# Dissecting the local

*Territorial Scale and the Social Mechanisms of Place*

Lund, Rolf Lyneborg

*Publication date:*
2019

*Document Version*
Publisher's PDF, also known as Version of record

# DISSECTING THE LOCAL

TERRITORIAL SCALE AND THE SOCIAL MECHANISMS OF PLACE

BY
ROLF LYNEBORG LUND

DISSERTATION SUBMITTED 2019

AALBORG UNIVERSITY
DENMARK

# DISSECTING THE LOCAL

## TERRITORIAL SCALE AND THE SOCIAL MECHANISMS OF PLACE

by

Rolf Lyneborg Lund

AALBORG UNIVERSITY
DENMARK

Dissertation submitted

.

*“A bad neighbor is a misfortune, as much as a good one is a great blessing.”*
(Hesiod, 700BC)

Dissertation submitted: 14/8-2018

PhD supervisor: Prof. MSO Anja Jørgensen
Aalborg University

Assistant PhD supervisor: Associate Prof. Claus D. Hansen
Aalborg University

PhD committee: Associate professor Mia Arp Fallov
Aalborg University

Professor Peter Hedstrøm
Linköping University

Professor Rob Atkinson
University of the West of England

PhD Series: Faculty of Social Sciences, Aalborg University

ISSN (online): 2246-1256
ISBN (online): 978-87-7210-472-0

Published by:
Aalborg University Press
Langagervej 2
DK – 9220 Aalborg Ø
Phone: +45 99407140
aauf@forlag.aau.dk
forlag.aau.dk



Printed in Denmark by Rosendahls, 2019

Standard pages: 59.07 pages (2,400 characters incl. spaces).

# ENGLISH SUMMARY

Place matters. In the light of the increasing spatial inequality, this thesis investigate methodologies to analyze sociospatial phenomena and presents a new methodology to better combine register data with geographical data. Furthermore, a new methodology to automate redistricting is proposed and used on different social phenomena.

Research paper 1 (From the Dark End of the Street to the Bright Side of the Road) examines the use of administrative borders when investigating spatial, socioeconomic inequality and proposes a new method of spatial division based on automated redistricting utilizing both register data and spatial data as roads, railways, rivers and other physical barriers. The results show that the use of smaller spatial units of measurement greatly improve the socioeconomic homogeneity compared to administrative units. Even when considering the concept of data smoothing and randomness, the model performs better not only at isolating small, homogenous spatial units but as units of analysis as well.

Research paper 2 (Moving to Prosperity?) presents analysis on effects of living in deprived areas in a life course perspective. Most literature on life course deprivation focus either on place of birth or duration of living in deprivation without consideration for the type of deprivation or if there are specific life periods where deprivation has a more negative effect. In this paper, I show that both exposure time and birth place matters but they cannot stand alone in explaining later life outcomes. This is done by accounting for direct, social effects, selection and districting in a combination of automated redistricting and counterfactual models. I show that one must consider precise measurements of place, a thorough consideration for neighborhood selection bias and also take into account the time of exposure to fully grasp the later effects of living in deprived neighborhoods. In the end, I show that especially the time from the age of 6 to 12 has a much more negative accumulative effect on later life outcomes than other times of life.

Research paper 3 (I Like the Way You Move) investigates the effects of living in deprived areas in a life course perspective as paper 2 but focus on the rural/urban divide and examines if deprivation has similar effects on residents living in rural areas compared to more urban areas. Furthermore, I present analysis on the effects of moving to and from urban and rural areas. By turning the focus from deprivation as a universal concept to a local concept, I show that the effects of deprivation on later life socioeconomic measurements differ widely between rural and urban areas. In general, both deprived and non-deprived urban areas show higher levels of educational attainment but have more unemployment and less income. Comparing those that were born and has lived their entire life in deprived areas in rural setting with their city counterparts, I show that the effects of deprivation on income and unemployment are

much less severe for the rural areas compared to the urban. I argue, that deprivation is a local phenomenon that cannot be said to have a universal effect on the residents, even if the measure of deprivation is the same.

Research paper 4 (Social Geographical Patterns in Membership of the Established Church in Denmark) examines the unequal geographical distribution of church membership in Denmark. A general theory of secularization is that rising levels of educational attainment decreases the overall level of religiosity in a society. In this paper, we show that not only is church membership rates very different in Denmark, where especially the Capitol area have very few members and Jutland many, we also show that the expected effects of education are much more prominent in the Capitol area. Areas in Jutland with high levels of educational attainment does not follow the same patters as the same type of areas in the Capitol. Thus, we argue that geography differentiate the effects of educational attainment on church membership and that universal effects must be differentiated.

The following summarizing chapters presented in this thesis is part a more thorough theoretical and methodological insight into the methodologies used in the research papers but also an expansion on the models for automated redistricting where I investigate optimization by machine learning and perspectives into satellite image recognition.

In this thesis, I point to the importance of asking not only questions about the neighborhood effects on the inhabitants that lives there but also asking more fundamental question about what a neighborhood is, how me measure it and what scale means to the way we process the effects. By using a combination of selection models and automated redistricting, I show that scale is very important when investigating neighborhood deprivation. Using administrative borders to isolate deprived areas are inadequate to reveal the intricate and often small clusters that are truly deprived. Furthermore, I show that deprivation is not one thing; deprivation in different geographical settings has a variety of different effects on later life outcomes of the residents. Thus, I argue that place is diverse and complex and that neighborhood research must account for the geographical difference between neighborhoods to fully understand the underlying mechanisms.

# DANSK RESUME

Sted har betydning. I lys af den stigende spatiale ulighed undersøger denne afhandling metodologier til at analysere sociospatiale fænomener og præsenterer nye metoder til bedre at kombinere registerdata med geografisk data. Ydermere præsenteres og anvendes en ny metode til automatisk områdeinddeling.

Forskningsartikel 1 (From the Dark End of the Street to the Bright Side of the Road) undersøger brugen af administrative områdeinddelinger til at belyse spatial socioøkonomisk ulighed og præsenterer en ny metode til automatisk områdeinddeling ved brug af både registerdata og spatial data som vejnet, vandløb og andre fysiske barrierer i landskabet. Resultaterne viser, at brugen af mindre spatiale enheder forbedrer den socioøkonomiske homogenitet betydeligt sammenlignet med administrative enheder. Selv ved tests for data-smoothing og tilfældighed i data skaber automatisk områdeinddeling i mindre områder mere homogene spatiale enheder og bedre enheder til senere analyse.

Forskningsartikel 2 (Moving to Prosperity?) præsenterer en analyse af effekten af at leve i depriverede områder i et livsforløbsperspektiv. Det meste litteratur om deprivation i et livsforløbsperspektiv fokuserer enten på fødselssted, eller på hvor længe en person har boet i et depriveret område uden at undersøge, hvilken type deprivation det handler om, eller om der er specifikke perioder i livsforløbet, der har større effekt end andre. I denne artikel viser jeg, at både eksponeringstid og fødselssted har betydning, men at de ikke kan stå alene som forklaring. Dette er gjort ved at tage hensyn til direkte, sociale effekter, selektion og områdeinddeling i en kombination af automatisk områdeinddeling og kontrafaktiske modeller. Jeg viser, at man må måle områder præcist, tage hensyn til områdeselektion og samtidig tage højde for eksponeringstidspunkt for at indfange senere livseffekter ved vokse op i depriverede nabolag. Til slut viser jeg, hvordan specielt tiden mellem 6 og 12 år har en betydelig mere negativ akkumuleret effekt på det videre livsforløb end andre tidspunkter.

Forskningsartikel 3 (I Like the Way You Move) undersøger effekten af at leve i depriverede nabolag, som forskningsartikel 2, men fokuserer i stedet på den urbane og rurale opdeling af deprivation. Effekten på det videre livsforløb undersøges dermed i et perspektiv, hvor både graden af ruralitet og det at flytte til og fra depriverede byområder og landområder tænkes at have en effekt. Ved at ændre forståelsen af deprivation fra et universelt begreb, der tænkes at have samme effekt alle steder til et lokalt fænomen, viser jeg, at effekten af deprivation er meget forskellig mellem by og land. Generelt viser resultaterne, at både depriverede og ikke-depriverede byområder har en højere grad af uddannelse, men har lavere indkomst og højere arbejdsløshed end lignende landområder. Sammenlignes dem der har boet hele livet i landlige, depriverede områder med lignende personer fra depriverede byområder, viser jeg, at graden af arbejdsløshed og indkomst er markant mindre i de landlige områder end de urbane. Jeg argumenterer her for, at deprivation skal forstås som et lokalt fænomen i stedet for et universelt problem og skal behandles forskelligt.

Forskningsartikel 4 (Social Geographical Patterns in Membership of the Established Church in Denmark) undersøger den geografiske fordeling af folkekirkemedlemskab i Danmark. Den generelle sekulariseringsteori argumenterer for, at en gennemsnitlig stigning i uddannelsesniveau i et land resulterer i ringere religiøs tilknytning. I denne artikel vises, at kirkemedlemsskabsraterne i Danmark varierer geografisk, hvor hovedstadsområdet har lav tilknytning mens der er høj tilknytning i Jylland. Samidigt vises det, at effekten af uddannelse på folkekirkemedlemsskab primært eksisterer i hovedstadsområdet og ikke i byområder på Jylland. Der argumenteres for, at sociale og økonomiske tendenser har forskellige effekter, når man samtidig kigger på den geografiske distribution og forholder sig til det lokale.

De følgende opsummerende kapitler præsenteret i denne afhandling er både en uddybning af de teoretiske og metodologiske elementer benyttet i forskningsartiklerne, men er ligeledes en udvidelse af modellen for automatiseret områdedannelse, hvor jeg undersøger optimering gennem machine learning og perspektiverer til automatisk billedgenkendelse af satellitbilleder.

I denne afhandling peger jeg på vigtigheden af at stille de rigtige spørgsmål om nabolagseffekter, men også på vigtigheden af at stille mere fundamentale spørgsmål som hvad områder er, hvordan vi måler dem og hvad skala betyder, for den måde vi bearbejder effekterne af områder. Ved at bruge en kombination af selektionsmodeller og automatiseret områdedannelse viser jeg, at skala er vigtigt for at kunne isolere effekterne af nabolagsdeprivation. Brugen af administrative områder til at undersøge disse effekter er ikke fyldestgørende. Jeg viser ligeledes, at deprivation ikke kun er én ting. Effekten af deprivation varierer med områder og har meget forskellig effekt på individers senere livsforløb. Jeg argumenterer for, at steder er forskellige og komplekse og at nabolagsforskning er nødt til at tage højde for geografiske forskelle mellem nabolag for at blotlægge de bagvedliggende mekanismer.

# ACKNOWLEDGEMENTS

Writing a thesis or doing any kind of research depends on more than a single individual. The Institute of Sociology and Social Work at Aalborg University excels in research collaboration and this is why I feel acknowledgements for this specific thesis is not only imperative but also hard; how does one synthesize the helpful interactions that has happened over a three year period? I will, none the less, try to reach out to the most important people and apologize if I have left someone out.

This thesis could not be accomplished without my supervisor Anja Jørgensen. From the beginning of the thesis and even before, Anja had a clear vision of what was missing in the field of neighborhood studies and had a firm belief in me and my ability to handle the subject. She has a brilliant eye for interesting research topics and I can only hope to one day reach a fraction of her insights. I am grateful for her many advice and for helping me find my way – both research wise and geographically. In this same paragraph, I would like to thank my co-supervisor Claus D. Hansen and the entire research group, SocMap. They have given me valuable feedback, excellent sparring and many great insights as well as invaluable social interactions that made stressful times easier.

At faculty and institute level, I want to thank our head of department Trond Beldo Klausen for being my guide and mentor from day one at AAU. He is one of my role models and a person I deeply admire. I also want to thank pro dean of research Søren Kristiansen for believing in my research and the many crazy ideas I have had. Many of the collaborations I have had would be impossible if not for the support of the faculty. Moreover, the Computational Analytics Laboratory for Digital Social Science (CALDISS) with Kristian Kjelmann at the helm has been invaluable for both brilliant feedback and facilitation of technical knowhow.

From abroad I want to thank the Institute for Computational Social Science for having me as a visiting researcher. I have met so many talented and inspiring people there and I very much look forward to keep the great cooperation in the future.

This acknowledgement would be empty without naming all the students I have taught and tutored over the years. Even though I have slightly surpassed my teaching obligations, I consider it a small price to pay for the excellent interactions I have had. One thing is to understand a concept, another is to understand it well enough to teach others about it and I have always had students curious about all the weird research ideas that I have. Søren Christian Krogh and Peter Clement Lund went from being my students to being my colleagues each in a PhD of their own and the sparring we have had over the years and still have is priceless.

Last but not least, I would like to thank my family. Living with me is not an easy feat and my work schedule is not always compatible with… any human interaction. Nonetheless, they have supported me even when I suddenly had an idea during dinner or wanted to attend a conference halfway around the world. My two-year old daughter has proclaimed that she also wants to be a "socologo" and I must admit that I understand why; it is a world full of amazement and constant learning. For her, it's mainly due to the constant availability of fruit and the occasional cake at the institute, but I consider it a win none the less.

# RESEARCH PAPERS AND THEIR PUBLICATION STATUS

Lund, R. L. (2018). From the dark end of the street to the bright side of the road – automated redistricting of areas using physical barriers as dividers of social space. *Methodological Innovations*, *11*(3), 1–15. Published

Lund, R. L. (2019). Moving to Prosperity? The Effect of Prolonged Exposure to Neighborhood Deprivation. *Scandinavian Journal of Educational Research, 0*(0), 1–17. Published.

Lund, R.L. (2019). I like the way you move - The effect of moving to and from different types of areas during adolescence on later socioeconomic outcomes. *City and Community.* In review.

Lund, R.L., Jørgensen, A. & Riis, O.P. (2019). Social geographical patterns in membership of the established church in Denmark. *Nordic Journal of Religion and Society*. Published.

# TABLE OF CONTENTS

# TABLE OF FIGURES



# TABLE OF EQUATIONS

# CHAPTER 1. INTRODUCTION

Whether it is a single electron's placement in an atom, the human's place in the world, or the earth's place in the galaxy, we are occupied with what placement means. In some sciences, such as geography, astronomy, or geology, space is all that matters, while in other sciences, such as mathematics, psychology, and computer science, space becomes less visible. Space and place are harder to understand within the domain of sociology. Some branches of sociology have almost no interest in how geography matters in how we understand the subject of our research. Looking at some of the most significant works of sociology throughout history, including *Social Systems* (Luhmann, 1995), *Interaction Ritual* (Goffman, 1967), and *The Structure of Social Action* (Parsons, 1968), we can see that people are understood as placeless beings who have the same function, no matter where the theory is applied.

Despite the importance place has in a theory, it is undeniable that people live in geographical space. Even grander theories, such as Zygmunt Bauman's theorizing about community, where he defines the new, global elite as truly exterritorial (Bauman, 2001) or in *The Consequences of Modernity*, where Anthony Giddens describes space as a phantasmagoric entity (Giddens, 1990), it is impossible to imagine human life without understanding it existing in a physical world.

This brings us to an important point. As Gieryn notes, space is not place (Gieryn, 2002). Where natural geographers and architects are mostly concerned with space in the geometric sense, many social sciences are more concerned with places that have people inside them. This is what Gieryn defines as the major difference between space and place. He also poses this question: Is it the focus on the structural and physical space, or is it on the social life that happens in places? He also offers a salient observation:

> Sociologists could become more adept with maps, floor plans, photographic images, bricks and mortar, landscapes and cityscapes, so that interpreting a street or forest becomes as routine and as informative as computing a chi-square. That visualizing (I think) is the next step. (Gieryn, 2002)

Most types of measurable social phenomena are somewhat easy to grasp. If we wish to measure income, we focus on currency, whether the currency has been adjusted for inflation or perhaps how precise we measure income. If we measure educational attainment, we need to know how education is categorized, whether the national educational system can be transformed to ISCED (International Standard Classification of Education) (Eurostat, 2014) ranking or maybe what overall field the education belongs to. The same can be said about unemployment, occupational type, health, and a plethora of variables frequently used either as outcomes or as causes. These variables are common, and they are used as measures of vertical differentiation in the sense that there are attributed low and high values that hold a meaning. However, social place is not as easily compartmentalized into good/bad or high/low.

We are used to seeing attributes ascribed to different places: urban, rural, suburban, and so on, but very few studies have tackled the question of place by looking at places either through maps or imagery (Gans, 2002). Where maps were crucial to many studies in the traditional Chicago School (Kelley & The Residents of Hull House, 1895; Park & Burgess, 1925) and before that to navigate through travel routes, they are of less importance in modern sociological literature (Gans, 1961, 2002; Gieryn, 2002). During the literature review of neighborhood effects and census tract studies, I encountered almost no maps or other spatial visualizations. This is, of course, not a sign of a lack of quality, but it is a sign that place has shifted from a spatial understanding to a more econometric one where the place becomes a number and not a recognizable physical feature.

Mapping, or cartography, is not often taught as part of a classical sociological training, but we are often taught about the importance of neighborhoods and how segregation is rooted in place (Gieryn, 2002). By combining geography and computer science with sociology, I have attempted to demonstrate how we can learn important things about the social world when we apply the element of physical geography as being equally important to the socioeconomic one. Where a place is located and not just how it is ranked in a non-spatial, socioeconomic sense is the center point of this thesis. The focus is on the link between the shape of cities, neighborhoods, streets, and buildings and how these objects serve not only as dividers of physical space but of social space as well.

During the development of the overall model of automated districting, a colleague jokingly said to me that this will be the first thesis he knows of that focuses purely on a single, independent variable. Even though it was said in jest, it struck me that he was in some way right. I do not have a comprehensive knowledge of PhD dissertations, and I have read even fewer, but I do know that the norm is to focus on a dependent variable and then use an assortment of independent variables to gain knowledge about how the dependent variable varies in some way or other. My hope is that the reason for a more detailed focus on place will become evident throughout this thesis.

This thesis is rooted in the theory, methodology, and literature about spatial effects, and one of the most common uses of spatial effects is found in the literature on neighborhood studies. In reduced form, neighborhood studies rely on the hypothesis that people living in places are affected either by the place itself or that the concentration of homogenous groups of individuals creates a feedback effect in terms of socioeconomic trajectories. It is the cornerstone of much segregation research and in the study of social and economic inequality. Nevertheless, we are not accustomed to fully grasping what inequality and segregation look like on a map. Plenty of studies have pointed to the fact that spatial inequality exists, but are much less concerned with precisely defining what a neighborhood is or why they use one specific geographical division instead of another (Gans, 2002).

The main point of this thesis is to utilize different methodologies to improve how we measure, see, and interpret social data in physical space. I do this very thing from the first paper where I outline an overall methodology to improve the usability of register data in a geographical setting (Lund, 2018) to the usage of this methodology in analyses that mainly revolve around geographical inequality in education (Lund, 2019a, 2019b) but also in social values as the geographical distribution of church membership rates (Lund, Jørgensen, & Riis, 2019). By doing so, I explore different facets of neighborhood and place effects and how a different understanding of the spatial aspect of neighborhoods can be used to isolate and better understand the actual mechanisms behind inequality and life course deprivation.

## 1.1. RESEARCH QUESTIONS

Place matters. Our place of birth, upbringing, and living defines what are in our reach, what we are presented with when we leave our home, and what immediate opportunities we have. This is why questions that revolve around the effects of place are important, but questions about the way we measure place are equally as important. This thesis revolves around multiple questions based on the same overall root question: *How can we modify the geography we measure, and what does that mean to the subject we measure?* Throughout this thesis, I try to address the overall question while asking more specific questions about inequality and how measuring inequality using a precise spatial methodology can help us better understand and analyze the underlying mechanisms of inequality and social differentiation.

The overall question is based on two different research questions. First, *what is place and how do we understand place when we wish to measure it?* As I will demonstrate throughout this thesis, measuring place is of utmost importance, but is often overlooked (Ferreira, Holan, & Bertolde, 2011; Leventhal & Brooks-Gunn, 2000; Petrovic, Ham, & Manley, 2018; Sampson, 2012). Place becomes a factor to control for or an entity one wishes to eliminate from a specific model, and thus, a great deal of research uses whatever geographical measurement is available in the data. In my first paper, I describe the difference between commonly used administrative areas and micro area models (Lund, 2018), and the research question is a methodological one, where I investigate how an algorithmic approach to scaling differs from purely administrative measures and what that means to our understanding of social phenomena.

For many, scale is a fixed point predefined in geography. In some instances, Danish parishes have not changed in almost 1,000 years, but we still use them without asking the obvious: *What does scale have to do with the phenomena I investigate?* This question is both methodological and theoretical in the sense that we need to be able to work with geography as a changeable and scalable entity, but we also need to understand how changing the scale changes the phenomena. The boundaries are the same, but the human content within the bordered areas change significantly. For example, measuring intergenerational social mobility between municipalities is one way to better understand how place affects individuals, but if the true mechanism of unequal social mobility lies in neighborhoods within municipalities, the analysis will be less valid.

The second part of the question is based on the first: *What happens if we try to understand where we live in combination with the way we act, live, and socialize?* By combining a well-defined base of geography as defined above, it becomes possible to better understand if we truly are, as Bauman notes, exterritorial (Bauman, 2001), or if we are affected by where we live and who we chose or are forced to live near. Three of my papers (Lund, 2019a, 2019b; Lund et al., 2019) deal with differentiated questions about the application of a special, small-scale geography and how different scales of measurement can lead to new conclusions.

This thesis is a methodological one where automation of districting is the key underlying goal. Nonetheless, using this methodology as a sociological tool is the main goal. Generating specialized methods to isolate specific spatial phenomena is only interesting if it is used to answer questions about inequality, segregation, or development we have been unable to answer earlier.

My main question and each of the two research questions outlined above is answered using Danish register data with geospatial links over time. For the first section, where a general model for geographical redistricting is performed, I use register data from the years 2000 to 2015 on the total Danish population (Lund, 2018) and a 100 x 100 meter square grid geographical referenced data file containing information on the number of inhabitants per 100 x 100 meter square. The second section relies on register data containing information on socioeconomic and social parameters from 1985 to 2016 used either as cohorts of children born from 1980 to 1986 (Lund, 2019a, 2019b) or as a full population (Lund et al., 2019) likewise linked to geography.[1]

[1] For a more thorough review of data, please see Lund, 2018, 2019a, 2019b; Lund, et.al, 2019

## 1.2. FROM LONDON THROUGH CHICAGO VIA L.A. TO BOSTON

The following subsection is designed to give a theoretical introduction to how spatial analysis and the understanding of place and neighborhoods have evolved over time and how this temporal change has formed the research questions and the methodological perspectives in my thesis.

The history of neighborhood research has taken many twists and turns within the different schools of sociology, ranging from purely descriptive to theoretical to extremely advanced methodological studies. The following is a review of the major changes in how we understand place and space in sociology. This is, of course, not a complete review, and as Robert Sampson wrote in his book *Great American City* when talking about a literature review process he and his colleagues conducted on neighborhood effects literature:

> By the dawn of the twenty-first century the literature on neighborhood effects was enormous. When my colleagues and I attempted a comprehensive review circa 2000 we discovered hundreds of studies and since then hundreds more have appeared. (Sampson, 2012)

His observation is solely related to studies with behavioral problems and health as outcome, and widening the scope to income inequality and social marginalization does not narrow down the list of articles. The following literature review should not be seen as a complete literature on neighborhood studies, but instead as an enumeration of the various foci on important changes in the way sociology articulates place theoretically. Most notably, this review purposely leaves out the works of theorists such as Jane Jacobs (Jacobs, 1992), David Harvey (Harvey, 2003), and John Urry (Urry, 2000). Common to the aforementioned authors is their devotion to geographical and spatial inequality, but also that their work is based on a general idea of place and space. The following section will primarily address research that works more closely with the place and especially research that revolves around the concept of neighborhoods.

### 1.2.1. THE BEGINNING OF NEIGHBORHOOD RESEARCH

Maps have been a crucial part of human history (Whitfield, 2005). From ancient times up through the time of the explorers, maps have been our way of defining where we are and where others are (Whitfield, 2005). We rely on them to navigate but also to draw borders and identify parts of the world, countries, cities, and local communities. It is one of the ways we understand geography: we draw it.

A common feature of most earlier historical maps was that they were designed to be descriptive. We needed maps mainly to find our way and to know if we were in one region or another. In some cases, maps were used to know what tribes or population groups were located where, but they were mainly used as an administrative tool and not as an analytical one (Sampson, 2008). Maps were also used to encompass parts of cities or to make decisions as to where specific population groups would be located. One could argue that the starting point of understanding space is where maps began. Dating from 24,000–25,000 BC (Wolodtschenko & Forner, 2007), the earliest maps were primarily very small-scale and were depictions of star clusters visible by night. The starting point of the transition from space to place as a more relevant setting was most probably Venice around the 16th century.

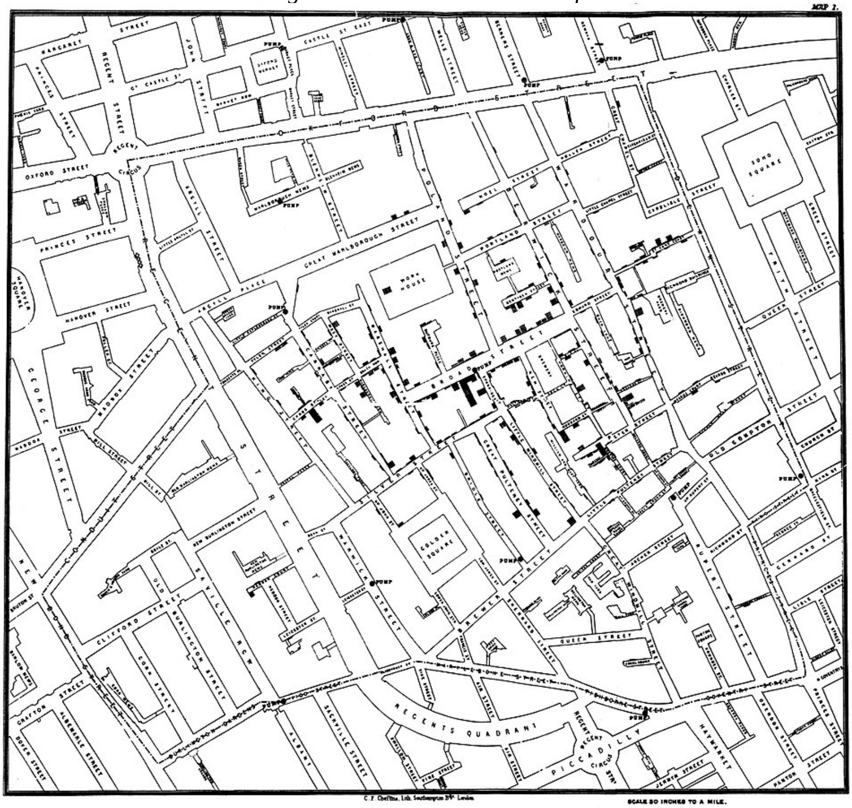
One of the most famous maps of Venice, the so-called de'Babari map (Figure 1) from the 16th century, is one of the only known and preserved examples of detailed woodcut maps from before the 17th century (Levenson, Oberhuber, & Sheehan, 1973). During that period, Venice became famous for something different but also interesting when researching spatial inequality; they invented the word "ghetto." The Venetian senate decided to move all the Jews located in the city to the Ghetto Nuovo (Calimani, 1987; Finlay, 1982) because of religious reasons. Even though this thesis does not have an explicit focus on ghettos, the example of Venice and this ghetto serves as one of the first maps that depicts deprivation.

*Figure 1 - The de'Barbari map of Venice*


The Ghetto Nuovo was surrounded by water and was only accessible by two bridges. This meant that the city could chose when to allow the Jewish population to enter the main city. Spatial segregation was nothing new during this age, but the detailed map historically documented the visual aspect, and it suggests that we might be able to see segregation on a map even when segregation is of no relevance to the cartographer. The political point of the ghetto at the time was to secure the segregation of one religion from another, but today it stands as a testament to how segregation and inequality were manifested in the city.

The change from an almost purely administrative tool to an analytical tool was primarily driven by the natural sciences, and one of the most famous uses was by epidemiologist John Snow in 1854 (S. Johnson, 2006; Snow, 1855). By mapping outbreaks of cholera, Snow was able to determine that cholera was spread by the city water supply and not by air, as earlier thought (S. Johnson, 2006), thus establishing that a map, in itself, can help understand a causal relationship.

*Figure 2 - Cholera outbreak map*

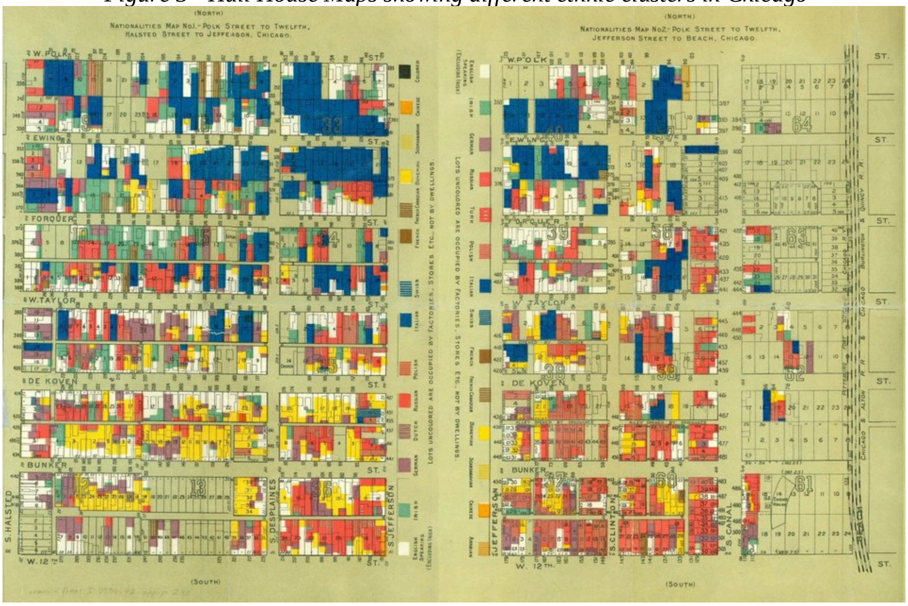



Source: (Lerner & Lerner, 2006; Snow, 1855).

