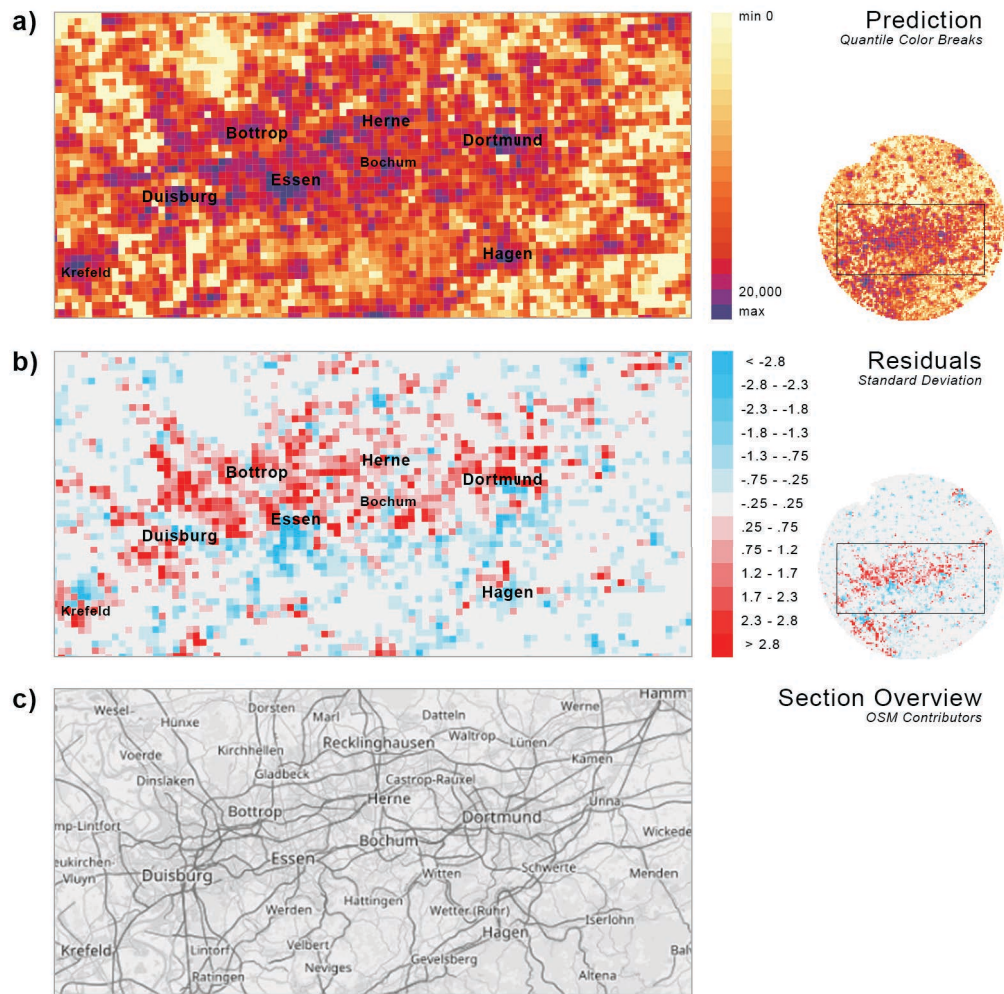
Figure 125: Visualisation of population prediction (a), difference between observed and predicted values (b), and overview map of streets, built up areas and city names (c) for the German model.
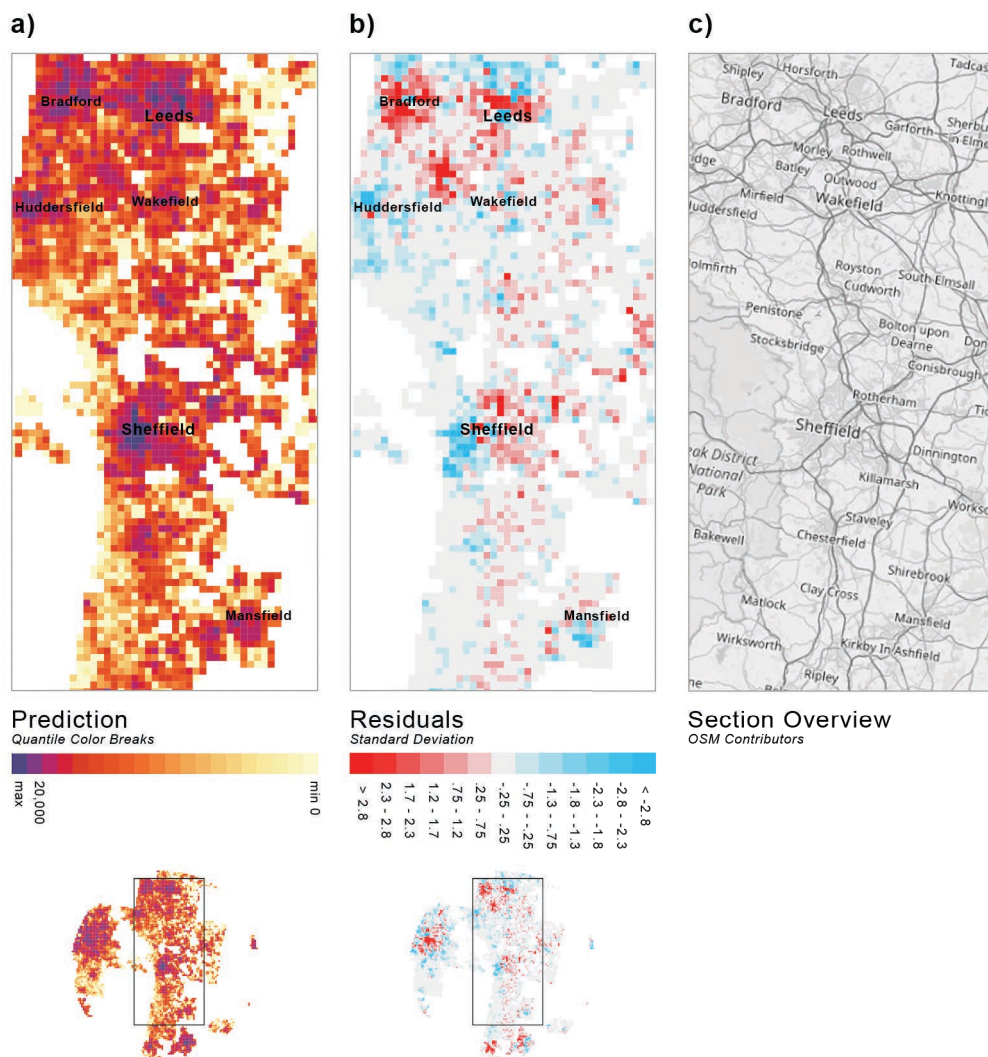
Figure 126: Visualisation of population prediction (a), difference between observed and predicted values (b), and overview map of streets, built up areas and city names (c) for the British model.

A comparison of the percentage differences highlights a reverse effect, where the percentage difference is relatively low for areas of high density (even in cases of residual outliers), while extreme outliers are present in rural and non-urbanised areas. Areas with low or small populations appear to be particularly prone to a high percentage error, which is an observation supported by previous research (ibid.; Brinegar and Popick 2010). Figure 127:a shows the spatial distribution of percentage errors over the two regions and those areas that are identified as outliers (highlighted with a black outline). The majority of errors are under-predictions due to areas with low population (in a range of 1-10) (see Figure 127:a GE: Detail Section). This is caused by the fitting process of the model function, which features a negative intercept resulting in negative predictions in extremely sparsely populated sample areas. Extremely positive percentage errors are approaching a much more random distributed across the region and are solely caused by building misclassification. Particularly industrial and agricultural buildings, e.g. gasholders, barns or other related buildings, can be wrongly classified as residential. In

areas where the observed total population is five and the actual number of residential buildings is two, a set of wrongly classified barns (usually four times the size of an average residential building) can already lead to an over-prediction of up to 800%. Few exceptions show extremely high negative percentage differences. These cases can solely be related to wrongly classified building volumes.
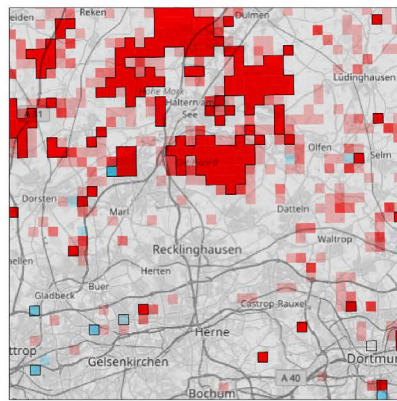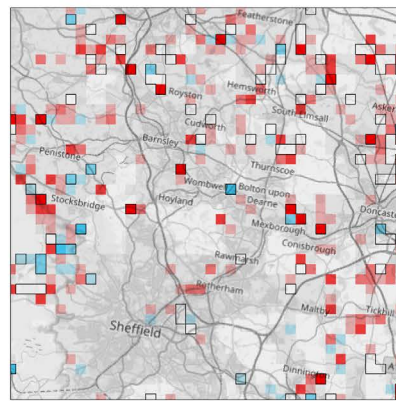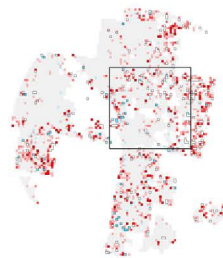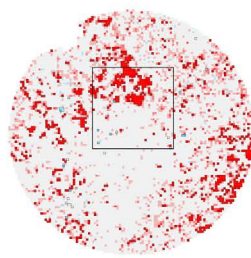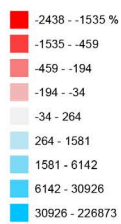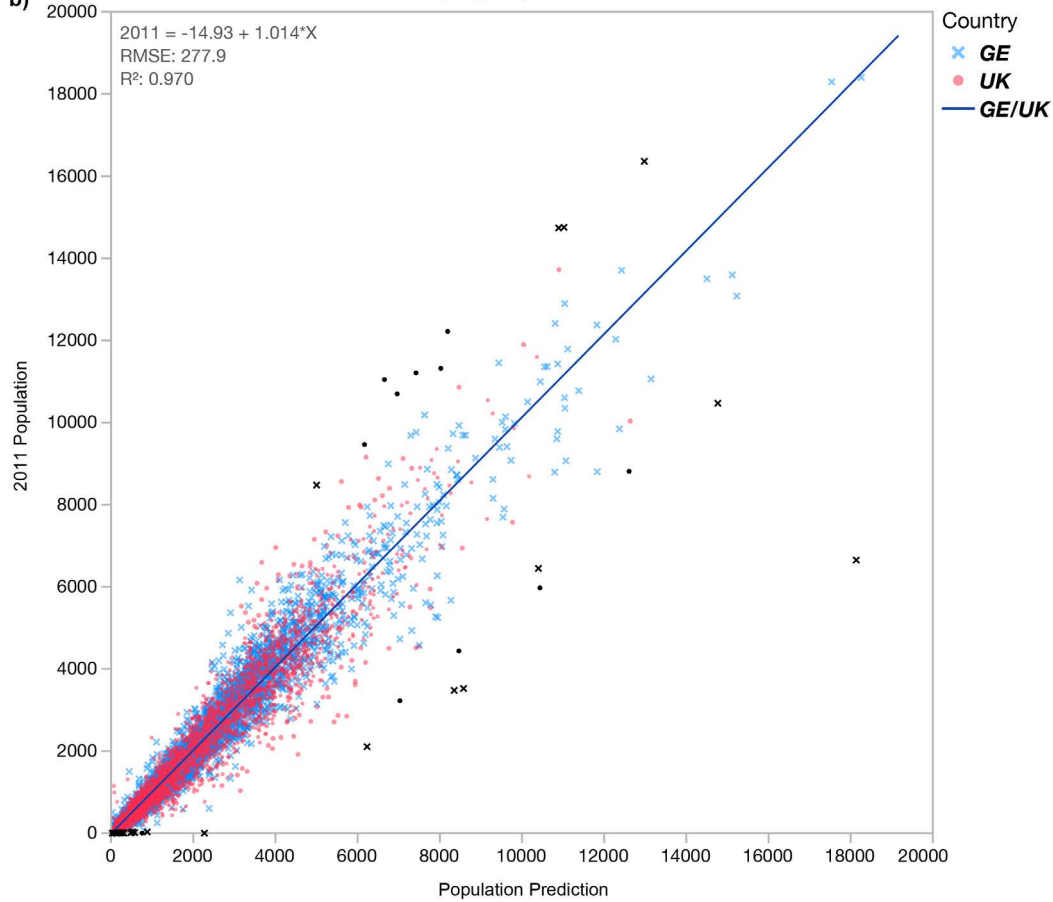
Figure 127: Spatial distribution of percentage differences between observed and predicted areas. Blue areas indicate over-prediction, red areas under-prediction (a). Actual vs. predicted plot for both models (n=16,076). Residual and percentage outlier (n=1,206) highlighted in black (b).

Both outlier behaviours, i.e. residual and percentage differences can be related to either model errors in extremely sparsely populated areas or wrongly classified building functions. Since the main aim of this predictive exploration is to test the quality of the data for the following comparison with network metrics, I will remove those sample areas from the dataset, which can be directly associated with either of the two error sources. I select residual outliers by identifying all values, which are part of the 0.05 tail quantile and surpass the interquartile range of the lower and upper quantiles by three times (20 cases). Percentage difference outliers are selected by identifying all values that are part of the 0.1 tail quantile and three times the interquartile range (1,286 cases). Other values that could be considered as outliers remain in the dataset, as it is not clear whether they are caused by misclassification or due to the differences in socio-economic characteristics of areas and buildings. Table 31, highlights the model improvement before and after the outlier adjustments.

Table 31: Differences between prediction and outlier adjusted prediction. Mean, median and standard error of overall percentage differences.

	Model	Mean	Median	Std. Err
All	GE (n=10,928)	-157.205	-0.546	21.970
	UK (n=5,148)	-97.752	-3.142	19.115
	Both (n=16,076)	-138.165	-1.323	16.141
Outlier Adjusted	GE (n=10,107)	-16.326	2.447	1.215
	UK (n=4,763)	-21.125	-0.991	2.793
	Both (n=14,870)	-17.863	1.412	1.217

The prediction accuracy has improved substantially after the removal of classification errors. The adjusted German model features a median over-prediction of 2.447% and a mean under-prediction of -16.326%, compared to the British model, where the population prediction has a median of -0. 991% and a mean that is predicted to be -17.863% lower than observed. The differences between median and mean are caused by substantial percentage difference outliers in very low populated sample areas (e.g. observed population = 3 vs. prediction = 248.685, leading to an over-prediction rate of 8189.531%). In such cases it is necessary to compare median differences rather than mean errors (Biljecki et al. 2016; Brinegar and Popick 2010). Compared to the median absolute percentage errors of the prediction results presented in Biljecki et al. (2016 p. 22), the results observed in both regions (GE and UK) are comparable to their most accurate prediction rates.

Another observation resulting from the cross-country comparison is that the model predictions for both regions are specific to the respective regional context and not transferable across different countries. This means, that the model for the German region cannot account for socio-economic characteristics present in the British region, causing substantial differences in volume per person ratios between the two areas. This becomes clearer with regards to the scatterplot and the linear relationship of the observed total population and optimised residential volumes for both models (Figure 128). In the British region, the per person habitable volume is much lower than in the German region. The German grid area 1km3129E4096 has an observed total population of 9,847 and a habitable volume of 3,290,165.487 m3 (334.129 m3 per person), a comparable British grid such as 1kmN3423E3560 has a total population of 9,864 but with 1,236,612.102 m3 (125.366 m3 per person) a 62.41% less habitable volume (Figure 128, highlighted in black).
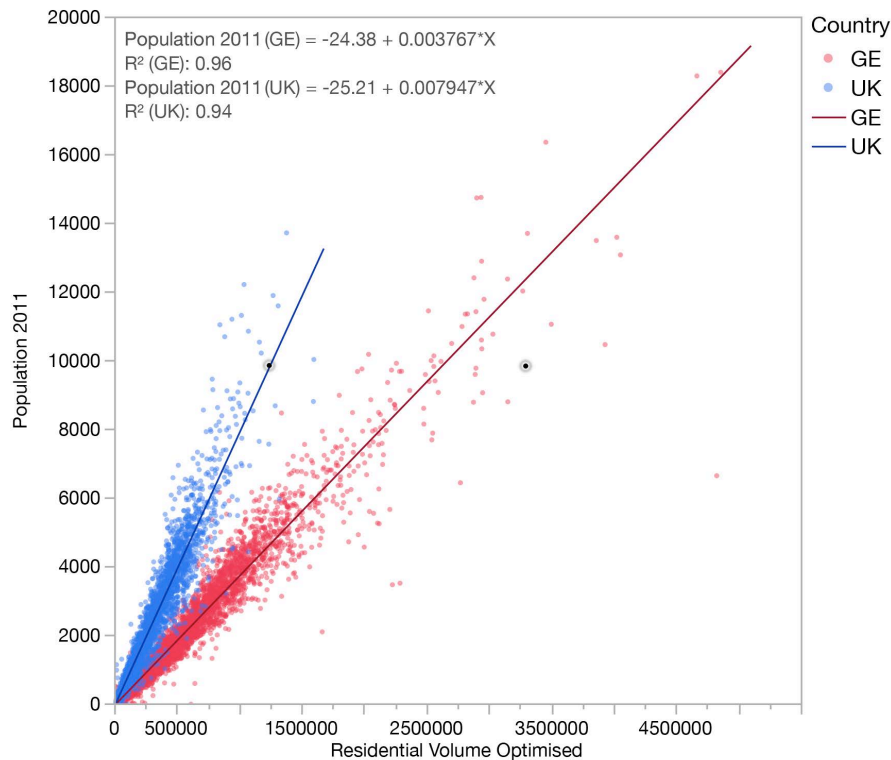
Figure 128: Scatterplot of the correlation of population data (total population) with UK and GE residential optimised volumetric information. Two comparable grid cells highlighted in black.

The tests and explorations have shown that semantically enriched 3D-building data present an effective method for the prediction of population data. However, the predictive accuracy highly depends on the quality of classified data, which is the case even if governmental datasets are employed (i.e. in the German case). Volumetric enrichment improves the prediction, but the results differ only marginally from those of non-volumetrically adjusted buildings. For the regional analysis, this means binary

classification of building functions (e.g. residential vs. non-residential) might already produce satisfactory predictions. The two predictive models cannot be applied independently from their cases. Even for two historically and socio-culturally highly comparable regions, such as the German Ruhr Valley and the British NDY region, differences of habitable volumes differ significantly (average habitable volume per person: GE = 272.469 m3, UK = 128.229 m3) rendering a single cross country modelling approach meaningless.

Nevertheless, the two models introduced feature a high predictive accuracy of 1.412 median absolute percentage error when combined. Due to this high accuracy level, I will argue that it is justified to distribute the population data to the level of individual buildings based on habitable building volumes in a disaggregation process. This will produce an appropriate total population per building for the comparison of network metrics.

8.1.3 BUILDING LEVEL POPULATION DISAGGREGATION

Following the selection of outlier-adjusted GEOSTAT grid cells, I conduct a data disaggregation procedure with the aim to calculate the total person per building (see 4.2.4, page 140 for an outline of the methodology). This calculation is done by, first, counting the total habitable volume per grid cell, second, by computing the ratio of person per habitable volume per grid cell, and third, by disaggregating the total population by a multiplication of the ratio and the respective habitable volume per building. Figure 129 shows the result of this process in the form of a histogram of the logarithmic distribution of building units and their computed population. Each building with a residential function in both regions has been assigned a population estimate with the total population resident at the very location. There are 2,629,059 buildings in the British model and 2,571,297 buildings in the German model. The German region has a bimodal distribution indicating a heterogeneous distribution of different density centres, whilst the British case has a high-peaked, clearly normal distribution. This is a characteristic that might be influenced by the differences in sample distributions, as the British model includes a smaller number of rural and agricultural areas due to the lack of available 3D-building information of the countryside in the UK.

a) UK Population per Building

b) GE Population per Building

Summary Statistics			Summary Statistics	
Mean	2.5988604		Mean	3.5135005
N	2629059		N	2571297
Minimum	0		Minimum	0
Maximum	1093.8699		Maximum	508.69623
Median	2.2015362		Median	2.4953256
Range	1093.8699		Range	508.69623
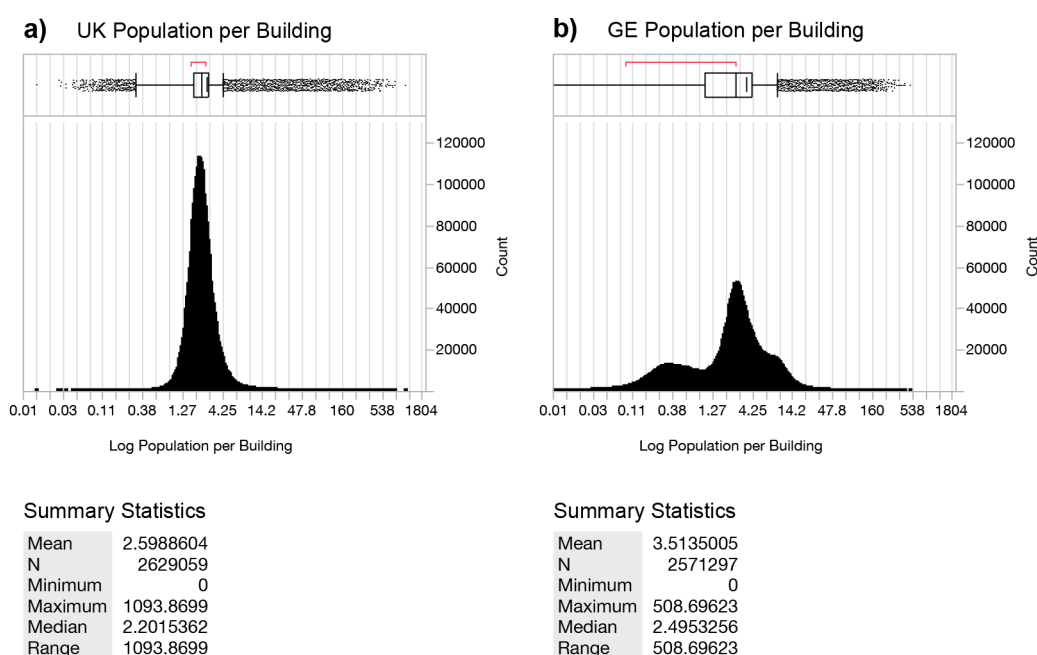
Figure 129: Histograms highlighting distribution of people per buildings after disaggregation to building level.