The above rendition of a cholera map by Snow shows how all the cholera outbreaks (marked by black dots) were centered around the water pump on Broad Street (Lerner & Lerner, 2006), and by further investigating the pump, he realized that the pump had been dug less than a meter from the ground surface, and thus the water was susceptible to outside contamination. He determined this by using the map and the geography of the city as analytical tools. Even though this study is far from what we think of as area studies in sociology today, the impact is evident in much of the later research in epidemiology. Maps changed from being a way of seeing the world to a way we understand it – from purely descriptive to models of social causality (Sampson, 2008).

From the cholera maps of London in 1854 to Hull House in Chicago in 1895 came one of the major steps in merging social information with cartography. With the publication of *Hull House Maps and Papers* (Kelley & The Residents of Hull House, 1895), maps became a better way understand the city, enabling the answers to essential questions: Who lives there? Where do they live?, as well as generating the questions and answers that were common knowledge to the inhabitants of the city of Chicago, but no one had ever looked at systematically. By using survey techniques, classical mapping tools, and observational studies, the residents of Hull House were able to map out not only how nationalities clustered together in the city but also how wages, living conditions, and general well-being differed widely across even smaller parts of the city of Chicago (Kelley & The Residents of Hull House, 1895).

*Figure 3 - Hull House Maps showing different ethnic clusters in Chicago*

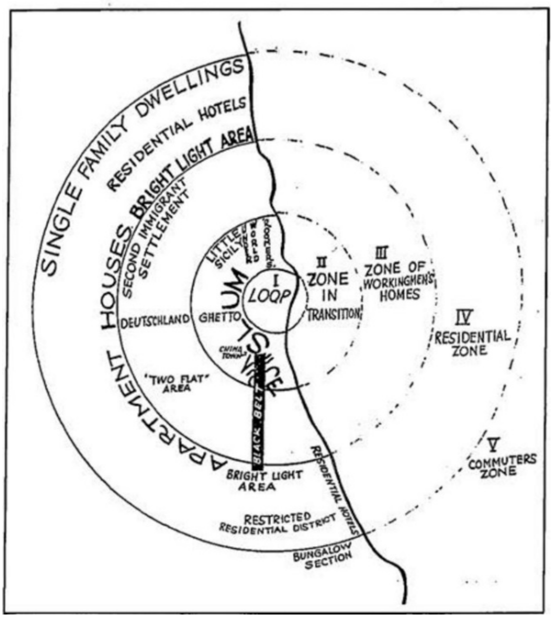



Source: (Kelley & The Residents of Hull House, 1895).

The above map is a representation of how different nationalities settled in Chicago in 1895. The study spanned only a few blocks in the city, but was rich in information on how the working wages differed between groups. In itself, the Hull House maps were interesting just to understand how immigrants formed neighborhoods that were rooted in their inherent nationality, but with the full collection of essays and secondary research on the neighborhoods, they used the map very much like Snow's cholera map of London: not in a purely descriptive way but as a way of creating analytical causality between language skills and wage gaps (Kelley & The Residents of Hull House, 1895).

The Hull House was one of the earliest groups to use maps as a way of analyzing ghettoization and represented, to some extent, the starting point for the Chicago School of Sociology. Few can deny the impact the Chicago School of Sociology has had on the way we understand neighborhood research. With the publication of *The City* (Park & Burgess, 1925) and *The Gang*: A Study of 1,313 Gangs in Chicago (Thrasher, 1927) in the mid-1920s, the researchers began experimenting with mapping not just the municipalities or the parishes and counting inhabitants, but instead looking at the city as a dynamic entity that had a history to tell beyond the administrative boundaries that were used. They looked at the city as more than a collection of randomly placed individuals and, to borrow the terminology of Peter Hedström (Hedstrom, 2005), tried to dissect the social aspect of the city to better understand how the city and the people living there might be the same and different all at once (Park & Burgess, 1925). One of the results, and perhaps the most famous one, was the concentric zones of the city (Waterman, Park, Burgess, & Boyd, 2006), shown in Figure 4.

*Figure 4 - The city*




Source: (Waterman et al., 2006).

What Robert Park and his colleague did was to use the earlier research of Hull House combined with many other mappings, ethnographic studies, and census surveys to propose a theory of place. The concentric zones of the city were designed inductively and explained how inhabitants had shaped the city through the lenses of observation and theory. Furthermore, the research highlighted the fight for urban space occurring during the early 20th century in Chicago among groups in ethnic enclaves. The findings indicated how zones push, pull, and reform their meaning through social interaction and through a social Darwinist evolution the city evolves (Park & Burgess, 1925).

The idea that maps were primarily for administrative purposes was changing. Maps were used to generate sociological theory and concepts and to better understand social phenomena and mechanisms.

### 1.2.2. NEIGHBORHOOD EFFECTS

Even though the journey through research into place and space is simplified in this short review, few can deny the impact of either the Chicago School or the more econometric period that followed. Where the earlier research was motivated by a very inductive way of looking at neighborhoods and areas, the next phase of more comprehensive neighborhood studies was focused on measuring neighborhoods in a more deductive way. Many of the studies that followed the more econometric way of thinking were motivated by the goal of identifying and isolating an “area effect” or “neighborhood effect” as an almost universal effect that a neighborhood would produce (Cox, 1969; Friedman, 1955; Miller, 1977; Schelling, 1971; Wilson, 1987). As Herbert Gans noted:

> Among sociologists, somewhat the same spatial notion has taken the form of looking for “neighborhood effects.” In this field of study, the neighborhood is conceived to have good or bad effects because of what it does for or to people, particularly the poor. (Gans, 2002, p. 334)

Even though Gans pointed out that the sociologists had adopted this train of thought, the beginning of this new way of understanding neighborhoods again grew from Chicago, but not, as one might think, from the School of Sociology, but instead from the Department of Economics. One of the first uses of the phrase “neighborhood effects” came from the Department of Economics at the University of Chicago and was coined by Milton Friedman in a book chapter titled “The Role of Government in Education” (Friedman, 1955). Here, Friedman explained how “neighborhood effects” can be thought of as a way to lift the socioeconomic state of neighborhoods by boosting educational attainment by subsidizing different types of education. For him, the social aspect of the neighborhood or, to be specific, the social interaction inside the neighborhood was less relevant than the overall effect of the “container” or the socioeconomically homogenous group of people living within a non-defined geographical space (Friedman, 1955).

This is what was later coined as the “concentration thesis,” where the main theoretical thinking was to understand how specific concentrations of groups of individuals were catalysts of the neighborhood. Deprivation and ghettoization became issues related to how many individuals with specific socioeconomic traits were located in limited space and not so much how the city was shaped. It all came down to number of individuals and how much the sum of individuals either lowered or raised the average income level of an area. The area existed, but the physical form was much less important than the sum of individuals.

This way of thinking was somewhat ahead of its time, but during the 1960s and 1970s, more and more researchers started to account for something that had been an unobserved entity earlier: the neighborhood. Along with this came the concept of the political thinking of Miller and Cox, who both focused on voting behavior in spatial models (Cox, 1969; Miller, 1977) and hypothesized that neighborhood dynamics would play a major role in the political decisions individuals took and not only because of socioeconomic division of geographical space.

Even with the emergence of these new waves of understanding space, it was not until 1987 with the book *The Truly Disadvantaged* by William J. Wilson that sociology would begin measuring, using, and working with space as an econometric concept (Wilson, 1987). With this book, Wilson took a step toward a new understanding of neighborhoods and used words such as “ghettos” and “deprivation” as collective terms for many different neighborhoods at different scales (Wilson, 1987). The geography in itself was of less importance than the individuals inside it; ghetto, slum, deprived, rich, poor – they all became types that were thought to have a sense of commonness and, thus, a collective effect on the individuals within (Wilson, 1987, 1996). Nonetheless, Wilson detailed important ways to measure and analyze neighborhoods and geographical inequality, and he underlined the differences in deprived areas, while still maintaining that the complexity of geography could be reduced. This reduction, however, came at the expense of the varying neighborhood and the many different effects that could be found within.

This way of thinking does have some common ground with the Chicago School. Both focused on understanding place as a concept, but they differed in the way they approached the place as a unit. The Chicago School was interested in understanding each element of the city as an entity that was unique. Each area had an essence, even though it shared many communalities with adjacent areas or even areas further away, as described in their use of concentric zones, so they understood the city as many different places that formed a whole. The more econometric approach was less focused on uniqueness and more on typology as a concept.

One of the main critiques of the neighborhood effects approach is the first part of the concept: neighborhoods. While the methodology is often sound and the idea behind the testable hypotheses interesting, the neighborhood can disappear and instead become an effect for which one wishes to add statistical control. Neighborhoods become somewhat random and only consist of a number for identification with very little information on what part of the city, municipality, or country each neighborhood is located in. As noted by Herbert Gans (2002):

> Although neighborhood effects researchers are working with a spatial concept, they do not always define neighborhood or report who and what in the neighborhood actually produces effects. Moreover, quantitative researchers too often use census tracts as proxies, as if a Bureau of the Census statistical artifact could have good or bad effects. (Gans, 2002, p. 334)

While it might seem excessive to lash out against clustering types as census tracts,[2] Gans's point is not so much the concept of neighborhoods as numbers but how these neighborhoods are used. The lines on the map are drawn somewhat randomly and without concern for the social life that takes place between the lines (Gans, 2002). Furthermore, he references an important question about the root of the effects: How can a census tract in itself have an effect? This is, of course, a simplified jest on the actual underlying problem, which is a theoretical one. The question Gans poses is based on the actual effects that we measure because, without a discussion about the neighborhood itself, it is easy to accidentally imply that the neighborhood, in itself, has a negative effect on the people living there. Where the Chicago School scholars were advocates for the ecological approach of neighborhood studies, the econometric methods of Friedman and Wilson had a focus on the individual gains of neighborhoods. On one hand, the neighborhood is an entity of ecological substance, while on the other hand, there is focus on individuals who happen to exist in a neighborhood.

This duality is what propels the research into neighborhoods further into an important discussion about what the neighborhood is and what it means to live in one.

### 1.2.3. LOS ANGELES, THE OUTCASTS, AND THE GODS OF CHAOS

Where the Chicago School was based on the local and observing the fight for space and the later econometric approaches focused on isolating effects, the L.A. School of Urbanism was driven by how materialistic entities such as buildings and city design

---

[2] Census tracts in Gans's perspective refer to the common administrative areas first developed in the United States of America (Krieger, 2006; U.S. Department of Commerce Economics and Statistics Administration Bureau of Census, 1994). The tracts were developed to aid and further cluster neighborhoods in smaller areas to collect survey data in regions where register data are not commonly available. In most cases, a census was somewhat arbitrarily constructed as the distance a surveyor can accomplish in a day (U.S. Department of Commerce Economics and Statistics Administration Bureau of Census, 1994).

were catalysts for the social struggle (Dear & Flusty, 2012; Scott & Storper, 2015). This neo-Marxist theory was based on the concept of the city as a literal battleground for class struggles and the constant conflict between social control and social life (Curry & Kenney, 1999; Monahan, 2002). Even though both schools had an interest in the social life of the city and the struggle for space, the L.A. School saw the struggle as primarily a horizontal one and not a intersocial struggle as the Chicago School did.

Some of the more prominent members of the school relied heavily on more post-structural and postmodern theories of space where the city architecture, in itself, was considered as a form of oppression or a conduit for power (Dear & Flusty, 2012; Soja & Gren, 1991). Other members of the school subscribed to a more post-Fordist tradition with a focus on both the human capital approach (Storper & Scott, 2009) and how capitalist systems have changed the way we shape our cities and our lives to a flexible specialization (Scott & Storper, 2003). In both schools, there was focus on inequality inherent in the city and how the inhabitants in the city are shaped and are shaping the space around them.

The concept of the shape of the city has been a very important part of the L.A. School of Sociology (Curry & Kenney, 1999). Drawing heavily on Foucault (Foucault, 1988) and Lefebvre (Lefebvre, 1991), the emphasis is on power inequality and, consequently, how the city is shaped to accommodate the flow of power (Davis, 2005). This marks one of the most significant tipping points when comparing the L.A. School of thought to most other schools because the city itself and its architecture, shape, and development are used to understand how power flows (Davis, 2005). As Mike Davis (2005) put it:

> Night after night, hornetlike helicopter gunships stalk enigmatic enemies in the narrow streets of the slum districts, pouring hellfire into shanties or fleeing cars. Every morning the slums reply with suicide bombers and eloquent explosions. If the empire can deploy Orwellian technologies of repression, its outcasts have the gods of chaos on their side. (Davis, 2005, p. 15)

Besides the very graphic imagery used by Davis, he also displays one of the key factors in the L.A. School’s way of thinking: oppression, power, and reaction. This way of critical thinking understood the city as a scene where a power struggle constantly plays out. This type of analysis relies much more on the Chicago School than the later econometric types of analyses where the overall shape and feel of the city matters more than the individuals that exist within it. By drawing on critical theory, they saw the city as a class struggle rather than an individual struggle (Curry & Kenney, 1999; Davis, 2005).

### 1.2.4. A NEW ERA OF NEIGHBORHOOD STUDIES

In parallel with the L.A. School grew another branch of spatial studies heavily inspired by analytical sociology and empirical analysis. Even though the main heading for this section makes a promise to end up in Boston after discussing the L.A. School, this last step is much less rooted in a geographical school and more in an idea that emerged in multiple places during the late 1980s. The reason for choosing Boston as the last step on the ladder is because of the work of Robert Sampson, but it could just as well have been ascribed to Oslo (Elster, 1989), Stockholm (Hedstrom, 2005; Hedström, 1994; Hedström & Swedberg, 1996), or New York (Hedström & Bearman, 2009). What separates these directions in the school of analytical sociology is the aim of the research, and Robert Sampson's aim was primarily the neighborhood (Sampson, 2008, 2012; Sampson, Morenoff, & Gannon-Rowley, 2002; Sampson & Sharkey, 2008). As noted by Hedström and Bearman:

> Analytical sociology is concerned first and foremost with explaining important social facts such as network structures, patterns of residential segregation, typical beliefs, cultural tastes, common ways of acting, and so forth. It explains such facts not merely by relating them to other social facts—an exercise that does not provide an explanation—but by detailing in clear and precise ways the mechanisms through which the social facts under consideration are brought about. In short, analytical sociology is a strategy for understanding the social world. (Hedström & Bearman, 2009, pp. 3-4)

What truly distinguishes analytical sociology, which often relies on quantitative and econometric methods, from the purely econometric way of thinking about neighborhoods is not just a change of methods, but instead is a way of thinking about the question one asks. It is preoccupied with the mechanisms that generate specific social phenomena as inequality but also spatial segregation (Hedström & Bearman, 2009; Hedström & Swedberg, 1996; Sampson, 2012), but it often considers that geography is more than just a container (Sampson, 2012). Sampson introduced the concept of collective efficacy first as a part of the self-efficacy theory (Bandura, Craighead, & Weiner, 2010; Sampson, Raudenbush, & Earls, 1997), but later as a more complete and independent theory to explain local crime rates and spatial inequality (Sampson, 2012).

Collective efficacy is a way of thinking about the local as a form of self-governing entity. A purely "container"-approach to the spatial element would argue that, if the state of the neighborhood is poor or if the housing is so cheap that it attracts a disproportionate amount of high crime and low-employed individuals, it would affect those living there as individuals in their later life outcomes. This approach allows us to explain why negative effects exist, but it cannot explain why different deprived neighborhoods reproduce different life trajectories. The theory of collective efficacy explains this phenomenon as something inherent in the neighborhood: a way the community is able to govern itself by social interaction (Sampson, 2012). Social cohesion, social interaction, and community are difficult to capture with traditional methods since we have very limited possibilities for understanding human interaction in large datasets that are not specifically designed to capture this and, thus, this is often overlooked.

To capture this sociality in quantitative data, we need to be more focused on place as an object that can differ even when comparing similar types of areas; ghettos are different, suburbs are different, and gated communities are different, even if they seem homogenous in their socioeconomic tendencies. Thinking not only directly about homogeneity within areas but also about how areas that are homogenous within can produce different outcomes allows us to better understand how different neighborhood effects can be. In the spatial branch of analytical sociology, this is an interesting subject and one that not only Robert Sampson has worked with (Hedström, 1994; Keuschnigg, Mutgan, & Hedström, 2019; Legewie, 2018; Legewie & Schaeffer, 2016; Petrovic et al., 2018).

The way of thinking about mechanisms in neighborhood research sprang up in many places at once. Even though Snow's study of cholera in London is more than 164 years old, the concept of thinking about ideas and social movements as contagious is timeless. Peter Hedström (1994) used this analogy to describe the spread of local unions in Sweden from 1890 to 1940, specifically to shed light on how social movements spread much like a disease[3] in the spatial realm, instead of just enumerating the number of movements and explaining how they evolved. By adding a spatial element, we are able to see more nuances of our data and thus better understand spatial movement in a social context with a strong theoretical framework.

Understanding the different mechanisms in spatial terms, while also thinking about community, is a key theme in analytical sociology, but it is slowly becoming more important in other areas of sociological thinking as well. Studies that have tackled questions about place in an analytical framework include those on immigration (Pais, South, & Crowder, 2012; Valdez, 2014), kinship (Dawkins, 2006; Fiscella & Fremont, 2006; Jarvis, Kawalerowicz, & Valdez, 2017), networks (Cullinan, 2011; Hoem, 2007; Nordlund, 2018; Zang, 2006) and overall segregation and inequality (Cohn & Jackman, 2011; Ferreira et al., 2011; Hillier, 2005; Humberd, Clair, & Creary, 2015; Keuschnigg et al., 2019; Petrovic et al., 2018). This list represents only a fraction of the body of work, but there is an important common denominator in all of these studies: scale.

To fully understand the geographic effects, or neighborhood effects, it is impossible to negate scale. Newer studies have gone into detail, and my paper, "From the Dark

[3] Here, disease is an analogy to the transmission of ideas and not to imply negative connotations about social movements.

End of the Street to the Bright Side of the Road" (Lund, 2018), discusses this within a methodological approach. Measuring residential phenomena requires a decision about scale, and newer studies are focusing on understanding what scale means when we measure the social world (Ferreira et al., 2011; Keuschnigg et al., 2019; Petrovic et al., 2018). The methodological literature on scale is a key part of the methodological discussion in the next section, but the theoretical point is just as important. When we measure social phenomena, we need to understand at what level we measure them and what we can expect when we chose the scale if we wish to understand the mechanisms that are key to our neighborhood questions.

## 1.3. THE JOURNEY

The theories about and empirical work done on neighborhoods and sociospatial phenomena have grown exponentially the last 150 years. From maps of the stars chiseled in caves to cholera outbreaks to a merger of the spatial and the social, we have moved from purely descriptive maps to analytical maps that convey a specific meaning. Even though the purely econometric container solution is still very much in favor because of the unique possibilities that lie in reducing something complex to a very narrow definition of neighborhood, many researchers point to the fact that the neighborhood is more than just the administrative number it has been given. In this thesis, I argue for a specialized geography that must follow the social element we wish to investigate, and I will, in the following sections, clarify how we can use modern methods and computer science to embrace data complexity instead of reducing it.

# CHAPTER 2. METHODOLOGY

Conducting neighborhood studies from a national, quantitative perspective can be challenging in numerous ways. Understanding the people that live in a neighborhood in a quantitative setting requires full information on who lives in the neighborhood and a wide range of information about what is specific about that group of individuals. In addition, the geography in itself can be challenging and poses essential questions: At what scale do we measure a neighborhood? What makes a neighborhood? If these questions are left unanswered or, worse, ignored, the empirical research into neighborhoods and neighborhood effects will be flawed.

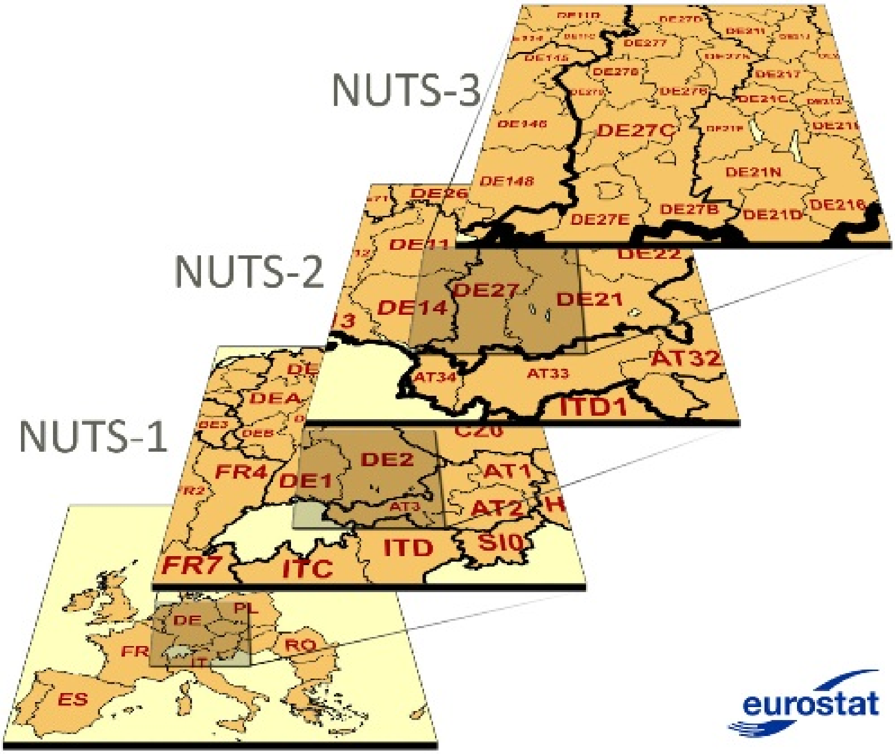
In the following chapter, I will outline how different types of neighborhood studies have explored the concept of the local and how the articles in this thesis have investigated different methodological perspectives to further grasp the essence of the neighborhood in connection with scale, homogeneity, redistricting, and selection with a special emphasis on measuring spatial inequality.

## 2.1. GEOGRAPHY AND DATA

The concept of measuring a group of individuals rooted in geography, as shown in the introductory part of this thesis, is nothing new. With the rise of digitalization and ability to gather data in different forms, however, the need to generalize and formalize how we measure social phenomena, in a geographical sense, is a newer trend.

In the 1970s the European Parliament assigned Eurostat with the task of standardizing area types for research purposes (European Comission, 2018). This standardization was a non-formal agreement among countries until 2003 when the European Parliament agreed on a specific form of division called the nomenclature of territorial units for statistics (NUTS). Each country would keep its local administrative units (LAUs) but would also be committed to deliver data on the three levels of NUTS: major socioeconomic regions, basic regions for the application of regional policies, and small regions for specific diagnoses as illustrated in Figure 5.

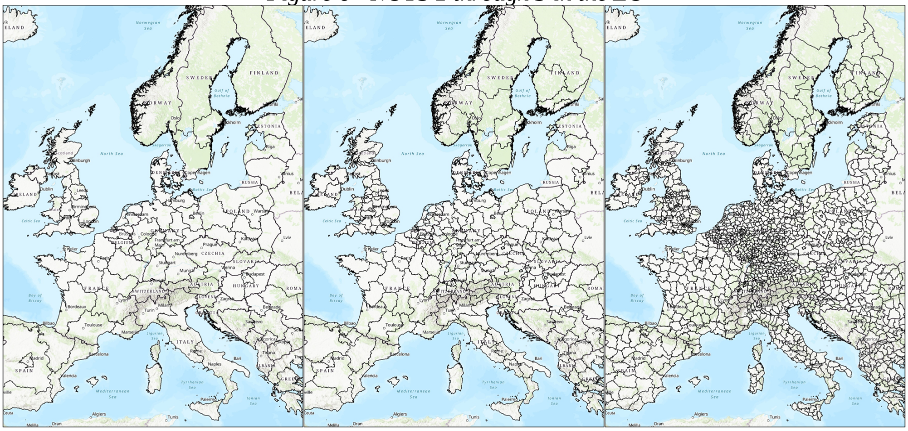
*Figure 5 - Division of NUTS levels 1 through 3*




Source: (Eurostat, 2018).

This standardization made it possible to compare countries with each other and was a big step in comparative studies among the EU membership countries. However, one problem remains unsolved in the use of NUTS: the scaling issue. Because of the overall divisions of NUTS levels 1–3, smaller countries are almost impossible to divide. Where countries such as France, Germany, and the United Kingdom are divided into usable subgroups of on average of 14 regions in NUTS 1, 35 in NUTS 2, and 227 in NUTS 3,  countries such as Denmark, Estonia, and many others have no NUTS level 1, an average of 3 regions in NUTS 2, and only 10 in NUTS 3. Compared with their standard LAUs, this is a very broad and problematic way of dividing the country from a national perspective. This is illustrated in Figure 6.

*Figure 6 - NUTS 1 through 3 in the EU*



This is why, at the same time of the implementation of data gathering with NUTS, the European Commission initiated the grid solution first proposed in the early 2000s by the European Communities (European Communities, 2002, 2003). The grid structure is a simple one in theory: using squares of varying sizes, we are able to encompass any type of geography easily. This was part of many individual EU countries' initiatives that saw the limitations of the NUTS and even the LAUs and required more specific ways to measure land coverage and inhabitants. The standard grid sizes vary from 10 x 10 kilometers to 100 x 100 meters (European Communities, 2002). This neutralizes the common problem of NUTS where comparative analysis of countries was desired but problematic due to unequal geographical divisions.

The power of the NUTS is that all EU membership countries are committed to deliver population data at all three levels. The NUTS follow the local LAUs, and thus, the country has an interest in gathering data in the areas. This is also where the grid structure struggles. None of the EU countries are required to deliver data on grid structures nor to actively utilize any of the common structures found in the grid. This is less problematic when used for the primary reason for its development, environmental research (European Environment Agency, 2019), because environmental data are more easily gathered than population data. However, only a few European countries utilize the square grid actively and these are almost exclusively those with a well-developed register data system.

In Denmark, Statistics Denmark adopted the square grid with divisions of 100 x 100 meters, 1 x 1 kilometers and 10 x 10 kilometers as the main geographical link among different types of registers, as well as the common administrative areas such as municipalities and parishes. This means, that a 100 x 100 meter square grid is the smallest geographical unit possible in the registers and can only be linked to other individual level data if the square has at least 100 inhabitants. The scalability of the grid is one of its main strengths but also increases the complexity of the geography immensely. If one chooses the 100 x 100 meter geography, as I have demonstrated (Lund, 2018), clustering is unavoidable, and if one chooses square kilometers, the range between the least and most inhabited areas would be large. In certain places in Copenhagen, a 1-km square can contain more than 10,000 inhabitants, while a 1-km square in selected places in the rural parts of Jutland would be uninhabited and surrounded by numerous uninhabited squares.

Significant portions of this thesis are devoted to the discussion of the following question: How do we measure place in a meaningful manner, and what are the consequences of different methodologies? Changing geography has consequences, not only in the way we observe the spatial distribution of individuals but also in the underlying questions we ask: Can we expect homogeneity? Are all social phenomena equal in spatial terms?

## 2.2. APPROACHES TO MEASURING NEIGHBORHOOD EFFECTS

Throughout my time writing this thesis, I have identified numerous studies that, in one way or another, make use of place when measuring effects. All the following studies try to overcome the problem of geography and comparability by utilizing different methodologies comparable to Eurostat. There are many traditions and many different ways of accomplishing an approximation of place, but for the purpose of clarification, I have reduced the newer methodological approaches of neighborhood effects to four different groups. Each group works with the concept of neighborhood in a radically different way. The four main groups are administrative areas, nearest neighbor clustering, small area clustering, and Bayesian clustering. In my paper, “From the Dark End of the Street to the Bright Side of the Road,” I have included a thorough discussion of the different methodologies (Lund, 2018), and I will only include a short summary here to connect the discussion to the subject of scalability and the importance of scale.

The first approach is the utilization of administrative areas. Common types of administrative areas are municipalities, parishes, and census tracts (Åslund & Fredriksson, 2009; Cutchin, Eschbach, Mair, Ju, & Goodwin, 2011; Eriksson, Hjalmarsson, Lindquist, & Sandberg, 2016; Flower, 1994; Jensen & Tienda, 1989; Rotger & Galster, 2019; Ruggles, 2014; Sadler & Lafreniere, 2017). These units of measurement are, as noted earlier, often used because nothing else is available. Unless the study specifically investigates how parish borders change over time, not many researchers would assume that areas with very little practical meaning in the present can be a proxy for a neighborhood. The scale of the geography is often large and the number of residents, in a Danish context, varies from 3 to 30,000. For a thorough analytical comparison, see Lund (2018).

The second type is nearest neighbor clustering or different versions of *k*-means clustering (Ferreira et al., 2011; Jhou, Chi, & Hsieh, 2010; Östh, 2018; Östh, Clark, & Malmberg, 2015; Vu, Lee, & Bui, 2014). K-means nearest neighbor clustering has its advantage in the way it uses data. Instead of using any geography, except for distance, it uses each individual to build likeness. The result is a large matrix where everybody essentially remains neighbors, but the distance between them weighs how much they can affect each other (Ying Zhang, 2006). This is also the greatest weakness; the clustering ignores geography all together and treats all types of land as equal, and the neighborhoods are less neighborhoods than they are vague groupings of individuals. In this case, scale becomes fuzzy; the neighborhoods are not contained, and even weaker effects from non-close neighbors are included. This is discussed further in Lund (2018, 2019b).

The third type is small area clustering (Lagerlund, Merlo, Vicente, & Zackrisson, 2015; Merlo et al., 2013). Small area clustering is used in many places and most noticeably in the USA with enumeration districts, also called census tracts (Bellavance, Normand, & Ruppert, 2007; Ruggles, 2014), and it is now used in Sweden with the development of the Small Area Market Statistics (SAMS) (Östh, Malmberg, & Andersson, 2014). In both cases, this clustering encompasses a relatively small area, but is created with a general purpose. In both cases, it is mostly created to divide geography into smaller entities, and in the Swedish case, it is created to generate homogenous areas on socioeconomic characteristics (Östh et al., 2014). The result is areas that, on average, have 1,100 (SAMS) and 5,500 inhabitants and very little concern especially for the more rural parts of the country (Lund, 2018).

The fourth is Bayesian clustering (Borgoni & Billari, 2003; Johnelle Sparks, Sparks, & Campbell, 2013; Logan, Spielman, Xu, & Klein, 2011; Ocana-Riola, Sanchez-Cantalejo, & Fernandez-Ajuria, 2006; Vinikoor, Kaufman, MacLehose, & Laraia, 2008). Bayesian clustering is much more complicated than the previous methods and relies on two things: already existing geographical units and *a priori* assumptions about these units. In general, this means that some form of neighborhood must be present in the data beforehand to completely model the *a priori* assumptions about the place. Bayesian clustering is by far the type of clustering that comes closest to securing homogeneity, but it cannot function as an inductive tool. This is also the biggest drawback of the method: it relies on already existing geography and assumptions revolving around the research question. Thus, considering widely different scales and types of neighborhoods is impossible with this methodology.

Each of these four groups of research handles both scale and grouping very differently. From almost no geographical anchoring in the case of nearest neighbor clustering to very static anchoring in the case of administrative borders, they all offer very distinct solutions to dividing individuals into geographical groups. In this thesis, I have worked with a new methodology to include both the geographical elements and different types of social data into a single method to automate small area generation for use in social science.

## 2.3. CONSIDERING AUTOMATED REDISTRICTING

Not using the common administrative areas requires working with geography before approaching the data one wishes to analyze. The larger the geography, the more work goes into designing the unique areas unless the creation of these areas is automated. This process is called automated redistricting (Altman, 1997).

Redistricting is the craft of circumscribing areas for different reasons, where the most common is political redistricting (Altman & Mcdonald, 2012). In the USA, redistricting is an ongoing process where, because of their election system, it can have a large impact on elections (Altman & Mcdonald, 2012). This is why the discussion of automated redistricting in the literature is primarily focused on how to make redistricting bipartisan (Altman & Mcdonald, 2012; Browdy, 1990; Fifield, Higgins, Imai, & Tarr, 2018; Gelman & King, 2006; Nagel, 2006; Zitlau, Weaver, Siegfeldt, Hess, & Whelan, 1965).

The first use of automation in redistricting came long before modern computers and not in the form we would recognize as automation (Zitlau et al., 1965). During the development of the first wave of automation pioneered by Zitlau et al. (1965), the research was centered on distances within a district. They proposed a measurement that would help understand a given area type as a population moment of inertia by utilizing the individual sum of square distances to the district center. In short, they used the compactness of the population in a given area to compare different types of areas in regard to the population living there. Even though this solution was very advanced for the time and much more data-driven than the standard solutions of cartographers and political scientists, it was relatively unused when forming districts for elections (Altman & Mcdonald, 2012).

From 1960 to 1990, the methods to develop precise districting tools became more and more advanced, but they became less used in the social sciences (Altman, Macdonald, & Mcdonald, 2005). The use of large quantities of data were more common in agricultural and non-human sciences than within social sciences, and thus, the social mapping became more *ad hoc* based on simpler cartography and less on specific automated processes. Considering the local election districts, the non-automated solution was often much faster that creating programs that could handle the large collection of different social data available at the time.

During the 1990s with the introduction of mapping software such as ArcGIS (ESRI, 2019), mapping and working with maps became more relevant, but was still mainly used for non-social data analysis (Mentese & Okuyucu, 2013). In the wake of software development, we have seen a surge in thinking about the combination of social science and geography, as noted above. Software relying on nearest neighbor clustering as Equipop (Östh, 2018) and methods that rely on geographical modeling are becoming more available.

With the availability of both data and processing power, geography and social data can be modeled, calculated, and altered in whatever way one wishes. The question then becomes this: What makes sense? There are many elements to consider when working with districting and many different and important technical and theoretical questions to ask before applying the actual code. Below, I outline different methods I have used in my papers.

### 2.3.1. MAKING STATIC CODE ARGUMENTS WITH THEORY

The first two papers in this thesis are based primarily on static code or static algorithms (Lund, 2018; Lund et al., 2019). The main difference between the primary algorithm developed in this thesis and the evolution of it used in the later papers (Lund, 2019a, 2019b) is the flexibility of the smallest units. As described above, one of the main theoretical and empirical aspects of neighborhood studies is that there are physically and geographically visible objects that divide the landscape and form the environment in which we chose to live. The main point here relies not only on Hull House (Kelley & The Residents of Hull House, 1895) and the way different nationalities slowly form enclaves and smaller groups within a city, but also on newer research that shows class formation and homogenized segregation in the city landscape (Feld, 1981; Krieger et al., 2017; Lamont & Monar, 2002; Legewie & Schaeffer, 2016; Petrovic et al., 2018; Riva, Gauvin, & Barnett, 2007; Smith, Benavides, Pariona, & Tuesta, 2003).

The main logic behind the overall algorithm to handle automated redistricting can be found in the paper “From the Dark End of the Street to the Bright Side of the Road – Automated Redistricting of Areas Using Physical Barriers as Dividers of Social Space” (Lund, 2018) where I develop the first implementation of a model for micro areas. The algorithm for generating micro areas works in two steps and is purely based on a geographic referenced file that covers Denmark in the 100 x 100 square grid and contains nothing more than the inhabitant count for each square in the years 2000, 2005, 2010, and 2015. The first step of the algorithm is designed to secure the minimum requirements from Statistics Denmark where all area statistics need to be aggregated to at least 100[4] inhabitants per areal unit. The algorithm uses different geographical layers provided by Danish Map Supply, which is a part of the Agency

[4] At the time of writing the first paper in 2017, the minimum requirement was 150, but was later changed to 100. The decision to make the change was somewhat informal and no direct communication was sent to inform concerned parties of this change. Most of the later research papers use 100 inhabitants as the smallest amount, but there might be instances where the minimum requirement is defined as 150.

for Data Supply and Efficiency, and they contain information about roads, railroads, water streams, forests, and other elements, as shown in Figure 7.

*Figure 7 - Initial map of geographical dividers*


These physical objects will serve as neighborhood borders; objects that divide the physical space and creates a visible division between one group of housing and the next. Since Statistics Denmark utilizes a square grid of varying sizes to couple geographical data to individual level data, the above map needs to be processed. Below are a depiction the Danish capitol Copenhagen with a square grid overlay (Figs. 8 and 9, left) and how the above map looks when borders are modified to fit the square grid (Figs. 8 and 9, right). This is done by assigning each grid cell to the area where the cell best belongs percentagewise. This changes the overlay to the geography slightly where roads go from being straight lines to more angular to accommodate the square grid.

*Figure 8 - Close-up of Copenhagen with square grid and initial merging*

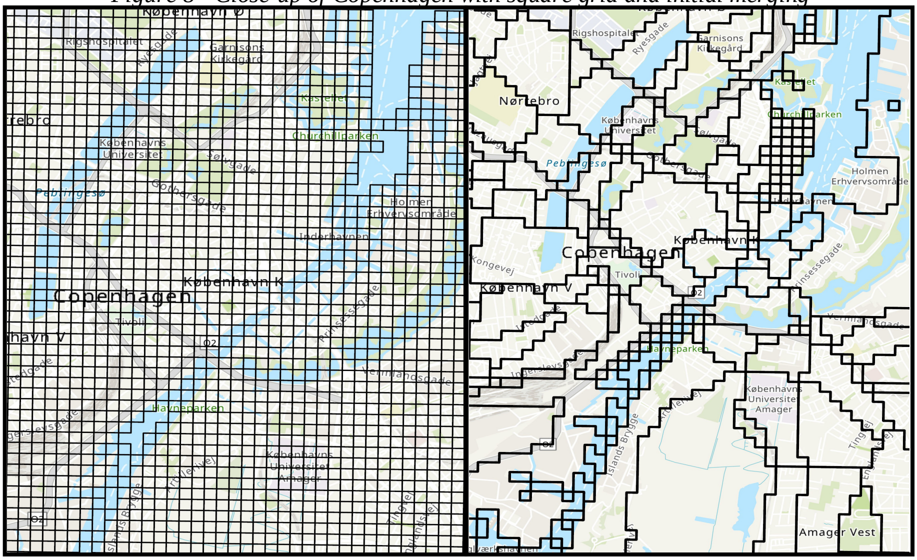
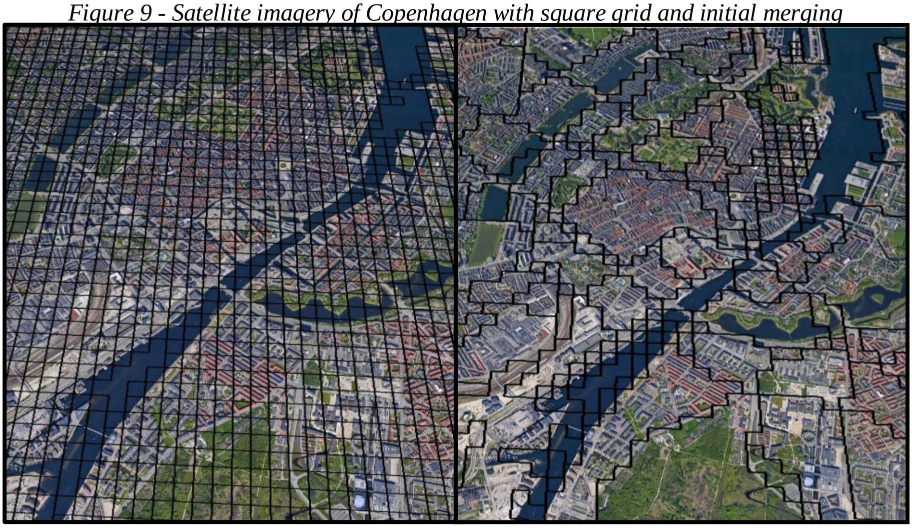



*Figure 9 - Satellite imagery of Copenhagen with square grid and initial merging*



One of the main problems in using only physical objects as barriers in a macro perspective is data cleaning. Consider a highway with multiple tracks or even a roundabout in a smaller city; they are technically barriers that contain space between them. By simply performing a geographic cut[5] at the borders shown in Figure 8, this results in an area count of more than 100,000 with almost 80% of them being either less than 10 $m^2$ in size or with an inhabitant count of 0. Roundabouts alone accounted for almost 5% of all areas created. This same tendency can be seen in both the center part and upper left quadrant of Figure 9 (right) where areas are cut off as single 100 x 100 meter squares. In Copenhagen, it is not impossible to find single squares that house more than 100 individuals, but since some of the squares are formed by bridges crossing waterways, it becomes important to implement further clustering before actual individual level data can be applied.

The second part of the algorithm is by far the most demanding, both in terms of algorithmic arguments and processing power. The problem with geographical clustering is adding a physical dimension simply because the complexity quickly exceeds what normal computers are able to handle, even when the data still only consist of the square grid with inhabitant count for each square and the physical barriers. The main goal is to reach 100 inhabitants per area to satisfy discretion criteria, but since a neighborhood exists as a non-divided, bounded entity, we rely on physical proximity as well, and thus must enforce localized clustering. After data cleaning, I found that there were 20,940 areas that had inhabitants present in the years 2000, 2005, 2010, and 2015 (Lund, 2018).

There are two overall solutions to make sure that the discretion criteria are met without adding any additional algorithmic arguments. The first solution is to start the clustering at a random geographical location and the move the clusters from that point, while the other is to calculate all possible merges to get as close to 100 inhabitants as possible. Regarding the first solution, there is one very problematic factor: the starting point. When selecting the starting point at random, one is left with a non-reproducible and ever-changing map that is heavily influenced by where the random point is located. Choosing a random polygon at the beginning also creates an avalanche effect; all areas merged from a single point will be completely different for each time the algorithm runs.

The second solution is more precise, since the mathematical best way of merging can be defined as the solution with as many areas with 100 inhabitants that would be reproducible. However, without adding any additional arguments, the pure mathematical complexity increases exponentially when trying to merge areas based solely on inhabitant count to reach 100 individuals. Each area has an average of 5.4 neighboring areas, and with no logical geographical starting point, all permutations of

[5] Specifically, this is done by changing multiple geographical features to polygons, so all closed circuits are defined as single polygons.

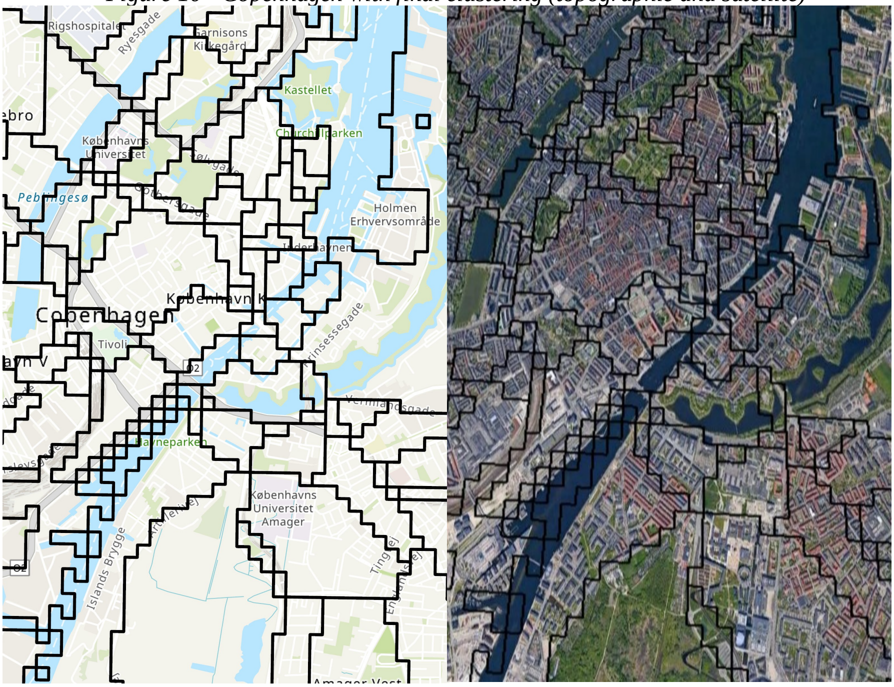
merges must be calculated by testing all starting points with all merges one by one in a very long and demanding iterative loop. The theoretical lowest number of calculations that needs to be made before a possible solution is more than $10^{105}$ and thus not a wise solution (Lund, 2019a) solely because of computer constraints.

Instead of using brute force to solve the problem, a more theoretically driven solution is put in place. Since the overall aim of clustering in this thesis is to capture the social element of neighborhoods, the arguments that follow are either directly driven by a theoretical understanding of neighborhoods, where the physical element of the neighborhood affects the way individuals inhabit an area, or by the desire to ensure that the solution is as comparable in the inhabitant count as possible. This follows algorithmic properties in the sense that the list of arguments is replacing matrices and thus this goes from a complex multidimensional problem to a sorting problem (Lund, 2018).

The second part of the algorithm follows a similar pattern as the first, but has added dimensions; first, it registers all areas as applicable for clustering. Then, it sorts all areas by percentage borders shared, by how many merges it would require to reach an end goal of at least 100 inhabitants in each, and by what solution creates the most areas with as many areas as close to 100 inhabitants without going too far above the 100 (Lund, 2018). By making it a sorting algorithm and not relying on approximation, I am able to reproduce the same grouping each time until I change the rules of sorting.

This results in 8,042 unique areas, all separated by physical borders and completely physically connected. In Figure 10, I show a topographic map (left) and satellite map (right) of Copenhagen created after the second algorithm was run. Looking specifically at the upper middle part, it becomes clear that the second part has taken care of much of the noise produced solely by barriers where many smaller squares were left as individual squares by the first clustering. By sorting, I have been able to keep the very densely populated areas as they are, while the less populated and those produced by noise in the geography have been eliminated by being joined by adjacent areas.

*Figure 10 - Copenhagen with final clustering (topographic and satellite)*




The average inhabitant count in the new areas is 537[6], while the standard deviation is 212.37 and the range is 100–1,487. Compared with parishes with an average of 2,097 inhabitants, standard deviation of 3,718.69 and a range of 25 to 45,187, smaller areas have improved the comparability of spatial analysis simply by narrowing the issues that arise when comparing unequally sized[7] areas.

In my paper "From the Dark End of the Street to the Bright Side of the Road," I test my assumption that small areas generated with a theoretical base are better at capturing homogeneity than small administrative areas such as parishes (Lund, 2018) and find that smaller areas are more homogeneous. To test how much of the increase in homogeneity is a result of data smoothing, I construct different types of similar, random reductions of neighborhood scale polygons. While there are signs that some of the increase in homogeneity is based on data smoothing, this is only part of the

---

[6] This is mainly due to especially larger cities such as Copenhagen, Aarhus, and Aalborg where even one 100 x 100 meter square is inhabited by many more than 100 inhabitants. In some larger apartment blocks, it is not uncommon to have as many as 1,000 inhabitants per square.

[7] Measured solely by inhabitant count, the total geographical size of small areas are, on average, much smaller than parishes as well, but since these areas are specifically designed to work with social data, geographical size is not a goal in itself.

increase. By looking at completely randomly generated clusters and comparing them with the small area model, I show that using physical barriers as dividers of space is more efficient at capturing homogeneity (Lund, 2018).

## 2.4. FURTHER CLUSTERING BASED ON LIKENESS

One of the main drawbacks of this methodology is that the search for the smallest, most comparable units sometimes makes for a confusing overview. This methodology ignores whether the areas are comparable in regard to social factors, which means that common neighboring areas measured on social characteristics are still counted as two separate areas. The scientific reason is obvious: the smallest units of measurement ensure precision measurements across different types of variables and scientific questions, but since there are more than 8,000 base areas created, the product can be problematic to understand visually.

Clustering is further possible in two very distinct ways: non-geographical clustering based on likeness and geographical clustering based on likeness and closeness. Both methodologies can teach us something important about the place, but choosing which type to use is detrimental to capturing the desired neighborhood effects. When questions about general policy making appear, there is a point to simplifying the results without compromising the integrity.

Considering the first type of clustering, the non-geographical clustering, we are able to learn much about similar types of areas. This does not reduce the number of areas, but it clusters them together in similar types. Using this methodology, it becomes possible to isolate specific types of areas based on either a theoretical typology or other isolated findings one wishes to locate on other geographic locations. Below, I outline the model used to cluster for deprivation in the article "I Like the Way You Move" (Lund, 2019a), which can be applied for finding similar areas that might not share a similar geographic relationship but instead similar social characteristics. First, I utilize a simple Euclidian distance matrix to handle multivariate distances. This takes the form of:

$$d_{ij} = d\big(\{X_i\},\{X_j\}\big) = \left\|X_i - X_j\right\|, \qquad 1$$

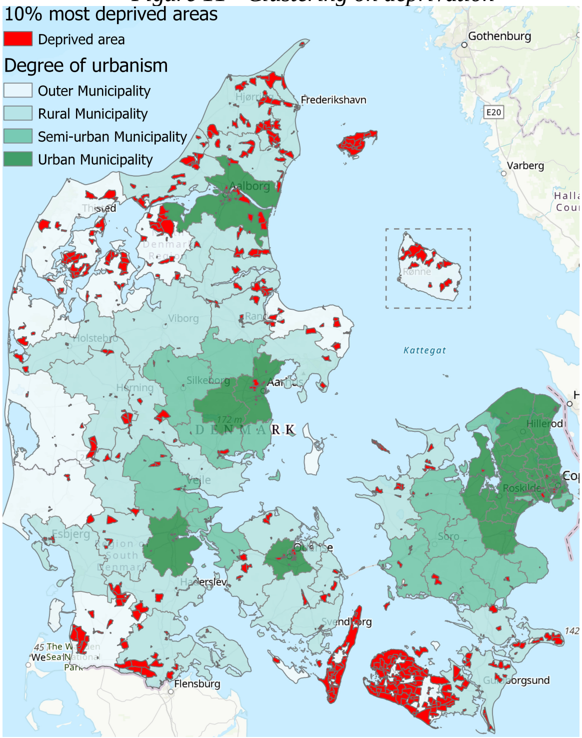
letting the cluster distances be the Euclidian distance between the i'th and j'th cluster for the complete matrix of the variables desired, and then followed by a recursive formula (Lance & Williams, 1967) defined as:

$$d_{(ij)k} = a_i d_{ik} + a_j d_{jk} + \beta d_{ij} + \gamma \left|d_{ik} - d_{jk}\right|, \qquad 2$$

while utilizing Ward's method by defining $a_i$ as $\frac{n_i+n_k}{n_i+n_j+n_k}$, $a_j$ as $\frac{n_j+n_k}{n_i+n_j+n_k}$, $\beta$ as $\frac{-n_k}{n_i+n_j+n_k}$ and $\gamma = 0$ (Lance & Williams, 1967). This method minimizes the error sum of squares and is further constrained to only let each cluster form in a first step, so that the total cluster size is aimed to generate the smallest clusters possible and to only allow areas with the most similar features to form new clusters. This also considers not just clustering on simple means but also the distribution of an area.[8]

This results in a similar type of geographical distribution, as seen in the Figure 11, which was created as a composite measure of income, unemployment, and educational attainment (Lund, 2019a). Areas marked in red show the bottom 1st decile in the overall index. This was used to isolate the 10% most deprived areas in Denmark.

*Figure 11 - Clustering on deprivation*




As shown in Figure 11, clustering is still concentrated in certain areas, but are much less sensitive to distance. Some locations have a dense cluster of deprivation, but are still classified as unique entities and not as uniform, geographical entities. In the above

[8] Distribution is a multitude of measurements for each variable included in the clustering. In the above example, I have included minimum and maximum values, skewness, median, mode, variance, and kurtosis.

case, it made sense to disentangle geography from deprivation because the overall point was to locate deprived neighborhoods no matter where they were located and later investigate the life course effect of these neighborhoods (Lund, 2019a). If the goal was instead to investigate how different segments of deprivation, such as unemployment, could be more problematic in rural areas compared to urban areas, this tells us very little.

Because the Euclidian distance is used to measure equality between social variables, geographical distance becomes harder to incorporate.[9] Geographical distance can be incorporated, but without modification, geographical distance becomes noise in the model instead of an actual measurement to help create similar clusters. The reason for this noise is because geographical distance will be measured, even in a standardized form, as part of $||X_i - X_j||$, which is the set of distances between non-geographical variables. If the distance is large between socially deprived areas, it will not negate the cluster, but will instead introduce unwanted heterogeneity, where geographical distance only in extreme cases will make the difference. If precision is wanted, distance must be introduced as a separate object not directly associated with the dissimilarity matrix of socioeconomic values. This is why geographically sensitive clustering must be considered if we follow an assumption that deprivation could be different at different locations, or if we are looking for specific clusters of deprivation set in a smaller, confined geographical space.

Most clustering based on geography follows a type of distance-weighted matrix to penalize observations that are geographically far from each other (Duque, Aldstadt, Velasquez, Franco, & Betancourt, 2011; Hong & Sadahiro, 2014; Russ & Kruse, 2011). Some researchers primarily make use of a density-based function, where the spatial cluster based on concentration in itself is the goal and not the underlying features that make it interesting in a geographical sense (Ester, Kriegel, Sander, & Xu, 1996; McInnes, Healy, & Astels, 2017). In other words, density-based clustering only clusters on spatial density and not on density of second-order values such as socioeconomic phenomena.

The second type of geographical clustering can handle both geography and second-order values. There is no single way the literature defines adding spatially constrained weights to a dissimilarity index, but in a generalized form, most modifications are made citing Oliver and Webster (1989), where they outline a model to modify a matrix to satisfy second-order values ascribed to the area in question (Guo, Ahn, & Zhu,

[9] In this example, I use Euclidian distance, which is the most commonly used distance matrix measurement, since it relies on a version of equal distances between measurements.

2015; Oliver & Webster, 1989; Warren, Metternicht, & Speijers, 2006). Below, I outline a standard form as:

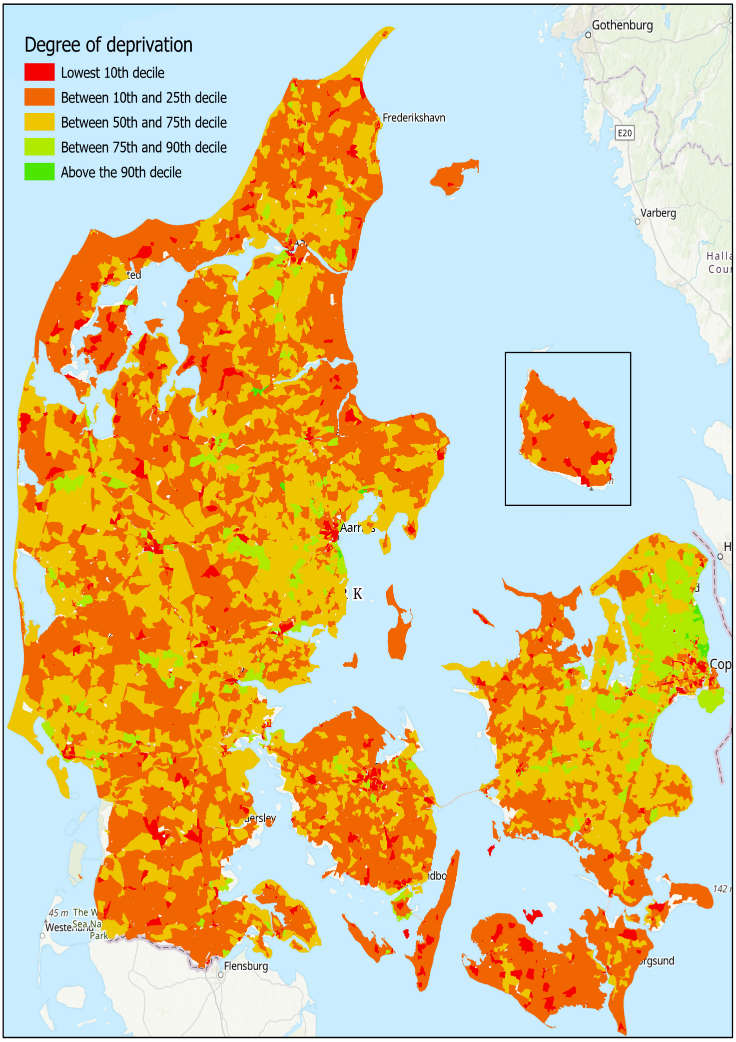
$$d_{ij}^{*} = d_{ij} \cdot f\left(z_i - z_j\right) = d_{ij} \cdot w_{ij}, \qquad 3$$

where $d_{ij}^{*}$ is the modified dissimilarity between neighborhoods, *i* and *j*, $z_i$ and $z_j$ are the locations of the values of interest,[10] and  is an evaluation of *f* at points *i* and *j*. This basically creates a weighted matrix where the distances are multiplied by the distance between an XY coordinate scheme (Warren et al., 2006). What this does is  it forces objects that are physically near and have shorter distances in the second order, or in this case, socioeconomic variables closer to each other, while values that have a longer distance will be pushed further apart when applying Eq. (2). This means, that the clustering will be less transferable in a purely non-geographical sense, where deprivation can change form or be the same in different places without it being put in the same type of cluster.

In praxis, this can result in a large reduction of individual areas and form larger, more easily understood groupings. In Figure 12, I show a model where I have used the full socioeconomic index formed with standardized values of area average income, unemployment, and educational attainment.

[10] Going with the notation of Eq. (2), this is the location of the sets of variables of interest captured in $X_i$ and $X_j$.

*Figure 12 - Further clustering on deprivation*




As shown, multiple clusters have formed, and we have a more general idea of the typologies in Denmark. This, however, does come with the price of transparency. Even though clusters around Copenhagen appear to be different, they are actually very much alike in the raw index, varying less than 1%. Because of the structure of weighted dissimilarity matrices, they are forced to be different in the final clustering. In some cases, this is wanted. If the question asked is based on a theory that different types of socioeconomic background will have different effects, it is useful to even have very similar clusters separated, but if the goal is to understand a uniform tendency, then this methodology will severely skew the results.

The problem with further clustering is twofold; first, the geography changes with what we are measuring. Even though it seems logical that people group differently on different parameters, the consequences could be counterproductive to the goal of bettering the visual understanding of the map. Secondly, scale will change massively, leaving some areas small in both inhabitant count and geographical size, while others are much larger. We have grown so accustomed to administrative areas, and especially when we interpret data, that constant changes in geography easily lead to confusion.

In a Danish setting, most policy making is being done within regions or municipalities, and each local municipality has its own way of using maps to understand its data. Small area models and further clustering might not help the individual municipality to make decisions, but where these types of analyses make sense is in the direct implementation of local policies. If social problems are located within a municipality, it will help immensely if the municipality is able to show how areas change and exactly where problems arise. By choosing the correct type of clustering, this task becomes easier and can help national policy makers to better understand where concentrations of deprivation are located.

### 2.4.1. PURE AUTOMATION WITH MACHINE LEARNING

The overall problem with the static code is inherent in the name; it is static. This means that there is complete control of the algorithm, and all parts are run for reproducibility. However, this also means that the model will never improve or otherwise challenge the inherent assumptions underlying the 100 x 100 square grid clusters. To incorporate an evolving environment and, at the same time, reduce complexity in a permutation setup, newer methodologies are needed.