An exemplary visualisation of the generated dataset can be seen in Figure 130. The complex pattern of small-scale spatial occupation features patches of high and low density, with a general tendency of higher population degrees in central areas (Figure 130:c). Population density varies according to three major factors: i) the number of covered footprint area along a street, ii) the difference of building heights and iii) the difference in the overall population density in the observation areas (GEOSTAT 1 x 1km grid). These factors result in a large-scale heterogeneous pattern, individual streets, however, are often characterised by homogenous building types. This holds despite of the previously mentioned visual tendency of a pattern that follows a process of centrality (see Figure 130:c and Figure 130:d).
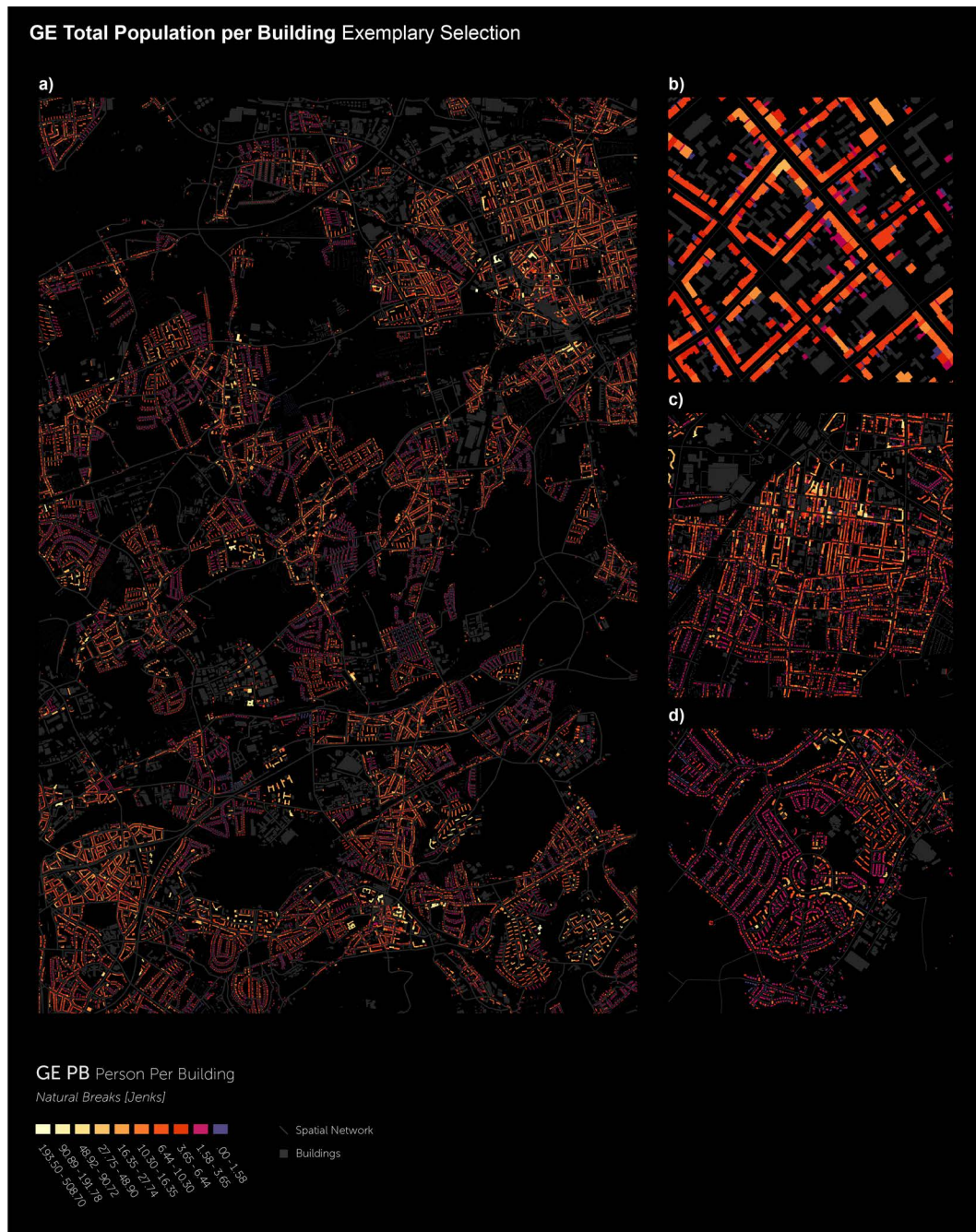
Figure 130: Total population per building for exemplary selected areas of the German model in different scales.

8.1.4 STREET LEVEL POPULATION AGGREGATION

Street-level aggregation of building-based information is ideally performed via a match of single streets with a building-based on building entrance information. Such building entrance information is not available for regional or countrywide datasets. In such cases alternative analysis approaches are necessary. I will briefly introduce two of these approaches and explain why a combinatory approach results in satisfactory data aggregation. Data aggregation on the street level can be achieved either via a near analysis or a buffer zone analysis (Figure 131).

Near Analysis   *Buffer Zone Analysis*
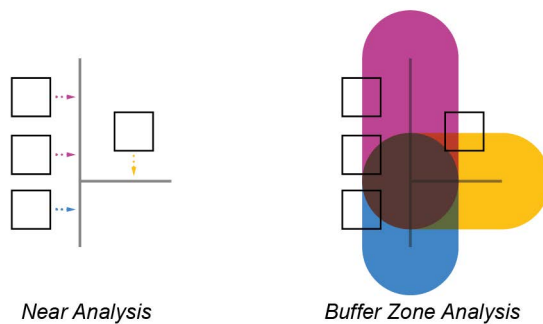
Figure 131: Visualisation of near and buffer zone analysis.

NEAR ANALYSIS calculates the distance of an input feature with the closest feature of a target dataset. In the application of building- to street-level aggregation, this entails calculating the closest street for each building. Such proximity information can then be used to create distance thresholds to exclude buildings that are not connected to the street network, i.e. connected through long private access routes that are not part of the public rights of way system. A disadvantage of near analysis approaches is that each building can only be related to one, single street. Buildings in city centres, however, have the size of entire blocks surrounded by sometimes four streets and often feature multiple entries. Moreover, buildings located at T-intersections have also got multiple spatial relationships and an unclear segment allocation. In these cases, a single street relationship fails to account for such morphologies and results in over-estimations of individual streets.

BUFFER ZONE ANALYSIS uses a previously defined search zone of equal distance from all sides of an input feature and selects all target features that fall within this search zone. Buffer zones for line features can either be generated in a circular manner around the line end or as a flat end with a cut-off at the endpoint of the line feature. The former is used in this analysis to only account for buildings that are along either side of a street as well as at the end of cul-de-sacs and T-intersections. This approach is expected to account for buildings with potential multiple entrances as well as unclear allocations at T-intersections. Since values of buildings, which fall into more than one buffer zone, are aggregated without taking the multi-relationship into account, it is necessary to pre-process and divide the values by the number of incidents where the buffer zones overlap prior to the aggregation process. A disadvantage of the buffer zone approach is the need to specify a distance buffer, rather than to explore the data distribution through a distance threshold analysis, as is the case for the near analysis. A single buffer distance approach can lead to substantial misallocations. Buildings that are in morphological settings where the distances from the street centre are high are not taken into consideration when the buffer distance is below the threshold. At the same time,

314

buildings that are in dense small-scale environments can fall into multiple wrongly allocated buffers if the distance is too high.

In order to overcome both issues I propose to employ a near analysis to identify the maximum distance for first order matches from all residential buildings to the respective closest street segment to generate individual distances for a dynamic buffer generation, where the distance of each buffer is generated based on the individual distance. The cut-off distance for this allocation is 50 metres, as buildings that are not in the proximity of 50 metres to a street segment can be considered as unrelated. All streets, which have not received a first order match, are then linked to the closest building with the average second order distance. The resulting distance value can then be used for the generation of a dynamic buffer zone (Figure 132). Values of buildings falling into multiple buffers are divided by the total number of buffer intersection, guaranteeing a proportional distribution of values on all surrounding segments of the spatial network. Buildings will always be joined to, at least, the closest streets, while the overlap effect at intersections is proportionally distributed depending on the number of intersections. This prevents an entirely wrong allocation of buildings and proportionally minimises the negative effect of wrong allocations.
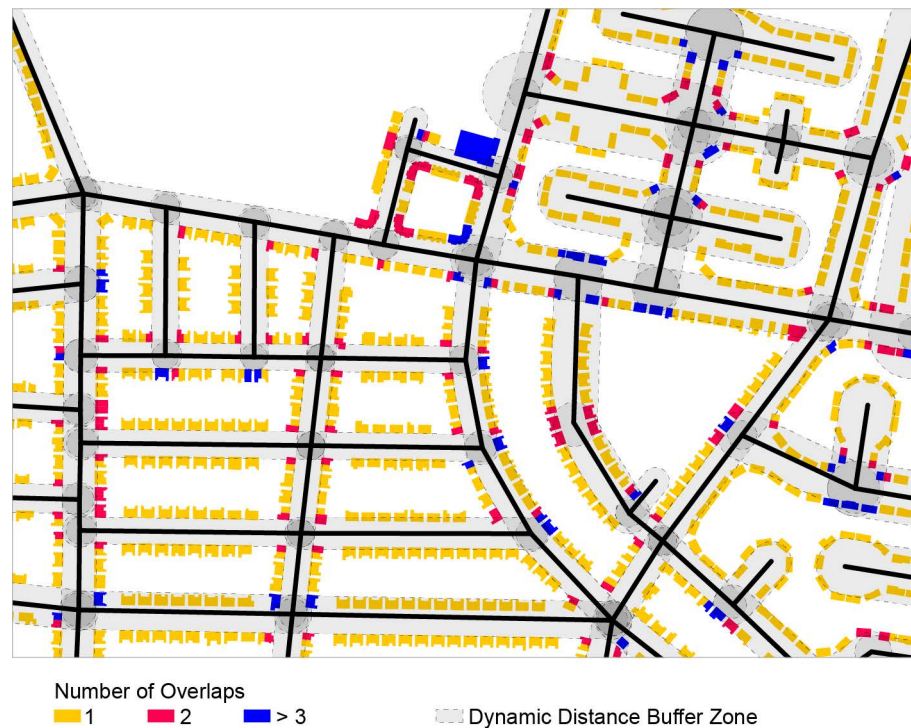
Figure 132: Example of proposed dynamic distance buffer zones approach and number of resulting building overlaps.

The result of the approach described above is a value for the total population per street segment. The total population per segment is highly influenced by the segment length because longer street segments have a higher probability of having a larger number of buildings adjacent to them. Due to this relationship, the total number of people per street needs to be adjusted by the street length. The following comparisons are based on total population per street segment length (*ps*) (8.1), where *pb* equals the total population per building and *sl* is segment length.

$$ps = \sum \frac{pb}{sl}$$

(8.1)

Figure 133 shows a visual representation of the generated data for the German region; each street holds a *ps* value and is coloured using a natural breaks colour gradient. This is the first time a dataset of this precision and extent has been produced. From the overall regional map (Figure 133:a) a clear pattern of agglomerations of high-value streets around the main cities (Duisburg, Dortmund, Essen, Krefeld, Herne and Bottrop) is visible, but once can also make out distinctive smaller patches in the North of Duisburg. This pattern is in both regions comparable to the observed population of the GEOSTAT 1 x 1km grid. On a larger scale, this pattern is characterised by centre-to edge distance decay. However, if observed on a smaller scale, a more distinctive pattern appears, where city centres feature low populations at their very core and high population densities in their immediate adjacency (Figure 133:b). This highlights the complex relationship between relative centrality, service and trade, as well as residential functions. The relationship between closeness centrality and residential spaces must be polynomial to a higher degree, where population per street segment growths with increasing centrality but reaches a point where the population degree justifies the establishment of service and trade functions competing with residential functions for space and hence leads to a population decrease in areas of maximum centrality.

Figure 133: Total population per street segment length for the German Ruhr Valley and the city of Krefeld in different scales.

However, correlation tests have shown no significant relationship between *ps* and any spatial metric. Neither closeness, nor betweenness centrality factors have shown to have any relationship of a linear or polynomial kind. No relationship could also be found between *ps* and any of the initial analytical radii (49 closeness centrality and 49 betweenness centrality measurements). Figure 134, shows an exemplary scatterplot of cubic correlations between *ps* of the German model and EFA CC factors. There are a series of potential reason for this lack of any observable relationships, such as a higher degree of variation in building types, which is disproportional to the relative centrality.

An example would be residential dwelling projects, e.g. social housing estates, whose network location is not only the result of a socio-economic process, but is more strongly influenced by political decision-making. Moreover, housing projects are increasingly becoming a part of economic speculation, in which a location is chosen based on an assumption on the future location within the overall city or region. Ultimately, centrality might only be important for the respective neighbourhood or area in general, rather than the individual street level. This means occupational decision-making processes are more strongly influenced by the relative centrality of an area, rather than the individual street. This would imply that such relationships could be observed on a larger aggregation level.
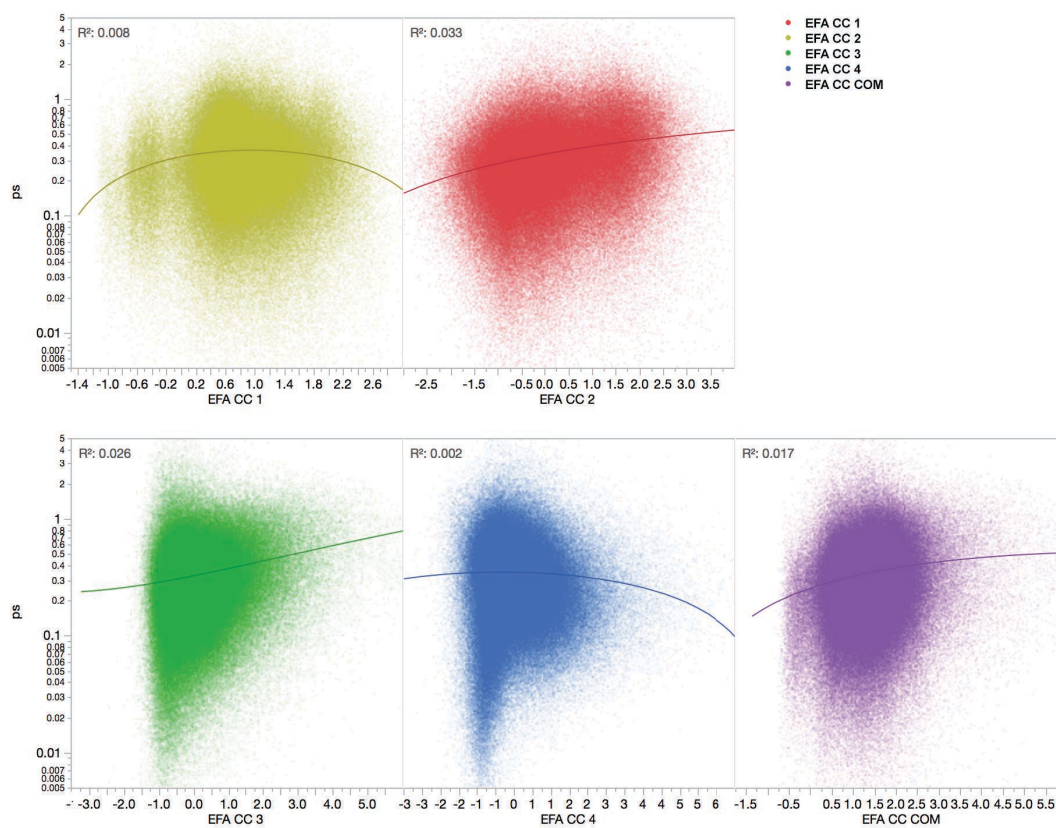
Figure 134: Scatterplots of Log Population per Street Segment and EFA CC Factors.

8.1.5 HIERARCHICAL SAMPLING MODEL

Hierarchical modelling is a method to test the relative importance of these kinds of neighbourhood effects described (see 4.2.2, page 133 for the used methodology). The approach is defined by making use of differently sized sample grids that can be consistently compared with each other through their hierarchical relationship. As previously discussed (Figure 30), I make use of three different sample grid sizes (i.e. 250 metres, 500 metres and 1,000 metres). In a sequential comparison, the total population for each sample grid is calculated and the number of streets that intersect with a sample

grid is aggregated. This is measured by summarizing the disaggregated population per building that falls into a grid. Figure 135, exemplifies the sample effect on the data resolution and the differences in the respective pattern for the total population and values of the EFA CC2 factor.
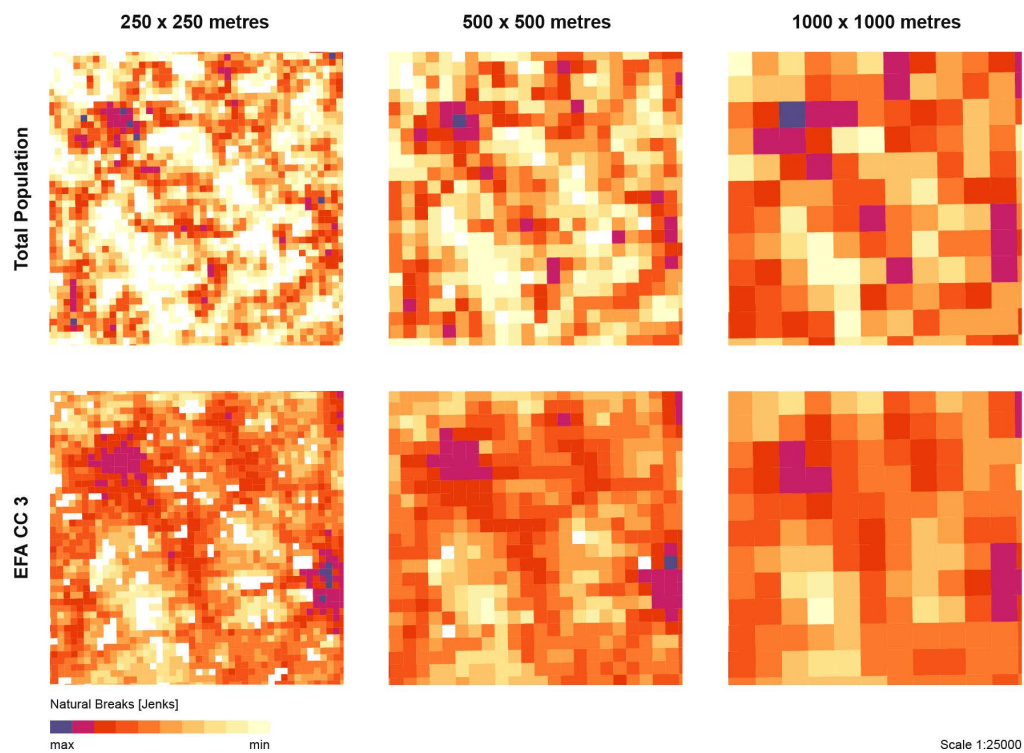
Figure 135: Hierarchical modelling, comparison of total population at the three different sample levels and EFA CC2.

Table 32, shows the results of this analysis. Each row indicates the R-squared value of correlations between the four latent centrality structures for closeness centrality and the total population per sample grid. The first and second factors show no relevant relationships. This result is in line with the observation made in the morphological comparison of the visualised factors in Chapter 7, where only the third and fourth factors contained meaningful structures for urbanised areas. This is the case for both regions and across all three sample grid sizes. A relationship can be reported for factors three and four (EFA CC 4 and EFA CC 3) with the highest correlation coefficient existing between the observed population and EFA CC 3 (r2 = 0.438 for UK and r2 = 0.562 for GE). There is a clear increase of the linear relationship with an increase in grid size, observable in both regions, e.g. for the UK model at 250 metres (r2 = 0.152), at 500 metres (r2 = 0.268) and at 1000 metres (r2 = 0.438). Overall, closeness centrality can explain only very little of the variance in population estimates.

Table 32: R-squared of latent centrality structures of closeness centrality and total population at different sample scales for both model areas.

	Grid Size	N	EFA CC4	EFA CC3	EFA CC2	EFA CC1
UK	250	50,360	0.383	0.152	0.023	0.027
UK	500	16,349	0.432	0.268	0.035	0.024
UK	1000	4,697	0.413	0.438	0.059	0.020
GE	250	161,712	0.071	0.242	0.033	0.029
GE	500 x	40,428	0.171	0.390	0.050	0.009
GE	1000 x	10,107	0.211	0.562	0.096	0.005

Comparing the results of closeness centrality to those of betweenness centrality, the latter proves to be a better estimator for population aggregation (Table 33). Again, small-scale latent centrality structures (i.e. EFA BC 4 and EFA BC 5) have a linear relationship with population data, whereas only a weak (EFA BC 3) and no relationship can be reported for intra- and inter-regional latent centrality structures (EFA BC 2 and EFA BC 1). The latent centrality structure that performs the best is EFA BC 4 on a 1 x 1 kilometres grid sample with an R-squared of 0.713 for the UK and a R-squared of 0.645 for GE. Similar to closeness centrality the correlation coefficient increases with increasing sample grid size regions, e.g. for the UK model at 250 metres ($r2 = 0.313$), at 500 metres ($r2 = 0.555$) and at 1000 metres ($r2 = 0.713$). The fifth latent centrality structure for betweenness centrality (EFA BC 5) maintains a relatively similar result across all grid sizes. Similar to CC, the correlations with EFA BC 5 show consistent correlation coefficients with only small deviations of up to 0.06 across all sample grid sizes.

Table 33: R-squared of latent centrality structures of betweenness centrality and total population at different sample scales for both model areas.

	Grid Size	N	EFA BC5	EFA BC4	EFA BC3	EFA BC2	EFA BC1
UK	250	50,360	0.547	0.313	0.040	0.000	0.008
UK	500	16,349	0.584	0.555	0.120	0.009	0.003
UK	1000	4,697	0.567	0.713	0.271	0.038	0.001
GE	250 x	161,712	0.419	0.391	0.091	0.006	0.000
GE	500 x	40,428	0.480	0.551	0.182	0.016	0.000
GE	1000 x	10,107	0.429	0.645	0.325	0.038	0.000

Only betweenness centrality on a sample grid of 1 x 1 kilometres constitutes an option for population estimation on a regional scale. However, if one compares these findings with correlations between network properties, i.e. the number of segments (or

theoretically spaces) and the total segment length per sample grid, it becomes clear that these properties alone can sufficiently account for the total population present at a given location. Table 34, exhibits the results of a comparison between network properties and total population. A strong relationship can be reported for the UK model with a maximum of 0.848 R-squared for total segment length and 0.833 for segment count. The results for the GE model show a weaker relationship of 0.674 total segment length and 0.683 segment count. The difference between both models might be caused by the fact that the German model incorporates many more rural and suburban areas than the British model, due to the differences in the availability of 3D-building information.

Table 34: R-squared of network properties and total population at different sample scales for both model areas.

	Grid Size	N	Total Segment Length	Segment Count
UK	250	50,360	0.355	0.644
	500	16,349	0.716	0.768
	1000	4,697	0.848	0.833
GE	250	161,712	0.283	0.512
	500	40,428	0.517	0.616
	1000	10,107	0.674	0.683

It becomes evident that the latent centrality structures found do not hold explanatory power of residential spatial occupation that goes beyond those of simple network properties. The differences between the UK and GE model, point to a substantial increase of explanatory power when the rural and agricultural areas are excluded prior to the analysis. If this is done, simple network properties in the form of the total segment length or total segment count can account for 85% of the variance in the data. This result shows that spatial networks can form a very cost-efficient alternative for locational population estimation in cases where no other information is available.

8.1.6 CENTRES HIERARCHY IDENTIFICATION THROUGH NETWORK CENTRALITIES

Finally, the last enquiry of this thesis focuses on the relationship of network centralities and the spatial location of commercial activity, i.e. service and trade functions (see 4.2.5, page 144 for the methodology employed). Hillier's notion of the movement economy lets us expect functions of service and trade at those locations, which hold a higher potential for random encounters. Here especially, a foreground network of linked centres reflects the process of modulating space in such a way that busier and quieter areas emerge (in previous studies operationalised as radius n). This network of busier spaces is expected to intersect with commercial functions as it has been shown in the context of the city.