What lies beyond the more static arguments of code is a more intuitive and purely automatic version of redistricting that applies a set of machine learning algorithms based on cluster analysis. Commonly, machine learning algorithms are divided into two different categories: supervised and unsupervised. The supervised category of algorithms is based on learning through a test dataset where the outcome is given. The algorithm trains with the training dataset and has the goal to predict the correct given outcome in the data and then evaluate the performance on the unknown test data (Hastie, Tibshirani, & Friedman, 2016). The problem with automating area clustering in supervised models is that there are no test data given in an unknown spatial clustering. The number of areas after clustering cannot be given beforehand, and even if it was possible by calculating the closest number to reach 100 inhabitants in each area, there would be no way for the model to evaluate specific area clusters.

Even though standard $k$-means clustering technically falls in the category of unsupervised machine learning, it becomes computer intensive as soon as $N$ increases, and in the case of register data level N, pure clustering is both problematic in processing power but also, as I outline above, in the way it handles geographic data. The solution, as outlined in the paper "I Like the Way You Move" (Lund, 2019a), was to introduce a stochastic gradient descent[11] in combination with spatially constrained weights. Because this is used to form areas of at least 100, there is relatively little information to take into account. The goal of the algorithm is still to isolate areas with more than 100 inhabitants that are in close connection with each other.

What differs from the static code is the relaxation of the sorting rules. Where the static code relies completely on a sorting mechanism to optimize the clustering solution, this method relies on a set of functions that seek to secure 100 inhabitants, minimize geographical size, and minimize number of merges. This is done by setting a cost function where inhabitant size below 100 is set infinitely large no matter if the result is 1 or 99 inhabitants. At 100 inhabitants, the cost is zero and then increases with each person above 100 per area (Lund, 2019a). This is done to make sure that there are no areas created with less than 100 inhabitants, while still maintaining 100 as the main goal. The other two functions are simply based on an incrementing counter; 0 merges have no cost and then costs increase in increments of 1 for each merge. Geographical size cost is measured based on the smallest possible geography, the 100 x 100 square. It introduces a penalty for larger areas, and since the goal simply is to minimize area size, this means that the model prefers instances where the final area is small.

A standard stochastic gradient descent utilizes a sum of cost as:

$$Q(w) = \frac{1}{n}\sum_{i=1}^{n} Q_i(w)\,, \tag{4}$$

where $w$ is the weight being applied to the loss $Q$ that must be optimized as the sum of costs at each $i$ over $n$ samples (Bose, Maheshwari, & Morin, 2003; Bottou & Bousquet, 2008; Lin, Lu, & Xiao, 2014). The difference between a standard gradient descent clustering and a stochastic gradient descent solution is the way the true gradient is calculated (Bottou & Bousquet, 2008; Ruder, 2016). As a standard gradient descent is designed to work in a training environment where batches of the data are used in 80% to 90% of the full data to train the model, the true gradient approximation

[11] Gradient descent is an optimization algorithm that helps find the minimum of a function by moving toward the steepest descent defined by the negative of the gradient (Bottou & Bousquet, 2008; Ruder, 2016). In a reduced form, gradient descent can be seen as $b = a - \gamma\nabla f(a)$, where b is the bias or next position of the gradient, a is the current position in the matrix, gamma is the waiting factor, while the true gradient, $\nabla f(a)$, is the current direction to reach the steepest gradient descent. While gradient descent is a relatively simple form of optimization, stochastic gradient descent utilizes processor power better because it runs at randomized single points at the data and not on the full batch data (Bottou & Bousquet, 2008).

requires substantial computational power (Bottou & Bousquet, 2008; Ruder, 2016). The differences in stochastic gradient descent are thus found in the way the gradient is calculated. Where gradient descent minimization algorithm comes in the form of:

$$w_{t+1} = w_t - n \sum_{i=1}^{n} \nabla Q_i(w_t)\,, \qquad 5$$

where the maximum gradient direction  that is being optimized for each  is calculated in a summation form for the full batch at each step of the clustering. The advantage of this method is the completeness of each step. The approximation of the true gradient is calculated for each $i$ in $n$ batches. This is, however, still too data intensive even for server scale computers when factoring the spatial element. The solution implemented here is the stochastic gradient descent that is:

$$w_{t+1} = w_t - n\nabla Q_i(w_t), \qquad 6$$

where the updated maximum direction of the gradient no longer takes the form of the total sum of the batch, but instead as a randomly chosen $i$ from the data. This introduces a clear weakness; the gradient is no longer based on large batches of data, but only on a randomly selected data point and thus sensitive to randomized skewness. This is also the method's advantage. Since the algorithm runs iteratively, it is able to learn what creates the best gradient for each iteration. Even with random data points, it is able to discern what makes sense to cluster and what is less important, so that the final clustering will define its own set of rules much like the ones presented in the static code section. The only difference is that this methodology is solely based on machine learning and not on predefined rules.

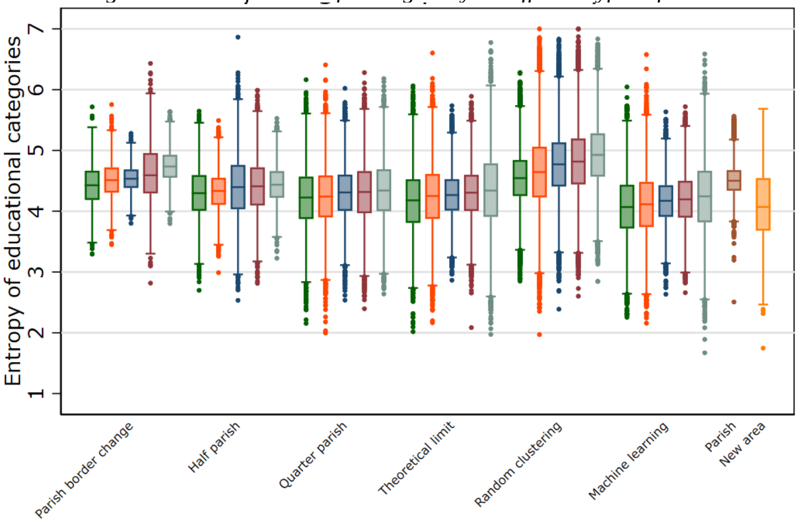
When gradient descent is used in other types of machine learning, it is often in a train/test data environment where the data is split into two parts. In a clustering environment where the goal is to cluster a specific geography and not to make a general model for other types of geography, there is no real train/test division, but instead randomly chosen starting points for each iteration of the algorithm on the full dataset.

While cluster analysis is most commonly used to reduce the complexity of data and can be compared with other commonly used methods of complexity reduction such as factor analysis, latent class analysis or correspondence analysis (Bien & Tibshirani, 2011; Gondek & Hofmann, 2007; Krymkowski, Sawinski, & Domanski, 1996; Rubinov, Soukhorokova, & Ugon, 2006), the use shown above is less focused on a heavy set of multidimensional reductions, but instead on the optimization of only a few dimensions but with large quantities of data. In practice, when the above stochastic gradient descent is used iteratively in the form of a Lance and Williams algorithm (Lance & Williams, 1967) shown in Eq. (2) and the cost function is modified using spatially constrained weights (Oliver & Webster, 1989) shown in Eq.

(3), the clustering is set to optimize only for spatial proximity and the optimal number of inhabitants in each cluster.

With stochastic gradient descent, the physical barriers become less relevant, and the clusters are purely based on the mathematical problem of reaching 100 inhabitants from a 100 x 100 meter grid. The interesting finding with this methodology is that the difference between a geographical model that takes physical borders into consideration and the above more automated model is not large. There are more areas generated, and the solution is closer to a mean of 100 inhabitants on average, but the homogeneity measured on common socioeconomic variables either stays the same or is slightly worse. This is depicted in Figure 13, where the homogeneity[12] of educational attainment is calculated. The other methodologies shown are described in full in (Lund, 2018).

*Figure 13 - Comparison of homogeneity in different types of area clusters*




Modified figure from (Lund, 2018).

[12] Homogeneity is measured as an exponential function to the Shannon entropy in the form of: $exp\left(\sum_{i=1}^{n} p(x_i) log_b \frac{1}{p(x_i)}\right)$, where $x$ represents the frequency of each of the seven educational categories in area $i$. The advantage of using this method is discussed in the paper "From the Dark End of the Street to the Bright Side of the Road" (Lund, 2018) and more in depth in *"Entropy and Diversity"* (Jost, 2006).

The machine-learning algorithm has been run without a seed to force some variation in the training and solution to compare different solutions. As in Lund (2018), the algorithm has been trained and clustered 100 times, and the solutions are ranked from best (left) to worst (right) in each of the different algorithmic categories shown in the bottom legend. Comparing the best solution in the machine learning category to the theoretically formed areas in the far right bar, it becomes clear that very little is gained by adding the more computationally heavy approach of a stochastic gradient descent learning environment.

This raises an interesting question about how we should approach the physical aspect of neighborhoods in general. Even though there is a point that municipalities and parishes are bad proxies for neighborhoods, there seems to be a limit to where further specification will accomplish more. In his paper “The Focused Organization of Social Ties,” Feld approaches this by looking at different social foci: “The focused organization of social ties implies that a researcher should understand and measure relations to foci in order to understand the structure of a network” (Feld, 1981).

With this, Feld points to not only the social aspect of human interaction but to the field in which the interaction takes place. He advocates for a theoretically grounded approach to data analysis and even suggests how larger roads can work as dividers of social space as well (Feld, 1981). Even though the automatic approach can reach almost the same levels of homogeneity and much smaller average inhabitant counts, it lacks other physical elements apparent in the environment in which we live. This is not evidence for choosing one method over the other, but it indicates the fact that neighborhood research needs to look not only at neighborhoods as a geographical, architectural, or social phenomenon but as a combination of the three.

## 2.5. THE PROBLEM WITH SELECTION AND TIME

Even though the discussion of causality is an always present constant in all research, the broader field of sociology of place is only partially committed to fully account[13] for the problems that arise when measuring change over time in systems or entities that have the potential to be auto-correlated or suffer from selection bias (Jovanovic, 1996). In this section, I discuss some of the methods I have used in my papers and elaborate more on both the methodological foundation as well as the theoretical framework behind it.

Neighborhood selection is always a problem as we cannot randomly assign neighborhoods at birth to investigate causal mechanisms in an experimental design.

[13] By “fully account,” I, of course, mean to the extent that we are able to isolate causal mechanisms in social systems.

Investigating mechanisms becomes more problematic when applying time, because not only is the placement of individuals not random, the individuals' placement in previous years is a very strong predictor for the following year. Moving from smaller rental places in the rural parts of Denmark to large owner-occupied villas in urban settings is not as likely as staying in the same place or moving to a similar place. This is what Wodtke et al. describe as *dynamic selection* (Wodtke, Harding, & Elwert, 2011). Without accounting for dynamic selection, which in a mathematical sense can be seen as autocorrelation, the estimates of neighborhood effects will be inflated (Wooldridge, 1995).

While dynamic selection is discussed in the paper "Moving to Prosperity" (Lund, 2019b), this following section is a discussion of causality and how a counterfactual framework can substitute non-experimental data. The counterfactual methodology relies on "what if" questions, such as: How long would person A be unemployed if person A, who is unemployed and has no education, has completed a specific educational program? The goal is to approximate the true effect of a given phenomenon by using instances in the data where other individuals share a similar socioeconomic background but differ in educational attainment. In the above example, the educational attainment becomes a "treatment" of which we can test the effect.

The counterfactual framework did not fully develop before 1973 when David Lewis proposed a proper theoretical base (Lewis, 1973). This was further applied in statistics by Donald Rubin where the first modern use of the potential outcomes was developed (Rubin, 1974). Two of my papers (Lund, 2019b, 2019a) rely heavily on this way of thinking because of the unique properties of the potential outcome methodology that allows one to isolate certain effects and to address selection more precisely. To fully understand and break apart the mechanisms inherent in neighborhood effects, one needs to first make sure that we can isolate specific mechanisms from each other in a practical and theoretical framework. This is a large part of the movement analytical sociology offers, and many both theoretical and empirical examples of mechanism isolation have been made (Buck, 2001; Demeulenaere, 2011; Hedström & Bearman, 2009; Hedström & Swedberg, 1996; Hedström & Ylikoski, 2010).

The potential outcome framework is based on an impossible question: What if more than one mutually exclusive outcome happens to the same individual? (Lewis, 1973; Rubin, 1974). The problem with this question is partly philosophical and partly statistical because most researchers agree that, if possibilities are mutually exclusive, they cannot occur for the same individual. The classical experiment is often regarded as the ideal design for identifying causal patterns because we are able to calculate a naïve individual or average causal effect simply represented as:

$$ICE = Y_1 - Y_0, \tag{7}$$

$$ACE = E[Y_1] - E[Y_1], \tag{8}$$

Here, the causal effect is the simple difference between the resulting outcome of the test and control. Since non-experimental studies rely on other methods than randomized control and test groups, the naïve approach will be heavily skewed. The reason behind this skewness is the selection that occurs in almost all instances where we could be interested in effects of “treatments.” We cannot observe an individual living in a deprived neighborhood and in a wealthy neighborhood simultaneously. Furthermore, we cannot separate the status of living in different types of neighborhoods from the selection to move into them. The composition of a neighborhood depends both on the motives of inhabitants for entering or leaving it, and on internal processes.

To approximate a causal relationship which can identify the effect of deprivation, we need to be able to compare people who were subject to the “treatment” of deprivation with similar types of people who did not receive the “treatment.” The only way to achieve this is to make sure that the selection problem is answered both theoretically and empirically. Methods such as propensity score matching and nearest neighbor matching were often thought of as solutions to these problems, but newer methods have been proven to more adequately handle treatment selection (Wodtke et al., 2011; Wooldridge, 2002, 2010).

Wodtke et al. (2011) handled dynamic selection in a very similar way using inverse-probability-weighted regression adjustment, and the same method has been used in the papers “Moving to Prosperity” and “I Like the Way You Move” to address treatment selection and to isolate very specific mechanisms inherent in the spatial perspective of upbringing. In this case, the counterfactual model becomes a methods of moments doubly robust estimator (Wooldridge, 2010) that has some very unique qualities. Wodtke et al. (2011) utilized a classical “treatment” of a dummy measurement of deprivation, while my papers instead utilize a multinomial logistic approach to account for multiple treatments. This is explained in depth in (Lund, 2019b, 2019a).

Causality is important in social science. However, causality should not be understood as the outcome of a simple, mechanical process. The issue is, therefore, affiliated with the issues of autocorrelation and selection, especially in neighborhood research. If left unattended, the results can be severely skewed and, in the worst case, wrong. Using methodologies such as potential outcomes and counterfactual frameworks, while a solution to these problems, imposes a different problem: complexity. In my search to fully account for selection, causality, and specificities, I argue that we need to apply different methodologies and scales to solve problems such as spatial deprivation, inequality, and ghettoization. Research should also work theoretically on how deprivation is understood and practically on how policy making can address changes and how can we only get closer to the true mechanisms by having proper specifications.

# CHAPTER 3. RESEARCH PAPERS

## 3.1. PUBLICATION LINKS AND LOCATION

This thesis is built on four different papers. Below are the full references and links (if available):

Lund, R. L. (2018). From the dark end of the street to the bright side of the road – automated redistricting of areas using physical barriers as dividers of social space. *Methodological Innovations*, *11*(3), 1–15.

Lund, R. L. (2019b). Moving to prosperity? The effect of prolonged exposure to neighborhood deprivation. *Scandinavian Journal of Educational Research*, 1–17.

Lund, R. L. (2019a). I like the way you move - The effect of moving to and from different types areas during adolescence on later socioeconomic outcomes. *Geoforum*.

Lund, R. L., Jørgensen, A., & Riis, O. P. (2019). Social geographical patterns in membership of the established church in Denmark. *Nordic Journal of Religion and Society*.

## 3.2. RESEARCH PAPERS AND MAIN FINDINGS

In the following section, I provide short summaries of the research papers.

In research paper one ("From the Dark End of the Street to the Bright Side of the Road"), I aimed to investigate the viability of automated redistricting in combination with register data using a static code, as described above, to generate small, comparable units of analysis. I used three different measures to test for homogeneity: accumulated education measured in months, income, and ethnicity. The main aim was to compare how well the algorithm captures homogenous neighborhoods in both administrative units, but also more randomized smaller units. In educational attainment, income, and ethnicity, smaller areas performed far better at capturing homogeneity than administrative, parish level areas.

One thing that could be problematic about making the comparison is the overall concept of data smoothing (Faraway & Simonoff, 2006; Petrovic et al., 2018; Petrović, Ham, & Manley, 2017). There is a chance that the reason for an increase in the level of homogeneity is purely that the average $N$ decreases as the areas grow smaller and thus occurs by random chance. To test for this, I created five different versions of a non-theoretical border algorithm that explores homogeneity when changing size and shape of parishes, when ignoring physical barriers completely, and when doing a completely randomized clustering (Lund, 2018). Comparing the theoretical small areas proposed in the thesis indicates that all other versions of clustering show the same result: theoretical small areas are better at capturing neighborhood-level homogeneity (Lund, 2018, p. 11). Besides finding evidence for the hypothesis that neighborhoods are more likely to form homogenously when there are physical separators present, it was also evident that, when clustering was completely segregated from geography and only aiming to secure 100 inhabitants per area, the homogeneity was much worse than all other solutions that were grounded in geography.

The second research paper, "Moving to Prosperity?", explored the effect of living in deprivation over time and if length and time of exposure to deprivation has an effect on later life educational attainment, using a cohort of all Danish children born from 1980 to 1986 ($N$ = 256,345). This paper built on the findings of Wodtke, Harding and Elwert (2011) where they determined that the previous literature interested in the effects of living in deprivation often only measured place of living once (e.g., place of living at birth). This introduces an interesting duality; the question becomes whether the duration of exposure to deprivation matters or if it is purely where one is born that has an impact on later life outcomes (Wodtke et al., 2011). Wodtke et al. relied on a counterfactual research design as well as utilizing methods to reduce the impact of dynamic selection as described earlier, but they failed to account for the area itself.

Thus, I identified three distinct problems to overcome before adequately capturing neighborhood effects: direct effects, selection and districting (Lund, 2019b). Where direct effects can be ascribed to the individual effects of socioeconomic background and other individual level effects as most forms of statistical control can account for, the other two are more complicated to account for. By applying both statistical control and inverse probability-weighted regression adjustments, I was able to capture the overall individual effects, and by combining this with redistricting to capture the truly deprived neighborhoods, I showed that birthplace and accumulation is only part of the story.

Even though the results showed clear evidence for the accumulation hypothesis, where the number of years that are spent in deprivation become increasingly more impactful on later life educational attainment, the story became less clear when I looked at what happens when one moves to and from areas of deprivation. The pure effects of living in deprivation until age 18 are by far the lowest on income and unemployment, but moving away from deprivation during early childhood will only counteract a fraction of the expected effect (Lund, 2019b, p. 14). These results are interesting because the most common way of measuring deprivation is that either a single measurement or a full account for accumulated deprivation is enough to explain the negative effects later in life, but this paper showed that a combination of exposure and timing explains more than either of the two paradigms alone.

The third research paper, “I Like the Way You Move,” followed closely in line with the second paper, but instead of a focus on exposure/timing, this paper investigated if the effects of deprivation on educational attainment, income, and unemployment are the same in different degrees of rurality (Lund, 2019a). One of the major discussion topics in sociology of place is the effect of urbanization and how the class divide is driving a rural/urban segregation (Savage & Warde, 1993; Urry, 2009). During the last decade, much of the EU policy making has been directed to the problems in the rural parts of all membership countries and how to counteract some of the problems that arise when urbanization draws people to the city (European Comission, 2007). In a Danish setting, policy implementations are focused on both urban deprivation and ghettoization, as well as rural deprivation (Ministry of Transportation, 2018). However, the main question these implementations are built upon has largely been left unanswered: What is the effect of growing up in rural and urban areas, and are the effects of deprivation the same no matter the degree of rurality?

In this third paper, I utilized the same overall counterfactual design as in paper two where I tested the overall outcome on income, educational attainment, and unemployment at age 30 and if this varies, based on place of upbringing. This means that the data were split into four different overall groups based on the urbanization classification used in Denmark: urban municipalities, semi-urban municipalities, rural municipalities, and outer municipalities (Ministeriet for Fødevarer Landbrug og Fiskeri, 2011). Using the same cohort as in paper two, I followed all children born from 1980 to 1986 ($N$ = 256,345) from birth to the age of 30, where I measured the aforementioned outcomes. The paper tested small areas both in the four overall classifications, but also tests a subsample of these groups by looking at the effects of growing up in deprived neighborhoods in these four degrees of urbanization. This means that, besides looking at degree of urbanization, I furthermore measured if deprivation is the same no matter where it is found.

There were two interesting points to be made. First, the individuals who gain most from moving are by far the people moving from outer municipalities to urban municipalities, both when considering the total population and the deprived decile. Compared with people living in urban municipalities their whole life, individuals who move from outer to urban municipalities have significantly less unemployment, higher incomes, and more education at age 30 (Lund, 2019a). This can be mostly explained by selection; moving that far away from a childhood home often requires an opportunity that the urban environment can offer. In most cases, this is educational institutions. It is impossible to get education higher than lower level tertiary and, in almost all cases, higher than secondary in outer municipalities. The only places in Denmark that offer higher tertiary education are urban municipalities (Lund, 2019a).

The second point were how staying in the specific municipality type affects the individual. Where the effects of staying are modest when looking at the full population, they become much more significant when comparing the deprived subgroup of the sample. Here, those that stay in deprivation in the outer municipalities fare better than their urban counterparts do. They have a 22%-point higher income, and they have seven days per year less unemployment. On the other hand, the deprived subgroup also suffers the largest gap in educational attainment with more than a full year less than the urban group. Where the first point was somewhat explainable by selection, this second point is more rooted in place-defined attributes. Educational attainment is explained by access, but the income and unemployment gap can be explained by how people living in deprivation differ.

This second point is also interesting from a segregation perspective. Almost all the literature regarding inequality in attained education and later life outcomes focuses on the universal benefit education has on all socioeconomic parameters (Andersson & Malmberg, 2013; Breen & Karlson, 2014; Davies et al., 2002; Erola, Jalonen, & Lehti, 2016; Holm & Jæger, 2008; Lund, 2019a). This is, to some extent, due to the prevalent theoretical discourse in sociology where the direct transmission of economic and cultural capital from parents to children is thought to have an almost universal effect (Bourdieu, 1998). My results indicate that it might be more complex than a direct transmission when adding place as a factor. The direct value of education and the indirect value of educational transmission between generations is less obvious in outer municipality deprived areas. The value of education there is far lower than any other place, and the income and unemployment rate is not directly correlated with educational attainment.

However, this does not negate the overall segregation problem with lack of education. Many other studies from Marx, Weber, and Durkheim to newer studies imply that education is much more than a means to higher income levels and upwards social mobility (Boltanski & Chiapello, 2005; Breen & Karlson, 2014; Durkheim, 1949; Horkheimer, 1972; Lenski, 1978; Scherger & Savage, 2010). The duality of these results requires more research because, if the transmission of cultural capital is limited in more rural areas without affecting income and unemployment rates, we lack information as to how this deprivation differs from that in urban environments.

The fourth paper, “Social Geographical Patterns in Membership of the Established Church in Denmark,” were somewhat different in scope than the other papers (Lund et al., 2019). While the focus still revolved around social inequality, the aim was to investigate membership rates of church membership in Denmark and how membership rates differ geographically. Even though the paper seems thematically different from other parts of the thesis, the underlying assumption that neighborhoods matter is still the same.

Using information from the church tax register, we were able to discern who members of the Danish State Church are and who are not. A unique thing in Denmark is that, as soon as one is baptized, one enters the Danish State Church. Baptism is a common practice in Denmark, and even though the overall percentages of yearly baptism have gone from more than 80% of all newborns in 1990 to little less than 60% in 2017 (Statistics Denmark, 2018), more than 75% of Danes are still members of the Danish State Church (Statistics Denmark, 2019). This number is closer to 82% when controlling for ethnicity (Lund et al., 2019).

Membership, of course, does not directly imply religious views or a change in overall religiousness in the Danish population since the foundation of the measurement is church tax.[14] Nonetheless, since income-adjusted church tax was introduced in 1920, changes in membership rates can be thought of not only as an economic speculation but also as a social and personal choice. During the literature review on religiousness and religious practice, we found that, while most research has investigated the direct link between socioeconomic trends and an outcome, it has disregarded the place (Glaeser & Sacerdote, 2008; Jelent, 1990; D. C. Johnson, 1997; Kirkpatrick & Kirkpatrickt, 1997; Rosa, 1998). Studies especially from British and American authors had a strong focus on how educational attainment reduced the chance of conforming to religious practice and church membership, but none of the studies in our review had tested a place-specific trend in religious activity.

[14] In a Danish setting, all members of the State Church are required to pay .87% (on average, but this varies slightly among different municipalities) of their income in church tax. Membership is thus only possible if one pays this tax, which is recorded in the Danish registers.

# CHAPTER 4. CONCLUSION, DISCUSSION, AND PERSPECTIVES

## 4.1. AUTOMATION IN SOCIOLOGY

Personalized medicine is a rapidly growing field in pharmacology and general medicine (Hamburg & Collins, 2010). It revolves around the idea that even though a specific illness is similar among individuals, the same treatment might yield different results (Hamburg & Collins, 2010). In short, it requires minute changes in medicine types among individuals to treat the same illness and reduce the potential side effects of a given treatment. The concept requires attention to not only the illness but also individual factors that vary among ethnicity, gender, age, and many other factors (Hamburg & Collins, 2010). The same can be said of the way we measure social phenomena. If we transfer this line of thought to issues of scale within urban sociology and social geography, the "person" in this analogy becomes the collective or the sense of the place for the often homogenous group of individuals living there. The collective is differentiated, and neighborhoods cannot always be treated as a common group. To capture the differences among neighborhoods, sociology needs to treat neighborhoods as unique entities; a common set of social problems among neighborhoods might not need the same "treatment."

Earlier in this thesis, I asked how we could modify the geography we measure and what it means to the subject we measure. I have shown how a differentiated way of treating the geographical division of neighborhoods has a considerable and real impact on the way we conceive of the individual and the relation between the neighborhood and the individual (Lund, 2018, 2019a, 2019b; Lund et al., 2019). Regarding modification of geography, I have compared the proposed version of this thesis with similar methods and shown that how we divide space can have a significant impact on what we measure and the results we interpret (Lund, 2018). Likewise, I have shown that life course inequality differs widely not only when considering a small scale, long-term exposure hypothesis (Lund, 2019b) but also when considering the context in which the deprivation is located (Lund, 2019a). Lastly, I have shown that a general spatial awareness when it comes to social characteristics such as church membership can help us better understand underlying mechanisms and tendencies in society (Lund et al., 2019).

Automation and advanced exploratory methods are still very much a niche discipline in sociology, and even though research groups and universities around the world are moving into the field, the implications of a more specialized sociology are still not clear to most. In this thesis, I have shown the importance of measuring social phenomena precisely in space and how different phenomena require different methodological and analytical approaches. If small area clustering is used to localize certain aspects of the population regarding social phenomena, be it concentration, homogeneity in general, or other parameters that are measurable in register data in a diagnostic, personalized approach, it becomes a very powerful tool. We can consequently see where deprivation, sickness, religious membership, and a broad assortment of socioeconomic variables concentrate and evolve.

Below, I summarize the methodological and theoretical implications this thesis offers. Furthermore, I outline the possibilities in a policy perspective and the ethical dangers this research poses.

### 4.1.1. METHODOLOGICAL IMPLICATIONS

In this thesis, I have demonstrated different methods to work with geography in connection to individual level data to further investigate how the link between the physical world and the way we inhabit it can be used as sensible entities for analysis. Using administrative borders to measure social phenomena fails to capture the very essence of neighborhood studies: the neighborhood itself. I have, throughout this thesis, put emphasis on the neighborhood, both in a purely geographical sense and from a social science perspective. If the goal is to understand neighborhood effects, measuring a neighborhood adequately should be of main concern, but often in research, the parish or municipality constitutes a proxy, which is a geography that has very little connection to anything other than the sum of neighborhoods within it. Asking what effect a neighborhood has on a given social phenomenon entails important questions, such as: What is a neighborhood, and what effect are we investigating?

Considering the first research question asked earlier, regarding modification of geography in a sociological sense, automated redistricting is the answer proposed in this thesis. In a purely methodological sense, automated redistricting using physical barriers as separators secures better social homogeneity on numerous social phenomena than using administrative areas (Lund, 2018). Even though automated redistricting is nothing new (Altman, 1997), the use of individual level register data, geography, and automation is. Using a static algorithm, as seen in (Lund, 2018), offers computational speed and reproducibility, but requires more subjective decisions about merges that could cloud the end result. Using a more advanced solution, as seen in (Lund, 2019a), offers less subjectivity and more precision when creating the geography, but the logic behind the final solution is harder to comprehend and, as I have shown in Section 2.4.1, offers little improvement over the static solution.

Using sociologically grounded logic to implement automation in social science brings better units of analysis. Utilizing better units of analysis solves the problem of having non-homogenous entities for analysis, but not the problem of isolating place-based effects. I have demonstrated how a counterfactual design, here in the form of inverse probability-weighted regression adjustments, can help to solve the selection problem in neighborhood research (Lund, 2019a, 2019b). The combination of small geographical units and counterfactuals when studying educational and socioeconomic inequality indicates two important results: life course effects and urbanization. Length of residence in deprivation is an important factor for reproducing deprivation, but when looking at deprivation in a life course setting, it becomes evident that time of exposure is equally as important. The same can be said in regard to the degree of urbanization; living in deprivation has widely different effects in the more rural parts of Denmark compared to urban settings.

One of the major problems with a geographic scale that changes based on the research topic is the comparison between studies and between different types of micro area models. If the overall geographic scale and how the division of inhabitants in this scale changes, it becomes harder to understand exactly where the location is or how to compare it to earlier results. We are used to reading maps, and we expect to be able to internalize where cities or municipalities are located after studying a map, and if we are local, we know where different neighborhoods are, but we are not able to understand a national scale geography broken into micro scale. This problem is mainly due to the fact that we think of geography as a static entity that cannot change, no matter how much the individuals that live there change. I argue for a much less rigid understanding of neighborhood borders and districts. Instead of thinking of inhabitants as defined by where they live, I argue that the symbiosis between place and inhabitants makes the place and, thus, the scale, and therefore the district must change accordingly.

Even though this thesis has placed an emphasis on how the methodologies work, the implication for how we understand neighborhoods and inequality is just as important. In Copenhagen, the physical distance between the richest decile and the poorest is, in some cases, less than 500 meters (Lund, 2019a); this is something that would be missed if using any of the existing administrative units. In summation, small-scale automated redistricting allows us to locate, isolate, and analyze these differences and investigate them without assuming that the large, unrelated area surrounding the isolated entity automatically shares the social characteristics within it.

### 4.1.2. THEORETICAL IMPLICATIONS

As stated in the "Introduction," this thesis has had a methodological aim. First, it aims to investigate how automation can improve how we measure neighborhoods and then apply this methodology to different aspects of inequality, while developing methods and testing different aspects of the neighborhood.

Transferring the methodological insights to a general understanding of deprivation and social science is a pivotal point in my research. The methodological goal is to secure areas that are homogenous, but in doing so, I point to how sociology understands deprivation and to a direction I think the field should move to. Deprivation can be hyper-local, and it can be general. Sometimes, it is isolated, and sometimes it is highly correlated with the areas in the proximity. Without first asking sociological questions, such as those posed by the Chicago School (Park & Burgess, 1925), we fail to isolate what a neighborhood is and how the interaction between the individual and neighborhood is essential to prioritize before we even begin to discuss general spatial effects.

I have earlier pointed to the work of Robert Sampson and especially his theoretical work on collective efficacy (Sampson, 2012). In essence, it points to the fact that a neighborhood can be many things, ranging from isolated houses with very little efficacy to tight-knit communities, and as such, neighborhoods should be treated differently. In my paper, penned with Anja Jørgensen and Ole Riis, we show that standard theories of secularization, where the general level of educational attainment is thought to have a negative impact on all church activities (Deb & Sinha, 2016; Dillon & Wink, 2007; Glaeser & Sacerdote, 2008; Hill, 2011), does not work when spatial relations are taken into account (Lund et al., 2019). Educational attainment and church membership are correlated in the capitol area of Denmark, but much less so in the rest of Denmark. The same can be said when looking at urbanization and the effect of educational attainment, where the general consensus in sociology, in the wake of Bourdieu (Bourdieu, 1998), is that accumulation of human capital will result in increased levels of income and educational attainment; considering place reveals that these universal effects are much more local than we earlier thought (Lund, 2019a) and are sometimes reliant on factors such as population density and how large the flux of inhabitants is (Lund et al., 2019).

Place demands a more specialized methodological and theoretical foundation. Educational attainment is often thought of as a main source of capital in sociology, and, simply put, when education increases, other life events will change as a result almost as a universal law (Bourdieu & Passeron, 1990; Hjellbrekke & Korsnes, 2004). In the case of church membership, education only correlates negatively with membership rates in Copenhagen, but not the rest of the country (Lund et al., 2019). In the case of comparing urban and rural life course events, it becomes clear that educational attainment is of little socioeconomic relevance in the most rural parts of

Denmark – even with significantly lower levels of educational attainment, individuals living in rural areas are shown to have higher incomes and less unemployment than their urban counterparts (Lund, 2019a). The social life and the many different social elements that exist within each neighborhood create not only different outcomes but also the different ways that various constellations of capital affect the neighborhood population.

This offers nuances to the discussion of how we think about universalism in sociology. Deprivation and human values is thought, in almost all the literature presented in this thesis, to have a universal effect on socioeconomic outcomes. Even studies that differentiate on place often do so with no concern for *where* that place is. When applying the *where* in a more descriptive way (Lund et al., 2019) or in a more narrow setting (Lund, 2019a), the effects differ widely, and to generate universally expected outcomes becomes much harder, but also much more precise. If our goal is to describe the mechanisms behind specific social phenomena as inequality, deprivation, or even church membership, we cannot expect direct, universal laws of causality to apply everywhere. As with personalized medicine, a personalized sociology must be able to leave behind grand theories when explaining small-scale, socioeconomic phenomena and embrace the social differences within seemingly similar neighborhoods.

This point is also pivotal in how this thesis expands on earlier work. Where many use specialized tools to analyze cities or smaller areas that often have many inhabitants (Lund, 2018), this thesis has developed tools to encompass all parts of a geography where humans live. Where Robert Sampson (2012) focuses heavily on single cities and the intricacies of smaller areas, this methodology allows an expansion into both rural and outer areas. In many cases, looking at these rural zones brings forth important knowledge of how place of residence matters (Lund, 2019a; Lund et al., 2019). Collective efficacy is not just one thing; the collective changes as the type of neighborhood changes. Even when neighborhoods are seemingly similar, the location still matters. I believe that the methodology developed here can help to expand on the theoretical and empirical work in neighborhood research, but can also serve as a tool to expand on how social efficacy can be understood and used as an explanatory societal phenomenon.

### 4.1.3. POLICY IMPLICATIONS

In this thesis, I show how inequality and deprivation are less municipality problems than localized problems within municipalities. Within different types of municipalities, deprivation has been shown to have very different intergenerational and direct effects on the residents within them (Lund, 2019a, 2019b). Even with the Danish ghetto list, where some form of micro scale has been introduced, it becomes a question of clustering on large entities with a general purpose of looking at clusters of immigrants and not as a tool to improve or better understand the development within it (Ministry of Transportation, 2018). In short, general methods, such as the ghetto list, still suffer from a very important problem: they do not differentiate among different places that all fall into the category of ghetto – either it is a ghetto or it is not.

Developing tools to analyze effects on a micro scale should not just be used to further the study of neighborhoods and neighborhood deprivation, but as a tool to focus on specific areas based on evidence instead of a general idea of where social problems exist. Specialized medicine is gaining momentum because the evidence suggests that it works (Hamburg & Collins, 2010), and place-based policies could be thought of in the same way; we generate models to encompass different social phenomena and use those models to make policy decisions specialized to the specific problem we encounter. This is not proposed as a complete solution to end all policy debate on spatial inequality. More specialized does not necessarily result in the right solution, but it will help decision-makers differentiate between spatial problems.

Part of the Finance Bill is the block subsidy, where the Danish state provides subsidies for each of the five Danish Regions and the 98 municipalities (Ministry of Finance, 2019). Each region and municipality is required to spend the block subsidy internally, but is free to use the block subsidy as it chooses (Ministry of Finance, 2019). In a Danish context, 150 million kroners are earmarked for improving social inequality (Ministry of Finance, 2019). During a hearing in 2016, it was made public that there were differences between how each region and municipality spent the money, and in some cases, only a fraction of the budget was used to improve social inequality (Frederiksen, 2016).

Each municipality is different, and there can be many reasons as to why subsidies are handled differently, but many of the problems could be counteracted simply by using specialized tools to insure better subsidizes. Instead of subsidizing a bulk amount of money based on the overall municipality and regional problems in the total population (Ministry of Finance, 2019), policy makers could subsidize based on specific areas. Inequality and social problems, as shown in this thesis, can be isolated without any knowledge of the municipality if the methodology to investigate is designed to handle small area statistics.

The problem with a changing geography, from a policy perspective, is that subsidies are thought of as a municipality issue. Policy makers expect to see statistics at the municipality level because the summation at that aggregate level is easy to work with and comparable to that of other municipalities. My point is that inequality and social issues in themselves have very little connection to arbitrary municipality borders. There are, no doubt, issues that are more prevalent in some municipalities, but most relate to a national level problem. When trying to isolate deprivation or inequality, we need to be able to investigate and understand the specific phenomenon without concern for administrative borders. From a policy perspective, the approach presented in this thesis can work as a way to disrupt the self-evident way we interpret spatial divisions and to help better understand different spatial dimensions of social phenomena.

### 4.1.4. ETHICAL IMPLICATIONS

With specialized tools come special ethical dilemmas. Capturing the local creates possibilities for both reducing inequality, but also the risk of increasing it. I have outlined how specialized policies can help reduce inequality, but there are instances where this information could be used to target specific groups of the population. During my thesis work, I have been approached by a large, Danish insurance agency, numerous marketing agencies, and a few investment funds that were interested in segmenting the Danish population to further their interests.

Both the insurance and marketing industry can use specialized geography to focus on a specific group of the population they want to target. On one hand, using personalized marketing to make sure funds are used best is not inherently negative, but if, on the other hand, this segregation is used to exclude parts of the population from opportunities others have, or, in the case of the insurance industry, it is used to adjust premiums not only because of individual histories with insurance but also that very specific, small area one inhabits, the consequences for especially low income, high crime areas will be severe.

In the case of investment funds, there are no immediate consequences because they have very little interest in the population as such. What an investment fund looks for is high-yield investments, and being first often means that, with even a smaller initial investment, the yield will be high. Being able to predict areas that are prone to growth before the market does means being able to buy land before prices increase. The problem arises when the non-personalized focus results in a direct negative impact on parts of the population. Being able to predict prosperity will increase the inequality in an area if the only people benefitting from the prediction are the ones with the means to buy large portions of land. On the other hand, if the model is turned around, it can also predict areas that are at risk of a value decrease. This could easily lead to a massive decline in investments and, thus, increase the probability of further deprivation.

I believe that research into neighborhoods and neighborhood inequality is important. Even with an empirically focused study that mainly revolves around isolating the driving mechanisms of social phenomena in neighborhoods, questions about research output arise and must be considered. That is why I believe it is important to understand what can come of the specialized methods and discuss if research into neighborhood inequality should be used to increase the overall rise of inequality happening in most of the world (Beckfield, 2019). I have been part of many research projects during my PhD work, but common to all of them is that the goal is to either further our understanding of differentiated neighborhoods or investigate how neighborhoods evolve. Thus, the program developed during this thesis is solely available for research and non-commercial use.

### 4.1.5. SPECIALIZED SOCIOLOGY?

Sociology can be a broad science that outlines grand theories of society, but it can also be a specialized science to capture social phenomena on a micro scale. Some of the most famous theories in sociology are the grand theories of society, and as I mentioned in the introduction, many of them deal with place in one way or the other, but common for all is that they deal with place as a concept of an overall idea and not as something unique and differentiable.

One of the things that stuck in my mind was a conversation I had with my PhD supervisor where she told me to “blow up the schools.” Without context, this could be read as a problematic statement, but what she meant was to not conform to just one sociological school of thought. With the rise of the Chicago School came the understanding of the city as an organism with different parts moving and interacting. With the rise of the econometric school came methods and concepts to process spatial data and the earliest development of neighborhood effects. With the LA School came the concept of power in design and in the bricks and mortar of the city, and with analytical sociology came a framework to fully interconnect the different elements of spatial sociology.

I have utilized parts of the schools mentioned above, understanding the city and the rural as organic with many different groups while still maintaining that the physical space plays a part in how we inhabit the place. To measure the place as something tangible that exists with precise methods, while acknowledging that the mechanisms I try to isolate can never be reduced to positivist laws of causation, is an approximation of underlying effects that requires constant methodological and theoretical development.

This thesis is not meant as an attack on grand theories, and I fully believe that grand theories can help us understand important aspects of our society. Specialized sociology does not imply that all sociology should be specialized and solely focused on isolating effects and mechanisms, but it does imply that empirical sociology within an analytical framework should stand as a field that dares to question how we measure social phenomena and constantly tries to improve the tools and methods we use to get closer to the underlying social mechanisms we seek to describe.

## 4.2. PERSPECTIVES WITH ARTIFICIAL INTELLIGENCE

One of the major unanswered questions throughout this thesis is *which* physical borders work as dividers of neighborhoods and how neighborhoods are born, evolve and perhaps even perish. Using register data, we are able to investigate how neighborhoods change; some experience heavy growth in housing and inhabitant count, while others stagnate or experience what the Danish media has coined as “village death,” where especially the more rural villages are left as ghost towns (Bangild, 2017). Even though data can show us the social indicators for spatial prosperity, stagnation, and death, we do not know how a neighborhood in prosperity looks.

Most have very specific ideas how a deprived neighborhood appears. Driving on a country road in Denmark brings you through many smaller villages whose names feel foreign but have existed for many years. I have always been curious about the very distinct change in the type of smaller villages as I drive through them. Some look like miniature versions of suburban neighborhoods, others have a very rural, rustic, but idyllic feel to them, while others are characterized by the distinct lack of maintenance.

Even though “lack of maintenance” could be a predictor of increasing neighborhood deprivation, there are very few ways we are able to quantify structural decay without looking at each specific neighborhood. This is one of the major points of Gieryn (2002): sociologists needs to be better equipped at looking at other objects than the purely social. So, a question arises: How do we look at all neighborhoods? Drawing on the Chicago School and much of the research that came from that period (Burton, 2004; Kasinitz, 2005; Park & Burgess, 1925; Weppner, 1977), one solution could be to simply look. What came from “looking” in Chicago became one of the most important contributions to theory of social place, but even though much of the theory has been applied in many places in the world, the fact is that the study was based on Chicago. Kelley, Park, Burges, and Thrasher lived in the city and felt the city (Kelley & The Residents of Hull House, 1895; Park & Burgess, 1925; Thrasher, 1927).

We cannot feasibly look at all the cities and villages, but we can use the technological advances we have made since the start of the Chicago School to employ machines to look for us. With the emergence of satellites and the ever growing amount of register data in Denmark, we now have the opportunity to work with a combination of methods.

Other sciences have taken steps to fully utilize the possibilities of more modern methods of understanding images with machines. In computer science, image recognition and the use of imagery is evident in the strides that have been made in face recognition in both the private sector and in government surveillance. In the health sciences where diagnostic work relies heavily on radiographic imagery, the use of machine learning combined with body scans has led to numerous important discoveries in early detection of malign diseases, such Alzheimer's and tumors, and preventable side effects of more common diseases (Isasi, Zapirain, & Zorrilla, 2011; Kominami et al., 2016; Ravishankar, Jain, & Mittal, 2009; Zubi & Saad, 2011). On a macro scale, the use of satellite imagery has grown from being almost solely for military or guidance to being used in the sciences to detect crop yields for farmers (Yang, Everitt, & Murden, 2011) and building detection systems to assess land usage in urban settings (Yun Zhang, 1999).

These overall developments have slowly begun to grow within the social sciences, but mainly in the intersection between economy and geography to identify sales prices or vulnerable areas in third world countries using satellite imagery (Watmough et al., 2019; Watmough, Atkinson, Saikia, & Hutton, 2016; Watmough, Palm, & Sullivan, 2017). The main breakthrough in the above examples is the fact that automation is finding its way into a combination of geographical non-social images to understand a social phenomenon such as third world extreme poverty. While sociology has been quick to apply modern network analysis tools (Borgatti, Mehra, Brass, & Labianca, 2009; Breiger, 2009; Butts, 2001; Hoff, Raftery, & Handcock, 2002; Nordlund, 2018) and simulation models (Benenson, Hatna, & Or, 2009; Schelling, 1971; Studer & Ritschard, 2016), the adaption of automated image analysis has taken much longer.

One of the main advantages of utilizing image recognition in sociology is that we are further able to combine the physical world we live in with our data. The following section will explore a branch of machine learning called convolutional neural networks that I have been pursuing in combination with my thesis. I hope this will be part of a broader sociological movement in the future – not only in spatial research but also as a way to understand larger social gatherings and other events that are too large to analyze by even a team of researchers without the help of machines.

### 4.2.1. DATA AND COMPUTING

Where register data can show us individual traits and accomplishments, they cannot tell us how physical objects look. The same can be said of imagery on its own; we are able to see what a neighborhood looks like, but not the people living there. Combining computer vision with register data adds another layer to the overall understanding of what a neighborhood is, but the requirements for data become much more demanding. The data sources used in this smaller example include register data on an individual level aggregated to small areas, register data on house level with information on age of house, and time since last renovation, as well as other building-specific information. This also includes information on exact coordinates of each individual house. The last dataset here is the visual one: satellite imagery.

Without even going into too much detail on different types of satellites and the overall evolution of satellite imagery, one realization when working with satellites is that commercial and commonly available datasets in acceptable image resolution were only available after the turn of the millennium with the Landsat 7 program (Loveland & Dwyer, 2012). Even with the launch of Landsat 7 in 1999, it was not until 2013 with Landsat 8 and 2015 with Sentinel-2 multispectral[15] that high-resolution imagery became available (Drusch et al., 2012).

As an example of satellite imagery on the neighborhood level, images of the southern part of the Danish city Silkeborg are used in Figure 14 as examples of three different band spectrums often used in computer vision recognition (Yang et al., 2011), ranging from standard true color Red-Green-Blue (RGB) (left), to Normalized-Difference Snow Index (NDSI) (mid), and then false color (right). The interesting part of image recognition is often the realization that human eyes and computer eyes work very differently.

*Figure 14 - Multiband spectrum satellite imagery of Silkeborg*

While we are accustomed to pictures where shapes and colors make sense, computers are much less concerned with how it looks to us. Computer vision works with contrasts, sizes, and shapes in pixels so that each pixel can work as an array or string of numbers. One way of working with pictures could be to measure contrast as a number, where 1 is white and 100 is black but since these pictures contain color, we need to be able to tell colors apart as well. The table below (table 1) is a representation of a hexadecimal color array of the 56 top right pixels in the true color image in Figure 14 generated with scikit-image (van der Walt et al., 2014).

[15] The main difference between ordinary satellite imagery and multispectral imagery is the availability of spectral bands that the satellite can detect. Human eyes identify most objects in a combination of red/green/blue, but multiple bands allow us to target specific objects easier than others. For an example, see Yang et al. (2011) where they use a combination of green, red, near-infrared, and short-wave infrared to detect crops.

*Table 1 - Array of small satellite image section*

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| 0x6B | 0x7A | 0xE9 | 0x30 | 0xA6 | 0xB0 | 0x2F | 0x6B |
| 0x5A | 0xA6 | 0xF2 | 0x9C | 0xD5 | 0x89 | 0xCD | 0x5A |
| 0x87 | 0x14 | 0x54 | 0x63 | 0xA5 | 0x14 | 0xEC | 0x87 |
| 0xA8 | 0x58 | 0x56 | 0x51 | 0x67 | 0x7B | 0x56 | 0xA8 |
| 0x47 | 0x4E | 0xE4 | 0x34 | 0x24 | 0x6A | 0x6A | 0x47 |
| 0x18 | 0x82 | 0x8A | 0x90 | 0x0A | 0x00 | 0xA7 | 0x18 |
| 0x43 | 0x5A | 0x23 | 0xCD | 0xA8 | 0xB5 | 0x18 | 0x43 |
| 0x33 | 0x14 | 0xE1 | 0x4B | 0x8A | 0x4A | 0x2E | 0x33 |

When the goal is to understand what a specific image means, an image reduction such as the one above reduces the visual complexity so computer programs are able to process it. Each hexadecimal represents an RGB color and thus has meaning for the computer. This, however, does not help the computer with *what* it sees. For this important part, we need labelling. The main issue for most research using satellite imagery is reliable labeling. For a computer to understand what an image means in the social world, it needs pictures with annotations to be able to train. In principle, the researcher needs to identify what it is in a picture that is important and what is less important. Looking at the satellite imagery below, a high resolution version of the image presented before from Silkeborg, we are able to discern what the pixels are.

*Figure 15 - Satellite imagery from part of Silkeborg*

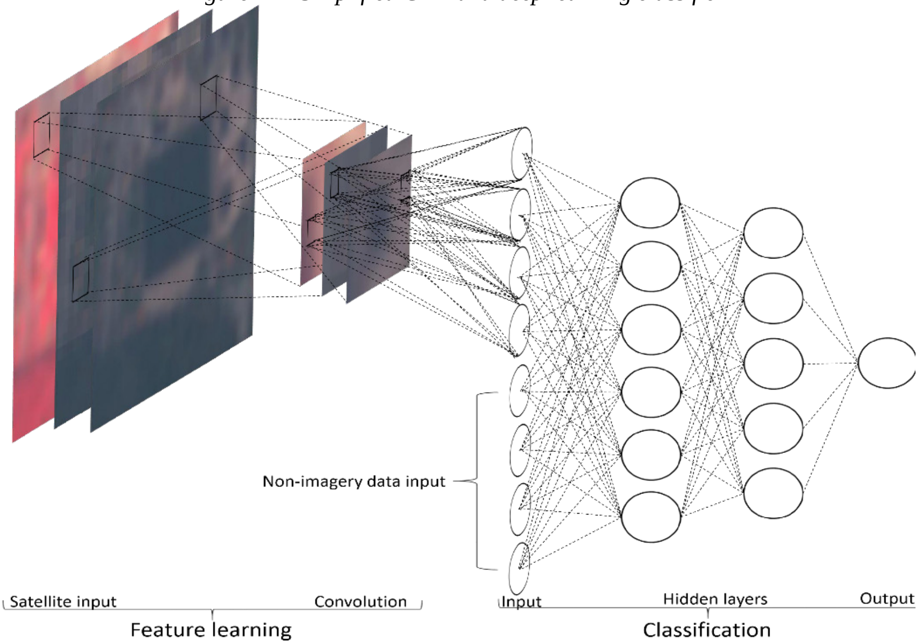
With Danish register data, we have a shortcut to labelling. Through the housing registers, almost all parts of the Denmark have been labeled and updated through time, so forests, lakes, houses, and almost all objects that fall into the category of physical have been geocoded and can be added to satellite data. With this information, the computer will be able to understand what it sees, and more importantly, it has the information needed when the final step is introduced: the social and socioeconomic measurements.

As an example, I have experimented with predicting socioeconomic status on the neighborhood level with a convolutional neural network to make a preliminary test of small area neighborhood borders. I have collected data on the city of Silkeborg and the closest suburbs within a 20-km radius from the city center, as seen at the top of Figure 16, and split the picture into 500 x 500 meter images as seen at the bottom of Figure 16. The data consist of sentinel-2 multispectral images from 2015 in true color, false color, and NDSI (Copernicus, 2019), register data from the year 2015 consisting of income, educational attainment, and unemployment indexed and aggregated to small areas, and data from the housing register with information on the age of the house, time since last major reconstruction, and number of rooms suitable for habitation.

*Figure 16 - Satellite imagery of Silkeborg and suburbs*


The network is designed to use the 500 x 500 meter images, combine them with the other data types at individual level, house level, and aggregate level, respectively, and make a prediction as to what parts of the city are deprived. Below is a simplified representation of a combination of a convolutional neural network (Chen & Peter Ho, 2008; Kubat, 2015; Längkvist, Kiselev, Alirezaie, & Loutfi, 2016) and a deep learning classifier (Berrar, 2016; Leban, Zupan, Vidmar, & Bratko, 2006; Matiolaski, Maksimova, & Dziech, 2016).

*Figure 17 - Simplified CNN and deep learning classifier*




On the left side are the multispectral satellite imagery sets which are the first part of an input. To keep the figure simple, only one convolutional layer has been added but in praxis, multiple layers may be added to effectively reduce what in the imagery predicts the outcome. When the imagery has been reduced, the non-imagery register data are added and fed into a classifier that uses both register data but also the output from the convolutional neural network to make a final prediction. In simple terms, the model trains thousands of times, testing different combinations of imagery and registers, each time learning through back propagation[16] and using the housing registers to understand what it is looking at.

The primary evaluation is done by the model "checking" how well it predicted the socioeconomic index compared to the known index at small area level, since this is the smallest scale I can measure socioeconomic status. Since the model essentially only predicts what we already know, the socioeconomic status in the primary evaluation, the secondary output where differentiation is done at house level is more interesting. Even though the model will believe that houses that are different than what

[16] Back propagation is the praxis of letting the data start at a random point, feeding the data forward through the network, testing the prediction to the known result, and then changing parameters before running again, thus slowly learning how different parameters affect the final prediction power (Hastie et al., 2016).

the socioeconomic index is in a specific area, this is because it is unable to predict precisely what this will help us learn. If secondary output predicts specific parts of an area as different from the area they belong to, this might be evidence that some border types are more suited than others.

Below is a representation of a secondary output from the model where correct guesses[17] are marked with a green dot at the center of the house while red dots represent wrong guesses. In this instance, the red dots were predicted to be 50% lower than the area socioeconomic index they were in.

*Figure 18 - Secondary output of individually predicted houses*

As the model is now, it can only let the user know that something went wrong in this area and not why. There are no physical barriers that my earlier models could detect

[17] In this case, the limit for a correct guess is +/–20% of each area's average socioeconomic status. If the guess falls within +/–20% of the tested area's average socioeconomic status, the dot becomes green. If the distance between prediction and average is larger than 20%, it becomes red. In a real world application, a 20% margin could be lowered considerably, but this is a simple, working prototype and not a fully trained model.

and use as dividers of space and nothing in particular to notice as to why the model guessed wrong. By looking at the housing register, there seems to be less renovation done in houses marked with red than the surrounding area, but for now, the model or the data cannot give us the answers other than we need to keep looking and we need to work on how we understand neighborhoods.

### 4.2.2. POSSIBILITIES (AND DANGERS) WITH AI

There is no doubt that the trend in AI and machine learning is currently peaking. Big companies, such as Google, Facebook, Amazon, and Apple, have invested billions in research. When big companies drive an agenda with AI and bet that it will create a bigger overturn, this influences many other elements of the world, as well. Vivek Wadhwa from Stanford, borrowing from Duke professor Dan Ariely, sums it up quite well: “Artificial intelligence is like teenage sex: everyone talks about it, nobody really knows how to do it, everyone thinks everyone else is doing it, so everyone claims they are doing it” (Wadhwa, 2017).

It is a fine line to walk, between advancing technologically driven analysis, because it is possible, and doing it because it makes it possible to answer important sociological questions. With the advancement of computational power in the last 30 years, we have witnessed how an iPhone from 2012 was capable of 2.7 times the processing power of the 1985, 32 million dollar Cray-2 supercomputer (Reisinger, 2016). Even when Moore’s law[18] seems to slow down, companies push the limit of the amount of data we can process each year (Morgan, 2018). We are no longer limited by questions such as how much data can we process? Or how large are the data files? Instead, we must ask ourselves what makes sense to analyze?

Neighborhood research now has the possibility to transcend what was earlier solely possible either with ethnographic methodologies where the neighborhood was seen, felt and observed or with quantitative studies where the total sum of individuals was accounted for. We are able to simultaneously measure a neighborhood as well as “see” it – not only as bricks and mortar like many others, but also as a unique piece that holds its own meaning in both social interaction and geographical placement as well as how it looks. This, combined with thorough qualitative studies at selected places, I believe, will advance the way we look at neighborhoods in the future.

Building a heavy duty, Graphical Process Unit (GPU) core intensive computer can be done for less than $500, and commercial satellite imagery is becoming more and more

[18] Moore’s law (Moore, 1965) is based on the observation that the number of transistors in a dense integrated circuit doubles every two years. This law has been stable for many years, and even when it slowly started to plateau, the rise of graphic processing units and their utilization in heavy duty calculations again propelled the possibilities of data processing further on. For more on this, please see (Morgan, 2018; Schaller, 1997)

easily available as is the software needed to analyze the data. The road I have outlined in this final part of my thesis is one I hope to travel in my future research. There is no doubt that this too will heighten our understanding of neighborhoods and the social life which is lived within them.

# LITTERATURE LIST

Altman, M. (1997). Is automation the answer: The computational complexity of automated redistricting. *Rutgers Computer and Law Technology Journal*, *23*(1), 81–141.

Altman, M., Macdonald, K., & Mcdonald, M. (2005). From crayons to computers: The evolution of computer use in redistricting. *Social Science Computer Review*, *23*(3), 334–346.

Altman, M., & Mcdonald, M. P. (2012). Case Western Reserve Law Review Redistricting Principles for the Twenty-First Century REDISTRICTING PRINCIPLES FOR THE TWENTY-FIRST CENTURY. *W. Res. L. Rev*, *62*(4).

Andersson, E., & Malmberg, B. (2013). *Contexual effects on educational attainment in individualized neighborhoods: Differences across gender and social class* (SPaDE No. 10). Stockholm.

Åslund, O., & Fredriksson, P. (2009). Peer Effects in Welfare Dependence. *Journal of Human Resources*, *44*(3), 798–825.

Bandura, A., Craighead, W. E., & Weiner, I. B. (2010). Self-Efficacy. In *The Corsini Encyclopedia of Psychology*. Hoboken, NJ, USA: John Wiley & Sons, Inc.

Bangild, C. (2017, February 25). Hvad kom først: Skolenedlæggelse eller landsbydød?

Bauman, Z. (2001). *Community - Seeking Safety in an Insecure World*. Cambridge: Polity Press.

Beckfield, J. (2019). *Unequal Europe: Regional Integration and the Rise of European Inequality*.

Bellavance, C., Normand, F., & Ruppert, E. S. (2007). Census in Context: Documenting and Understanding the Making of Early-Twentieth-Century Canadian Censuses. *Historical Methods*, *40*(2), 92–103.

Benenson, I., Hatna, E., & Or, E. (2009). From Schelling to Spatially Explicit Modeling of Urban Ethnic and Economic Residential Dynamics. *Sociological Methods & Research*, *37*(4), 463–497.

Berrar, D. (2016). Learning from automatically labeled data: case study on click fraud prediction. *Knowledge and Information Systems*, *46*(2), 477–490.

Bien, J., & Tibshirani, R. (2011). Hierarchical Clustering With Prototypes via Minimax Linkage. *Journal of the American Statistical Association*, *106*(495), 1075.

Boltanski, L., & Chiapello, E. (2005). The new spirit of capitalism. *International Journal of Politics, Culture, and …*.

Borgatti, S. P., Mehra, A., Brass, D. J., & Labianca, G. (2009, February 13). Network analysis in the social sciences. *Science*.