The following section presents a comprehensive analysis of the intersection of the locations of commercial activity and the latent centrality structures identified. The analysis highlights that centres are formed not in relation to one particular latent centrality structure, but to a more complex combination of different latent centrality structures.
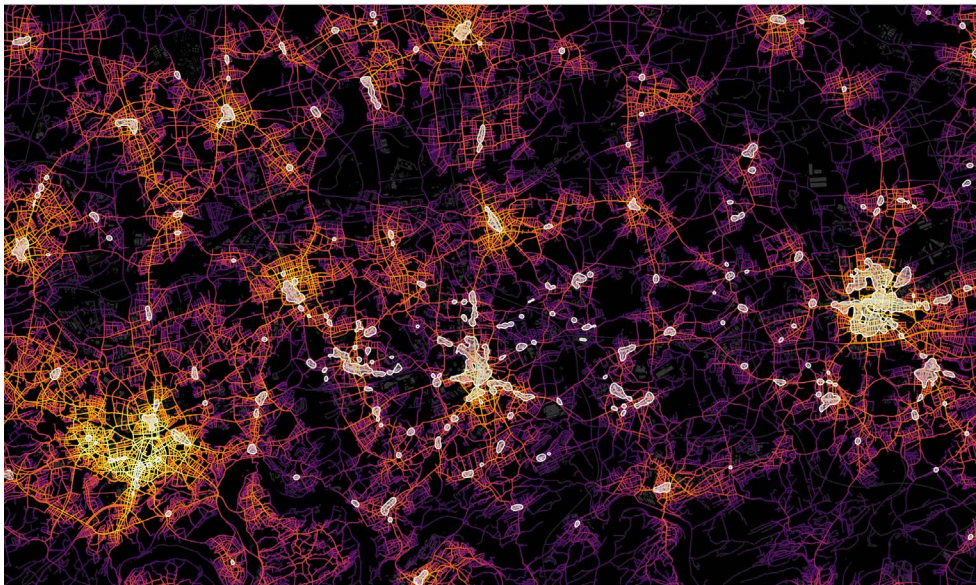
Figure 136: Combined latent centrality model for betweenness centrality (a), and the third latent centrality factor for closeness centrality (b) superimposed on the location of identified agglomerations of commercial activity for the German model.

Figure 136 shows a superposition of the previously identified centres (see 4.2.5, page 144) with the visualisation method of the combined latent centrality model for betweenness centrality introduced in Chapter 7, as well as a visualisation of the third

latent centrality factor for closeness centrality. In both maps, a) and b), we can see the visual relationship between the extracted latent centrality factors and commercial activity. Three fundamental types of relationships between latent centrality factors and commercial activity can be identified: i) commercial activity located immediately on a latent centrality structure, ii) commercial activity in proximity to a latent centrality structure and iii) commercial activity at the beginning of a latent centrality structure (Figure 137).
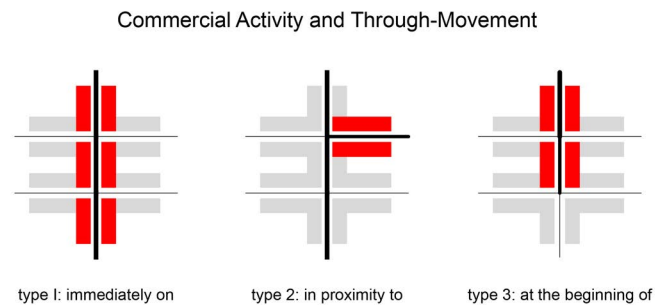
Figure 137: Three relationship types of commercial activity and latent centrality structure for betweenness centrality i.e. through movement.

The magnitude of the relationship between commercial activity and latent centralities for betweenness centrality can also be statistically demonstrated. For this purpose, the maximum value of those streets, which are in direct vicinity to a commercial activity, is spatially joined at the boundary of each agglomeration. Of 1,060 agglomerations of commercial activity identified, all but 60 clusters are located at a segment that is part of at least one of the 5 latent centrality clusters identified (i.e. have an EFA value above 1). This means that 94.3% of all centres are at locations that are part of the latent centrality structure. This is of particular relevance because all EFA BC factors together make up only 10 percent of the entire spatial network. Evidently, there is a strong relationship between through-movement and commercial activity on a regional extent. A deeper insight into this relationship can be gained by comparing at which latent centrality structure, and hence which particular scale, a centre is located. By comparing the pattern of values of each of the 1,060 centres for each latent centrality structure, one can infer which of the scales is supported most by the respective centre.

Such a comparison is made by means of a parallel coordinate plot, which is a common way of visualizing high-dimensional or multivariate data. Figure 138:a shows such a parallel plot for each centre as well as its respective loading on each of the latent centrality structures. The plot's y-axis is scaled relative to the minimum and maximum of each latent centrality factor. Each line represents a commercial agglomeration. The plot allows us to identify general tendencies and groups within the data. If centres share

a common relationship with latent centrality structures, then this is reflected in concurrent line developments. An initial observation made from Figure 138:a is that very few centres feature extremely high values, while the majority of centres have a high value on EFA BC 4 and almost all feature a relatively low value on EFA BC 1. At the same time, the distribution of very high values is similar across all EFA BC factors.
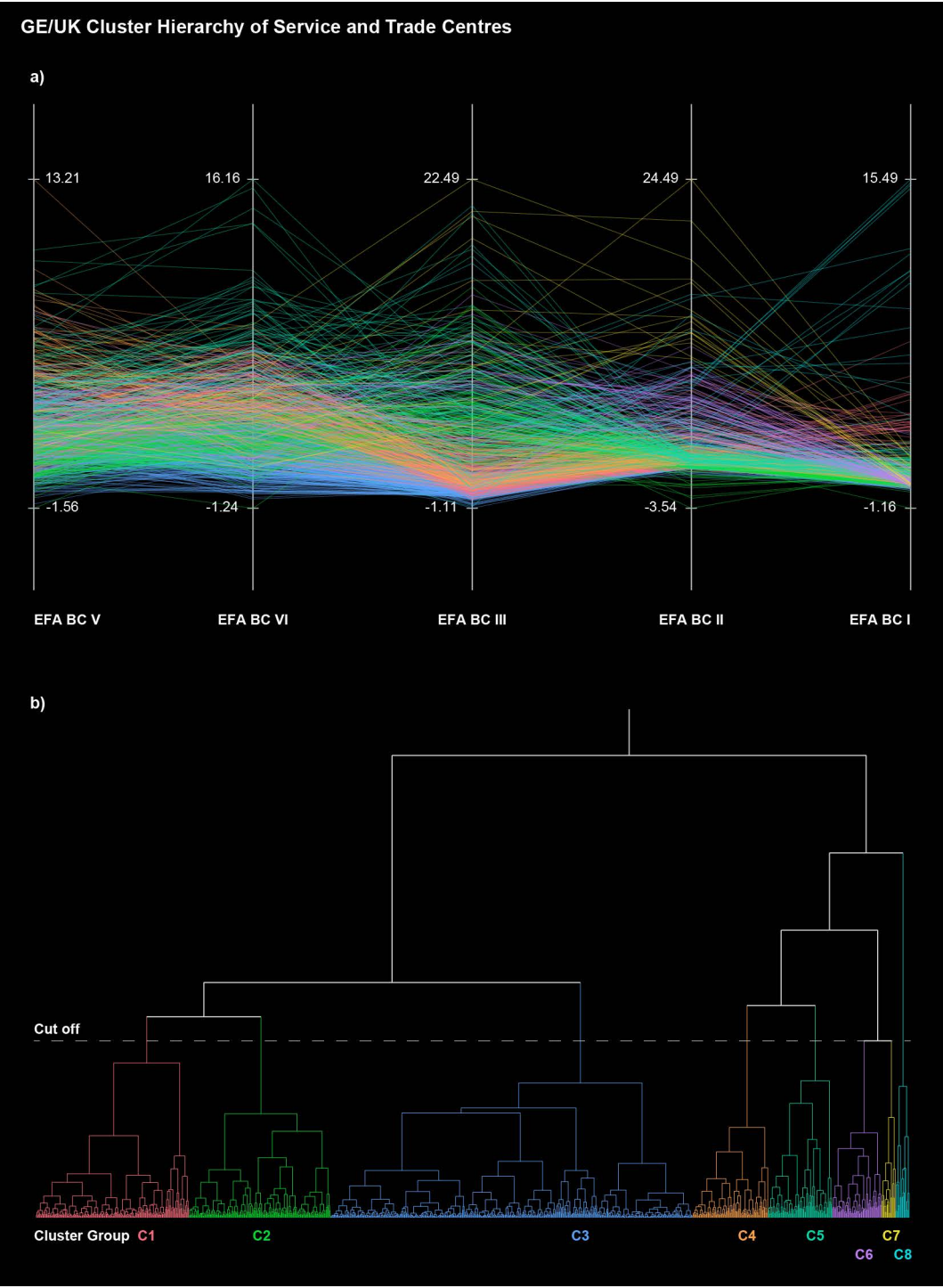
Figure 138: Parallel plot for all service and trade agglomerations and their maximum value of each latent centrality (a). Dendrogram showing the results of the hierarchical cluster analysis.

To investigate whether there are particular trends or groups within the dataset, I employ a cluster hierarchy analysis, which is a multivariate statistical technique that groups together those observations that share similar values across a number of variables. The result of this can be seen in the dendogram (tree diagram) in Figure 138:b. The dendogram allows us to identify eight distinctive clusters. It also shows how these groups relate to each other hierarchically: Clusters C1, C2 and C3 are part of a potential larger group, whereas cluster C4, C5, C6, C7 and C8 could constitute another larger group. Each group contains centres, which share commonalities between the patterns of high and low values on each EFA factor. If for example centre 1 has a high value on EFA BC4 and EFA BC3 and low values on EFA BC5, BC2 and BC1 and centre 2 features a similar distribution of high and low values, then both centres are allocated to cluster 5. Table 35, shows the properties of the eight clusters identified, as well as the number of centres that each cluster constitutes, along with the mean value for each EFA BC of all centres within a cluster. A first observation made from these cluster divisions is that there is a scalar relationship between the number of centres and their respective clusters. The largest cluster has twice as many centres (440) as the second and third largest cluster (186, 172), whereas the third largest cluster has twice as many centres than the fourth, fifth and sixth cluster (92, 78 and 60). The smallest group (cluster 8), loads the highest on the large-scale latent centrality structure (EFA BC 1), while the largest cluster loads the highest on small-scale latent centrality structures (EFA BC 5 and EFA BC 4). We can already infer from this table that the pattern of relationships is much more complex than the CPT hierarchy proposed would expect.

Table 35: Cluster hierarchy and mean value of each latent centrality factor of each cluster.

Cluster	Count	EFA BC5	EFA BC4	EFA BC3	EFA BC2	EFA BC1
1	186	1.695	**4.733**	1.447	0.818	*0.790*
2	172	1.145	2.755	**5.780**	*0.205*	0.407
3	440	1.459	**1.818**	0.676	0.666	*0.089*
4	92	**5.299**	4.315	1.761	1.275	*0.484*
5	78	3.849	**7.571**	**7.911**	1.111	*0.869*
6	60	2.578	4.122	**6.730**	5.441	*0.377*
7	17	2.938	4.529	**11.770**	13.655	*0.627*
8	17	2.261	4.233	4.983	7.609	**10.027**

The following eight figures (Figure 139–Figure 146) present each of the clusters. Each figure contains a parallel plot showing only those centres, which are part of the respective cluster group. I then conduct a visual analysis of morphological commonalities between centres within each cluster, based on a random selection of four centres for each group.
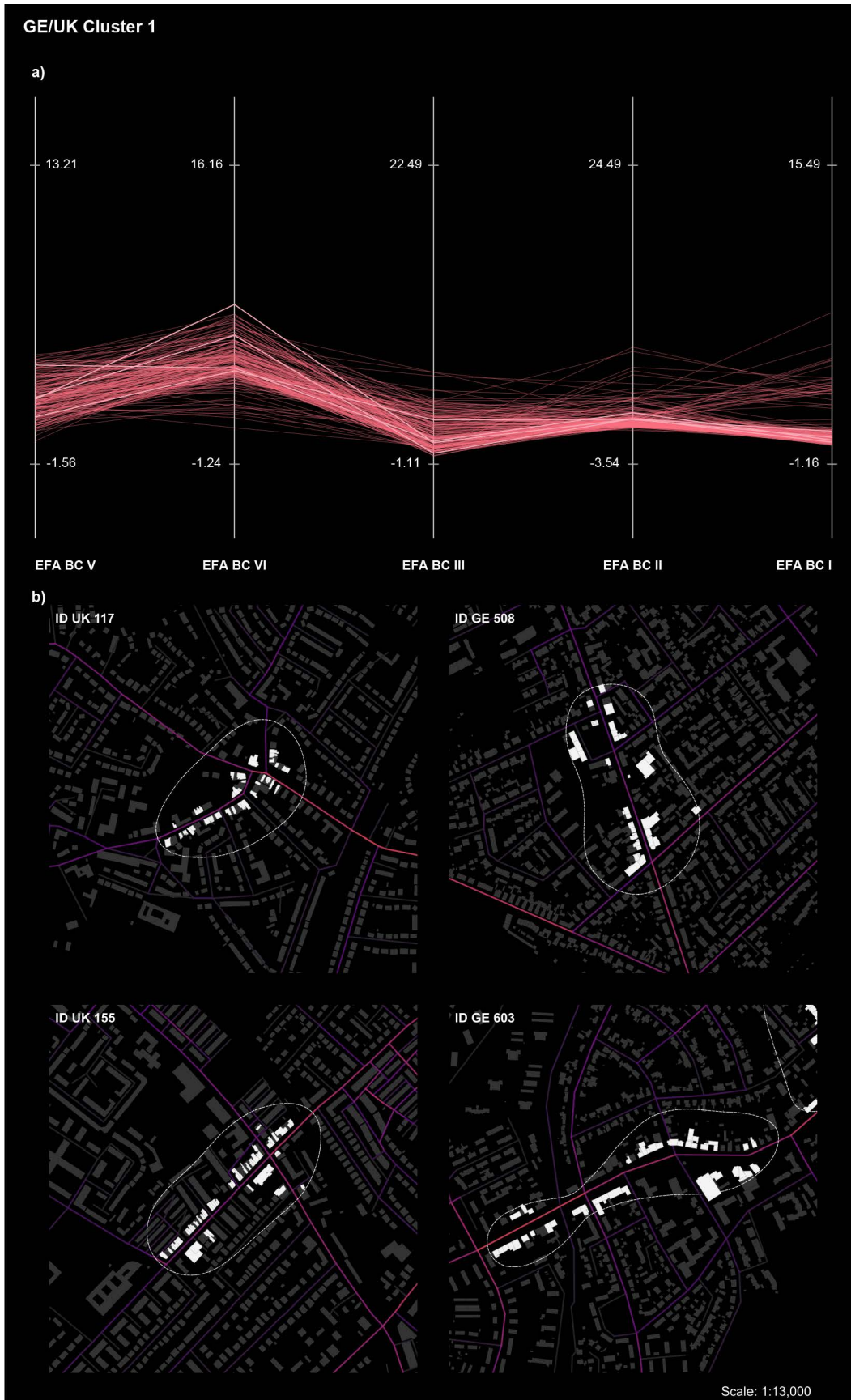
Figure 139: Parallel plot of the first cluster of service and trade agglomerations and its maximum value of each latent centrality (a). Mapping of commercial activity and the combined EFA BC model (b).

GE/UK Cluster 2

a)

EFA BC V	EFA BC VI	EFA BC III	EFA BC II	EFA BC I
13.21	16.16	22.49	24.49	15.49
-1.56	-1.24	-1.11	-3.54	-1.16

b)

ID UK 272

ID GE 155

ID UK 196

ID GE 468

Scale: 1:19,000
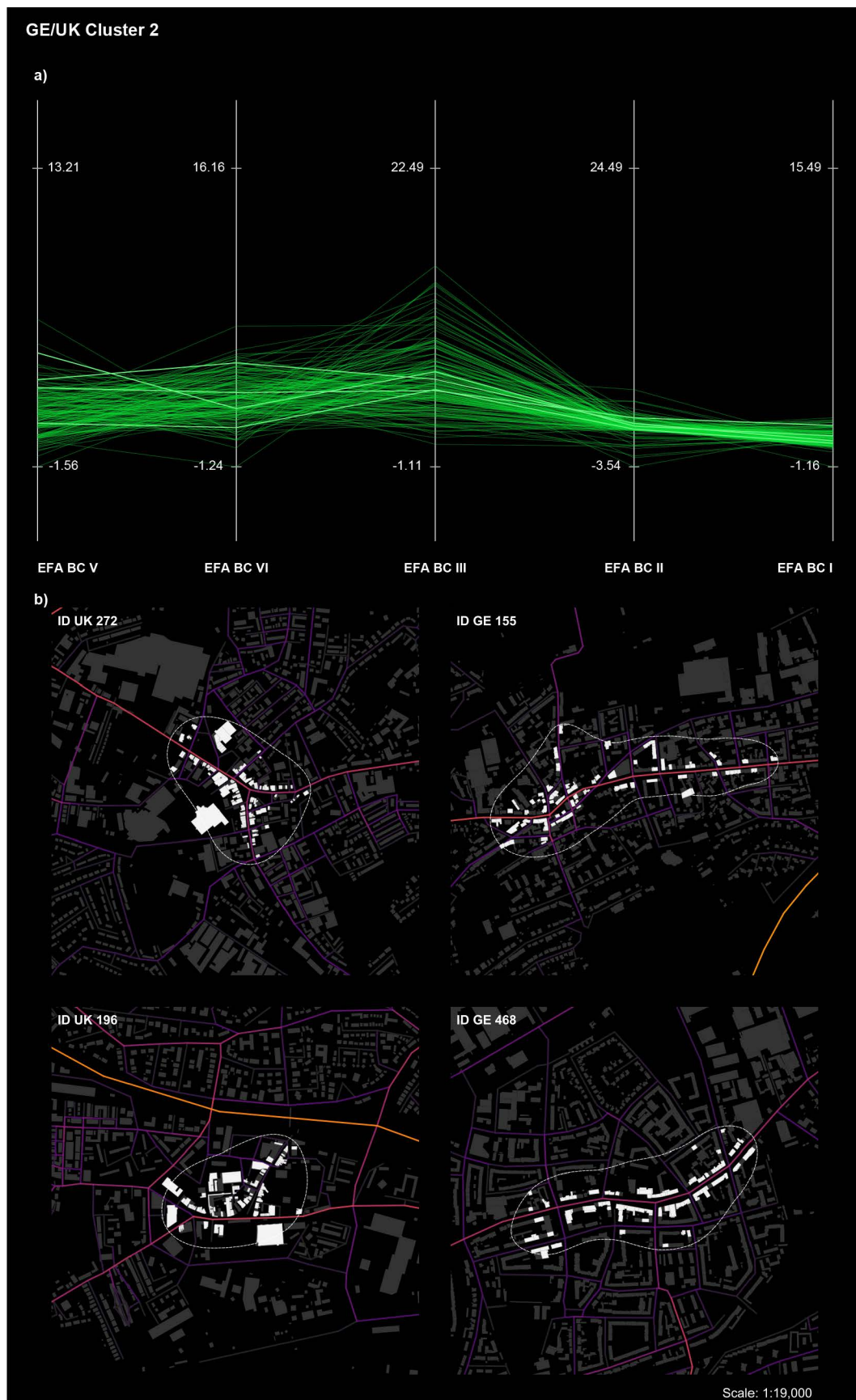
Figure 140: Parallel plot of the second cluster of service and trade agglomerations and its maximum value of each latent centrality (a). Mapping of commercial activity and the combined EFA BC model (b).

GE/UK Cluster 3

a)

| 13.21 | 16.16 | 22.49 | 24.49 | 15.49 |
| -1.56 | -1.24 | -1.11 | -3.54 | -1.16 |

EFA BC V   EFA BC VI   EFA BC III   EFA BC II   EFA BC I

b)

ID UK 138   ID GE 623

ID UK 238   ID GE 656

Scale: 1:8,000
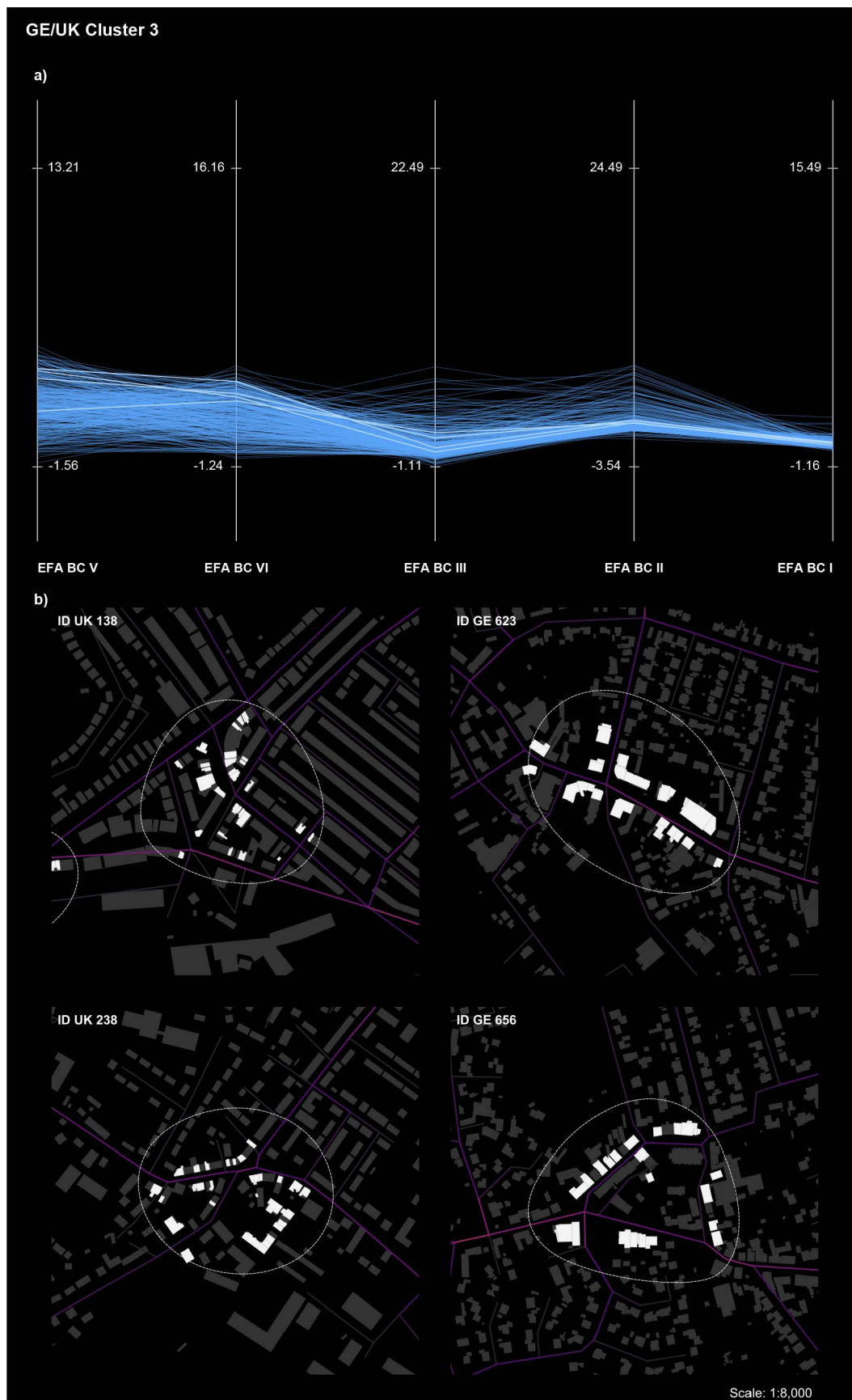
Figure 141: Parallel plot of the third cluster of service and trade agglomerations and its maximum value of each latent centrality (a). Mapping of commercial activity and the combined EFA BC model (b).