Borgoni, R., & Billari, F. (2003). Bayesian Spatial Analysis of Demographic Survey Data: An Application to Contraceptive Use at First Sexual Intercourse. *Demographic Research*, *8*.

Bose, P., Maheshwari, A., & Morin, P. (2003). Fast approximations for sums of distances, clustering and the Fermat-Weber problem. *Computational Geometry: Theory and Applications*, *24*(3), 135–146.

Bottou, L., & Bousquet, O. (2008). The Tradeoffs of Large Scale Machine Learning. *Advances in Neural Information Processing Systems*, *20*, 161–168.

Bourdieu, P. (1998). *Practical Reason: On the Theory of Action*. Stanford University Press.

Bourdieu, P., & Passeron, J.-C. (1990). *Reproduction in Education, Society and Culture*. SAGE Publications.

Breen, R., & Karlson, K. B. (2014). Education and social mobility: New analytical approaches. *European Sociological Review*, *30*(1), 107–118.

Breiger, R. L. (2009). Models and Methods in Social Network Analysis. *Contemporary Sociology: A Journal of Reviews*, *37*(5), 481–482.

Browdy, M. H. (1990). Simulated annealing: an improved computer model for political redistricting. *Yale Law & Policy Review*, *8*(1), 163–179.

Buck, N. (2001). Identifying Neighbourhood Effects on Social Exclusion. *Urban Studies*, *38*(12), 2251–2275.

Burton, L. M. (2004). Ethnography and the Meaning of Adolescence in High-Risk Neighborhoods. *Ethos*, *25*(2), 208–217.

Butts, C. T. (2001). The Complexity of Social Networks: Theoretical and Empirical Findings. *Social Networks*, *23*(1), 31–71.

Calimani, R. (1987). *The ghetto of Venice*. M. Evans.

Chen, C. H., & Peter Ho, P. G. (2008). Statistical pattern recognition in remote sensing. *Pattern Recognition*, *41*(9), 2731–2741.

Cohn, M. J., & Jackman, S. P. (2011). A Comparison of Aspatial and Spatial Measures of Segregation. *Transactions in GIS, 15*, 47–66.

Copernicus. (2019). Sentinel 2 data.

Cox, K. R. (1969). The voting decision in a spatial context. *Progress in Geography*, *22*(1), 81–117.

Cullinan, J. (2011). A Spatial Microsimulation Approach to Estimating the Total Number and Economic Value of Site Visits in Travel Cost Modelling. *Environmental and Resource Economics*, *50*(1), 27–47.

Curry, J., & Kenney, M. (1999). The Paradigmatic City: Postindustrial Illusion and the Los Angeles School. *Antipode*, *31*(1), 1–28.

Cutchin, M. P., Eschbach, K., Mair, C. A., Ju, H., & Goodwin, J. S. (2011). The socio-spatial neighborhood estimation method: An approach to operationalizing the neighborhood concept. *Health & Place*, *17*(5), 1113–1121.

Davies, R., Heinesen, E., Holm, A., Davies1, R., Heinesen2, E., & Holm3, A. (2002). The Relative Risk Aversion Hypothesis of Educational Choice. *Source Journal of Population Economics J Popul Econ*, *15*(15).

Davis, M. (2005). The Urbanization of Empire. *Social Text*, *22*(4), 9–15.

Dawkins, C. J. (2006). Are Social Networks the Ties that Bind Families to Neighborhoods? *Housing Studies*, *21*(6), 867–881.

Dear, M., & Flusty, S. (2012). The Postmodern Urban Condition. *Spaces of Culture: City, Nation, World*, 65–85.

Deb, M., & Sinha, G. (2016). Impact of culture on religiosity, cosmopolitanism and ethnocentrism. *Asia Pacific Journal of Marketing and Logistics*, *28*(1), 56–72.

Demeulenaere, P. (2011). *Analytical Sociology and Social Mechanisms*. Cambridge University Press.

Dillon, M., & Wink, P. (2007). *In the course of a lifetime : tracing religious belief, practice, and change*. University of California Press.

Drusch, M., Del Bello, U., Carlier, S., Colin, O., Fernandez, V., Gascon, F., … Bargellini, P. (2012). Sentinel-2: ESA's Optical High-Resolution Mission for GMES Operational Services. *Remote Sensing of Environment*, *120*, 25–36.

Duque, J. C., Aldstadt, J., Velasquez, E., Franco, J. L., & Betancourt, A. (2011). A computationally efficient method for delineating irregularly shaped spatial clusters. *Journal of Geographical Systems*, *13*(4), 355–372.

Durkheim, E. (1949). *The Division of Labor in Society*. (G. Simpson, Ed.). Glencoe, Illinois: The Free Press.

Elster, J. (1989). *Nuts and Bolts for the Social Sciences*. Cambridge University Press.

Eriksson, K. H., Hjalmarsson, R., Lindquist, M. J., & Sandberg, A. (2016). The importance of family background and neighborhood effects as determinants of crime. *Journal of Population Economics*, *29*(1), 219–262.

Erola, J., Jalonen, S., & Lehti, H. (2016). Parental education, class and income over early life course and children's achievement. *Research in Social Stratification and Mobility*, *44*, 33–43.

ESRI. (2019). ArcGIS Pro. Redlands: Environmental Systems Research Institute.

Ester, M., Kriegel, H.-P., Sander, J., & Xu, X. (1996). *A Density-Based Algorithm for Discovering Clusters in Large Spatial Databases with Noise*.

European Comission. (2007). *Regions in the European Union*. *Statistics*.

European Comission. (2018). History of NUTS - Eurostat.

European Communities. (2002). *TANDEM_GIS I A (feasibility) study towards a common geographical base for statistics across the European Union*. Luxembourg.

European Communities. (2003). *Feasibility Study – Comparable Statistics in the Area of Care of Dependent Adults in the European Union*. *Office*.

European Environment Agency. (2019). *Mapping Europe's ecosystems*.

Eurostat. (2014). Tertiary students (ISCED 5-6) by field of education and sex (educ_enrl5).

Eurostat. (2018). Background - Eurostat.

Faraway, J. J., & Simonoff, J. S. (2006). Smoothing Methods in Statistics. *Journal of the American Statistical Association*, *92*(439), 1214.

Feld, S. L. (1981). The Focused Organization of Social Ties. *The American Journal of Sociology*, *86*(5), 1015–1035.

Ferreira, M. A. R., Holan, S. H., & Bertolde, A. I. (2011). Dynamic multiscale spatiotemporal models for Gaussian areal data. *Journal of the Royal Statistical Society. Series B, Statistical Methodology*, *73*(5), 663–688.

Fifield, B., Higgins, M., Imai, K., & Tarr, A. (2018). *A New Automated Redistricting Simulator Using Markov Chain Monte Carlo.*

Finlay, R. (1982). The Foundation of the Ghetto: Venice, the Jews, and the War of the League of Cambrai. *Proceedings of the American Philosophical Society*, *126*(2), 140–154.

Fiscella, K., & Fremont, A. M. (2006). Use of Geocoding and Surname Analysis to Estimate Race and Ethnicity. *Health Services Research*, *0*(0), 060720074824039-???

Flower, P. (1994). The effects of refineries on neighborhood property values. *Journal of Real Estate Research*, *9*(3), 319-.

Foucault, M. (1988). *Madness and civilization : a history of insanity in the Age of Reason.* Vintage Books.

Frederiksen, C. H. (2016). *Answer to Finance Committee question 1*. Copenhagen: Ministry of Finance.

Friedman, M. (1955). The Role of Government in Education. In R. S. Solo (Ed.), *Economics and the Public Interest* (1st ed., pp. 123–144). New Brunswick: Rutgers University Press.

Gans, H. J. (1961). The Balanced Community: Homogeneity or Heterogeneity in Residential Areas? *Journal of the American Planning Association*, *27*(3), 176–184.

Gans, H. J. (2002). The Sociology of Space: A Use-Centered View. *City & Community*, *1*(4), 329–339.

Gelman, A., & King, G. (2006). Enhancing Democracy Through Legislative Redistricting. *American Political Science Review*, *88*(03), 541–559.

Giddens, A. (1990). *The consequences of modernity*. Stanford University Press.

Gieryn, T. F. (2002). A Space for Place in Sociology. *Annual Review of Sociology*, *26*(1), 463–496.

Glaeser, E. L., & Sacerdote, B. I. (2008). Education and Religion. *Source: Journal of Human Capital Journal of Human Capital*, *2*(2).

Goffman, E. (1967). *Interaction Ritual Essays on Face-To-Face Behavior*. Doubleday.

Gondek, D., & Hofmann, T. (2007). Non-redundant data clustering. *Knowledge and Information Systems*, *12*(1), 1.

Guo, R., Ahn, M., & Zhu, H. (2015). Spatially Weighted Principal Component Analysis for Imaging Classification. *Journal of Computational and Graphical Statistics*, *24*(1), 274.

Hamburg, M. A., & Collins, F. S. (2010). The Path to Personalized Medicine. *The New England Journal of Medicine*, *363*(4), 301–304.

Harvey, D. (2003). *The spaces of global capitalism.* Verso.

Hastie, T., Tibshirani, R., & Friedman, J. (2016). *The Elements of Statistical Learning* (2nd ed.). Springer.

Hedstrom, P. (2005). *Dissecting the social: On the principles of analytical sociology. Dissecting the Social: On the Principles of Analytical Sociology*. Cambridge: Cambridge University Press.

Hedström, P. (1994). Contagious Collectivities: On the Spatial Diffusion of Swedish Trade Unions, 1890-1940. *American Journal of Sociology*, *99*(5), 1157–1179.

Hedström, P., & Bearman, P. (2009). The Oxford Handbook of Analytical Sociology (p. 1611). OUP Oxford.

Hedström, P., & Swedberg, R. (1996). Social Mechanisms. *Acta Sociologica*, *39*(3), 281–308.

Hedström, P., & Ylikoski, P. (2010). Causal Mechanisms in the Social Sciences. *Annual Review of Sociology*, *36*(1), 49–67.

Hesiod. (700BC). *Theogony ; and, Works and days*. (M. L. West, Ed.). Oxford University Press.

Hill, J. P. (2011). Faith and Understanding: Specifying the Impact of Higher Education on Religious Belief. *Journal for the Scientific Study of Religion, 50*(3), 533–551.

Hillier, A. E. (2005). Residential Security Maps and Neighborhood Appraisals: The Home Owners' Loan Corporation and the Case of Philadelphia. *Social Science History*, *29*(2), 207–233.

Hjellbrekke, J., & Korsnes, O. (2004). Educational Mobility Trajectories and Mobility Barriers in the Norwegian Social Space. *International Journal of Contemporary Sociology*, *41*(2), 163–189.

Hoem, J. M. (2007). Demography, present and future. *Population*, *62*(1), 53–56.

Hoff, P. D., Raftery, A. E., & Handcock, M. S. (2002). Latent space approaches to social network analysis. *Journal of the American Statistical Association, 97*(460), 1090–1098.

Holm, A., & Jæger, M. M. (2008). Does Relative Risk Aversion explain educational inequality? A dynamic choice approach. *Research in Social Stratification and Mobility*, *26*(3), 199–219.

Hong, S., & Sadahiro, Y. (2014). Measuring geographic segregation: a graph-based approach. *Journal of Geographical Systems*, *16*(2), 211–231.

Horkheimer, M. (1972). Traditional and critical theory. *Critical Theory: Selected Essays*, *1*.

Humberd, B. K., Clair, J. A., & Creary, S. J. (2015). In our own backyard: when a less inclusive community challenges organizational inclusion. *Equality, Diversity and Inclusion: An International Journal*, *34*(5), 395–421.

Isasi, A. G., Zapirain, B. G., & Zorrilla, A. M. (2011). Melanomas non-invasive diagnosis application based on the ABCD rule and pattern recognition image processing algorithms. *Computers in Biology and Medicine*, *41*(9), 742–755.

Jacobs, J. (1992). *The death and life of great American cities*. Vintage Books/Knopf Doubleday Publishing Group.

Jarvis, B., Kawalerowicz, J., & Valdez, S. (2017). Impact of ancestry categorisations on residential segregation measures using Swedish register data. *Scandinavian Journal of Public Health*, *45*(17_suppl), 62–65.

Jelent, T. G. (1990). Religious Belief and Attitude Constraint. *Society for the*

*Scientific Study of Religion*, *29*(1), 118–125.

Jensen, L., & Tienda, M. (1989). Nonmetropolitan Minority Families in the United States: Trends in Racial and Ethnic Economic Stratification, 1959-1986. *Rural Sociology*, *54*(4), 509–532.

Jhou, M.-S., Chi, Y.-L., & Hsieh, Y.-S. (2010). Spatial Analysis of Suicide in Taiwan: Spatial Imitation or Structural Similarity TT - Transliterated title not available. *Journal of Population Studies*, (41), 1–65.

Johnelle Sparks, P., Sparks, C. S., & Campbell, J. J. A. (2013). An application of Bayesian spatial statistical methods to the study of racial and poverty segregation and infant mortality rates in the US. *GeoJournal*, *78*(2), 389–405.

Johnson, D. C. (1997). Formal education vs. religious belief: soliciting new evidence with multinomial logit modeling. *Journal for the Scientific Study of Religion*, *36*(2), 231–246.

Johnson, S. (2006). *The Ghost Map: The Story of London's Most Terrifying Epidemic-and How it Changed Science, Cities, and the Modern World*. Riverhead Books.

Jost, L. (2006, May). Entropy and diversity. *Oikos*. Blackwell Publishing Ltd.

Jovanovic, V. (1996). The Relationship between Migration, Genetic and Sociocultural Variables: A Log-Linear Approach. *Collegium Antropologicum*, *20*(2), 251–262.

Kasinitz, P. (2005). Bringing the neighborhood back in: The new urban ethnography. *Sociological Forum*, *7*(2), 355–363.

Kelley, F., & The Residents of Hull House. (1895). *Hull House Maps and Papers*. (R. Ely, Ed.). Boston: Thomas Y. Crowell and Co.

Keuschnigg, M., Mutgan, S., & Hedström, P. (2019). Urban scaling and the regional divide. *Science Advances*, *5*(1), eaav0042.

Kirkpatrick, L. A., & Kirkpatrickt, L. A. (1997). A Longitudinal Study of Changes in Religious Belief and Behavior as a Function of Individual Differences in Adult Attachment Style. *Source Journal for the Scientific Study of Religion*, *36*(2), 207–217.

Kominami, Y., Yoshida, S., Tanaka, S., Sanomura, Y., Hirakawa, T., Raytchev, B., … ChayaMA, K. (2016). Computer-aided diagnosis of colorectal polyp histology by using a real-time image recognition system and narrow-band imaging magnifying colonoscopy. *Gastrointestinal Endoscopy*, *83*(3), 643–649.

Krieger, N. (2006). A century of census tracts: health & the body politic (1906-2006). *Journal of Urban Health : Bulletin of the New York Academy of Medicine*, *83*(3), 355–361.

Krieger, N., Feldman, J. M., Waterman, P. D., Chen, J. T., Coull, B. A., & Hemenway, D. (2017). Local Residential Segregation Matters: Stronger Association of Census Tract Compared to Conventional City-Level Measures with Fatal and Non-Fatal Assaults (Total and Firearm Related), Using the Index of Concentration at the Extremes (ICE) for Racial, Econ. *Journal of Urban Health*, *94*(2), 244–258.

Krymkowski, D. H., Sawinski, Z., & Domanski, H. (1996). Classification Schemes and the Study of Social Mobility: A Detailed Examination of the Blau-Duncan Categories. *Quality and Quantity*, *30*(3), 301–321.

Kubat, M. (2015). *An Introduction to Machine Learning*. Cham: Springer International Publishing.

Lagerlund, M., Merlo, J., Vicente, R. P., & Zackrisson, S. (2015). Does the neighborhood area of residence influence non-attendance in an urban mammography screening program? A multilevel study in a Swedish city. *PLoS ONE*, *10*(10).

Lamont, M., & Monar, V. (2002). The Study of Boundaries in the Social Sciences. *Annual Review of Sociology*, *28*(2002), 167–195.

Lance, G. N., & Williams, W. T. (1967). A General Theory of Classificatory Sorting Strategies: 1. Hierarchical Systems. *The Computer Journal*, *9*(4), 373–380.

Längkvist, M., Kiselev, A., Alirezaie, M., & Loutfi, A. (2016). Classification and segmentation of satellite orthoimagery using convolutional neural networks. *Remote Sensing*, *8*(4).

Leban, G., Zupan, B., Vidmar, G., & Bratko, I. (2006). VizRank: Data Visualization Guided by Machine Learning. *Data Mining and Knowledge Discovery*, *13*(2), 119–136.

Lefebvre, H. (1991). *The production of space*. Blackwell.

Legewie, J. (2018). Living on the Edge: Neighborhood Boundaries and the Spatial Dynamics of Violent Crime. *Demography*, *55*(5), 1957–1977.

Legewie, J., & Schaeffer, M. (2016). Contested Boundaries: Explaining Where Ethnoracial Diversity Provokes Neighborhood Conflict. *American Journal of Sociology*, *122*(1), 125–161.

Lenski, G. (1978). Marxist Experiments in Destratification: An Appraisal. *Social Forces*, *57*(2), 364–383.

Lerner, L. K., & Lerner, B. W. (2006). *Medicine, health, and bioethics: essential primary sources. Social issues primary sources collection.*

Levenson, J. A., Oberhuber, K., & Sheehan, J. L. (1973). *Early Italian engravings from the National Gallery of Art.* Washington: National Gallery of Art.

Leventhal, T., & Brooks-Gunn, J. (2000). The neighborhoods they live in: The effects of neighborhood residence on child and adolescent outcomes. *Psychological Bulletin*, *126*(2), 309–337.

Lewis, D. (1973). Causation. *The Journal of Philosophy*, *70*(17), 556.

Lin, Q., Lu, Z., & Xiao, L. (2014). APCG, An Accelerated Proximal Coordinate Gradient Method. *Advances in Neural Information Processing …*, 1–9.

Logan, J. R., Spielman, S., Xu, H., & Klein, P. N. (2011). Identifying and Bounding Ethnic Neighborhoods. *Urban Geography*, *32*(3), 334–359.

Loveland, T. R., & Dwyer, J. L. (2012). Landsat: Building a strong future. *Remote Sensing of Environment*, *122*, 22–29.

Luhmann, N. (1995). *Social systems*. Stanford University Press.

Lund, R. L. (2018). From the dark end of the street to the bright side of the road – automated redistricting of areas using physical barriers as dividers of social space. *Methodological Innovations*, *11*(3), 1–15.

Lund, R. L. (2019a). I Like the Way you Move - The effect of moving to and from different types areas during adolescence on later socioeconomic outcomes. *Geoforum*.

Lund, R. L. (2019b). Moving to Prosperity? The Effect of Prolonged Exposure to Neighborhood Deprivation. *Scandinavian Journal of Educational Research*, *0*(0), 1–17.

Lund, R. L., Jørgensen, A., & Riis, O. P. (2019). Social geographical patterns in membership of the established church in Denmark. *Nordic Journal of Religion* *and Society*, *32*(01), 55–70.

Matiolaski, A., Maksimova, A., & Dziech, A. (2016). CCTV object detection with fuzzy classification and image enhancement. *Multimedia Tools and Applications*, *75*(17), 10513–10528.

McInnes, L., Healy, J., & Astels, S. (2017). hdbscan: Hierarchical density based clustering. *The Journal of Open Source Software*, *2*(11), 205.

Mentese, S., & Okuyucu, A. (2013). Geo-statistical methods in investigation of spatial distribution of population of the Bilecik province TT - Bilecik Ilinde Nufusun Mekansal Dagilisinin Jeoistatistiksel Yontemlerle Incelenmesi. *UluslararasI Sosyal ArastIrmalar Dergisi/The Journal of International Social Research*, *6*(24), 258–267.

Merlo, J., Ohlsson, H., Chaix, B., Lichtenstein, P., Kawachi, I., & Subramanian, S. V. (2013). Revisiting causal neighborhood effects on individual ischemic heart disease risk: A quasi-experimental multilevel analysis among Swedish siblings. *Social Science and Medicine*, *76*(1), 39–46.

Miller, W. L. (1977). *Electoral Dynamics in Britain Since 1918*. Palgrave Macmillan Limited.

Ministeriet for Fødevarer Landbrug og Fiskeri. (2011). *Landdistriktredegørelse 2011*.

Ministry of Finance. (2019). *Finance Bill 2019*.

Ministry of Transportation. (2018). Liste over ghettoområder pr. 1. december 2017 - Transportministeriet.

Monahan, T. (2002). Los Angeles Studies: the emergence of a specialty field. *City and Society*, *14*(2), 155–184.

Moore, G. E. (1965). Cramming more components onto integrated circuits. *Electronics*, *38*(8).

Morgan, T. P. (2018). Nvidia’s Datacenter Growth Beats Moore’s Law Big Time.

Nagel, S. S. (2006). Simplified Bipartisan Computer Redistricting. *Stanford Law Review*, *17*(5), 863.

Nordlund, C. (2018, September 2). Power-relational core-periphery structures: Peripheral dependency and core dominance in binary and valued networks. *Network Science*.

Ocana-Riola, R., Sanchez-Cantalejo, C., & Fernandez-Ajuria, A. (2006). Rural Habitat and Risk of Death in Small Areas of Southern Spain. *Social Science & Medicine*, *63*(5), 1352–1362.

Oliver, M. A., & Webster, R. (1989). A geostatistical basis for spatial weighting in multivariate classification. *Mathematical Geology*, *21*(1), 15–35.

Östh, J. (2018). Equipop.

Östh, J., Clark, W. A. V., & Malmberg, B. (2015). Measuring the scale of segregation using k-nearest neighbor aggregates. *Geographical Analysis*, *47*(1), 34–49.

Östh, J., Malmberg, B., & Andersson, E. K. (2014). Analysing segregation using individualised neighbourhoods. In Christipher D. Lloyd, Ian Shuttleworth, & David W. Wong (Eds.), *Social-spatial segregation: Concepts, processes and outcomes* (pp. 135–162). Clifton: Policy Press.

Pais, J., South, S. J., & Crowder, K. (2012). Metropolitan Heterogeneity and Minority Neighborhood Attainment: Spatial Assimilation or Place Stratification? *Social Problems*, *59*(2), 258–281.

Park, R. E., & Burgess, E. W. (1925). *The city*. Chicago: University of Chicago Press.

Parsons, T. (1968). *The structure of social action*. Free Press.

Petrovic, A., Ham, M. Van, & Manley, D. (2018). Multiscale measures of population: Within- and between-city variation in exposure to the sociospatial context. *Annals of the American Association of Geographers*, *108*(4), 1057–1074.

Petrović, A., Ham, M. van, & Manley, D. (2017). Size matters? Exploring scale for neighbourhood effects. In *AGILE conference 2017*. Springer. Springer.

Ravishankar, S., Jain, A., & Mittal, A. (2009). Automated feature extraction for early detection of diabetic retinopathy in fundus images. In *2009 IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops, CVPR Workshops 2009* (Vol. 2009 IEEE, pp. 210–217). IEEE.

Reisinger, D. (2016). Why Your iPhone Could Have Cost $32 Million in 1985.

Riva, M., Gauvin, L., & Barnett, T. A. (2007). Toward the next generation of research into small area effects on health: a synthesis of multilevel investigations published since July 1998. *Journal of Epidemiology and Community Health*, *61*(10), 853–861.

Rosa, H. (1998). On Defining the Good Life : Liberal Freedom and Capitalist Necessity, *5*(2).

Rotger, G. P., & Galster, G. C. (2019). Neighborhood peer effects on youth crime: natural experimental evidence. *Journal of Economic Geography*, *19*(3), 655–676.

Rubin, D. B. (1974). Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of Educational Psychology*, *66*(5), 688–701.

Rubinov, A. M., Soukhorokova, N. Y., & Ugon, J. (2006). Classes and clusters in data analysis. *European Journal of Operational Research*, *173*(3), 849.

Ruder, S. (2016). An overview of gradient descent optimization algorithms.

Ruggles, S. (2014). Big Microdata for Population Research. *Demography*, *51*(1), 287–297.

Russ, G., & Kruse, R. (2011). Machine learning methods for spatial clustering on precision agriculture data. In *Frontiers in Artificial Intelligence and Applications* (Vol. 227, pp. 40–49).

Sadler, R. C., & Lafreniere, D. J. (2017). Racist housing practices as a precursor to uneven neighborhood change in a post-industrial city. *Housing Studies*, *32*(2), 186–208.

Sampson, R. (2008). Moving To Inequality: Neighborhood Effects and Experiments Meet Structure. *AJS; American Journal of Sociology*, *114*(11), 189–231.

Sampson, R. (2012). *Great American City - Chicago and the Enduring Neighborhood Effect* (1.). Chicago: The University of Chicago Press.

Sampson, R., Morenoff, J. D., & Gannon-Rowley, T. (2002). Aseesing "Neighborhood Effects": Social Processes and New Directions in Research. *Annual Review of Sociology*, *28*(1), 443–478.

Sampson, R., Raudenbush, S. W., & Earls, F. (1997). Neighborhoods and violent crime: a multilevel study of collective efficacy. *Science*, *277*(August), 918.

Sampson, R., & Sharkey, P. (2008). Neighborhood Selection and the Social Inequality. *Demography*, *45*(1), 1–29.

Savage, M., & Warde, A. (1993). Cities and Uneven Economic Development. In *Urban Sociology, Capitalism and Modernity* (pp. 34–62). London: Macmillan

Education UK.

Schaller, R. R. (1997). Moore's law: past, present, and future. *IEEE Spectrum*, *34*(6), 52–55, 57.

Schelling, T. C. (1971). Dynamic models of segregation. *The Journal of Mathematical Sociology*, *1*(2), 143–186.

Scherger, S., & Savage, M. (2010). Cultural transmission, educational attainment and social mobility. *The Sociological Review*, *58*(3), 406–428.

Scott, A. J., & Storper, M. (2003). Regions, globalization, development. *Regional Studies*, *37*(6–7), 579–593.

Scott, A. J., & Storper, M. (2015). The nature of cities: The scope and limits of urban theory. *International Journal of Urban and Regional Research, 39*(1), 1–15.

Smith, R. C., Benavides, M., Pariona, M., & Tuesta, E. (2003). Mapping the Past and the Future: Geomatics and Indigenous Territories in the Peruvian Amazon. *Human Organization*, *62*(4), 357–368.

Snow, J. (1855). *On the Mode of Communication of Cholera* (2nd ed.). John Churchill.

Soja, E. W., & Gren, M. (1991). *Postmodern geographies - the reassertion of space in critical social theory*. *Geografiska annaler* (Vol. 73).

Statistics Denmark. (2018). Baptism in numbers.

Statistics Denmark. (2019). Membership of The Danish State Church in numbers.

Storper, M., & Scott, A. J. (2009). Rethinking human capital, creativity and urban growth. *Journal of Economic Geography*, *9*(2), 147–167.

Studer, M., & Ritschard, G. (2016). What matters in differences between life trajectories: a comparative review of sequence dissimilarity measures. *J. R. Statist. Soc. A*, *179*(2), 481–511.

Thrasher, F. M. (1927). *The gang: a study of 1,313 gangs in Chicago* (1st ed.). Chicago: University of Chicago Press.

U.S. Department of Commerce Economics and Statistics Administration Bureau of Census. (1994). *Census tracts and block numbering areas. The Geographic Areas Reference Manual (GARM).*

Urry, J. (2000). *Sociology Beyond Societies: Mobilities for the Twenty-First Century*. Psychology Press.

Urry, J. (2009). Speeding up and Slowing Down.

Valdez, S. (2014). Visibility and votes: A spatial analysis of anti-immigrant voting in Sweden. *Migration Studies*, *2*(2), 162–188.

van der Walt, S., Schönberger, J. L., Nunez-Iglesias, J., Boulogne, F., Warner, J. D., Yager, N., … Yu, T. (2014). scikit-image: image processing in Python. *PeerJ*, *2*, e453.

Vinikoor, L. C., Kaufman, J. S., MacLehose, R. F., & Laraia, B. A. (2008). Effects of racial density and income incongruity on pregnancy outcomes in less segregated communities. *Social Science and Medicine*, *66*(2), 255–259.

Vu, T. H. N., Lee, Y. K., & Bui, T. D. (2014). A technique for extracting behavioral sequence patterns from GPS recorded data. *Computing. Archives for Informatics and Numerical Computation*, *96*(3), 163–188.

Wadhwa, V. (2017). Don’t believe the hype about AI in business.

Warren, G., Metternicht, G., & Speijers, J. (2006). Use of a spatially weighted multivariate classification of soil properties, terrain and remote sensing data to form land management units. *American Society for Photogrammetry and Remote Sensing - Annual Conference of the American Society for Photogrammetry and Remote Sensing 2006: Prospecting for Geospatial Information Integration*, *3*.

Waterman, W. C., Park, R. E., Burgess, E. W., & Boyd, W. K. (2006). *The City.The Story of Durham*. (M. Janowitz, Ed.), *Social Forces* (Vol. 5). Chicago: The University of Chicago Press.

Watmough, G. R., Atkinson, P. M., Saikia, A., & Hutton, C. W. (2016). Understanding the Evidence Base for Poverty–Environment Relationships using Remotely Sensed Satellite Data: An Example from Assam, India. *World Development*, *78*, 188–203.

Watmough, G. R., Marcinko, C. L. J., Sullivan, C., Tschirhart, K., Mutuo, P. K., Palm, C. A., & Svenning, J.-C. (2019). Socioecologically informed use of remote sensing data to predict rural household poverty. *Proceedings of the National Academy of Sciences*, *116*(4), 1213–1218.

Watmough, G. R., Palm, C. A., & Sullivan, C. (2017). An operational framework for object-based land use classification of heterogeneous rural landscapes. *International Journal of Applied Earth Observation and Geoinformation*, *54*, 134–144.