GE/UK Cluster 4

a)

13.21	16.16	22.49	24.49	15.49
-1.56	-1.24	-1.11	-3.54	-1.16
EFA BC V	EFA BC VI	EFA BC III	EFA BC II	EFA BC I

b)

ID UK 276   ID GE 2

ID UK 268   ID GE 100

Scale: 1:20,000
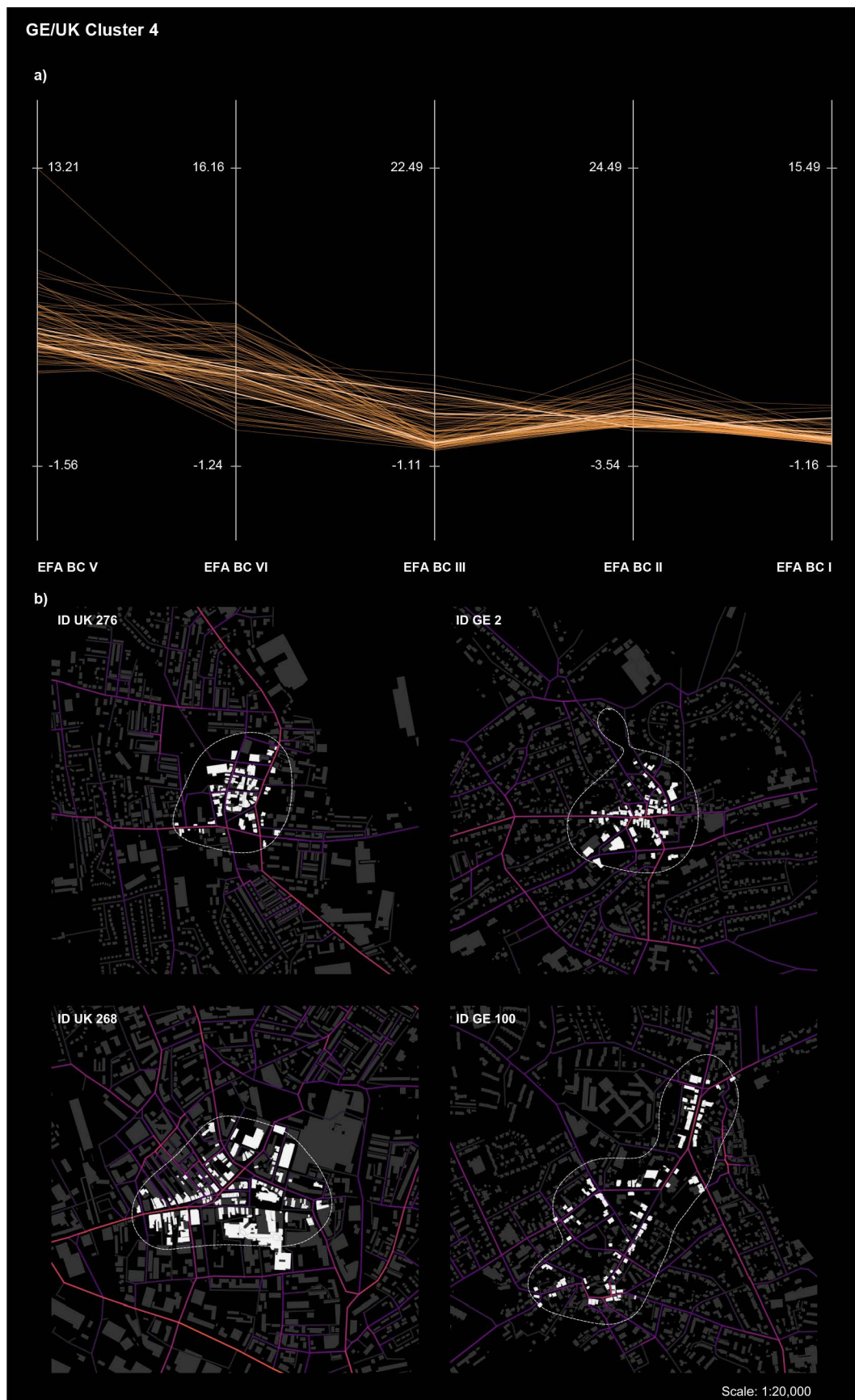
Figure 142: Parallel plot of the fourth cluster of service and trade agglomerations and its maximum value of each latent centrality (a). Mapping of commercial activity and the combined EFA BC model (b).

Figure 143: Parallel plot of the fifth cluster of service and trade agglomerations and its maximum value of each latent centrality (a). Mapping of commercial activity and the combined EFA BC model (b).

GE/UK Cluster 6

a)

EFA BC V	EFA BC VI	EFA BC III	EFA BC II	EFA BC I
13.21	16.16	22.49	24.49	15.49
-1.56	-1.24	-1.11	-3.54	-1.16

b)

ID UK 270

ID GE 97

ID UK 222

ID GE 208

Scale: 1:29,000
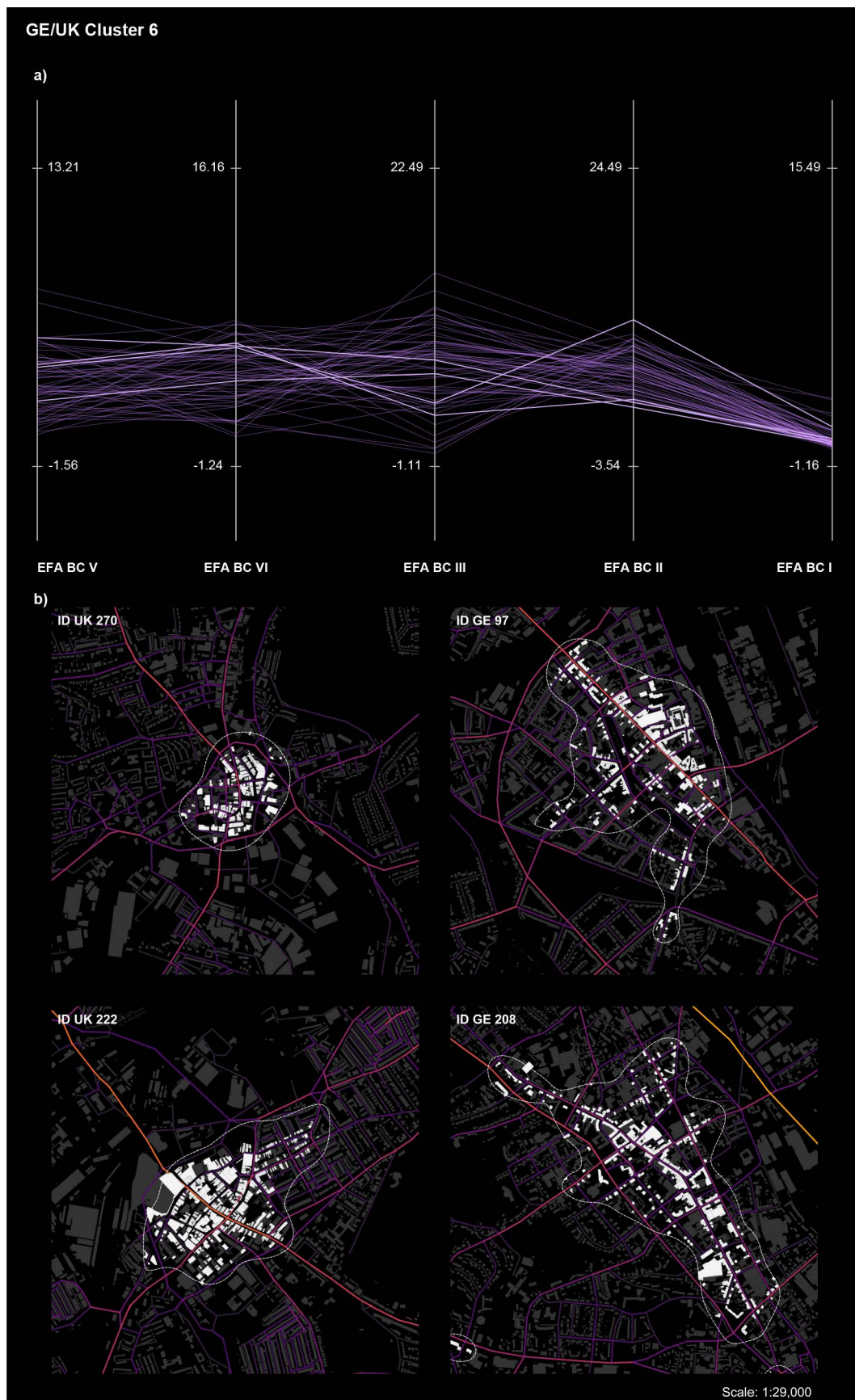
Figure 144: Parallel plot of the sixth cluster of service and trade agglomerations and its maximum value of each latent centrality (a). Mapping of commercial activity and the combined EFA BC model (b).
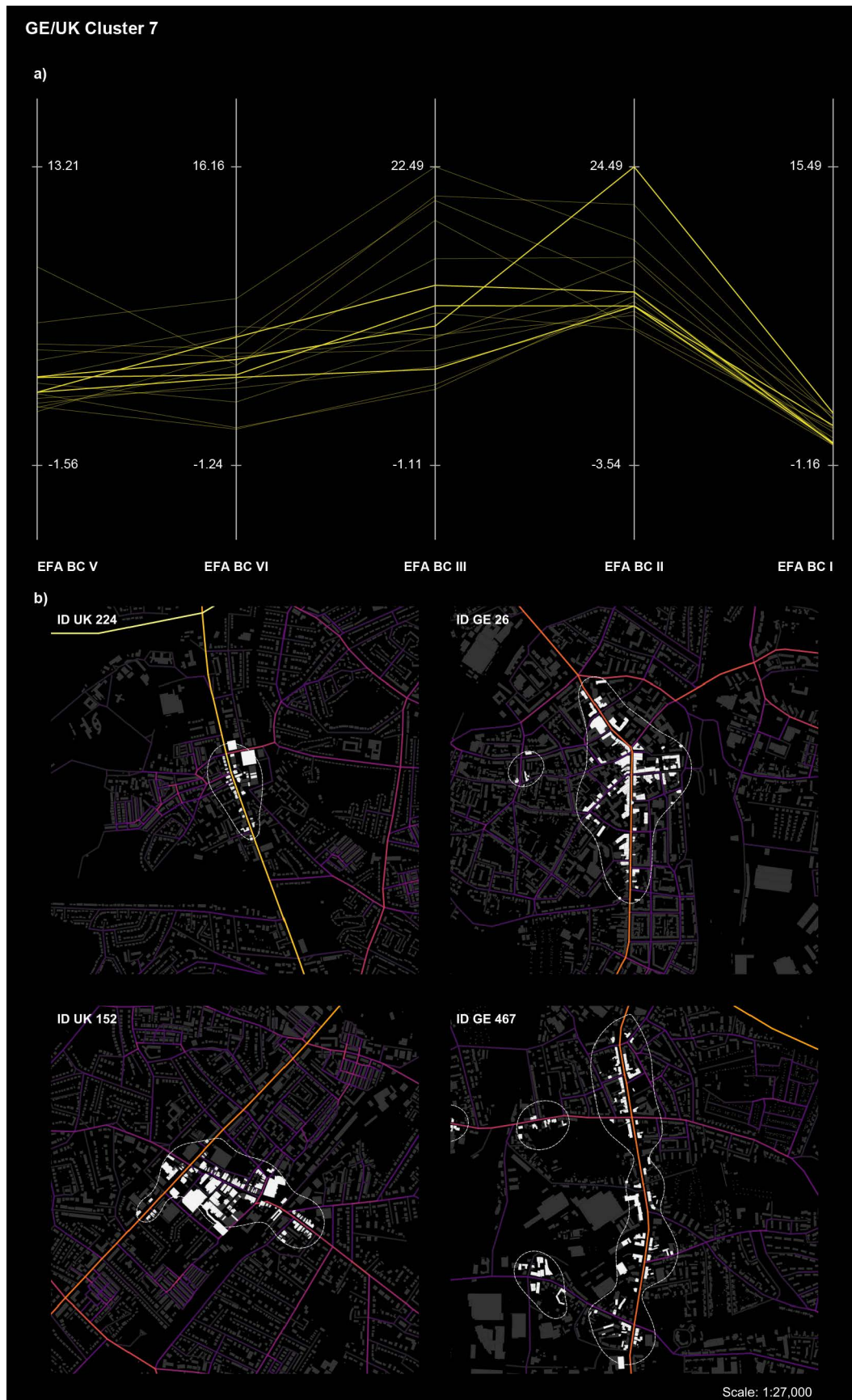
Figure 145: Parallel plot of the seventh cluster of service and trade agglomerations and its maximum value of each latent centrality (a). Mapping of commercial activity and the combined EFA BC model (b).

GE/UK Cluster 8

a)

| 13.21 | 16.16 | 22.49 | 24.49 | 15.49 |
| -1.56 | -1.24 | -1.11 | -3.54 | -1.16 |

EFA BC V   EFA BC VI   EFA BC III   EFA BC II   EFA BC I

b)

ID UK 109   ID GE 430

ID UK 86   ID GE 150

Scale: 1:19,000
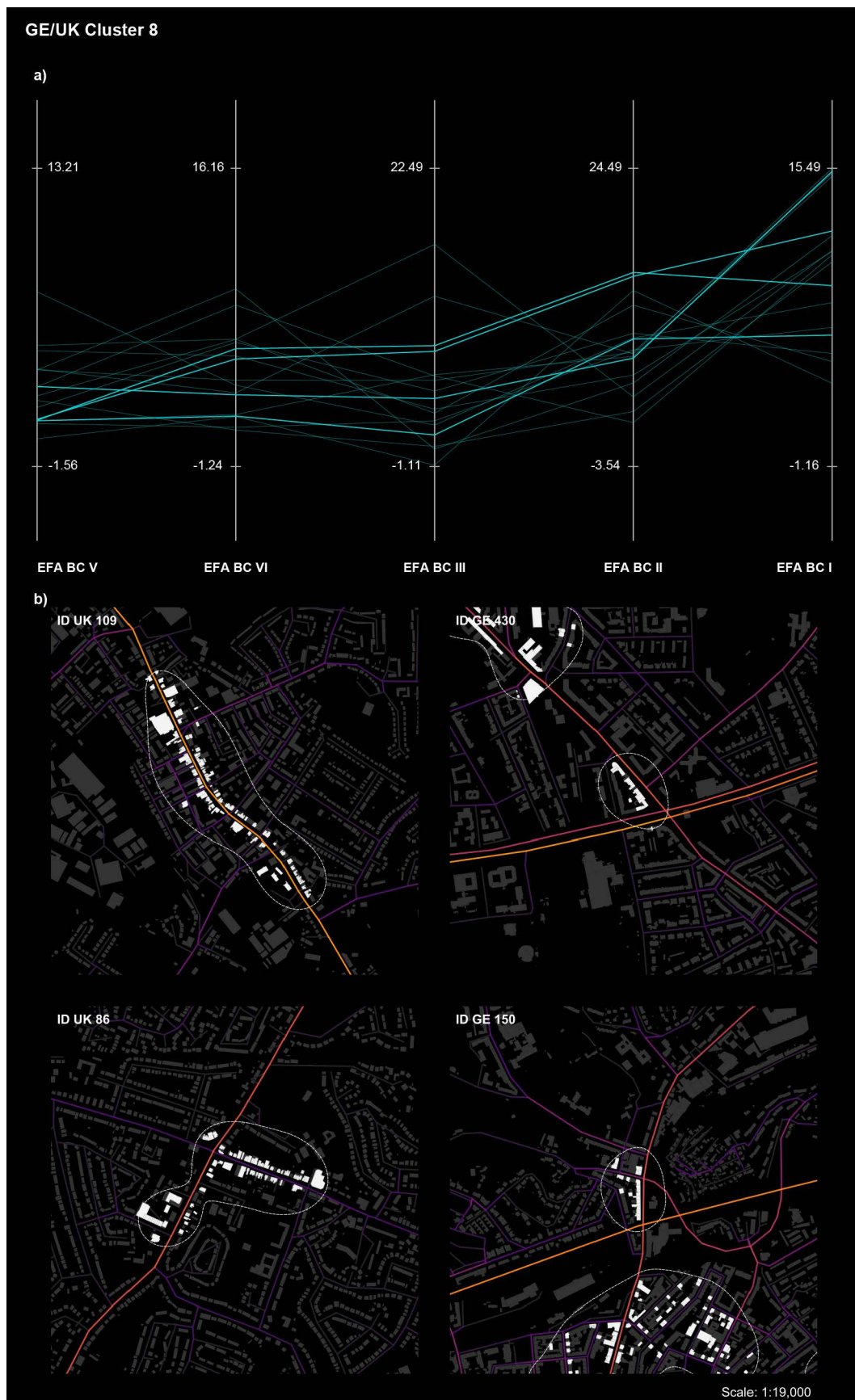
Figure 146: Parallel plot of the eighth cluster of service and trade agglomerations and its maximum value of each latent centrality (a). Mapping of commercial activity and the combined EFA BC model (b).

Cluster 1 (Figure 139) contains 186 centres and loads the highest on EFA BC 4 and the lowest on EFA BC 3. The centres are relatively small in size, which might reflect higher loadings on EFA BC 4, which is the latent centrality structure that can be related the most to inner-city relationships. The four randomly selected centres are all of the second relationship type where the centre is located at the beginning of a latent centrality structure (visible especially in Figure 139:b ID UK 117 and ID UK 155). Overall, C1 appears to characterise mostly centres that serve inner-city neighbourhoods.

Cluster 2 contains 172 centres and loads the highest on EFA BC 3, with some noteworthy loadings on EFA BC 4. The lowest relationship between cluster 2 and EFA BCs can be reported for EFA BC 1. EFA BC 4 is the latent centrality structure that is associated the most to relationships between urban areas as well as inter-city relationships. This could indicate that these centres profit particularly from journeys that do not choose these places as specific destinations. The morphological arrangement of buildings with commercial activity in cluster 2 is mostly linear (type 1: 'immediately on a latent centrality structure segment') (visible especially in Figure 140:b ID GE 468 and ID GE 155).

Cluster 3 is the largest cluster with 440 centres and exhibits overall the lowest EFA BC loadings of all clusters, with the maximum being EFA BC with a mean of 1.818. The centres in this cluster are all relatively small in size and feature all three types of building arrangements with regards to latent centrality structures. Commercial activity is rather scattered along the road, which might be an indicator of a developmental stage in centre formation. It is reasonable to expect this cluster is negligible if the initial selection is limited to those centres that have more than the set maximum of 10 different service and trade functions.

Cluster 4 (Figure 142) contains 92 centres and correlates strongly with the first and second EFA BC factor. All clusters are relatively large and exhibit a distinctive pattern of short and densely arranged segments. Commercial activity is located not only on roads of latent centrality structures but also on adjacent roads, sometimes covering entire blocks. Centres correlating strongly with EFA BC 1, are often characterised by historical urban cores, as these feature a denser spatial configuration than those centres that developed after the mid 19th century.

Cluster 5 (Figure 143) contains 78 centres that correlate the most with EFA BC 4 and EFA BC 5. This group includes most of the major city centres of cities such as Essen, Dortmund, Leeds or Sheffield. Cluster 5 centres are often in close proximity to centres of smaller sizes and feature large agglomerations of buildings with commercial functions. For this cluster type, it is characteristic that latent centrality structures of

high values surround the centre (see Figure 143:b ID GE 675, or ID UK 113). This enclosure by latent centrality structures might constitute an additional subtype to the previously identified three fundamental relationship types of commercial use and latent centrality structures.

Cluster 6 contains 60 centres and loads the highest on factor EFA BC 4 and EFA BC 2. The curve development of the factor correlation is comparable with cluster 3, yet the mean values are substantially higher. In terms of the overall centre morphology, the centres are comparable in size to those of cluster 5, yet they are not surrounded by smaller sub-centres and instead form rather large single agglomerations.

Cluster 7 and cluster 8 are very comparable in their general morphology. Cluster 7 contains 17 centres correlating strongly with EFA BC 2 and EFA BC 3, whereas cluster 8 contains 17 centres correlating strongly with EFA BC 2 and EFA BC 1. Both clusters are characterised by long linear building arrangements along segments of regional or intra-regional latent centrality structures. The centres of both cluster groups can also be compared to cluster 2 due to their type 1 relationship with the respective latent centrality structure. Centres of cluster 7 show overall a strong correlation with two latent centrality structures (EFA BC 2 and EFA BC3), while cluster 8 is characterised by a monotonous increase of correlation with latent centrality structures from EFA BC 5 to the peak at EFA BC 1.

Overall, the analysis shows that there is a differentiated morphology between all centres with three reoccurring relationship types, and/or a combination of those three. Moreover, there are specific cluster types that share common morphologies, as well as commonalities in their relation to latent centrality structures. Centres do not relate only to a single latent centrality, but to at least two latent centrality structures. This multi-relationship highlights that centres fostered by multi-scalar movement patterns, rather than being fostering by a single scale. Moreover, centres are located at locations that are beneficial for movement potential across scales. Evidently, centres can relate to either neighbourhood, city, inter-city, regional or intra-regional relationships, instead of a clear centre to city or centre to neighbourhood relation. The intensity of the relationship between centre and latent centrality structure differs with some centres relating to only a weaker part of the structure (lower EFA value), while others relate to strong latent centrality structures (higher EFA value).

Table 36 shows the total count of centres relevant to each latent centrality structure, as well as the percentage of the total centres. Each row represents a count of all centres that feature a latent centrality factor value of equal or above 1, 2 and 3. The higher the EFA value is, the stronger the relevance for the overall pattern. The majority of all centres

(87.7%) relate to EFA BC 4, the city-scale latent centrality structure of betweenness centrality. EFA BC 5, the neighbourhood scale, is the second most relevant latent centrality structure with 72.2% of all centres loading on it. The latent centrality structure with the fewest centres is EFA BC 1. Nevertheless, more than 10% of all centres are related to this inter-regional scale.

With regard to closeness centrality this picture changes slightly, as EFA CC 1 and EFA CC 2 are somewhat difficult to interpret as their extent goes beyond what can be considered a human scale centrality. The factor that is related to existing city centres (EFA CC 3) also contains the highest number of overlaps with centres, with 66.8% of all centres being located at locations that are related to EFA CC 3. The second highest relationship is between 52.9 per cent of all centres and the neighbourhood centrality structure (EFA CC 4). Overall, closeness centrality covers fewer centres in total compared to betweenness centrality.

Table 36: Total count and percentage of total centres with loadings on EFA BC and EFA CC latent centrality factors equal or above 1, 2 and 3.

EFA value	EFA BC 5	EFA BC 4	EFA BC 3	EFA BC 2	EFA BC 1	EFA BC COM
≥ 1	765 / 72.2%	**930 / 87.7%**	593 / 55.9%	298 / 28.1%	118 / 11.1%	**1020 / 96.2%**
≥ 2	449 / 42.4%	749 / 70.7%	466 / 44.0%	186 / 17.5%	52 / 4.9%	914 / 86.2%
≥ 3	251 / 23.7%	533 / 50.3%	372 / 35.1%	135 / 12.7%	35 / 3.3%	742 / 70.0%

EFA value	EFA CC 4	EFA CC 3	EFA CC 2	EFA CC 1	EFA CC COM
≥ 1	561 / 52.9%	**708 / 66.8%**	390 / 36.8%	418 / 39.4%	**996 / 94.0%**
≥ 2	267 / 25.2%	373 / 35.2%	164 / 15.5%	60 / 5.7%	624 / 58.9%
≥ 3	109 / 10.3%	157 / 14.8%	30 / 2.8%	0	252 / 23.8%

Finally, Figure 147 shows how the data is distributed across all EFA BC factors and EFA CC factors. Each diagram includes all 1,060 centres and their maximum betweenness latent centrality factor value and their mean closeness latent centrality factor value. The colours correspond to the eight clusters introduced earlier. All of the 40 highlighted points represent those centres that do not score on any of the latent centrality factors. The dashed line indicates the equal or above 1 threshold value. The diagram provides insights into the overall distribution: most of these 40 cases are part of cluster 3, which had previously been identified as the weakest of all clusters.
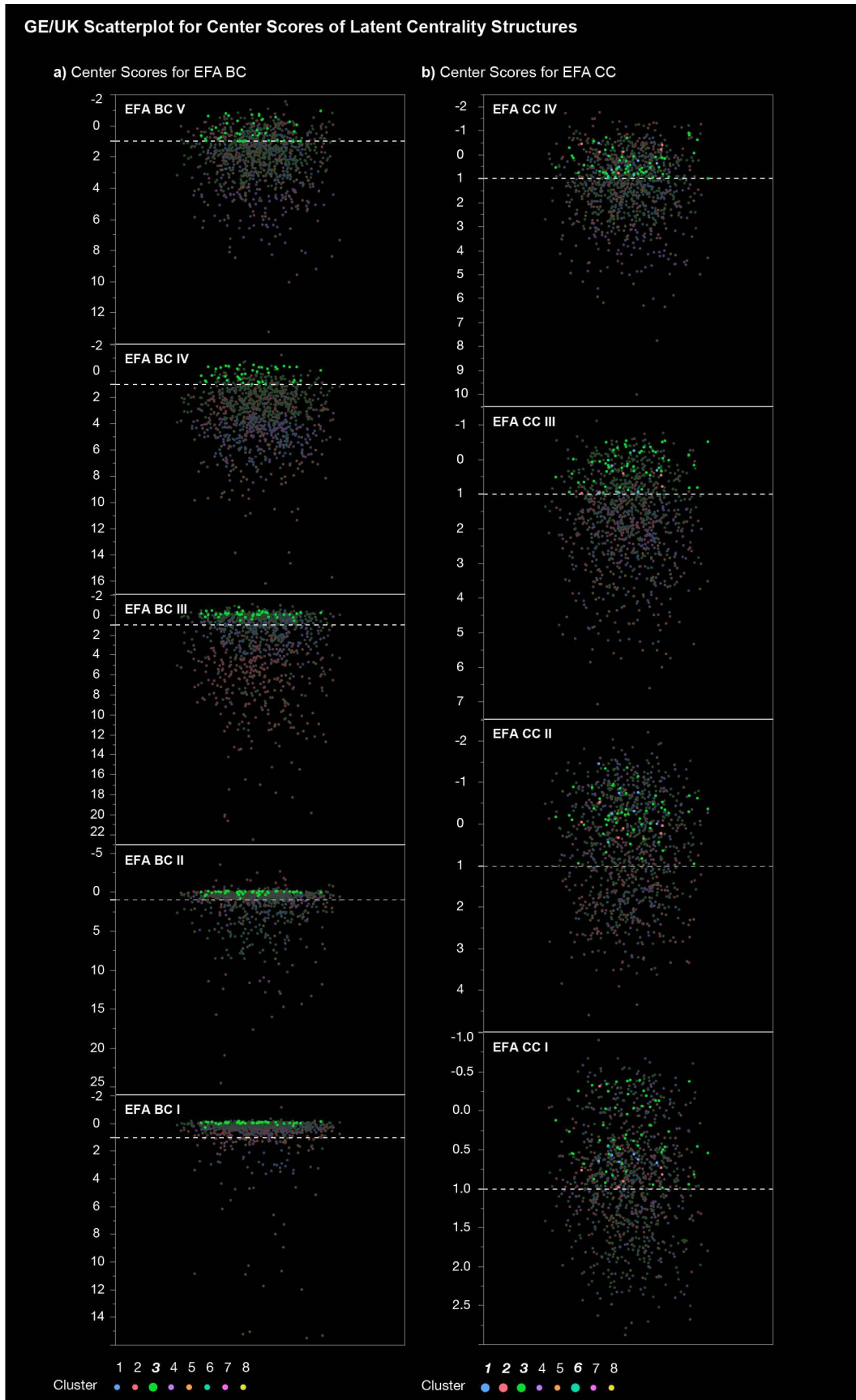
Figure 147: GE/UK scatterplots for centres and their respective score on each latent centrality factor. Highlighted by a dotted line is the value of 1-threshold, values above 1 show the relative relevance to the latent centrality structure.

8.2 SUMMARY

This chapter has elaborated on the relationship of socio-economic variables and latent centrality structures in the two case study regions, the German Ruhr Valley and the British NYD region. The main purpose of these enquiries is a validation test to scrutinise whether fundamental network structures hold explanatory power for patterns shaped by human behaviour. The results show that there are relationships between latent centrality structures and three socio-economic variables: regional movement flows, the location of commercial activity and residential occupation. The intensity of these relationships differs depending on the socio-economic variable and the latent centrality structure. The explanatory power differs substantially between the individual latent centrality structures.

The estimation of regional traffic flows show that the combined latent betweenness centrality model (EFA BC COM) can account for the majority of the overall variance in the data. The correlations between regional movement flows and latent centrality structures, as well as the 49 individual radii for betweenness centrality and closeness centrality highlight that a combination of all scales to a multi-scalar dataset (EFA BC COM) improves traffic flow estimates and can remove radius selection bias in space syntax analysis. Individual latent centrality structures prove to be weak in their explanatory power, however, differences in correlation coefficients between modes of travel and latent centrality structures indicate that each latent centrality structure can be related to a specific traffic flow. Passenger and all vehicular flows are best explained by the intra-regional scale (EFA BC 2). Heavy goods and freight traffic is best explained by the inter-regional scale (EFA BC 1).

The estimation of population densities through spatial metrics shows that 3D-building geometries in combination with semantic information are a reliable and accurate source for small-scale estimations of population density. This study is the first to use 3D-building information for regional population estimations in a cross-country comparison and shows that the relationship between habitable volume and total population at a given location differs substantially depending on the socio-cultural variables. The results of the proposed data disaggregation method demonstrate that the method is appropriate to bridge data differences in scale, resolution and precision. I have generated a new dataset of building- and street-level population of a hitherto unseen precision and scale, which can inform future research in this field. Relationships between latent centrality structures and population data on the street-level have not yielded any statistical relevance. However, the hierarchical model approach highlighted that such relationships exist at the level of neighbourhood aggregation. The best performing estimates are achieved at the sample resolution of

1000 by 1000 metres. Both neighbourhood and city latent centrality scales of closeness and betweenness centrality (EFA CC 4, EFA CC 5, EFA BC 4 and EFA BC 5) show to be able to explain relevant degrees of the variance in the data. Overall, compared to simple network properties latent centrality structures are less effective in such estimations. Total segment length alone could account for 84.4% of the variance in the data. The performance of estimates increases when the model is restricted to areas, which can be categorised as urban, rather than incorporating rural and agricultural areas in the model too. This alone could prove to be of relevance for countries with insufficient or incomplete census data. Moreover, it could form the basis for a method of low-cost and efficient population estimates.

Finally, regarding the location of commercial activity and spatial metrics, the results show that there are strong relationships between the location of service and trade functions and latent centrality structures of betweenness and closeness centrality. The strongest relationship can be reported for commercial activity and latent betweenness centrality structures. More than 96% (precisely 1020 centres) of all 1060 agglomerations of service and trade functions identified are located on roads of high EFA BC values, which make up only 10% of the entire spatial network. The results show that centres share a more complex relationship than previously thought. The hierarchical cluster approach demonstrates that commercial agglomerations feature a multi-combinatorial relation with independent scales. The analysis identified 8 distinctive scalar combinations, showing that centres relate to scales not in an ordered hierarchy, but in a more complex and less ordered hierarchy where agglomerations can be central to a neighbourhood latent centrality structure (EFA BC VI) while simultaneously being central to a regional latent centrality structure (EFA BC II). The visualisation of combined latent centrality scales that I proposed enables direct interpretations of such scalar relationships among commercial activity. The morphological description presented differences in the spatial arrangement of commercial activity between different clusters, indicating the existence of distinctive spatial forms that might emerge in response to the scale relationship.

The following chapter will synthesise these findings with the overall contribution made in this thesis and discuss their implications for future research and the field in a broader context.

CHAPTER 9

DISCUSSION: CENTRALITY A RELATIVE CONCEPT

9 CHAPTER

9.1 DISCUSSION: CENTRALITY A RELATIVE CONCEPT

This thesis looked at two polycentric urban regions, the Ruhr Valley region and the Nottinghamshire, Derbyshire and Yorkshire region to examine the role centrality plays in their spatial organisation. The research focused, on the one hand, on the development of a set of methods for the analysis of regions by spatial network metrics, and, on the other hand, on the empirical evidence for the prediction of socio-economic variables on the basis of observed centrality patterns. While doing so, a series of methodological issues for the analysis of regional networks where raised and their solutions were proposed and tested.

The theoretical foundation of these explorations is extensively discussed in Chapter 1 and Chapter 2, along with elaborations on differences between theories aiming to explain the spatial organisation of regions and cities, as well as the implication of these positions for empirical research of spatial organisation. The focus lay specifically on disentangling the complex relationship of the emergence of centres in settlement patterns that are caused by the geographical location of geological resources, and the contrasting notion of central place theory that argues in favour of the emergence of centres through microeconomic processes. The theoretical conclusion raised issues of the definition of the region as an entity itself and suggested that research interested in PURs needs to take into consideration human agency in the real space in order to identify complex scalar relationships. Research concerned with cities within PURs, on the other hand, needs to take into consideration their intrinsic patterns of multi-scalar centres, which can only be revealed by their regional embedding. This research has demonstrated that a space syntax approach brings valuable insights into such multi-scalar relationships and more importantly to the scale of regional analysis.

Chapter 3 offered the reasoning for the selection of the two case studies and a historical analysis of the development of the regional spatial organisation and its main driving forces. Chapter 4 proposed a methodological approach and the data needed for the analysis of such PURs. Further methodological issues were addressed in Chapter 5, where a workflow for the usage of voluntary geographic information for the analysis of regional street networks was proposed. Chapter 6 provided fundamental insights into the forms and characteristics of regional street networks and the implications for the generations of randomised regional street networks. Chapter 7 presented the analytical explorations of regional street networks and specific latent centrality structures within them. Here, I argued that these centrality structures reflect fundamental network

properties, whose predictive power for socio-economic variables in PURs is extensively tested in chapter 8.

This chapter recapitulates the core concepts and issues this thesis had focused on; it further aims to offer a synthesis of the findings and an embedding of these findings into a wider theoretical context.

9.1.1 REVIEW AND SYNTHESIS

POLYCENTRIC URBAN REGION: THE CASE STUDIES. The Ruhr Valley and the NDY region are two examples of an industrially induced growth of polycentric urban regions.

VGI DATA IN NETWORK ANALYSIS. Space syntax as a tool for spatial network analysis has traditionally focused on axial line representations for a citywide analysis. The method transforms axial lines into a dual-graph and performs network metrics on this graph. Traditionally, such axial line maps are produced by cartographers or algorithmic generation methods. Both approaches are insufficient for research interested in large-scale regional or countrywide analyses. This thesis has shown that OpenStreetMap data can be used to emulate axial line representations through a specific simplification workflow. The theoretic reasoning behind such an application has been discussed in Chapter 3. Moreover, the necessary steps in order to employ such data have been highlighted and a new GIS-workflow for the simplification and removal of excessive network information has been proposed and tested against different sources of road network data as well as traditional axial line representations. The result of this comparison shows that such data constitutes an appropriate source for spatial network analysis. In summary, these comparisons have led to the following findings:

- Neither OSM, nor any other road network data source should be employed in space syntax analysis without a rigorous network simplification process. Differences in the level of detail and resolution of such datasets make the results of subsequent network analyses in terms of their comparison to other spatial networks derived from different source datasets unreliable.
- A simplification method was proposed, based on three principles: i) different functional networks (e.g. dual-carriage ways, pavements, cycle roads, tracks) are treated as single spaces, simplified to single-line segments, ii) curved roads are simplified based on visual field thresholds, equivalent to the street width, and iii) intersections that are visibly connected (i.e. connected to a street within a visual threshold) are topologically joined at a single connection point. The method has been extensively tested and results have shown, that simplified OSM road

networks are highly comparable to axial line representations and constitute an appropriate alternative for a large-scale analysis.

- VGI information includes spatial information produced and updated by the very users of the spaces mapped and, thus, embed local knowledge of the spaces that goes beyond governmental datasets.

The high correlation coefficients between network centrality measurements of simplified OSM data and traditionally produced axial line models raise questions about the fundamental differences between these two spatial representations. On one hand, road centre line data can feature excessive nodal information and traffic management details, axial line maps, on the other hand, are simplified representations of a network of convex spaces. Theoretically, these are two intrinsically different representations of the environment; yet, statistically speaking these differences are not as pronounced as expected. The simplification workflow results in a more abstract version of a road centre line map. Due to the reduction of traffic road elements, the combinations of streets that are in close vicinity to each other and the nodal reduction process that employs the visual field, the results come very close to what the early axial line intended, namely to represent space in a way so that it reflects the differences in path-decision making. In this regard, the simplified OSM model, might in fact be a more precise representation, as angles of street intersections are not met at potentially arbitrary angles, but at those that reflect the actual experience of the space in real-world scenarios. However, the aim of the simplification workflow is not to replace the axial line representation as such, as it features a well-founded theoretical rationale, but rather to open up the method of space syntax and spatial network analysis to regional inquires beyond the scope of the city.

RANDOM REGIONAL STREET NETWORKS. The field of random network generation is well established with a large set of different algorithms that can compute networks of different magnitude and characteristics (e.g. high-dimensional networks, scale free networks, small world networks). With regards to random spatial networks, where the topology alone does not contain sufficient information about the structure and form of the network, and with valuable information held in the network's geometric properties (i.e. location, length, angle), the situation is a different. Only one model of complete spatial randomness is available (ERPGr), which was employed in this thesis. The core issue in this context is, that in spatial statistics, observed patterns are often compared against the null hypothesis of complete spatial randomness. In real-world scenarios, this state of complete spatial randomness is incredibly rare and, hence, unsuitable due to its unlikely nature. Instead, similar elements in space often form clusters, or as Tobler puts it in his first law of geography 'everything is related to everything else, but near things are more related than distant things' (1970 p. 236). This situation led to the

development of a new algorithm for the random generation of street networks that matches the core characteristics of existing street networks, except the spatial configuration. Moreover, little has been written about the characteristic and form of regional spatial networks. Since regions cannot be described simply as larger versions of cities it was necessary to contribute to knowledge on the statistical properties of such regional networks. The main contributions of this inquiry were the following:

- Regional street networks feature cluster processes of varying degrees.
- Segment length distributions in regional street networks follow specific double Pareto or log-normal distributions. This is different to the previously reported power law distributions in the context of cities, which might be caused by observing only a fraction of the network in the past.
- Traditional random spatial network generation sets out with a Poisson point distribution. Under this premise, and with the aim to generate a connected graph, a double Pareto segment length distribution cannot be achieved. The fundamental mechanism influencing whether a randomly generated spatial network can follow the same segment length distribution is the density distribution of networks nodes. This distribution can be emulated by a kernel density function of existing street networks.
- A Variance-Gamma Planar Graph with radius restriction has been proposed by making use of the kernel density function for k-nearest neighbours. As a part of this work, an algorithmic realisation for the generation process has been developed, which can be applied to various scenarios (see appendix I).

The algorithm proposed is free of scale and can also be applied in the context of the city or the neighbourhood. It, first, identifies the Kernel density of road intersections, second, a point pattern of the stochastic variance gamma process is fitted and a randomised realisation of this process is generated in space. Third, based on a radius restriction edges between these points are generated until the graph is fully connected or the maximum number of segments has been generated.

SPATIAL ORGANISATION OF PURs. The thesis set out with the notion that human action in space leads to the emergence of specific patterns of spatial configurations, which reciprocally influences the future potential of human action. Christaller's CPT, is based on a microeconomic assumption that human action in space is highly influenced by the aim to minimise distance cost (here economic distance), and ultimately leads to a spatial realisation of hierarchically ordered centres fostering commercial activity. Following this CPT notion, a particular spatial realisation of such hierarchical order, or in other words, a particular spatial configuration that fosters such hierarchical relationships must be observable in the regional spatial organisation. Chapter 6 made

an enquiry, to which extent this pattern of hierarchical order is retraceable in space and detectable by its spatial configuration. As presented in Chapter 2, both regions under investigation share an industrial heritage that has strongly influenced their development. CPT is a theory that explicitly excludes industrial factors. Furthermore, post-industrial regions are commonly thought of as polycentric and are, hence, not expected to follow a CPT hierarchical order. This situation brought up the following questions: what are the driving forces for the emergence of the spatial organisation in such PURs and what fosters commercial activity? This inquiry proposed to make use of latent factors derived from an exploratory factor analysis to statistically reveal underlying patterns and to investigate what the spatial organisation of PURs' underlying structure is as well as whether PURs feature a hierarchical pattern nevertheless.

To explore the role of the spatial network in fostering such hierarchical patterns of human spatial organisation, this thesis employed a comparative approach using a spatial network model of regional scope that incorporates all publicly accessible spaces as the first of its kind (from small pedestrian alleys to crosscountry motorways). The analysis compared two randomly generated street networks with the two case studies and analysed them on 49 different radii of betweenness and closeness centrality. The results were visualised using a multi-classification symbology in GIS allowing morphological interpretations of the structures and a theorisation of their scale function. This exploration based its findings on a very large dataset, with more than 400 million centrality values for two centrality measures and four different models. The results of these explorations can be summarised in the following findings:

- All spatial networks tested in this project have exhibited distinctive latent centrality structures, highlighting that every spatial network inherently possesses such a structure that can be revealed by an EFA.
- Betweenness centrality data from the four tested regional spatial networks (ERPGr, VPGr, GE and UK) resulted in 5 latent centrality factors derived from the EFA analysis, exhibiting a clear and comparable, simple structure. BC is a more robust spatial metric when compared across different regional spatial networks, than CC.
- Closeness centrality data from the two tested randomly generated regional spatial networks (ERPGr and VPGr) resulted in 5 and 6 latent centrality factors, while both real-world regional spatial networks (GE and UK) resulted in 4 latent centrality factors extracted from the EFA analysis. All regions showed a clear, simple structure but only the two real-world models were highly comparable. CC shows a higher degree of variability between

- randomly generated and real-world models, with the latter being highly comparable.

- The latent centrality factors extracted from an EFA for BC, partly match with those radii reported by Christaller's central place theory. Spatial representations of these 5 LCFs show distinctive spatial patterns that match existing semantic descriptions for specific scales, i.e. factor V neighbourhood, factor IV city, factor III metropolitan, II intra-regional and I inter-regional scale.

- The latent centrality factors derived by an EFA for CC, also partly match with those CPT radii that BC did not describe. Spatial representations of these 4 LCFs show distinctive spatial patterns that match existing semantic descriptions for specific scales, i.e. factor IV neighbourhood, factor III between city/metro, II between-metro/intra-regional and I inter-regional scale.

- The spatial configuration of both regions (UK and GE) exhibited strikingly similar latent centrality structures.

The results have shown that geometric properties of street networks play a fundamental role in determining a network's behaviour. The analysis of two different centrality measures and the latent centrality structures obtained point to specific phase transitions between different distance patterns. These distance patterns highlight the relative importance of each street for the overall spatial structure at each respective scale. With regards to betweenness centrality, LCFs can be employed for a hierarchy classification of each network segment. This classification provides not only a group membership of independent scale groups but by providing a value for the relative importance of each segment within the scale group itself. Together all five BC LCFs together form the fundamental frame of the overall network. BC LCFs have shown patterns that are interpretable and can be joined with existing semantics for each scale, of which all result in meaningful patterns. With regard to CC LCFs, only the IV and III factor offer such a direct interpretation. The visualisation of the first two factors (I and II) do not provide a direct real-world interpretation. This is due to the very meaning of closeness centrality; spaces that are nearest to all other spaces on radii that go beyond the average city sizes of those cities contained in the model (in the two real-world examples this is the case for radii ≥20,000 metres) tend to highlight areas between agglomerations, rather than the agglomeration itself. This is of course a matter of the properties of the system under investigation; in regional networks where individual cities are of larger extents (e.g. ≥20,000 metres) these scale structure will become meaningful as well. Whereas in regions, such as the observed ones, a potential interpretation for such scales is that those segments of the network that are part of one

of the first factors (I and II) give an indicator for the direction of further urbanisation processes.

NETWORK CENTRALITIES FOR URBAN AND REGIONAL PLANNING. This inquiry set out from the theoretical positioning that networks can be thought of as environments for simulations for human behaviour. In this study, network centralities are calculated based on an angular segment analysis, which calculates the shortest path based on human-decision making and the way humans move in space, i.e. by minimising angular differences on their way through space. One interpretation is that those calculated centralities form an appropriate simulation of an agent-environment relationship by a) representing the environment by a network of all accessible spaces (SIMP model) and b) by an agent-based behaviour rule that reacts to this environment (minimisation of angular distance). Centrality patterns are, beyond the value of the spatial metric, indicators for the potential of human movement. Those network segments, which exhibit high centrality values are hence indicators for higher potential of human encounter depending on the trip type. Based on this premise, this study needs to be understood in the context of human behavioural modelling and the observed patterns as a result of agglomerated human action. In this context, the last part of this work constituted an investigation into the relationship of human behaviour and the patterns produced by this.

The latent centrality structures have been tested for their predictive power for the estimation of socio-economic variables. Specifically, three main angles of inquiry were selected: i) the estimation of traffic flows, ii) the estimation of population and iii) the relationship between the location of commercial activity through spatial metrics. With a focus on the core notions on the spatial organisations of regions (e.g. CPT), these three angles aimed to identify methods to predict, where humans move and where humans occupy regional spaces. These tests are a way to demonstrating potential mechanisms in the complex spatial organisation of polycentric urban regions. These verifications demanded a sophisticated approach of several GIS methods and techniques that go far beyond generic GIS applications.

The following findings can be reported with regards to,

i) estimation of traffic flows:

- Betweenness centrality, showed the strongest explanatory power for the overall variance in the data of regional traffic flows. Hence, regional movement appears to be better estimated by the concept of through-movement.

- Individual latent centrality factors proved to be weak in their explanatory power, but specific types or classes of traffic correlated stronger with particular LCFs (EFA BC1 and EFA BC2). Passenger and all vehicular traffic is best explained by the intra-regional scale, whereas heavy goods and freight traffic is best explained by the inter-regional scale.

- The combination of all extracted latent centrality scales for betweenness centrality, in the form of the proposed combined model (EFA BC COM) can form an appropriate method for the estimation of regional traffic flows. A combination of all scales to a multi-scalar dataset improves traffic flow estimates and removes radius selection bias in a space syntax analysis.

ii) estimation of population:

- 3D building geometries enriched by semantic information proved to be a reliable and accurate source for the small-scale estimation of population densities. This is also the case, for cross-country comparisons when the sources of semantic information differ significantly in the form and level of detail.

- Adjustments of habitable volume, based on service and trade 3D-building information improved the explanatory power of the variance in the data only marginally. Improvements can be reported specifically for inner city areas, which constitute nevertheless difficulties in their estimation, due to the complex superposition of different functions and the lack of precise volumetric or floor-level functional information.