Weppner, R. S. (1977). *Street Ethnography: Selected Studies of Crime and Drug Use in Natural Settings*. *SAGE Annual Reviews of Drug and Alcohol Abuse*. Thousand Oaks: Sage Publications.

Whitfield, P. (2005). *Cities of the World: A History in Maps*. *University of California Press*. University of California Press.

Wilson, W. J. (1987). *The truly disadvantaged : the inner city, the underclass, and public policy*. Chicago and London: The University of Chicago Press.

Wilson, W. J. (1996). *When Work Dissapears: The world of the new urban poor* (1.). New York: Alfred A. Knopf, Inc.

Wodtke, G. T., Harding, D. J., & Elwert, F. (2011). Neighborhood effects in temporal perspective: The impact of long-term exposure to concentrated disadvantage on high school graduation. *American Sociological Review*, *76*(5), 713–736.

Wolodtschenko, A., & Forner, T. (2007). Prehistoric and Early Historic Maps in Europe: Conception of Cd-Atlas, *2*(2), 114–116.

Wooldridge, J. M. (1995). Selection corrections for panel data models under conditional mean independence assumptions. *Journal of Econometrics*, *68*(1), 115–132.

Wooldridge, J. M. (2002). *Inverse probability weighted M-estimators for sample selection, attrition, and stratification*. *Port Econ J* (Vol. 1). Springer-Verlag.

Wooldridge, J. M. (2010). *Econometric Analysis of Cross Section and Panel Data*. *MIT Press* (2nd ed., Vol. 58).

Yang, C., Everitt, J. H., & Murden, D. (2011). Evaluating high resolution SPOT 5 satellite imagery for crop identification. *Computers and Electronics in Agriculture*, *75*(2), 347–354.

Zang, X. (2006). Social Resources, Class Habitus and Friendship Ties in Urban China. *Journal of Sociology*, *42*(1), 79–92.

Zhang, Ying. (2006). Nonparametric k-sample tests with panel count data. *Biometrika*, *93*(4), 777.

Zhang, Yun. (1999). Optimisation of building detection in satellite images by combining multispectral classification and texture filtering. *ISPRS Journal of Photogrammetry and Remote Sensing*, *54*(1), 50–60.

Zitlau, P. A., Weaver, J. B., Siegfeldt, H. J., Hess, S. W., & Whelan, J. N. (1965). Nonpartisan Political Redistricting by Computer. *INFORMS*, *13*(6), 998–1006.

Zubi, Z. S., & Saad, R. A. (2011). *Using some data mining techniques for early diagnosis of lung cancer Using Web Mining for User Behavior Understanding View project Using Some Data Mining Techniques for Early Diagnosis of Lung Cancer*.

MI

*Original Article*

Methodological Innovations
September-December 2018: 1–15
© The Author(s) 2018
Article reuse guidelines:
sagepub.com/journals-permissions
DOI: 10.1177/2059799118814386
journals.sagepub.com/home/mio
SAGE

# From the dark end of the street to the bright side of the road: Automated redistricting of areas using physical barriers as dividers of social space

**Rolf Lyneborg Lund**

**Abstract**

This study examines the properties of administrative areas compared to a new method of automated redistricting when measuring social differentiation and segregation. Using physical barriers, such as roads, railways, streams, areas of uninhabited nature, and the like as dividers of social space, this study explores alternative ways of thinking social belonging and social cohesion that are beyond standard measures of geography and utilize areas of smaller size and population count. The geographical data are linked to Danish register data of the total Danish population in 2015, N = 4,986,125 on key variables of income, months of completed education, and ethnicity. The overall findings in this study suggest that rethinking geography when localizing social enclaves and segregated communities yields better results than using the more illogical administrative areas. The visual inspection, entropy levels of homogeneity, and intraclass correlation suggest that smaller areas that are divided by physical objects serve as a better reservoir of social cohesion and therefore better measurement of social inequality.

**Keywords**

Social segregation, inequality, GIS, neighborhood studies

## Introduction

Numerous studies have investigated the formation and effect of neighborhoods on both individual and structural levels on a wide variety of measures (Damm and Schultz-Nielsen, 2008; Ministry for City, Habitation and Rural Districts, 2014; Galster, 1989, 2010; Grannis, 1998; Lee and Campbell, 1997; Logan et al., 2011; Massey and Denton, 1988; Sampson, 2008). The goals of these studies vary in both how they perceive neighborhoods and how they conceptualize space. Some focus especially on segregation and to explain segregation inside areas (Bower et al., 2014; Breetzke and Horn, 2006; DeSilva et al., 2012; Grannis, 1998; Johnson et al., 2004; Zingher and Thomas, 2014), while others seek to explain social outcomes as effected by the total amount of neighborhoods (Buck, 2001; Fone et al., 2007; Pattison and Robins, 2002; Pickett and Pearl, 2001; Veenstra et al., 2005).

What they all share is the neighborhood as an entity to contain the people of interest. This container can be any entity the researcher chooses and quite often the data limitations restrict research to a predefined set of administrative areas. Earlier studies in sociology, especially the work of the Chicago school (Park, 1928; Park and Burgess, 2007; Park et al., 1967), revolutionized the way we understand neighborhoods, but before the emergence of computers, the general and macro level statistical analyses were impossible. With the emergence of the first personal computer software designed for Geographical Information System (GIS) in 1986 (Clifford et al., 2010), it was still only a select few in sociology that worked with neighborhoods as a non-predefined entity. It was not before the end of the 1990s that access to Microsoft Windows–driven GIS-editing software became widely available but still mostly limited to geographical sciences (Clifford et al., 2010). With the evolution in computers and computational power, larger macro-models for GIS take

Department of Sociology and Social Work, University of Aalborg, Aalborg, Denmark

**Corresponding author:**
Rolf Lyneborg Lund, Department of Sociology and Social Work, University of Aalborg, 9220 Aalborg, Denmark.
Email: rolfll@socsci.aau.dk

less time and have become easier to utilize. This advancement has paved the way for a wide array of models and subdivisions of geography to contain social data.

Even with the advancements of GIS and geo-referenced data, a lot of research still uses administrative borders (parishes or municipalities) as their smallest unit of reference when trying to understand the inhabitants inside (Andersson and Malmberg, 2013; Åslund and Skans, 1985; Cunha et al., 2009; Fischer et al., 2004; Söderström and Uusitalo, 2010; Zingher and Thomas, 2014). Even when utilizing smaller areas, as Census Tracts in the United States, the usefulness or validity of the areas is very rarely questioned (Bower et al., 2014; Krieger et al., 2017a, 2017b). As Lee et al. (2008) notes, "Most studies implicitly assume that the tract constitutes an appropriately-sized spatial unit for capturing segregation." This raises some very fundamental questions about the understanding of place and living: How do we know that the areas we use to contain the social aspects of its inhabitants make sense? What is the effect of using only predetermined administrative areas as opposed to exploring the possibilities of GIS-coded data?

The overall concept of a neighborhood is more than just the size and the qualitative feeling of being in a neighborhood and a specific definition of how to capture neighborhoods can easily end up being wrong when considering what the goal of capturing neighborhoods is. "Capturing segregation," as noted in the quote above, implies that the neighborhood in question has a very specific composition and that, to be segregated from other areas, it must be somewhat homogeneous before it truly captures the social differences between one area and the neighboring ones. Especially, the human ecology tradition with roots in the Chicago School has worked with understanding neighborhoods as something that creates some form of unity (Buttimer and Seamon, 1980; Gans, 1961; Hwang, 2015; McIntosh, 1986; Newton and Johnston, 1976; Taylor, 1997) which later spurred the concept of social efficacy in the work of Robert Sampson (2012). The concept of social efficacy is especially interesting when trying to understand local communities; proximity is only interesting if it brings on some form of social efficacy inside the neighborhood. This efficacy can either be in the form of social coherence or as an unspoken way to define the neighborhood as something uniform (Sampson, 2008, 2012; Sampson et al., 2002). In the center of efficacy is proximity; without closeness there can be very little dynamic social efficacy. This, Sampson notes, does not mean that there will be social efficacy solely based on proximity but that this is a factor that needs to be present.

This article presents a methodological approach to redistricting with special focus on homogeneity and measuring small-scale neighborhoods in comparison with administrative areas on key variables as income distribution, educational attainment, and ethnic composition. The goal is not to explain the root cause of the segregation or any direct causal link between settlement and segregation level but instead point out that level of measurement matters when it comes to geographical distribution.

## The methodological understanding of neighborhoods

There are studies that utilize geographical information more refined than just the administrative areas. The point of proximity to define neighborhoods is becoming more common when trying to understand smaller areas of living (Damm and Schultz-Nielsen, 2008; Feld, 1981; Freisthler et al., 2016; Grannis, 1998; Jones and Huh, 2014; Jones and Pebley, 2014; Kwan, 2013; Lee and Campbell, 1997; Lee et al., 2008; Logan et al., 2011; Patterson and Farber, 2015). Many of the papers try to go further than to use general administrative areas, but because of either data limitations or problems in linking this to geography, they struggle to either propose a general model that can be utilized on a macro scale or produce areas that follow a specific logic.

They all follow the same criteria at a varying rate, which are proximity, small size, homogeneity, and geography. Proximity, here understood as people living close together, is often understood as a way of securing homogeneity; that the people living close to one another also share similar beliefs and socioeconomic status. The overall problem with proximity and thereby homogeneity is also inherent in the way we understand geography and social life; the center of an area can only appear once the area is present and not the other way around. The question then becomes, "A proximity to what?" This is also important to conceptualize the size of the area because proximity and homogeneity can only appear once the entity that holds these things does not suffer from the generalization of aggregation too severely.

Some of the newer methods that offer more detailed area definitions vary in how they prioritize the above criteria. Commonly used methods include nearest neighbors in different ways, small area statistics (like the Swedish Small Areas for Market Statistics (SAMS)) that are focused on market statistics, and Bayesian spatial models. These methods all offer improved use of space but do so at the cost of precision when it comes to understanding the neighborhood as a useful entity.

Studies that focus on nearest neighbors are becoming more and more frequent especially since the freeware program Equipop, which utilizes K-nearest neighbors, has grown in popularity (Andersson and Malmberg, 2013; Dawkins, 2006; Lee et al., 2008; Östh et al., 2015). One of the major advantages of the nearest neighbors' approach is the inherent use of the population as a complete set of neighbors. As Equipop uses whatever clustering base one chooses to generate the neighbor connections, many other forms of overlapping neighborhood measurements, as health status in neighborhoods as a measure of "fuzzy" health in areas (Propper et al., 2007; Veldhuizen et al., 2013), socioeconomic status in voter behavior (Johnston et al., 2005; Macallister et al., 2001), distance to one another or to economic centers (Kryvobokov, 2013), or even moving patters to kinship (Clark, 2017).

This allows for more intricate connections between individuals when applying the data to the geography and does not require any hard borders since everyone are, in some sense, connected; the methods rely on fuzzy borders instead of firm. This is, though, also the main problem with the method. K-nearest neighbors do not abide by geographical borders or physical objects but instead rely on numbers of people. This could be improved if the researcher has complete individual level data connected to a specific coordinate but since almost all register data require some sort of anonymity, the concept of connecting each individual person to geography makes it impossible to uphold the discretion criteria. This type of method can often generate some very homogeneous areas but at the cost of the geographical sense of place.

The geographical sense of place is much more in focus when utilizing other generic geographical units like the SAMS in Sweden (Brydsten et al., 2017; Carlsson et al., 2017; Lagerlund et al., 2015; Merlo et al., 2013; Östh et al., 2014; Sundquist et al., 2016). These areas are constructed especially for homogeneity in smaller units since they are often purchased by commercial organizations to better focus their marketing at the correct demographic. The problem with these units is that they change in form and shape over time and that the small size is of less concern. This means that they are designed for encompassing very class-specific entities as income and education but can easily miss more subtle signs of segregation. This is, of course, to make the areas more attractive to companies, since areas only containing 100 persons might be too small of a focus group. The average unit of SAMS contains 1100 inhabitants with only a slightly lower median of 1062, where the new areas created in this article have a mean inhabitant count of 537 and a median of 249. The change over time and the still relatively large size of units makes the SAMS very attractive to companies but makes the use in demographical and sociological research much more limited.

The last method to be touched upon in this article is the Bayesian approach (Fiscella and Fremont, 2006; Johnelle Sparks et al., 2013; Law et al., 2015; Vinikoor et al., 2008). Many studies focusing on Bayesian methods use hotspot analysis to locate areas and then use already existing blocks or smaller areas to cluster and get a more homogeneous clustering. This method allows a very high amount of homogeneity as well as a direct way to control size and population count. This does, however, require a priori assumptions about the population distribution. As in the work by Johnelle Sparks et al. (2013), they model infant mortality rates with a priori distributions of means equal to the average risk of the neighboring counties and draw subsamples from this to predict racial and poverty segregation. This means that a Bayesian model can inherently account for a very high amount of homogeneity, but it is also a very specific model; it can account for specific social phenomena but changes with the subject at hand. Assumptions must change as the phenomena change.

This article proposes a new method to generate areas that are more grounded in the physical barriers and areas that are much smaller than the widely available administrative areas as well as utilize administrative data to fully understand the complexity of these areas.

## Data

This article utilizes two different types of data: geo-referenced data and registers for the Danish population. The first segment of data, the geo-referenced data, consists of The National Square Grid and a large collection of topographical vector-based object maps that contain roads, streams, lakes, forests, and most other place-specific objects found in Denmark. The National Square Grid is a national system of vector grids constructed by The Danish Geodata Agency and Statistics Denmark that measure 100 m × 100 m and have unique identifications and spatial reference. This is, by itself, not very interesting but because The National Square Grid is linked to each person in the Danish registers, this makes it possible to place each person living in Denmark inside a square that is 100 m × 100 m. When considering redistricting, it is very valuable to have the smallest units of measurement as possible and being able to modulate areas in cells that are no larger than 100 m × 100 m makes for ideal clustering. One could argue that the most ideal form would be to keep the smallest unit of measurement and not cluster the square grid in any way but because Statistics Denmark operate with very strict confidentiality requirements that require at least 100 households per geographical unit and taking into consideration that, in 2017, less than 1% of the squares are inhabited by more than 100 households, this makes using only square grids impossible.

Another reason for not using only the 100 m × 100 m cells is of a theoretical perspective; what area do we interact with each other and how do we define the social barriers that consists of the feeling of "us" and "them"? A lot of research has pointed to some sort of cohesion inside areas and has tried to define what makes a neighborhood (Damm and Schultz-Nielsen, 2008; Deng, 2016; Freisthler et al., 2016; King et al., 1994). Scott L Feld even points to the fact that even though we live in specific areas, these are often divided by specific physical barriers like roads, railways, and other objects commonly found in both the urban and rural landscapes (Feld, 1981). By this logic, the square grid, by itself, will be as illogical as other administrative area divisions.

The other set of data consist of register data for the total of the Danish population over 18 years of age in 2015. The registers are a compilation of individual level information about education measured in full months of total education, primary school included, income measured as gross income per year, age, gender, and ethnicity. All data on interval level have been utilized when mapping but categorized into ordinal measures for the entropy measurement. Furthermore, the data consist of other geographical information like parish and municipality. All of this is linked to the square grid after the clustering has taken place.

## Methodology

As stated in the introduction, most studies that investigate the effect of neighborhood or residential area use predefined and often administrative geographical units of measurement. The overall problem with administrative areas is, especially in a Danish context, that even the smallest areas of measurement, parishes, are very poor indicators of the types of people who live there. The Danish parishes are, in most cases, many hundreds of years old and have not been updated or redistricted, as new settlements have taken place. This perhaps makes sense in a religious perspective, since most parishes still belong to a specific church but when interested in sociodemographic areas and social segregation, this type of measurement is lacking. What this article proposes is another way of thinking place of living. These next sections will outline the process of setting up criteria for the automated redistricting algorithm and show how measurements of area homogeneity are set up.

### *Inductive automated redistricting—criteria*

Considering the theoretical and practical foundation presented earlier, the algorithm to handle the automated redistricting is based on inductive reasoning. The overall criteria were as follows:

- Are separated by physical barriers;
- Are contained within a single polygon and not separated by other polygons;
- Have at least 100 households present in the years 2000, 2005, 2010, and 2015.

The algorithm works in two steps; first step is to apply the barriers in question, which are highways, roads broader than 6 m, rivers and streams broader than 3 m, railways, lakes, forests, coastlines, and intakes. This is also the reason for labeling the algorithm as inductive. Since there can be no preconception about what areas should be formed, all areas are defined by the criteria and emerge solely because of physical barriers that are thought to create not only a visible barrier but also a social barrier that establishes a sense of "the people on the other side of the road" (Feld, 1981). Using this algorithm also implies that there can be no real preconception about how many inhabitants can be present in one polygon. Earlier research has applied a divider once the number of inhabitants has been reached but this goes beyond the logic of using physical barriers as the most important social divider in regard to neighborhoods (Damm and Schultz-Nielsen, 2008). This has shown to be a very small problem since more than 90% of the areas are smaller than 1000 inhabitants and less than 1% bigger are inhabited by more than 5000. From a purely methodological standpoint, it would be simple to divide those larger areas into smaller areas, but this would also result in a radical break with the barrier criteria. For this reason, areas are not manipulated if they contain more than 100 inhabitants.

After the initial first step, the square grid is applied. The square grid, in this case, contains not only information about square location but also number of households in each square. Since the smallest possible division of inhabitants is the square grid, the grids are dissolved into the areas where the largest part of the square is located. The borders of the areas are then formed after the squares so that the smooth borders are replaced with the borders of the squares in each area. By doing this, it is possible to calculate how the population is distributed into the first array of areas (Figure 1).

As can be seen in Table 1, the total amount of new areas is 20,940, and of these areas, only 28% of the areas meet the minimum requirements of 100 households. What is also quite evident is that it is impossible to secure large enough areas by only using barriers. Furthermore, a valid point would also be that a neighborhood with only four residents would be very poor at capturing the neighborhood effects.

To remain true to the criteria that all barriers must be kept as separators would mean that further clustering would stop at this point. This is, however, not possible because of the discretion criteria of Statistics Denmark so another algorithm performs the second clustering. The criteria set here are as follows (Figure 2):

- All areas must be applicable for a clustering;
- Areas must share borders;
- Areas with the largest borders measured in percentage shared will be considered for clustering first;
- Selection of areas to the clustering process is based on the least possible amount of merges;
- Selection of areas to the clustering is second based on resulting in the smallest possible number of inhabitants in the merged areas if there are more than one way to obtain the least available merges;
- Areas must be merged until 100 inhabitants are reached.

The main point in the above criteria is to make the algorithm work in a way that results in the least amount of area merges. The problem in selecting a specific point to start the selection process is that the final merge would vary extremely and would be different each time a different starting polygon was selected. This still holds true for this method in the way that a different polygon would result in a different merge. Because the algorithm initially calculates, how the merge

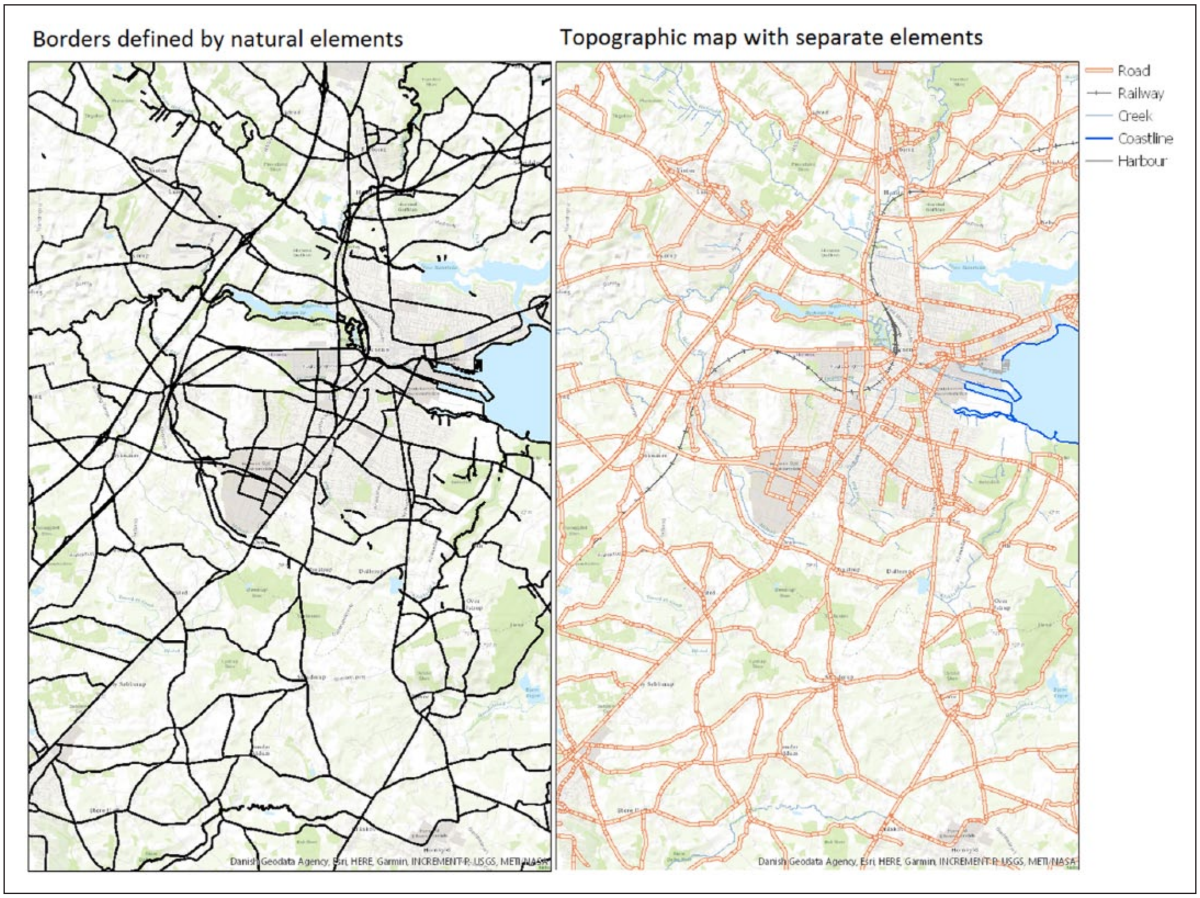



**Figure 1.** First implementation of algorithm.

**Table 1.** First step to cluster grid in areas defined by barriers.

| Households | 2000 | 2005 | 2010 | 2015 |
|---|---|---|---|---|
| 1–4 | 4.27 | 4.47 | 4.47 | 4.81 |
| 5–9 | 5.88 | 5.50 | 5.66 | 5.57 |
| 10–19 | 10.43 | 10.53 | 10.38 | 10.65 |
| 20–49 | 23.18 | 23.00 | 22.99 | 23.55 |
| 50–99 | 19.14 | 19.11 | 18.81 | 18.32 |
| 100–149 | 9.07 | 9.02 | 9.06 | 8.55 |
| 150< | 28.02 | 28.37 | 28.63 | 28.55 |
| Total N | 20,940 | 20,940 | 20,940 | 20,940 |
| Total % | 100.00 | 100.00 | 100.00 | 100.00 |

would be if the least possible merges is the main criteria, and getting the least inhabitants in each area, the algorithm consequently creates the same merges if the process was to be repeated.

The reasoning behind the criteria that all areas must be applicable is twofold; first, it is to make sure that the algorithm has enough adjacent polygons to select for merges even if a specific area holds more than 100 households, but second it secures that if a large border is shared, the smaller area does not merge with a more marginal area because of restricted areas. But securing the largest shared borders does not help with the fact that neighboring areas that should not be merged end up being merged; since the only way to apply data to the model is to secure 100 inhabitants in each area, this criterion at least secures a proximity so that social interaction inside areas is more plausible than if they were divided by large areas.

After applying the second step, all areas are above the discretion criteria. The only thing the algorithm does not solve is the problem with islands. There are in total eight islands inhabited that do not meet the minimum requirements for Statistics Denmark. Since the point of this algorithm is to utilize physical borders, these few islands have been removed. Later, research could consider implementing these in some form (Table 2).

## *Measurement of area homogeneity*

Since one of the overall theoretical ideas presented in this article is based on the social classes' physical settlement, it is of importance to measure the overall homogeneity of the inductive areas. The main problem with the standard measures of segregation is how to work around multiple categories. Many researchers are interested in minorities compared to majorities inside given areas (Barone, 2011; Charles and

Bradley, 2009; Charles and Grusky, 1995; Damm and Schultz-Nielsen, 2008), but because the aim of this article does not only encompass diversity between groups without an inherent minority but it also needs to be able to compare many different categorical variables with a varying set of categories. To account for the categorical elements in the article, I have used Shannon's entropy and this takes the form of

$$H(X)=\sum_{i=1}^{n}p(x_i)\log_b\frac{1}{p(x_i)}$$

where $x$ represents the frequency of a given educational group represented in the $i$th area. As Jost (2006) points out, there is an overall problem using entropy as a measure for diversity, since entropy and diversity do not contain the same properties (Jost, 2006; Rao and Thomas, 1988; Ricotta and Szeidl, 2006). Where the interpretation of entropy can be thought of as a measure of uncertainty, diversity is a more intuitive measurement to understand because it contains the effective number of groups observed in $i$. To address this problem, Jost points out that when comparing effective number of species over different aerial units, the form would be $\exp(\sum_{i=1}^{n}p(x_i)\log_b(1/p(x_i)))$. This not only considers how many educational groups are present inside each area but also weighs each category to their respective probabilities and makes it intuitively easier to read. Another property of having the exponential function is to be able to use means and other parametric measurements in a meaningful way.

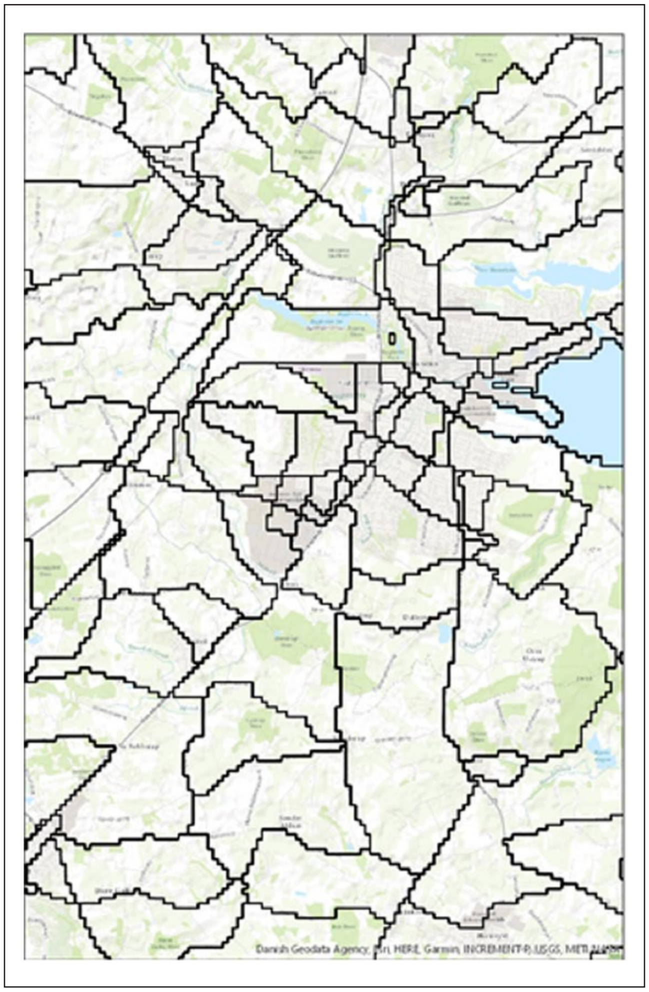



**Figure 2.** Final step of algorithm.

**Table 2.** Percentiles of residents in areas after final clustering.

| | 2000 | 2005 | 2010 | 2015 |
|---|---|---|---|---|
| 0%–10% | 208 | 207 | 204 | 196 |
| 10%–20% | 222 | 225 | 224 | 218 |
| 20%–30% | 240 | 243 | 245 | 240 |
| 30%–40% | 262 | 265 | 268 | 264 |
| 40%–50% | 289 | 294 | 300 | 299 |
| 50%–60% | 326 | 335 | 343 | 345 |
| 60%–70% | 382 | 399 | 414 | 420 |
| 70%–80% | 511 | 528 | 554 | 565 |
| 80%–90% | 764 | 798 | 824 | 847 |
| 90%–100% | 1362 | 1393 | 1444 | 1487 |

## Scaling

One of the main issues about comparing different methods to secure area homogeneity is to understand how one method differs from others. Most of this article focus on the difference between this new proposed method of area division compared with administrative areas as parishes since this is the most widely used scale but one could argue that if one reduces N in areas, general data smoothness would ensue a greater homogeneity. As noted by Samardzic-Petrovic et al. (2016), scale matters when wanting to encompass subgroups in the population. To account for this and to fully investigate the physical barrier approach compared to similar approaches, I have applied a wide set of moderations and simulations.

As can be seen in Table 3, five different versions of algorithms have been run to test how much of the increased homogeneity is due to data smoothing and how much is due to the actual method.

Each of the above methods is run as loops 100 times to compare differences in simple chance divisions. Because of the computational requirements to run these, and especially the last two, only 100 runs have been performed.