- The proposed method of disaggregation for large-scale population data can be employed for the precise calculation of population per building, street segment and street length. The proposed method provides future research with a large-scale dataset of street level population estimates.

- Relationships between centrality measurements and population data on the street-level have not shown any statistical relevance. The hierarchical model approach highlighted that such relationships exist at the level of neighbourhood aggregation, with the strongest relationship at the sample resolution of 1000 by 1000 metres. Both neighbourhood and city latent centrality scales of closeness and betweenness centrality (CC4, CC5, BC4 and BC5) showed the be able to explain relevant degrees of data variance.

- Contrary to the explanatory power of centrality measures is the simple network property of total segment length able to account for 84.4% of the variance in the population data. The performance of estimates increases when the model is restricted to areas, which can be categorised as urban,

rather than incorporating rural and agricultural areas in the model too. This alone could prove to be of relevance for countries with insufficient or incomplete census data. Moreover, it could form the basis for a method of low-cost and efficient population estimates.

iii) location of commercial activity and spatial metrics:

- Latent centrality scales for betweenness centrality proved to be an appropriate tool for the classification of centres and their relation to respective scales. Of all 1,060 identified agglomerations of commercial activity are 96 percent directly next to a street of at least one of the 5 identified latent centrality scales of betweenness centrality. This is of particular relevance because the BC latent centrality structure makes up only 10 percent of all segments in the system.

- Two latent local centrality scales for closeness centrality (EFA CC VI and EFA CC III) proved to be of a further meaningful application, with 83% of all centres located directly next to a segment relevant to at least one of the 2 latent centrality scales of closeness centrality.

- The hierarchical cluster approach highlighted that commercial agglomerations feature a multi-combinatorial relation with independent scales. The method identified 8 distinctive scalar combinations, showing that centres relate to scales not in an ordered hierarchy, but in a more complex disordered hierarchy where agglomerations can be central to a neighbourhood latent centrality structure (EFA BC VI) while simultaneously being central to a regional latent centrality structure (EFA BC II).

- The proposed visualisation of combined latent centrality scales enables direct interpretations of the scalar relationships and commercial activity.

Finally, this work forms a novel contribution by testing traffic count data against a complete spatial network representation of a region for the very first time. Furthermore, by estimating population data through 3D-building geometries in combination with semantic information and disaggregation to the level of the street an innovative new population dataset has been produced.

What implications do these findings have on the broader context of polycentric urban regions? First of all, this thesis has demonstrated that PURs exhibit a spatial organisation of multi-dimensional complexity. Such regions are simply too complex to be fruitfully analysed solely through qualitative or simple descriptive methods. The sheer magnitude of spatial regional networks, the representation of the physical reality

with several million streets and spaces exceeds by far what the human mind can process. This is why we need advanced statistical methods and models that allow us to grasp and structure the complexity of these entities. Such methods help to reveal patterns that otherwise would have been inaccessible for the external observer and enable us to develop fitting concepts and fill them with meaning.

The results have also highlighted that the rigid hierarchical structure of Christaller's CPT cannot account for the complexity encountered in PURs. If Christaller's CPT is transferred into the actual physical space and its network terminologies, his notion of economic accessibility becomes one of closeness centrality. According to Hillier, this is related to to-movement potential, which is the probability that a space is selected as a destination. Christaller's central places would be located at such configurationally beneficial locations if economic distance were transferred into the physical distance. However, only 84% of all centres are in the proximity to a street with a relevant value of one of two EFA CCs. Contrary to this, the location of centres is much better explained by betweenness centrality. The vast majority (96%) of all 1,060 centres is in an immediate proximity to an EFA BC structure. This implies that the location of centres can be better described by the concept of betweenness centrality, or in other words by the potential of through-movement, i.e. the probability of being on a segment in the network that is more likely to be part of shortest paths in the system. Previous studies have already highlighted strong relationships between BC and commercial functions. This relationship highlights a fundamental shortcoming of the CPT theory, namely that commercial activity overall is based more strongly on random encounters on journeys (betweenness centrality) than on destination-based movement (closeness centrality). The investigations have shown how the combination of specific scale structures fosters commercial activities and revealed that the fundamental microeconomic mechanism of random encounter is of particular relevance in PURs.

This highlights that the expectation of the order of hierarchy in regions must be flexible to account for the multifaceted patterns that can be observed in polycentric urban regions. This implies a fundamental impact for policies aiming to stimulate growth in polycentric urban regions, as they need to take into account a balanced competitiveness on the level of the neighbourhood. This becomes even more relevant with regard to the present speed of global urbanisation processes. In the light of this development we can expect that in the future, polycentric urban regions will not be the exception, but will become a much more common regional morphology. This emphasises the importance for understanding polycentric urban regions, particularly as this study is hitherto the only one that provides systematic and large-scale evidence on the internal geographies of this regional type.

The advancement of the space syntax methodology from the context of the city to the regional continuum has shown how the use of relational epistemologies, such as the theory of the spatial configuration, provides valuable insights into PURs. It also highlights, however, that there is a need for a broader conceptualisation of large urban agglomerations. Polycentric urban regions constitute a morphological type *sui generis*. As I demonstrated empirically, theories developed in the context of the city need to be extended and conceptually re-adjusted in order to bring meaning to larger regional constructs such as PURs. The identification of scale as a critical theoretical and methodological issue plays a fundamental role in such conceptualisations. I have proposed how the concept of scale can be operationalised in PURs by the identification of latent centrality structures. This contribution is not only of use for cases of polycentric urban regions, but will prove fruitful in the analysis of any region or even city.

This study focused solely on the regional spatial network and therefore takes a 'space first' approach. There are multiple, additional layers that should and need to be taken into account for a comprehensive understanding of the spatial organisation of PURs. To name only few, these layers can be found in the complex temporal networks of urban infrastructure, networks of public transport, networks of communication and technology and networks of the dynamic relationship between firms.

9.1.2 FURTHER RESEARCH AND POTENTIAL APPLICATIONS

This thesis has opened up the possibility to analyse any form of large-scale agglomeration using a network-based approach by employing open source data. Issues related to differences in the scale, the aggregation and the precision of regional datasets form one of the main obstacle for scientific enquiries on large-scale urban entities. From 2017 to 2050, the cities of the world will have to accommodate an additional 2.5 billion people. This projection presents us with the challenge on how to develop the inevitable massive urban expansion that will be brought about in this process to provide the population with access to health service, education and elementary infrastructure. For this, we urgently need to equip urban planners and researchers with concrete tools to better understand and manage this expansion. The methods proposed in this thesis aim to contribute to the development of tools required to steer these developments. Many of the points made in this thesis will be of relevance in the future, when the definition of cities moves away from the current Western example of city sizes and shifts towards new metropolitan mega regions. This work forms a particular contribution in presenting a series of approaches to bridge scale and geographical entity differences between datasets.

The proposed model for disaggregation of population data using 3D-building information in combination with semantic information on building usage has shown a superior performance in the prediction of building-based population as well as of street-based population. Population data on the level of the street segment is of great interest for a large variety of different fields, such as epidemiology, econometrics and transport planning. Potential applications are, for example, the estimation of the population affected by air pollution caused by traffic or the potential spread of infectious diseases through the spatial network, based on precise street-level population estimates. An interesting inquiry might also involve explorations of network centralities that are based on the actual street-level population. As such, rather than computing the simple spatial potential based on all-to-all relationships on the dual graph, such an enquiry could focus on an origin-weighted map making use of the estimated population per street segment.

The estimation of population data by the sole usage of simplified OSM data and centrality measures has demonstrated to be able to explain up to 90% of the variance in population. If this method proves to hold similar explanatory power in less developed areas, it could work as a cost-effective and simple method for small-scale population estimates (250 x 250, 500 x 500 or 1000 x 1000 metres grids). This is of particular relevance for less developed countries with insufficient census data and large-scale informal developments.

9.2 CONCLUSIONS

The two polycentric urban regions, the Ruhr Valley and the Nottinghamshire, Derbyshire and Yorkshire region exhibit extraordinary spatial organisations. This thesis provided a fundamental contribution to understanding such spatial organisations. It revealed the particular polycentric spatial organisation by advancing the network-based approach of space syntax to a regional continuum. The research demonstrated that insights into the functioning of the spatial organisation of PURs can be gained by taking into account human agency embedded in physical space, as well as the reciprocal effect of the spatial organisation for the emergence of centralities. It sought to make a foundational methodological contribution by joining space syntax and CPT in the definition of polycentric urban regions, pioneering the use of complex and highly messy datasets exposing the inadequacy of existing polycentric models.

The thesis set out from the argument that the effects of globalisation have provided a fresh impetus to revitalising the tradition of regional scale in the academic debate that Walter Christaller and August Lösch initiated almost a century ago. In this context, space syntax has been found to provide an unexpected contribution to the arena of

regional studies. By examining potential points of connection between CPT and space syntax, this thesis identified scale as a critical theoretical and methodological issue. It proposed a conceptualisation of scale in the form of latent centrality structures and tested this concept empirically on socio-economic variables of human movement, occupation and economic activity. The results of these tests have shed light on the complexity of relationships between the spatial configuration and society in a regional context.

This thesis is located at the intersection of the local street and the region as a whole. Emphasising the importance of the regional embedding, the study aspires to show that human action in space forms a fundamental mechanism in understanding large-scale urban agglomerations. This advances Michael Batty's agenda for the 'New Science of Cities' for large urban regions in the context of globalisation. Ultimately, this thesis is a contribution towards what could constitute a novel branch of regional spatial analysis.

BIBLIOGRAPHY

Abler, R., Adams, J. S., & Gould, P. (1971). *Spatial Organization: The Geographer's View of the World*. Englewood Cliffs: Prentice-Hall.

AdV. (2015). *Dokumentation zur Modellierung der Geoinformationen des amtlichen Vermessungswesens (GeoInfoDok) Version 7.0.2. ALKIS-Objektartenkatalog DLKM*. Retrieved from <c:%5CUsers%5CEste%5CDocuments%5CDiss%5C2010%5CALKIS-OK-5_1.pdf%5Cnrothbergerg 12.01.2009 - ERl_AdV_11-12-2008_1_mit_Anlagen.doc.pdf>

Agnew, J. (1993). Space, scale and culture in social science. *Place/culture/representation*, 251.

Alahmadi, M., Atkinson, P., & Martin, D. (2013). Estimating the spatial distribution of the population of Riyadh, Saudi Arabia using remotely sensed built land cover and height data. *Computers, Environment and Urban Systems*, 41: 167–76. Elsevier Ltd. DOI: 10.1016/j.compenvurbsys.2013.06.002

Alexander, C. (1966). A City is Not a Tree. *City*, 122: 46–55.

Allen, P. M., & Sanglier, M. (1981a). Urban evolution, self-organization, and decisionmaking. *Environment and Planning A*, 13/January 1980: 167–83.

Allen, P. M., & Sanglier, M. (1981b). A Dynamic Model of a Central Place System - II. *Geographical Analysis*, 13/2: 149–164. DOI: 10.1111/j.1538-4632.1981.tb00722.x

Andersen, M. (1992). Spatial analysis of two-species interactions. *Oecologia*, 91/1: 134–40. DOI: 10.1007/BF00317252

Anderson, W., Guikema, S., Zaitchik, B., & Pan, W. (2014). Methods for estimating population density in data-limited areas: Evaluating regression and tree-based models in Peru. *PLoS ONE*, 9/7. DOI: 10.1371/journal.pone.0100037

Arentze, T., Hofman, F., & Timmermans, H. (2004). Predicting multi-faceted activity-travel adjustment strategies in response to possible congestion pricing scenarios using an Internet-based stated adaptation experiment. *Transport Policy*, 11/1: 31–41. Elsevier. DOI: 10.1016/S0967-070X(03)00016-7

Arthur, W. B. (1994). *Increasing returns and path dependence in the economy*. Ann Arbor: University of Michigan Press.

Arya, S., Mount, D., Kemp, S. E., & Jefferis, G. (2017). RANN: Fast Nearest Neighbour Search (Wraps ANN Library) Using L2 Metric.

Auerbach, F. (1913). Das Gesetz der Bevölkerungskonzentration. *Petermanns Geographische Mitteilungen*.

Awrangjeb, M., Ravanbakhsh, M., & Fraser, C. S. (2010). Automatic detection of residential buildings using LIDAR data and multispectral imagery. *ISPRS Journal of Photogrammetry and Remote Sensing*, 65/5: 457–67. Elsevier B.V. DOI: 10.1016/j.isprsjprs.2010.06.001

Babu, G. J., & Feigelson, E. D. (1996). Spatial point processes in astronomy. *Journal of Statistical Planning and Inference*, 50/3: 311–26. DOI: 10.1016/0378-3758(95)00060-7

Baddeley, A., Gregori, P., Mateu, J., Stoica, R., & Stoyan, D. (Eds). (2006). *Case studies in spatial point process modeling*, 1st ed. New York: Springer. DOI: 10.1007/0-387-31144-0

Baddeley, A., Rubak, E., & Turner, R. (2016). *Spatial Point Patterns Methodology and Applications with R*. Boca Raton, Florida: CRC Press.

Baker, A. R. H. (1969). Reversal of the Rank-Size Rule: Some Nineteenth Century Rural Settlement Sizes in France∗. *The Professional Geographer*, 21/6: 386–92. DOI: 10.1111/j.0033-0124.1969.00386.x

Bakillah, M., Liang, S., Mobasheri, A., Jokar Arsanjani, J., & Zipf, A. (2014). Fine-resolution population mapping using OpenStreetMap points-of-interest. *International Journal of Geographical Information Science*, 28/9: 1940–63. Taylor & Francis. DOI: 10.1080/13658816.2014.909045

Barabási, A.-L. (2009). Scale-Free Networks: A Decade and Beyond. *Science*, 325/5939: 412–3. DOI: 10.1126/science.1173299

Barros, A. P., Silva, P., & Holanda, F. (2007). Exploratory study of space syntax as a traffic assignment tool. *6th International Space Syntax Symposium*, 1–14.

Barthélemy, M. (2011). Spatial networks. *Physics Reports*, 499/1–3: 1–101. DOI: 10.1016/j.physrep.2010.11.002

Batten, D. F. (1995). Network Cities: Creative Urban Agglomerations for the 21st Century. *Urban Studies*, 32/2: 313–27. DOI: 10.1080/00420989550013103

Batty, M. (2005). Cities and complexity: Understanding cities with cellular automata, agent-based models, and fractals. Cambridge: The MIT Press.

Batty, M. (2006). Hierarchy in Cities and City Systems. Pumain D. (ed.) *Hierarchy in Natural and Social Sciences*, Vol. 44, pp. 143–68. Springer-Verlag: Berlin/Heidelberg. DOI: 10.1007/1-4020-4127-6_7

Batty, M. (2007). *Complexity in City Systems: Understanding, Evolution, and Design* (No. 117). Working Paper Series, Vol. 44. London.

Batty, M. (2013). *The New Science of Cities*. Cambridge, Massachusetts: MIT Press.

Batty, M., & Rana, S. (2004). The automatic definition and generation of axial lines and axial maps. *Environment and Planning B: Planning and Design*, 31/4: 615–40. DOI: 10.1068/b2985

Beatty, C., & Fothergill, S. (1996). Labour Market Adjustment in Areas of Chronic Industrial Decline: The Case of the UK Coalfields. *Regional Studies*, 30/7: 627–40. DOI: 10.1080/00343409612331349928

Beatty, C., Fothergill, S., & Powell, R. S. (2007). Twenty Years On: Has the Economy of the UK Coalfields Recovered? *Environment and Planning A*, 39/7: 1654–75. DOI: 10.1068/a38216

Beckmann, M. J. (1986). The location of production activities. Nijkamp P. (ed.) *Handbook of regional and urban economics*, pp. 21–95. Amsterdam.

Berman, M., & Diggle, P. (1989). Estimating Weighted Integrals of the Second-Order Intensity of a Spatial Point Process. *Journal of the Royal Statistical Society. Series B (Methodological)*, 51/1: 81–92.

Biljecki, F., Ledoux, H., & Stoter, J. (2016). An improved LOD specification for 3D building models. *Computers, Environment and Urban Systems*, 59: 25–37. DOI: 10.1016/j.compenvurbsys.2016.04.005

Biljecki, F., Ledoux, H., Stoter, J., & Zhao, J. (2014). Formalisation of the level of detail in 3D city modelling. *Computers, Environment and Urban Systems*, 48: 1–15. Elsevier Ltd. DOI: 10.1016/j.compenvurbsys.2014.05.004

Biljecki, F., Ohori, K. A., Ledoux, H., Peters, R., & Stoter, J. (2016). Population estimation using a 3D City Model: A multi-scale country-wide study in the Netherlands. *PLoS ONE*, 11/6: 1–22. DOI: 10.1371/journal.pone.0156808

Birch, C. P. D., Oom, S. P., & Beecham, J. A. (2007). Rectangular and hexagonal grids used for observation, experiment and simulation in ecology. *Ecological Modelling*, 206/3–4: 347–59. DOI: 10.1016/j.ecolmodel.2007.03.041

Bivand, R. S., Pebesma, E. J., & Gómez-Rubio, V. (2008). *Applied Spatial Data Analysis with R*. New York, NY: Springer New York. DOI: 10.1007/978-0-387-78171-6

Bloch Holst, V. (2011). *GEOSTAT 1A – Representing Census data in a European population grid*. *The European Forum for GeoStatistics*.

Blum, A., Hopcroft, J., & Kannan, R. (2013). Foundations of Data Science, 407. DOI: 10.13140/2.1.5115.0726

Blumenfeld, H. (2007). The Economic Base of the Metropolis: Critical remarks on the 'basic - nonbasic' concept. *Journal of the American Planning Association*, 21/4: 114–32. DOI: 10.1080/01944365508979342

Brinegar, S. J., & Popick, S. J. (2010). A Comparative Analysis of Small Area Population Estimation Methods. *Cartography and Geographic Information Science*, 37/4: 273–84. DOI: 10.1559/152304010793454327

Buja, A., & Eyuboglu, N. (1992). Remarks on Parallel Analysis. *Multivariate Behavioral Research*, 27/4: 509–40. DOI: 10.1207/s15327906mbr2704_2

Camagni, R., & Salone, C. (1993). Network Urban Structures in Northern Italy: Elements for a Theoretical Framework. *Urban Studies*, 30/6: 1053–64. DOI: 10.1080/00420989320080941

Carranza, P. M., & Koch, D. (2013). A Computational Method for Generating Convex Maps. Kim Y. O., Park H. T., & Seo K. W. (eds) *Proceedings of the Ninth International Space Syntax Symposium*, p. 064: 1-11. Seoul: Sejong University.

Castells, M. (2000). *The Rise of the Network Society*, 2nd ed. London: Wiley-Blackwell.

Cermasi, O., & Psarra, S. (2013). Space Syntax , Landscape Urbanism And The Peri-Urban Condition : The case of Bologna and Modena in Italy. Kim Y. O., Park H. T., & Seo K. W. (eds) *Proceedings of the Ninth International Space Syntax Symposium*, pp. 1–18. Seoul: Sejong University.

Chen, G., Esch, G., Wonka, P., Müller, P., & Zhang, E. (2008). Interactive procedural street modeling. *ACM Transactions on Graphics*, 27/3: 1. DOI: 10.1145/1360612.1360702

Christaller, W. (1933a). *Die zentralen Orte in Süddeutschland: Eine ökonomisch-geographische Untersuchung über die Gesetzmäßigkeit der Verbreitung und Entwicklung der Siedlungen mit städtischen Funktionen*, 1st ed. Jena.

Christaller, W. (1933b). *Central Places in Southern Germany.*, Translatio. Englewood Cliffs, New Jersey: Prentice-Hall.

Christian, D. (2004). *Maps of time*. Berkeley: University of California Press.

Clark, P. J., & Evans, F. C. (1954). Distance to Nearest Neighbor as a Measure of Spatial Relationships in Populations. *Ecology*, 35/4: 445–53. DOI: 10.2307/1931034

Coffey, W. J. (1998). Urban Systems Research: Past, Present and Future | An Overview. *Canadian Journal of Regional Science*, 21/3: 327–64.

Costello, A. B., & Osborne, J. W. (2005). Best Practise in Exploratory Factor Analysis: Four Recommendations for Getting the Most From Your Analysis. *Practical Assessment, Research & Evaluation*, 10/7: 1–9.

County Council Derbyshire. (2011). *Analysis of Commuter Patterns in Derbyshire 2011*.

Dall, J., & Christensen, M. (2002). Random geometric graphs. *Physical Review E - Statistical, Nonlinear, and Soft Matter Physics*, 66/1: 1–9. DOI: 10.1103/PhysRevE.66.016121

Dalton, N. S., Peponis, J., & Dalton, R. (2003). To tame a TIGER one has to know its nature: Extending weighted angular integration to the description of GIS road-centerline data for large scale urban analysis. *4th International Space Syntax Symposium*, 1–10.

Dash Nelson, G., & Rae, A. (2016). An Economic Geography of the United States: From Commutes to Megaregions. (J. L. Rosenbloom, Ed.)*PLOS ONE*, 11/11: e0166083. DOI: 10.1371/journal.pone.0166083

Demographia. (2018). *Demographia World Urban Areas & Population Projections*. *Demographia*.

Department for Transport. (2016). *Road Traffic Estimates Methodology Note*. London.

Department for Transport. (2017). *Road Lengths in Great Britain 2016*. London, Great Britain.

Desyllas, J., & Duxbury, E. (2001). Axial Maps and Visibility Graph Analysis. A comparison of their methodology and use in models of urban pedestrian movement. *Proceedings of the Third International Space Syntax Symposium*, p. 27.1-13. Atlanta, USA.

Dhanani, A., Vaughan, L., Ellul, C., & Griffiths, S. (2012). From the axial line to the walked line: Evaluating the utility of commercial and user-generated street network datasets in space syntax analysis. *Proceedings of the Eighth International Space Syntax Symposium*, 1–32. Santiago de Chile: PUC.

Diggle, P. J. (1983). *Statistical Analysis of Spatial Point Patterns*. London: Academic Press.

Diggle, P. J. (2014). *Statistical Analysis of Spatial and Spatio- Temporal Point Patterns*, 3rd ed. Boca Raton, Florida: CRC Press.

Diggle, P. J., & Gratton, R. J. (1984). Monte Carlo Methods of Inference for Implicit Statistical Models. *Journal of the Royal Statistical Society. Series B*, 46/2: 193–227.

Doncaster Council. (2013). *Doncaster's Economic Growth Plan 2013-2018*.

Douglas, D. H., & Peucker, T. K. (1973). Algorithms for the Reduction of the Number of Points Required for Represent a Digitzed Line or its Caricature. *Cartographica (The Canadian Cartographer)*, 10/2: 112–122.

Eaton, B. C. (1972). Spatial Competition Revisited. *The Canadian Journal of Economics*, 5/2: 268–78.

Eaton, B. C., & Lipsey, R. G. (1976). The Non-Uniqueness of Equilibrium in the Loschian Location Model. *American Economic Review*, 66/1: 77–93.

Eaton, B. C., & Lipsey, R. G. (1979). Comparison Shopping and the Clustering of Homogeneous Firms. *Journal of Regional Science*, 19/4: 421–35. DOI: 10.1111/j.1467-9787.1979.tb00610.x

Eisenberg, B. (2007). Calibrating axial line maps. *Proceedings of the Sixth International Space Syntax Symposium*, pp. 1–14. Istanbul, Turkey.

Erdős, P., & Rényi, A. (1959). On random graphs I. *Publicationes Mathematicae (Debrecen)*, 6: 290–7. DOI: 10.2307/1999405

Fabrigar, L. R., & Wegener, D. T. (2011). *Exploratory Factor Analysis*, Online. Oxford: Oxford University Press. DOI: 10.1093/acprof:osobl/9780199734177.001.0001

Fan, H., Zipf, A., & Fu, Q. (2014). Estimation of Building Types on OpenStreetMap Based on Urban Morphology Analysis. *Connecting a Digital Europe Through Location and Place*, pp. 19–35. DOI: 10.1007/978-3-319-03611-3_2

Figueiredo, L., & Amorim, L. (2007). Decoding The Urban Grid: Or Why Cities Are Neither Trees Nor Perfect Grids. *6th International Space Syntax Symposium*, p. 006:1-006:16. Istanbul, Turkey. DOI: 10.1017/CBO9781107415324.004

Fik, T. J. (1988). Spatial Competition and Price Reporting in Retail Food Markets. *Economic Geography*, 64/1: 29–44.

Fik, T. J., & Mulligan, G. F. (1990). Spatial flows and competing central places: Towards a general theory of hierarchical interaction. *Environment and Planning A*, 22/4: 527–49. DOI: 10.1068/a220527

Flanagin, A. J., & Metzger, M. J. (2008). The credibility of volunteered geographic information. *GeoJournal*, 72/3–4: 137–48. DOI: 10.1007/s10708-008-9188-y

Florinsky, I. V., & Kuryakova, G. A. (2000). Determination of grid size for digital terrain modelling in landscape investigations—exemplified by soil moisture distribution at a micro-scale. *International Journal of Geographical Information Science*, 14/8: 815–32. DOI: 10.1080/136588100750022804

Fotheringham, A. S. (1983). A New Set of Spatial Interaction Models: The Theory of Competing Destinations. *Environment and Planning A*, 15/1: 15–36.

Freeman, L. C. (1977). A Set of Measures of Centrality Based on Betweenness. *Sociometry*. DOI: 10.2307/3033543

Friedmann, J. (1986). The World City Hypothesis. *Development and Change*, 17/1: 69–83. DOI: 10.1111/j.1467-7660.1986.tb00231.x

Gao, S., Wang, Y., Gao, Y., & Liu, Y. (2013). Understanding urban traffic-flow characteristics: A rethinking of betweenness centrality. *Environment and Planning B: Planning and Design*, 40/1: 135–53. DOI: 10.1068/b38141

Gatrell, A. C., Bailey, T. C., Diggle, P. J., & Rowlingson, B. S. (1996). Spatial point pattern analysis and its application in geographical epidemiology. *Transactions of the Institute of British Geographers*, 21: 256–74. DOI: 10.2307/622936

Gerke, S., Schlatter, D., Steger, A., & Taraz, A. (2008). The Random Planar Graph Process. *Random Structures and Algorithms*, 32/2: 236–61. DOI: 10.1002/rsa.20186

Ghorbani, M. (2013). Cauchy cluster process. *Metrika*, 76: 697–706. DOI: 10.1007/s00184-012-0411-y

Gil, J. (2013). Analysing the configuration of integrated multi-modal urban network. *Geographical Analysis*, 46: 368–91. DOI: 10.1111/gean.12062

Gil, J. (2015). Examining 'Edge Effects': Sensitivity of Spatial Network Centrality Analysis to Boundary Conditions. Kayvan K., Vaughan L., Sailer K., Palaiologou G., & Bolton T. (eds) *Proceedings of the 10th International Space Syntax Symposium*, p. 147:1-147:16. London: Space Syntax Laboratory, The Bartlett School of Architecture, University College London.

Glaeser, E. L., & Maré, D. C. (2001). Cities and Skills. *Journal of Labor Economics*, 19/2: 316–42. DOI: 10.1086/319563

Goodchild, M. F. (2007). Citizens as sensors: The world of volunteered geography. *GeoJournal*, 69/4: 211–21. DOI: 10.1007/s10708-007-9111-y

Green, N. E. (1956). Aerial Photographic Analysis of Residential Neighborhoods: An Evaluation of Data Accuracy. *Social Forces*, 35/2: 142–7. DOI: 10.2307/2573361

Griffiths, S. (2011). Temporality in Hillier and Hanson's Theory of Spatial Description: Some Implications Of Historical Research For Space Syntax. *The Journal of Space Syntax*, 2/1: 73–96.

Griffiths, S., Vaughan, L., Haklay, M. (Muki), & Emma Jones, C. (2008). The Sustainable Suburban High Street: A Review of Themes and Approaches. *Geography Compass*, 2/4: 1155–88. DOI: 10.1111/j.1749-8198.2008.00117.x

Gröger, G., Kolbe, T., Nagel, C., & Häfele, K.-H. (2012). *OGC City Geography Markup Language (CityGML) Encoding Standard. OGC 12-019*. Retrieved from <https://portal.opengeospatial.org/files/?artifact_id=47842>. DOI: OGC 12-019

Haggett, P. (1965). *Locational Analysis in Human Geography*, 1st ed. London: Edward Arnold (Publishers) Ltd.

Haggett, P., & Chorley, R. J. (1967). *Models in Geography*. London: Methuen.

Haggett, P., & Chorley, R. J. (1969). *Network Analysis in Geography*, 1st ed. London: Edward Arnold (Publishers) Ltd.

Haining, R. (1983). Modeling Intraurban Price Competition: an Example of Gasoline Pricing. *Journal of Regional Science*, 23/4: 517–28. DOI: 10.1111/j.1467-9787.1983.tb01007.x

Haklay, M. (2009). OpenStreetMap and Ordnance Survey Meridian 2 – Progress maps. Retrieved June 10, 2016, from <https://povesham.wordpress.com/2009/11/14/openstreetmap-and-ordnance-survey-meridian-2-progress-maps/>

Haklay, M. (2010). How good is volunteered geographical information? A comparative study of OpenStreetMap and ordnance survey datasets. *Environment and Planning B: Planning and Design*, 37/4: 682–703. DOI: 10.1068/b35097

Haklay, M., Basiouka, S., Antoniou, V., & Ather, A. (2010). How Many Volunteers Does it Take to Map an Area Well? The Validity of Linus' Law to Volunteered Geographic Information. *The Cartographic Journal*, 47/4: 315–22. DOI: 10.1179/000870410X12911304958827

Hall, P. (1999). Planning for the Mega-City: A New Eastern Asian Urban Form? Brotchie J., Newton P., Hall P., & Dickey J. (eds) *East West Perspectives on 21st Century Urban Development: Sustainable Eastern and Western Cities in the New Millennium*, pp. 3–36. Ashgate: Aldershot.

Hall, P., & Pain, K. (2006). *The Polycentric Metropolis: Learning from Mega-City Regions in Europe*. (P. Hall & K. Pain, Eds). London: Earthscan.

Hall, P., & Pain, K. (2012). *The polycentric metropolis: Learning from mega-city regions in Europe. The Polycentric Metropolis: Learning from Mega-City Regions in Europe*. DOI: 10.4324/9781849773911

Hanna, S. (2009). Spectral Comparison of Large Urban Graphs. Koch D., Marcus L., & Steen J. (eds) *Proceedings of the 7th International Space Syntax Symposium*, p. 039:1-039:11. Stockholm: KTH Royal Institute of Technology.

Hanna, S., Serras, J., & Varoudis, T. (2013). Measuring the structure of global transportation networks. Kim Y. O., Park H.-T., & Seo K. W. (eds) *Ninth International Space Syntax Symposium*. Seoul: Sejong University.

Harvey, D. (1969). *Explanation in Geography*. London: Edward Arnold.

Hengl, T. (2006). Finding the right pixel size. *Computers and Geosciences*, 32/9: 1283–98. DOI: 10.1016/j.cageo.2005.11.008

Henn, A., Römer, C., Gröger, G., & Plümer, L. (2012). Automatic classification of building types in 3D city models. *GeoInformatica*, 16/2: 281–306. DOI: 10.1007/s10707-011-0131-x

Hill, A. (2000). *The South Yorkshire Coalfield: A History and Development*. The History Press.

Hillier, B. (1996a). *Space is the Machine: a configurational theory of architecture*. Cambridge: Cambridge University Press.

Hillier, B. (1996b). Cities as movement economies. *URBAN DESIGN International*, 1/1: 41–60. Cambridge.

Hillier, B. (1999a). Centrality as a process: accounting for attraction inequalities in deformed grids. *Urban Design International*, 4/3–4: 107–27. Routledge. DOI: 10.1080/135753199350036

Hillier, B. (1999b). The hidden geometry of deformed grids. *Environment and Planning B: Planning and Design*, 26: 169–91.

Hillier, B. (2002). A Theory of the City as Object: or, how spatial laws mediate the social construction of urban space. *Urban Design International*, Vol. 3–4, pp. 153–79. Atlanta, USA.

Hillier, B. (2004). *Space is the machine*. Cambridge: Cambridge University Press.

Hillier, B. (2008). Space and spatiality: what the built environment needs from social theory, 3218/November 2013: 37–41. DOI: https://doi.org/10.1080/09613210801928073

Hillier, B. (2009). Spatial Sustainability in Cities: Organic Patterns and Sustainable Forms. Koch D., Marcus L., & Steen J. (eds) *Proceedings of the 7th International Space Syntax Symposium*, pp. 1–20. Stockholm, Sweden: KTH.

Hillier, B. (2014). The Generic City and its Origins. *Architectural Design*, 84/5: 100–5. DOI: 10.1002/ad.1815

Hillier, B., & Hanson, J. (1984). *The Social Logic of Space*. Cambridge: Cambridge University Press.

Hillier, B., & Iida, S. (2005). Network and psychological effects: a theory of urban movement, 475–90.

Hillier, B., Leaman, A., Stansall, P., & Bedford, M. (1976). Space Syntax. *Environment and Planning Series B*, 3: 147–85.

Hillier, B., Penn, A., Hanson, J., Grajewski, T., & Xu, J. (1993a). Natural Movement: or configuration and attraction in urban pedestrian movement. *Environment and Planning B: Planning and Design*, 20: 29–66.

Hillier, B., Penn, A., Hanson, J., Grajewski, T., & Xu, J. (1993b). Natural Movement: or, configuration and attraction in urban pedestrian movement. *Environment and Planning B: Planning and Design*, 20: 29–66.

Hillier, B., & Vaughan, L. (2007). The City as One Thing. *Progress in Planning*, 3/67: 205–30.

Hillier, B., Yang, T., & Turner, A. (2012). Normalising least angle choice in Depthmap and how it opens up new perspectives on the global and local analysis of city space. *The Journal of Space Syntax*, 3/2: 155–99.

Hirst, P., Thompson, G., & Bromley, S. (2009). *Globalization in Question*, 3rd ed. Cambridge, UK / Malden, MA, USA: Polity.

Horn, J. L. (1965). A rationale and test for the number of factors in factor analysis. *Psychometrika*, 30/2: 179–85. DOI: 10.1007/BF02289447

Hotelling, H. (1929). Stability in Competition. *The Economic Journal*, 39/153: 41–57.

Huang, L., Zhu, X., Ye, X., Guo, W., & Wang, J. (2016). Characterizing street hierarchies through network analysis and large-scale taxi traffic flow: a case study of Wuhan, China. *Environment and Planning B: Urban Analytics and City Science*, 43/2: 276–96. DOI: 10.1177/0265813515614456

Illian, J., Penttinen, A., Stoyan, H., & Stoyan, D. (2008). *Statistical Analysis and Modelling of Spatial Point Patterns*. Chichester, England: John Wiley & Sons Ltd. DOI: 10.1111/j.1751-5823.2008.00062_23.x

Jacobs, J. (1961). *The Death and Life of Great American Cities. New York*, Vol. 71. New York: Vintage Books. DOI: 10.2307/794509

Jalilian, A., Guan, Y., & Waagepetersen, R. (2013). Decomposition of Variance for Spatial Cox Processes. *Scandinavian Journal of Statistics*, 40: 119–37. DOI: 10.1111/j.1467-9469.2012.00795

Javadi, A., Emo, B., Howard, L. R., Zisch, F. E., Yu, Y., Knight, R., Silva, J. P., et al. (2017). Hippocampal and prefrontal processing of network topology to simulate the future. *Nature Communications*, 8: 1–11. Nature Publishing Group. DOI: 10.1038/ncomms14652

Jayasinghe, A., Sano, K., & Nishiuchi, H. (2015). EXPLAINING TRAFFIC FLOW PATTERNS USING CENTRALITY MEASURES. *INTERNATIONAL JOURNAL FOR TRAFFIC AND TRANSPORT ENGINEERING*, 5/2: 134–49. DOI: 10.7708/ijtte.2015.5(2).05

Jefferson, M. (1939). The Law of the Primate City. *Geographical Review*, 29/2: 226–32.

Jennings, M., & Lewis, R. (2017). *House price per square metre and house price per room , England and Wales : 2004 to 2016*. London. Retrieved from <https://www.ons.gov.uk/economy/inflationandpriceindices/datasets/housepricepersquaremetreandhousepriceperroomenglandandwales>

Jensen-Butler, C. (1972). Nearest neighbour analysis of a central place system. *Tijoschrift voor Econ. en Soc. Geografie*, September: 353–9.

Jiang, B. (2009). Street Hierarchies: A Minority of Streets Account for a Majority of Traffic Flow. *International Journal of Geographical Information Science*, 23/8: 1033–48. DOI: 10.1080/13658810802004648

Jiang, B., & Liu, C. (2009). Street-based topological representations and analyses for predicting traffic flow in GIS. *International Journal of Geographical Information Science*, 23/9: 1119–37. DOI: 10.1080/13658810701690448

Karimi, K., Parham, E., & Acharya, A. (2015). Integrated sub-regional planning informed by weighted spatial network models : The case of Jeddah sub-regional system. Karimi K., Vaughan L., Sailer K., Palaiologou G., & Bolton T. (eds) *Proceedings of the 10th International Space Syntax Symposium*, p. 71:1-71:16. London: Space Syntax Laboratory, The Bartlett School of Architecture, University College London.

Kathmann, T., Ziegler, H., & Thomas, B. (2009). *Staßenverkehrszählung 2005 Methodik - Verkehrstechnik Heft V 179*. Bergisch Gladbach.

Keil, A., & Wetterau, B. (2013). *Metropolis Ruhr. A Regional Study of the New Ruhr*. (Regionalverband Ruhr Department for Strategic Development and Communication, Ed.). Essen.

Kersting, V. (2009). Ruhratlas. *Atlas der Metropole Ruhr. Vielfalt und Wandel des Ruhrgebiets im Kartenbild*, pp. 142–5. Emons: Köln.

Keßler, C., & de Groot, R. T. A. (2013). Trust as a Proxy Measure for the Quality of Volunteered Geographic Information in the Case of OpenStreetMap. *Lecture Notes in Geoinformation and Cartography*, pp. 21–37. DOI: 10.1007/978-3-319-00615-4_2

Kloosterman, R., & Lambregts, B. (2001). Clustering of Economic Activities in Polycentric Urban Regions: The Case of the Randstad. *Urban Studies*, 38/4: 717–32. DOI: 10.1080/00420980120035303

Kraus, S., Senger, L., & Ryerson, J. (1974). Estimating Population from Photographically Determined Residential Land Use Types. *Remote sensing of Environment*, 42/5: 35–42.

Krenz, K. (2015). Capturing Patterns of Shrinkage and Growth in Post-Industrial Regions: A Comparative Study of the Ruhr Valley and Leipzig-Halle. Karimi K., Vaughan L., Sailer K., Palaiologou G., & Bolton T. (eds) *Proceedings of the 10th International Space Syntax Symposium*, p. 72:1-72:18. London: Space Syntax Laboratory, The Bartlett School of Architecture, University College London.

Krenz, K., Kostourou, F., Psarra, S., & Capille, C. (2015). Understanding the City as a Whole: An Integrative Analysis of Rio de Janeiro and its Informal Settlements. *ISUF 2015 XXII international Conference: City as organism. New visions for urban life*, 647–60.

Krugman, P. (1991). Increasing Returns and Economic Geography. *Journal of Political Economy*, 99/3: 483–99.

Kunze, C., & Hecht, R. (2015). Semantic enrichment of building data with volunteered geographic information to improve mappings of dwelling units and population. *Computers, Environment and Urban Systems*, 53: 4–18. Elsevier Ltd. DOI: 10.1016/j.compenvurbsys.2015.04.002

Lam, N. S.-N., & Quattrochi, D. A. (1992). On the Issues of Scale, Resolution, and Fractal Analysis in the Mapping Sciences. *Professional Geographer*, 44/1: 88–98.

Law, S., & Versluis, L. (2015). How do UK regional commuting flows relate to spatial configuration ? Karimi K., Vaughan L., Sailer K., Palaiologou G., & Bolton T. (eds) *Proceedings of the 10th International Space Syntax Symposium*, p. 74:1-74:21. London: Space Syntax Laboratory, The Bartlett School of Architecture, University College London.

Lensing, N. (2013). *Staßenverkehrszählung 2010 Methodik - Verkehrstechnik Heft V 234*. Aachen.

Lieshout, M. N. M., & Baddeley, A. J. (1996). A nonparametric measure of spatial interaction in point patterns. *Statistica Neerlandica*, 50/3: 344–61. DOI: 10.1111/j.1467-9574.1996.tb01501.x

Lilliefors, H. W. (1967). On the Kolmogorov-Smirnov Test for Normality with Mean and Variance Unknown. *Journal of the American Statistical Association*, 62/318: 399–402.

Loader, C. (1999). *Local Regression and Likelihood*. Springer New York.

Lösch, A. (1940). *Die räumliche Ordnung der Wirtschaft. Eine Untersuchung über Standort, Wirtschaftsgebiete und internationalem Handel.* Jena: Fischer. DOI: 10.2307/2573862

Ludwig, I., Voss, A., & Krause-Traudes, M. (2011). A Comparison of the Street Networks of Navteq and OSM in Germany. Geertman S., Reinhardt W., & Toppen F. (eds) *Advancing Geoinformation Science for a Changing World*, Lecture Notes in Geoinformation and Cartography, Vol. 1, pp. 65–84. Springer Berlin Heidelberg: Berlin, Heidelberg. DOI: 10.1007/978-3-642-19789-5_4

MacLeod, G., & Jones, M. (2011). Renewing Urban Politics. *Urban Studies*, 48/12: 2443–72. DOI: 10.1177/0042098011415717

Marshall, S., & Sutton, M. (2013). Explorations in Generative Street Layouts. Carmona M. (ed.) *Explorations in Urban Design: An Urban Design Research Primer*, pp. 181–92. Ashgate: Farnham.

Marston, S. A. (2000). The social construction of scale. *Progress in Human Geography*, 24/2: 219–42. DOI: 10.1191/030913200674086272

Marston, S. A., Jones, J. P., & Woodward, K. (2005). Human geography without scale. *Transactions of the Institute of British Geographers*, 30/4: 416–32. DOI: 10.1111/j.1475-5661.2005.00180.x

Massey, D. B. (1984). *Spatial divisions of labor: Social structures and the geography of production*. New York: Routledge.

Masucci, A. P., Smith, D., Crooks, A., & Batty, M. (2009). Random planar graphs and the London street network. *European Physical Journal B*, 71/2: 259–71. DOI: 10.1140/epjb/e2009-00290-4

Masucci, A. P., Stanilov, K., & Batty, M. (2013). Limited Urban Growth: London's Street Network Dynamics since the 18th Century. *PLoS ONE*, 8/8: 1–10. DOI: 10.1371/journal.pone.0069469

McLafferty, S., & Gosh, A. (1984). Model Multipurpose. *Geographical Analysis*, 16/3: 244–9.

Meentemeyer, V. (1989). Geographical perspectives of space , time , and scale, 3: 163–73.

Meijers, E. (2005). Polycentric urban regions and the quest for synergy: Is a network of cities more than the sum of the parts? *Urban Studies*, 42/4: 765–81. DOI: 10.1080/00420980500060384

Meijers, E. (2007). *Synergy in Polycentric Urban Regions*. DOI: 10.1017/CBO9781107415324.004

Mitzenmacher, M. (2003). A brief history of generative models for power law and lognormal distributions A brief history of generative models for power law and lognormal distributions. *Internet Mathematics*, 1/2: 226–51. DOI: 10.1080/15427951.2004.10129088

Mohajeri, N., Poursistany, P., Poursistany, P., & Gudmundsson, A. (2013). Quantitative Analysis of Structural Changes during Rapid Urban Growth : Case Study of Kerman , Iran. *American Society of Civil Engineers*, 141/3: 1–10. DOI: 10.1061/(ASCE)UP.1943-5444.0000213.

Mohamed, A. A., van Nes, A., A.Salheen, M., Kohlert, C., & Schwander, C. (2013). The Socio-Economic Implications of the Spatial Configuration in Greater Cairo Metropolitan Area. *Proceedings of the Ninth International Space Syntax Symposium*, p. 095.1-095.18. Seoul: Sejong University.

Møller, J., & Waagepetersen, R. P. (2004). *Statistical inference and simulation for spatial point processes*, Vol. 23. Boca Raton: CRC Press. DOI: 10.1201/9780203496930

Morrill, R. L. (1970). *The Spatial Organization of Society*. Belmont, California: Duxbury Press.

Mulligan, G. F., Partridge, M. D., & Carruthers, J. I. (2012). Central place theory and its reemergence in regional science. *Annals of Regional Science*, 48/2: 405–31. DOI: 10.1007/s00168-011-0496-7

Münter, A. (2011). Germany's Polycentric Metropolitan Regions in the World City Network. *Raumforschung und Raumordnung*, 69/3: 187–200. DOI: 10.1007/s13147-011-0090-6

Nations, U. (2013). *World Population Prospects: The 2012 Revision, Volume II, Demographic Profiles*, Vol. II.

Neal, Z. P. (2011). From Central Places to Network Bases: A Transition in the U.S. Urban Hierarchy, 1900-2000. *City & Community*, 10/1: 49–75. DOI: 10.1111/j.1540-6040.2010.01340.x

Neis, P., Zielstra, D., Zip, A., & Strunck, A. (2010). Empirische Untersuchungen zur Datenqualität von OpenStreetMap - Erfahrungen aus zwei Jahren Betrieb mehrerer OSM-Online-Dienste. *Angewandte Geoinformatik 2010*, p. 6. Salzburg, Austria.

Neis, P., Zielstra, D., & Zipf, A. (2011). The Street Network Evolution of Crowdsourced Maps: OpenStreetMap in Germany 2007–2011. *Future Internet*, 4/1: 1–21. DOI: 10.3390/fi4010001

van Nes, A. (2007). Centrality and Economic Development in the Rijnland Region: social and spatial concepts of centrality. *Proceedings of the Sixth*

International Space Syntax Symposium6th International Space Syntax Symposium, p. 015.1-16. Istanbul, Turkey.

van Nes, A. (2009). Analysing Larger Metropolitan Areas. On Identification Criteria for Middle Scale Networks. Koch D., Marcus L., & Steen J. (eds) *Proceedings of the Seventh International Space Syntax Symposium International Space Syntax Symposium*, p. 121.1-13. Stockhilm: KTH Royal Institute of Technology.

Neyman, J., & Scott, E. L. (1958). Statistical Approach to Problems of Cosmology. *Journal of the Royal Statistical Society. Series B*, 20/1: 1–43.

North, J., & Spooner, D. J. (1978). The Geography of the Coal Industry in the United Kingdom in the 1970s: Changing Directions? *GeoJournal*, 2/3: 255–72.

Nottingham Economic Strategy Research Bureau. (2012). *Nottingham City Economic Review*.

O'Reilly, T. (2005). What Is Web 2.0: Design Patterns and Business Models for the Next Generation of Software. *O'Reilly Media, Inc.* Retrieved March 1, 2016, from <http://www.oreilly.com/pub/a//web2/archive/what-is-web-20.html>

OpenStreetMap Wiki contributors. (2016). Editing Standards and Conventions. *OpenStreetMap Wiki*. Retrieved June 3, 2016, from <http://wiki.openstreetmap.org/wiki/Editing_Standards_and_Conventions >

OpenStreetMap Wiki contributors. (2017). Key:highway. *OpenStreetMap Wiki*. Retrieved January 30, 2017, from <http://wiki.openstreetmap.org/w/index.php?title=Key:highway&oldid=13 32527>

Ordnance Survey. (2013). *AddressBase products Classification scheme*.

Orford, S., & Radcliffe, J. (2007). Modelling UK residential dwelling types using OS Mastermap data: A comparison to the 2001 census. *Computers, Environment and Urban Systems*, 31/2: 206–27. DOI: 10.1016/j.compenvurbsys.2006.08.003

Parish, Y. I. H., & Müller, P. (2001). Procedural Modeling of Cities. *28th Annual Conference on Computer Graphics and Interactive Techniques*, pp. 301–8. DOI: 10.1145/383259.383292

Parr, J. (2004). The Polycentric Urban Region: A Closer Inspection. *Regional Studies*, 38/3: 231–40. DOI: 10.1080/003434042000211114

Parr, J. (2008). Cities and regions: Problems and potentials. *Environment and Planning A*, 40/12: 3009–26. DOI: 10.1068/a40217

Parr, J., & Suzuki, K. (1973). Settlement Populations and the Lognormal Distribution. *Urban Studies*, 10/3: 335–52. DOI: 10.1080/00420987320080471

Patterson, J. L., & Jones, P. (2016). Traffic modelling in cities - Validation of space syntax at an urban scale. *Indoor and Built Environment*, 25/7: 1163–78. DOI: 10.1177/1420326X16657675

Penn, A. (2003). Space Syntax And Spatial Cognition. *Environment and Behavior*, 35/1: 30–65. DOI: 10.1177/0013916502238864

Penn, A., Conroy, R., Dalton, N., Dekker, L., Mottram, C., & Turner, A. (1997). Intelligent Architecture: New Tools for the Three Dimensional Analysis of Space and Built Form. *Proceedings of the First International Space Syntax Symposium*, Vol. 2, p. 30.1-19. London: The Bartlett School of Architecture.

Penn, A., Hillier, B., Banister, D., & Xu, J. (1998a). Configurational modelling of urban movement networks. *Environment and Planning B: Planning and Design*, 25/1: 59–84. DOI: 10.1068/b250059

Penn, A., Hillier, B., Banister, D., & Xu, J. (1998b). Configurational modelling of urban movement networks. *Environment and Planning B: Planning and Design*, 25/1: 59–84. DOI: 10.1068/b250059

Peponis, J., Allen, D., Haynie, D., Scoppa, M., & Zhang, Z. (2007). Measuring The Configuration Of Street Networks. *Proceedings of the Sixth International Space Syntax Symposium*, p. 002:1–002:16. Istanbul, Turkey.

Peponis, J., Wineman, J., Bafna, S., Rashid, M., & Kim, S. H. (1998). On the generation of linear representations of spatial configuration. *Environment and Planning B: Planning and Design*, 25/4: 559–76. DOI: 10.1068/b250559

Peponis, J., Wineman, J., Rashid, M., Kim, S. H., & Bafna, S. (1997). On the description of shape and spatial configuration inside buildings: convex partitions and their local properties. *Environment and Planning B*, 24: 761–81.

Petrov, A. (2012). One Hundred Years of Dasymetric Mapping: Back to the Origin. *The Cartographic Journal*, 49/3: 256–64. DOI: 10.1179/1743277412Y.0000000001

Psarra, S., Kickert, C., & Pluviano, A. (2013). Paradigm lost: Industrial and post-industrial Detroit – An analysis of the street network and its social and economic dimensions from 1796 to the present. *URBAN DESIGN International*, 1–25. Nature Publishing Group. DOI: 10.1057/udi.2013.4

Qiu, F., Zhang, C., & Zhou, Y. (2012). The Development of an Areal Interpolation ArcGIS Extension and a Comparative Study. *GIScience and Remote Sensing*, 49/5: 644–63. DOI: 10.2747/1548-1603.49.5.644

Quattrochi, D. A., & Goodchild, M. F. (Eds). (1997). *Scale in Remote Sensing and GIS*. Boca Raton,: Lewis Publishers.

R Development Core Team. (2016). R: A language and environment for statistical computing. Vienna, Austria: R Foundation for Statistical Computing.

Ratti, C. (2004). Space syntax:some inconsistencies. *Environment and Planning B: Planning and Design*, 31: 487–99.

Read, S. (2013). Intensive urbanisation: Levels, networks and central places. *The Journal of Space Syntax*, 4/1: 1–17.

Reed, W. J., & Jorgensen, M. (2004). The Double Pareto-Lognormal Distribution—A New Parametric Model for Size Distributions. *Communications in Statistics - Theory and Methods*, 33/8: 1733–53. DOI: 10.1081/STA-120037438

Reicher, C., Zimmermann, K., Schwarze-Rodian, M., Jansen, H., Schlisio, K., Wagner, A., Polivka, J., et al. (2015). *Polycentric City Regions in Transformation – The Ruhr Agglomeration in International Perspective*. Dortmund.

Rotherham Council. (2008). *The Economic Plan for Rotherham*.

Sabidussi, G. (1966). The centrality index of a graph. *Psychometrika*, 31/4: 581–603. DOI: 10.1007/BF02289527

Salheen, M., & Forsyth, L. (2001). Addressing distance in the space syntax syntactical model, 93–110.

Sassen, S. (1991). The Global City: New York, London, Tokyo, p. 480. Princeton University Press: Princeton, New Jersey.

Sassen, S. (2005). The Global City : introducing a Concept. *The Brown Journal of World Affairs*, XI/2: 27–43.

Scott, A. J. (1999). The Geographic Foundations of Industrial Performance. *The Dynamic Firm*, pp. 384–401. Oxford University Press. DOI: 10.1093/0198296045.003.0016

Scott, A. J. (Ed.). (2002). *Global City-Regions: Trends, Theory, Policy*. Oxford: Oxford University Press.

Sehra, S. S., Singh, J., & Rai, H. S. (2013). Assessment of OpenStreetMap Data - A Review, 76/16: 17–20. DOI: 10.5120/13331-0888 10.5120/13331-0888

Senaratne, H., Mobasheri, A., Ali, A. L., Capineri, C., & Haklay, M. (Muki). (2016). A review of volunteered geographic information quality assessment methods. *International Journal of Geographical Information Science*, 8816/August: 1–29. DOI: 10.1080/13658816.2016.1189556

Serra, M., Hillier, B., & Karimi, K. (2015). Exploring countrywide spatial systems: Spatio-structural correlates at the regional and national scales. Karimi K., Vaughan L., Sailer K., Palaiologou G., & Bolton T. (eds) *Proceedings of the 10th International Space Syntax Symposium*, p. 84.1-84.18. London: Space Syntax Laboratory, The Bartlett School of Architecture, University College London.

Serra, M., & Pinho, P. (2013). Tackling the structure of very large spatial systems - Space syntax and the analysis of metropolitan form. *Journal of Space Syntax*, 4/2: 179–96.

Shapiro, S. S., & Wilk, M. B. (1965). An Analysis of Variance Test for Normality (Complete Samples). *Biometrika*, 52/3/4: 591. DOI: 10.2307/2333709

Sieverts, T. (1997). *Zwischenstadt zwischen Ort und Welt, Raum und Zeit, Stadt und Land. Bauwelt Fundamente*, 3. Berlin: Bauverlag BV.

Smith, D. A., & Crooks, A. T. (2010). *From Buildings to Cities: Techniques for the Multi-Scale Analysis of Urban Form and Function* (No. 155). Working Paper Series. London.

Smith, G., Gidlow, C., Davey, R., & Foster, C. (2010). What is my walking neighbourhood? A pilot study of English adults' definitions of their local walking neighbourhoods. *International Journal of Behavioral Nutrition and Physical Activity*, 7/1: 34. DOI: 10.1186/1479-5868-7-34

Soja, E. W. (1989). *Postmodern Geographies: The Reassertion of Space in Critical Social Theory* .

Soja, E. W. (2011). Regional Urbanization and the End of the Metropolis Era. Bridge G. & Watson S. (eds) *The New Blackwell Companion to the City*, 1st ed., pp. 679–89. DOI: 10.1002/9781444395105.ch59

Steger, M. (2013). Globalization and history: is globalization a new phenomenon? *Globalization*, 3rd ed., pp. 17–36. Oxford University Press. DOI: 10.1093/actrade/9780199662661.003.0002

Strohmeier, K. P., Gehne, D. H., Kurtenbach, S., & Förster, M. (2015). *Kleinräumige Segregationeffekte von Bergwerkstillegungsen im Ruhrgebiet am Besipiel ausgewählter Kommunen*. Bochum.

Strohmeier, K. P., & Kersting, V. (1992). Sozialraum Ruhrgebiet - Stadträumliche Differenzierungen von Lebenslagen, Armut und informelle Solidarpotentiale. *Das Ruhrgebiet - ein starkes Stück Nordrhein-Westfalen*, pp. 451–75. Essen.

Strohmeier, K. P., Neubauer, J., & Prey, G. (2002). Demografischer Wandel im Ruhrgebiet - Bevölkerungsentwicklung und Sozialraumstruktur im Ruhrgebiet, September.

Sun, J., Yu, X., Baciu, G., & Green, M. (2002). Template-based generation of road networks for virtual city modeling. *Proceedings of the ACM symposium on Virtual reality software and technology - VRST '02*, p. 33. New York, New York, USA: ACM Press. DOI: 10.1145/585740.585747

Swyngedouw, E. (2004). Scaled Geographies: Nature, Place, and the Politics of Scale. Sheppard E., Mcmaster R. B., & Swyngedouw E. (eds) *Scale and Geographic Inquiry: Nature, Society, and Method*, pp. 129–153. Blackwell Publishing Ltd. DOI: http://dx.doi.org/10.1002/9780470999141.ch7

Taylor, P. J. (2004). *World City Networks, A Global Urban Analysis*. New York: Routledge.

Thill, J. (1986). A Note on Multipurpose and Multistop Shopping, Sales, and Market Areas of Firms. *Journal of Regional Science*, 26/4: 775–84. DOI: 10.1111/j.1467-9787.1986.tb01074.x

Thomson, R. (2003). Bending the axial line: smoothly continuous road centre-line segments as a basis for road network analysis. *Proceedings of the Fourth International Space Syntax Symposium*, p. 10. London.

von Thünen, J. H. (1826). *Der isolierte Staat in Beziehung auf Landwirtschaft und Nationalökonomie*, 1st ed. Stuttgart.

Tomás, L., Fonseca, L., Almeida, C., Leonardi, F., & Pereira, M. (2016). Urban population estimation based on residential buildings volume using IKONOS-2 images and lidar data. *International Journal of Remote Sensing*, 37/sup1: 1–28. Taylor & Francis. DOI: 10.1080/01431161.2015.1121301

Trippl, M., Maier, G., & Tödtling, F. (2012). *Regional- und Stadtökonomik 2: Regionalentwicklung und Regionalpolitik*, Vol. 2. Springer-Verlag.

Turner, A. (2000). *Angular Analysis: A Method For The Quantification Of Space* (No. 23). Working Paper Series. London.

Turner, A. (2001). Angular Analysis. *Proceedings of the Third International Space Syntax Symposium*, p. 30.1-30.11. Atlanta.

Turner, A. (2005). Could A Road-centre Line Be An Axial Line In Disguise? *Proceedings 5th International Space Syntax Symposium*, 145–59.

Turner, A. (2007). From axial to road-centre lines: a new representation for space syntax and a new model of route choice for transport network analysis. *Environment and Planning B: Planning and Design*, 34/3: 539–55. DOI: 10.1068/b32067

Turner, A. (2009). Stitching Together the Fabric of Space and Society: An Investigation into the Linkage of the Local to Regional Continuum. Daniel K., Marcus L., & Steen J. (eds) *Proceedings of the Seventh International Space Syntax Symposium*, pp. 1–12. Stockholm: KTH Royal Institute of Technology.

Turner, A., Penn, A., & Hillier, B. (2005). An algorithmic definition of the axial map. *Environment and Planning B: Planning and Design*, 32/3: 425–44. DOI: 10.1068/b31097

Ugalde, C. M. de, Fujita, C., Bauermann, C. N., & Jobim, G. M. F. (2015). Identifying city-regional structures in Rio Grande do Sul, Brazil. Karimi K., Vaughan L., Sailer K., Palaiologou G., & Bolton T. (eds) *Proceedings of the 10th International Space Syntax Symposium*, p. 86:1-86:14. London: Space Syntax Laboratory, The Bartlett School of Architecture, University College London.

Varoudis, T., Law, S., Karimi, K., Hillier, B., & Penn, A. (2013). Space Syntax Angular Betweenness Centrality Revisited. Kim Y. O., Park H. T., & Seo K. W. (eds) *Proceedings of the Ninth International Space Syntax Symposium*, pp. 1–16. Seoul: Sejong University.

Waagepetersen, R. P. (2007). An estimating function approach to inference for inhomogeneous Neyman-Scott processes. *Biometrics*, 63/1: 252–8. DOI: 10.1111/j.1541-0420.2006.00667.x

Watson, M. K. (1978). The Scale Problem in Human Geography. *Grafiska Annaler*, 60B/1: 36–47.

West, D. S., & Von Hohenbalken, B. (1984). Spatial Predation in a Canadian Retail Oligopoly. *Journal of Regional Science*, 24/3: 415–27. DOI: 10.1111/j.1467-9787.1984.tb00812.x

White, R. W. (1974). Sketches of a Dynamic Central Place Theory. *Economic Geography*, 50/3: 219–27. DOI: 10.2307/142860

White, R. W. (1977). Dynamic Central Place Theory: Results of a Simulation Approach. *Geographical Analysis*, 9/3: 226–43. DOI: 10.1111/j.1538-4632.1977.tb00576.x

White, R. W. (1978). The Simulation of Central Place Dynamics: Two-Sector Systems and the Rank-Size Distribution. *Geographical Analysis*, 10/2: 201–8. DOI: 10.1111/j.1538-4632.1978.tb00011.x

Widaman, K. (2007). Common Factors Versus Components: Principals and Principles, Errors and Misconceptions. Cudeck R. & MacCallum R. C. (eds) *Factor Analysis at 100: Historical Developments and Future Directions*, pp. 177–204. Lawrence Erlbaum Associates, Publishers: Mahwah, New Jersey.

Wiegand, T., Martínez, I., & Huth, A. (2009). Recruitment in Tropical Tree Species: Revealing Complex Spatial Patterns. *The American Naturalist*, 174/4: E106–40. DOI: 10.1086/605368

WTO Secretariat. (2016). WTO | Regional Trade Agreements - Facts And Figures. Retrieved August 18, 2016, from <https://www.wto.org/english/tratop_e/region_e/regfac_e.htm>

Wu, S. S., Qiu, X., & Wang, L. (2005). Population Estimation Methods in GIS and Remote Sensing : A Review. *GIScience and Remote Sensing*, 42/1: 58–74.

Wu, S. S., Wang, L., & Qiu, X. (2008). Incorporating GIS building data and census housing statistics for sub-block-level population estimation. *Professional Geographer*, 60/1: 121–35. DOI: 10.1080/00330120701724251

Zielstra, D., & Zipf, A. (2010). A Comparative Study of Proprietary Geodata and Volunteered Geographic Information for Germany. *13th AGILE International Conference on Geographic Information Science*, Vol. 1, pp. 1–15. Guimarães, Portugal. DOI: 10.1119/1.1736005

Zipf, G. K. (1932). *Selected Studies of the Principle of Relative Frequency in Language.* Cambridge, Massachusetts: Harvard University Press.

Zipf, G. K. (1949). *Human behaviour and the principles of least effort.* Cambridge, Massachusetts: Addison-Wesley.

APPENDIX

R script for generating random planar graph with pre-generated point pattern:

```
require(spatstat)
require(maptools)
require(maps)
library(sp)
require(plyr)
require(dplyr)
require(RANN)
require(reshape2)
require(data.table)
require(rgdal)

#load SHP file with point pattern
Y <- readOGR("C:/ ... .shp")
#Define Owin of pattern area
w <-
as.owin(list(xrange=c(32275175,32475175),yrange=c(5611910,5811910)))

YD <- data.frame(Y)
YD <- YD[,-1]
YD <- YD[,-3:-5]
YD <- plyr::rename(YD, c("Simulation"="x", "Simulati_1"="y"))

#count number of points in Y for sampling
maxp <- nrow(YD)

#create k nearest neighbour matrix of points
#Define maximum radius
radius <- 6311
NNa1 <- nn2(YD, query = YD, k=30, treetype = c("bd"))

#rearrange matrix from column to rows
require(reshape2)
NNa2 <- melt(NNa1$nn.idx)
NNa3 <- melt(NNa1$nn.dists)
#merge matrices together
NNa4 <- cbind(NNa2,NNa3)
#add rownames to data.frame
NNa4$Var1 <- row.names(YD36)
#remove unecessary columns
NNa5 <- NNa4[ -c(2,4:5)]
NNa5 <- NNa5[NNa5$value.1 > 0, ]
rm(NNa2, NNa3, NNa4)
max(NNa5$value.1)
min(NNa5$value.1)
X <- max(NNa5$value.1)
hist(NNa5$value.1, breaks=X)

#change column name
NN5 <- plyr::rename(NNa5, c("Var1"="p1", "value"="p2",
"value.1"="length"))
rm(NNa5)
```

```
#as data frame
NN6 <- data.frame(NN5)
#create logic
NN6$Dup <- NN6$p1 < NN6$p2
#subset data based on logic
NN7 <- NN6[NN6$Dup == TRUE,]
NN8 <- NN6[NN6$Dup == FALSE,]
#swap columns
NN8 <- plyr::rename(NN8, c("p1"="p2", "p2"="p1", "length"="length"))
NN9 <- rbind(NN7,NN8)
#remove logic column
NN9 <- NN9[-c(4)]
#remove duplicate
rm(NN5, NN6, NN7, NN8)
NN10 <- NN9[!duplicated(NN9), ]
rm(NN9)
#Reindex row names
row.names(NN10) <- 1:nrow(NN10)
#drop distance column
NN10 <- NN10[,-3]

YD <- sample_n(NN10, 17265053, replace=F)
NN5a <- YD
NN10 <- NN5a[!duplicated(NN5a), ]
rm(NN5a)
#Reindex row names
row.names(NN10) <- 1:nrow(NN10)

#NN12 <- rbind(NN10a,NN12)
NN12 <- NN10
rm(NN10)
n <- 20000
nr <- nrow(NN12)
NN13 <- split(NN12, rep(1:ceiling(nr/n), each=n, length.out=nr))
NN12 <- NN13
rm(NN13)

YDT <- data.table(YD, keep.rownames=TRUE)
setkey(YDT,rn)

##### Generate first line #####
br <- 1
l1 <- NN12[[br]][1,]
NN12[[br]] <- NN12[[br]][-1,]
l1 <- c(t(l1))
l2 <- l1[2]
l1 <- l1[1]
l1a <- YDT[rn==(l1),2:3, with=FALSE]
l2a <- YDT[rn==(l2),2:3, with=FALSE]
x <- append(l1a$x,l2a$x, after = length(l1a$x))
y <- append(l1a$y,l2a$y, after = length(l1a$y))
YDZ <- data.frame(x,y)
#divide into two point sets
Z1 <- YDZ[1,]
Z2 <- YDZ[2,]
#create Line Segment Pattern
Z <- psp(Z1[1, 1], Z1[1, 2], Z2[1, 1], Z2[1, 2], window=owin(w))

l1 <- NN12[[br]][1,]
```

```
NN12[[br]] <- NN12[[br]][-1,]
l1 <- c(t(l1))
l2 <- l1[2]
l1 <- l1[1]
l1a <- YDT[rn==(l1),2:3, with=FALSE]
l2a <- YDT[rn==(l2),2:3, with=FALSE]
x <- append(l1a$x,l2a$x, after = length(l1a$x))
y <- append(l1a$y,l2a$y, after = length(l1a$y))
YDZ <- data.frame(x,y)
Z1 <- YDZ[1,]
Z2 <- YDZ[2,]
H <- psp(Z1[1, 1], Z1[1, 2], Z2[1, 1], Z2[1, 2], window=owin(w))
J <- crossing.psp(Z,H,fatal=TRUE,details=FALSE)
H1 <- as.vector(H$ends, mode="numeric")
J$x <- setdiff(J$x, H1)
J$y <- setdiff(J$y, H1)
J$y1 <- data.frame(J$y)
J$x1 <- data.frame(J$x)
J$Ally <- count(J$y1)
J$Allx <- count(J$x1)
J$Ally == 0 & J$Allx == 0

x0 <- append(Z$ends$x0,H$ends$x0, after = length(Z$ends$x0))
y0 <- append(Z$ends$y0,H$ends$y0, after = length(Z$ends$y0))
x1 <- append(Z$ends$x1,H$ends$x1, after = length(Z$ends$x1))
y1 <- append(Z$ends$y1,H$ends$y1, after = length(Z$ends$y1))
Z$ends <- data.frame(x0,y0,x1,y1)

#LOOP START

repeat {
  if ( br==205 ) { break }
  else
    if ( sum(!is.na(x0))==1030000 ) { break }
  else
    l1 <- as.vector(t(NN12[[br]][1,]))
  NN12[[br]] <- NN12[[br]][-1,]
  l2 <- l1[2]
  l1 <- l1[1]
  l1a <- YDT[rn==(l1),2:3, with=FALSE]
  l2a <- YDT[rn==(l2),2:3, with=FALSE]
  x <- append(l1a$x,l2a$x, after = length(l1a$x))
  y <- append(l1a$y,l2a$y, after = length(l1a$y))
  YDZ <- data.frame(x,y)
  Z1 <- YDZ[1,]
  Z2 <- YDZ[2,]
  H <- psp(Z1[1, 1], Z1[1, 2], Z2[1, 1], Z2[1, 2], window=owin(w))
  J <- crossing.psp(Z,H,fatal=TRUE,details=FALSE)
  H1 <- as.vector(H$ends, mode="numeric")
  J$x <- setdiff(J$x, H1)
  J$y <- setdiff(J$y, H1)
  J$y1 <- data.frame(J$y)
  J$x1 <- data.frame(J$x)
  J$Ally <- count(J$y1)
  J$Allx <- count(J$x1)
  if ( J$Ally == 0 & J$Allx == 0 ) {

    x0 <- append(x0,H$ends$x0, after = length(x0))
    y0 <- append(y0,H$ends$y0, after = length(y0))
```

```
      x1 <- append(x1,H$ends$x1, after = length(x1))
      y1 <- append(y1,H$ends$y1, after = length(x1))
      Z$ends <- data.frame(x0,y0,x1,y1)

    }
    if (is.data.frame(NN12[[br]]) && nrow(NN12[[br]])==0) { br <- br+1 }
    else{}
}


plot(owin(w))
plot(Z, add=TRUE)

require(foreign)
write.dbf(Z$ends, "C:/ ... .dbf", factor2char = TRUE, max_nchar = 254)
```