The first three types are based on parishes to investigate how much more homogeneity one can accomplish if one adjusts the parishes. They start with the parish as is and then a randomization is applied. The border change changes the circumference of the parishes dynamically so that no inhabitants fall into no man's land—this also means that parishes are being shrunk or enlarged at random. The second moderation is dividing the parish into two equally large half-parishes—each run is a different division at random. Quarter

**Table 3.** Moderations to the zonation on different scalars.

| Type of moderation | Description | $\bar{N}_i$ | $N_j$ |
|---|---|---|---|
| Parish border change | Keeping the parish placements but let the parish border vary with 1 km at random unless the parish discriminates the minimum inhabitants requirement | 2095 | 2190 |
| Half parish | Keeping parish borders as is but reducing parish to half size with the border drawn as a straight line from end to end unless the parish discriminates the minimum inhabitants requirement | 1035 | 4380 |
| Quarter parish | Keeping parish borders as is but reducing parish to quarter size with the borders drawn as straight lines from end to end where the theoretical angle can be anywhere between 1° and 179° as long as the minimum inhabitants requirement is met | 559 | 8760 |
| Theoretical limit | Removing physical barriers as separators and enforces a straight rule about meeting 100 inhabitants. Algorithm still enforce a rule where squares should share borders to cluster | 398 | 9107 |
| Random clustering | Completely random clustering where physical barriers and proximity is removed and the only considerations that the algorithm secures are areas with 100 inhabitants inside Danish borders | 156 | 29,874 |
| Parish | Unchanged parish areas | 2097 | 2190 |
| New area | New method areas | 537 | 8043 |

parishes follow the same logic except that this allows for oblique divisions—each quarter does not need to be exactly 25% of the parish if all four parts have met the requirements for a number of inhabitants.

The last two moderators are more in line with the idea of the method proposed in this article; they still work with smaller areas, but they ignore physical barriers. The theoretical limit moderator abandons barriers and instead focuses on reaching 100 inhabitants with squares sharing borders—this results in very small areas with no more than a mean inhabitant count of 156. The last moderation is a test to see whether geography matters at all; is it possible to generate homogeneity by pure chance?

## The concept of inductive neighborhoods in a Danish context applied

To better understand how these new areas work compared with the alternative parishes, a series of comparisons are made. The following section will try to show how smaller areas differ in understanding common socioeconomic and demographic trends in a geographical setting. The analysis will focus on educational attainment in months, yearly income, and ethnicity.

### *Educational attainment and the place we live*

Education in a Danish setting has undergone an expansion during the past 70 years. Educational attainment has seen a massive upswing and many political goals have been set to see this trend continuing. One thing that is especially important to understand in the progress of the educational attainment goals is the geographical dispersion of educational segregation, to pinpoint what areas are attaining education, and more importantly, which ones that do not. Many policymakers inform themselves using maps showing mean educational attainment in areas, but most of the time, these maps only tell very little about the actual segregation in a geographical perspective because the attainment means are being aggregated to either municipality or regional level.

A simple visual comparison of the mean of education length in months in new areas compared to parishes shows a very interesting trend; even at parish level, the localized educational segregation is being masked by aggregation compared to the new areas (Figure 3).

Small pockets of very low educational attainment are showing inside parish level data, and in some very specific cases, the variation inside a parish is so big that the attainment compared to neighboring areas on the left figure misses three whole levels of education.

Further investigating the difference between the new areas and parishes on a national level, with educational attainment at a categorical level, reveals that there is general problem with masking localized problematic areas within parishes (Figure 4).

Comparing entropy in the boxplot above shows that the median number educational categories present inside the same areas are close to 4, while the median categories inside parishes are 4.6. What is even more interesting is that from the 25th to the 75th percentile is generally much lower than that of the parishes. Unsurprisingly though, it is evident that the spread outside the 25th to the 75th percentile to the upper and lower adjacent values is larger for the new areas than for parishes, but since entropy is a measure of probabilities, it is expected that areas with only 100 residents are more sensitive to small changes in area composition than parishes are.

Another way of looking at the difference between parishes and new areas is the variation inside and between each

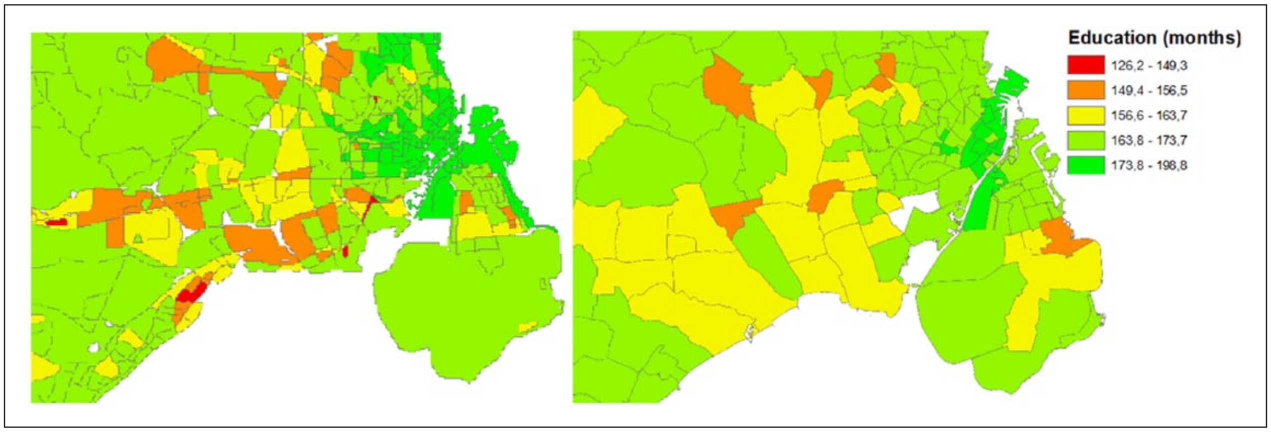



**Figure 3.** Smaller areas (left) and parishes (right) with average educational length in months.

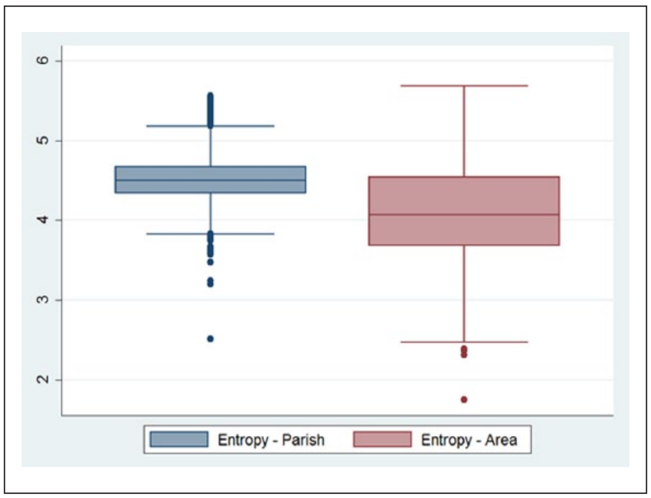



**Figure 4.** Entropy of educational groups in parishes and new areas.

aggregated measure. By utilizing an analysis of variance (ANOVA) on new areas and parishes, it is possible to fully grasp how the aggregation measures differ (Table 4).

Much like Figure 4, it is evident that there is less variance on average inside areas than there are in parishes. The mean sum of squares within areas is 1177, while the same measure is 1199 in parishes, but what is even more interesting is how much they differ in their between variation, with areas having a mean square of sum of 53,394, while parishes only have 49,885. This indicates that areas differ more between them than parishes and that areas are more homogeneous. When considering homogeneity, it is also worth noting that the intraclass correlation is 4.2 times larger in areas compared to parishes.

## *Examples: ethnicity*

One of the core concepts of residential segregation often centers on ethnicity and racial segregation. The goal of most of the research is to understand how segregated we are in our residential patters when it comes to race and to better understand how enclaves appear in closed geographical form. Research is often limited in the access to understand this segregation on national level because of data availability.

Figure 5 shows the percentage of first- and second-generation immigrants compared to the native population inside new areas (left) and parishes (right). As with education, the general racial compositions of the Capitol Area suffer from heterogeneity when only looking at data aggregated to parish level. Looking at the center of Copenhagen, a lot of areas emerge that are almost exclusively dominated by native Danes, whereas the southwest part of the map reveals enclaves that consist of areas that have more than 50% first- or second-generation immigrants (Figure 5).

Comparing this to the overall entropy on national level, as seen in Figure 6, these findings are consistent with the maps above. The median for new areas is 1.27, whereas the median for parishes is 1.34. This measure of entropy ranges from 1, where all residents inside a specific area are of either only native Danes or only immigrants, whereas an entropy of 2 is an equal part of both. Not surprisingly, this measure does not amount to many areas where the distribution is close to 2, since especially the Western areas of Denmark have a very low overall proportion of immigrants.

As with education, the 25th to the 75th percentile for areas is lower than that of the parish and the upper and lower adjacent values are bigger.

When investigating the mean sum of squares in Table 5, the pattern of more homogeneity within areas and more heterogeneity between areas than parishes can be seen. Likewise, the intraclass correlation is 2.6 times larger for the new areas than it is for parishes.

## *Examples: income*

Income redistribution is a large part of the Danish welfare state and thus the understanding of where the wealth accumulates

**Table 4.** Educational attainment.

| | SS | df | MS | Intraclass correlation | Mean, SD | Min | Max |
|---|---|---|---|---|---|---|---|
| Education (A) | | | | | | | |
| Between area | 4.294e+08 | 8042 | 53,394.821 | – | – | – | – |
| Within area | 4.584e+09 | 4,102,867 | 1117.1522 | 0.08391 | 33.34 | 18.76 | 44.21 |
| Education (P) | | | | | | | |
| Between parish | 1.085e+08 | 2174 | 49,885.951 | – | – | – | – |
| Within parish | 5.200e+09 | 4,334,210 | 1199.7594 | 0.01996 | 34.61 | 26.38 | 45.78 |

SS: sum of squares; MS: mean of squares; SD: standard deviation.

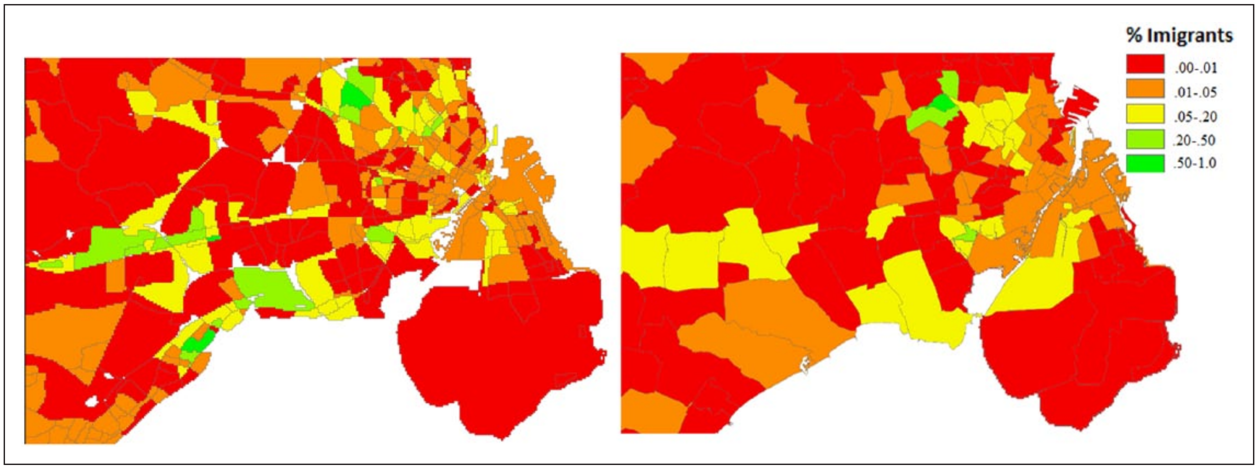



**Figure 5.** Smaller areas (left) and parishes (right) with percentage non-native residents.

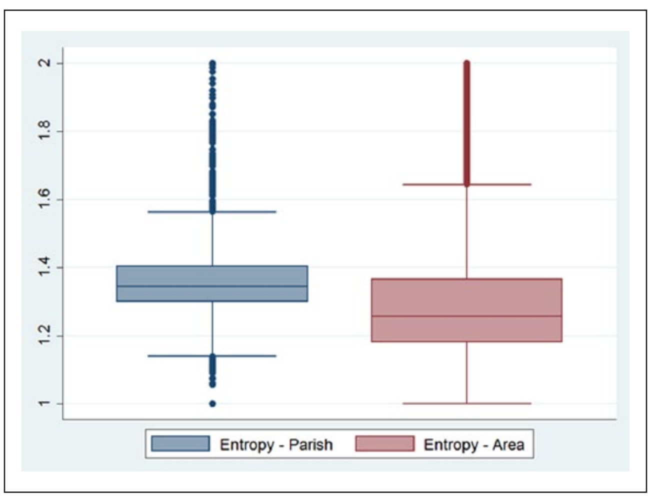



**Figure 6.** Entropy of ethnic heritage groups in parishes and new areas.

is important to understand how the redistribution should be performed. Looking at the map of the Capitol Area, a somewhat disturbing distribution appears when comparing new areas with parishes. Where both educational status and share of immigrants give some interesting insight into distribution and smaller enclaves, income distribution is a very different story.

Figure 7 shows income quintiles with red being low income and green being high income. What becomes very apparent is that the parishes on the right show almost no variation in the categories. Not a single parish consists of the highest income grouping when aggregating even though the wealthiest Danish areas are located just north of the capital, which is on the top of the map; we see only the second highest quintile range located there. Looking at the areas on the left, it is evident that there is a concentration of wealth just north of the Capitol.

The same, but not as extreme, goes for the most income deprived areas where only the center of Copenhagen is depicted as the lowest quintile, while this distribution is very different when considering the areas on the left. Much of especially the lower income areas are being obscured by aggregating data to parish level where many parishes to the west depict an average income level and not pointing out the more deprived areas that emerge on the left map.

Considering the entropy of both parishes and areas, where I categorized income in 12 groups, the same pattern is present. New areas hold a much lower median number of income groups, while the new areas have higher and lower adjacent values. This is further explained by the tendency where the between variation is larger for areas than for parishes and within variation is smaller for areas than for parishes, as explored in Table 6 with count data (Figure 8).

**Table 5.** Migrant groups.

| | SS | df | MS | Intraclass correlation | Mean, SD | Min | Max |
|---|---|---|---|---|---|---|---|
| Migrant (A) | | | | | | | |
| Between area | 40,692.014 | 8042 | 5.0599371 | – | – | – | – |
| Within area | 403,411.35 | 4,316,512 | 0.09345772 | 0.08998 | 0.29 | 0 | 0.50 |
| Migrant (P) | | | | | | | |
| Between parish | 16,461.467 | 2174 | 7.5719721 | – | – | – | – |
| Within parish | 455,227.67 | 4,557,417 | 0.09988721 | 0.03448 | 0.31 | 0 | 0.50 |

SS: sum of squares; MS: mean of squares; SD: standard deviation.

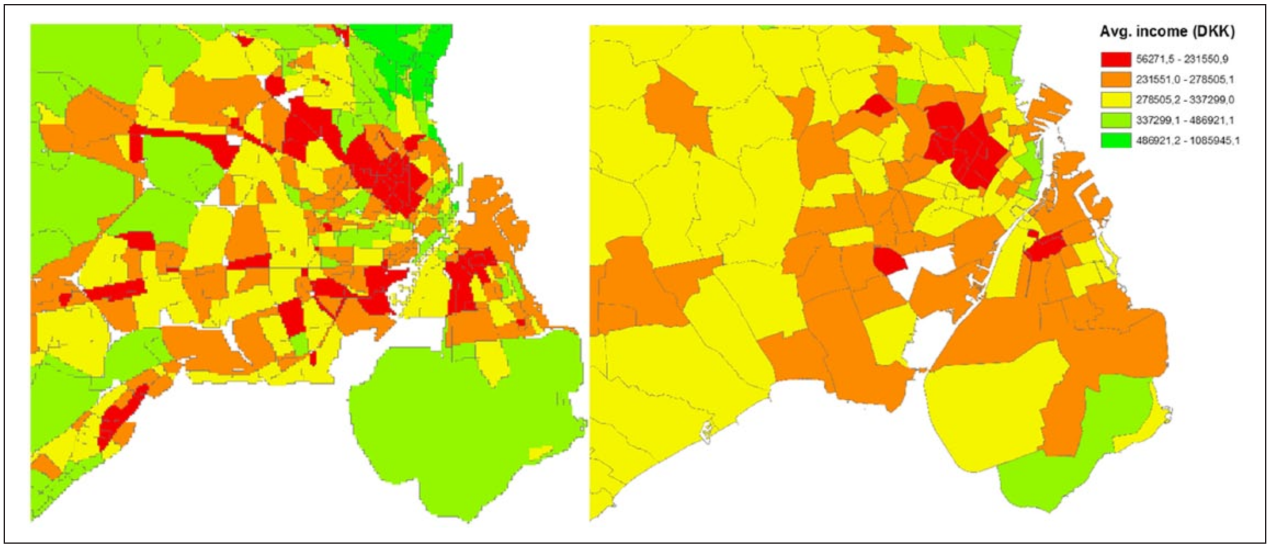



**Figure 7.** Smaller areas (left) and parishes (right) with average income.

**Table 6.** Education.

| | SS | df | MS | Intraclass correlation | Mean, SD | Min | Max |
|---|---|---|---|---|---|---|---|
| Income (A) | | | | | | | |
| Between area | 1.618e+16 | 8042 | 2.012e+12 | – | – | – | – |
| Within area | 3.383e+17 | 4,138,031 | 8.176e+10 | 0.04380 | 229,065.9 | 49,403 | 4,293,534 |
| Income (P) | | | | | | | |
| Between parish | 4.125e+15 | 2174 | 1.897e+12 | – | – | – | – |
| Within parish | 3.942e+17 | 4,230,119 | 9.411e+10 | 0.00986 | 265,932.6 | 125,765 | 6,419,540 |

SS: sum of squares; MS: mean of squares; SD: standard deviation.

One thing to note in the above table is the relatively low intraclass correlation. Even though it is 4.5 times larger for smaller areas than it is for parishes, it is still only 0.04. This could be explained by the fact that income is the measurement with the largest overall range of values, and that since it is a true ratio variable, it simply has too much variation to further a better correlation. This is supported using the ordinal variable used in the entropy measurement as replacement, which yields an intraclass correlation of 0.12 instead, but retains its relative difference of 4.5 from parishes.

## *Exploring scalars—education as perspective*

As described earlier, data smoothing could easily be responsible for most of the variation in homogeneity. Figure 9 introduces the moderations shown in Table 3 in simulated loops of 100 per type of moderation and ranks the runs from the best to the worst in terms of median entropy on education. To improve on readability, only five different distributions are shown for each moderations, 100 runs: the lowest median, the 25th percentile lowest median, the 50th percentile lowest

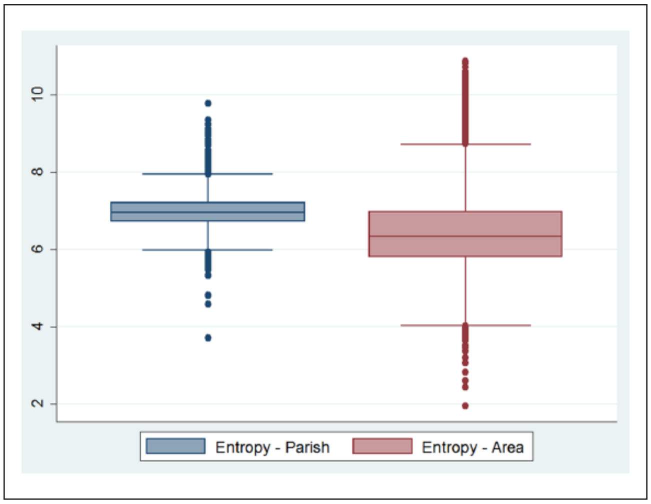



**Figure 8.** Entropy of income groups in parishes and new areas.

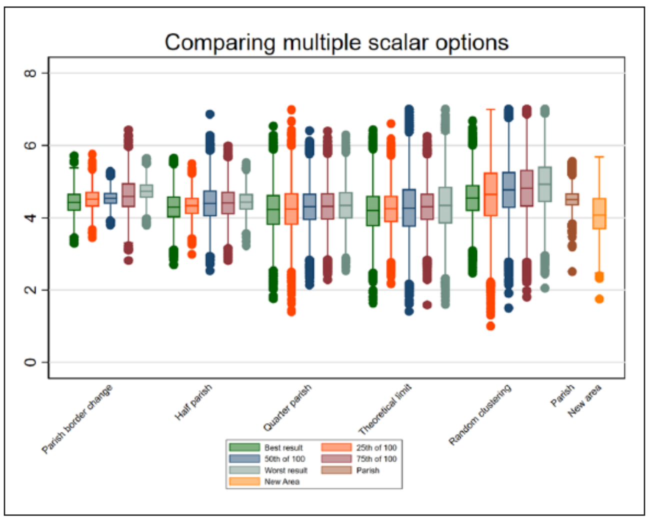



**Figure 9.** Loops of different moderation types from lowest to highest median in each moderation type.

median, the 75th lowest median, and the highest median entropy. As a reference, the original parish distribution and the new area distribution have been placed furthest to the right.

The above figure has a few interesting differences. First, it is worth noting that the new areas are more homogeneous in their inhabitant base compared to all other moderators even though there is some evidence of data smoothing. Considering the differences between half and quarter parishes compared to the theoretical limit division, there seems to be a limit to homogeneity purely based on reducing number of inhabitants. The difference between quarter parishes and the theoretical limit is basically non-existing even though the average inhabitant count has been reduced from 559 to 398 which is effectively a lower N than the proposed new areas. The best run of the theoretical limit moderation is closing in on the new areas, but it has a very logical drawback; the standard deviation in entropy is much higher. A simple explanation could be that non-barrier clustering is unable to take into account the housing prices and general neighborhood characteristics that could be factors in homogeneity and personal preferences in respect to housing.

Not surprisingly, the homogeneity as well as the standard deviation is by far the worst when performing the random clustering moderation. This is the smallest inhabitant average but it fails to account for both the physical and the local policies that could affect homogeneity.

What this implies is twofold; yes, size matters but the logic behind the scaling does as well. People do seem to adhere to some sort of logic when deciding where to live and that logic does not seem to only apply if we rescale to very small areas. Physical proximity does increase homogeneity but it does seem that this proximity is based on physical environment as well.

## Discussion

This article has shown that using other geographical divisions than administrative ones—even if they are relatively small—differs in the way we are able to perceive social and economic segregation and distribution. One discussion that is of utmost importance in this regard is, "Is this method better than many other methods designed to investigate non-administrative areas?"

This question is often not only the most pressing one but also the least interesting. How we define "better" changes in connection to what we want to understand and how we want to understand it. Most of the non-administrative areas are better at understanding local characteristics and inequality than administrative areas simply because they are smaller and therefore more likely to locate social enclaves. When it comes to the logic of non-administrative areas, the question to ask is no longer: "Are they better?" but instead "How are they different?" In this article, I propose a method to understand areas that differ from the commonly used methods and has both advantages and disadvantages. The main problem with this method is the border problem, where it becomes unclear whether people closer to the area border share increasingly more characteristics with people with adjacent areas. This is where especially K-nearest neighbors offer an advantage over the proposed method since the container presented here assumes that the area is uniform and that the border is the divider from one type of neighborhood to another. This could be considered not only a problem but also a strength in this method, since this hard division of neighborhoods allows for transferable and easy-to-understand area divisions. This is also necessary to investigate how streets

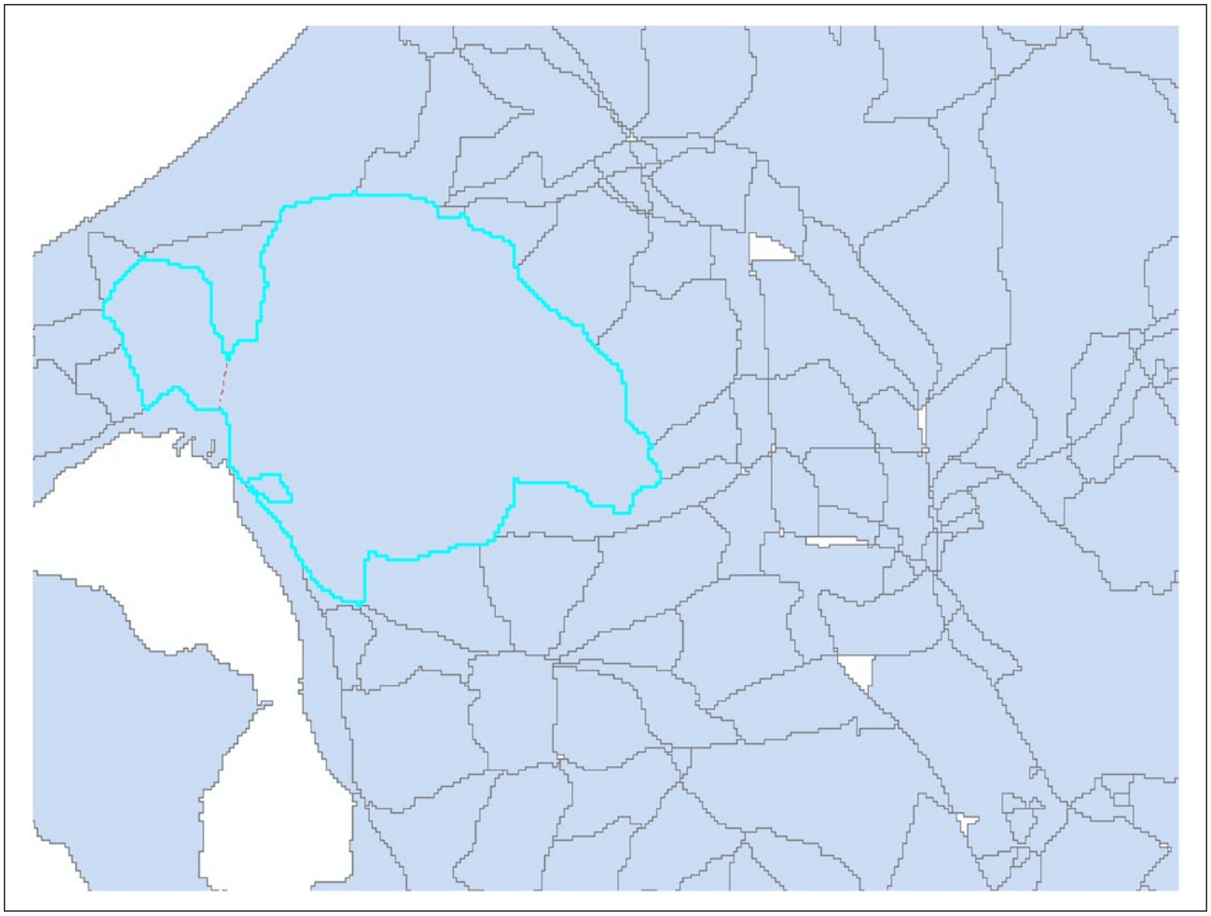


**Figure 10.** Extreme case of large area.

and natural barriers act as social barriers which are of less importance with fuzzy borders. One thing that would improve the method proposed in this article would be the ability to test how the borders function; if people change drastically at the physical border or if the change is graduate. The data limitations of Statistics Denmark render this impossible to test, but it would greatly improve the certainty of the border hypothesis. Nevertheless, this method is grounded in the logic behind settlement and how people inhabit areas and offer a much more logical way of redistricting than many other methods that rely solely either on geography or on social characteristics.

This problem arises with Bayesian methods as well. The a priori assumptions change the areas and require decisions made from the research to constantly take into account how the changes occur. Bayesian methods also require very specific knowledge and discussion of the a priori assumptions, which makes the method complicated and requires a new model for each research question. If research is to offer informed answers especially regarding policy and action-based decisions, a general model for segregation and area division is more applicable. The method proposed here can be used without fear of breaking data discretion requirements and can be easily adjusted in types of borders and number of inhabitants with the only a priori assumption being what borders to use and how large the clusters should be.

It is worth discussing the assumption that is the center of this method; areas can only be divided by physical barriers. In some cases, it would be logical that areas are too large to contain only one neighborhood, or one enclave of inhabitants would benefit from a division. Even though only less than 10% of the areas consist of more than 1000 inhabitants, it could perhaps solve the outlier problem when looking at the various entropies. This is, however, a discussion between logical perception and methodological purity. To what extent should the borders function as separators? In the case of the most extreme cases in Figure 9, which is one of the largest areas when considering both size, 84.6 km$^2$, and inhabitants, $N = 14{,}509$, one could argue that there might be something else than physical barriers to contain the social life. However, considering that the entropy of education in this area is 4.1, which is almost the median, it is difficult to pinpoint how to make this divide. Area size only correlates with educational entropy at 0.05, while the number of inhabitants correlates at 0.31. This indicates that most diversity measures would increase with number of people no matter the size of the place of interest. This, of course, is logical since the probability of a wider diversity increases with numbers, but it also complicates the logic of physical barriers in the case of heterogeneous areas (Figure 10).

Further adjustment of the overall algorithm could include a softer version of non-barrier divisions that consider area size, inhabitant count, and standard deviation in specific measurements and automatically divide at the areas' narrowest point. In the example above, this would only somewhat solve the problem, since this area would be divided where the red line is proposed.

## Conclusion

The literature on area effects and neighborhoods has long been focused on the effects first and the areas second. This article proposes a new method as an alternative to not only the administrative areas but also the non-administrative methods of geographical division if the main goal is to achieve homogeneity. The main point is to create areas that have a simple logic in their creation and offer a much better model to locate microsocial enclaves in a wide variety of social measurements thus focusing on homogeneity. The main problem with many other methods of automated redistricting is that the formation process is very complicated and requires either massive computational power or many deductive decisions before the formation. This method offers a high level of control over area formation and a highly logical interpretation of data assigned to the areas.

Comparing entropy, within/between variation and intraclass correlations between the areas proposed in this article compared to administrative parishes show not only a much higher homogeneity but also a better overall between variation. From a purely descriptive angle, the maps generated for educational attainment, ethnicity, and income reveal some very interesting subgroups of the population that would otherwise have been overlooked—when focusing on not only the deprived but also the wealthy areas.

One thing to consider is the application of this methodology; when comparing the proposed methodology to variations of zonation, it is evident that this method offers homogeneity above all else. This is often a main premise when trying to understand a neighborhood and how the inhabitants choose where to relocate but is, of course, only a smaller part of the complete neighborhood constructing literature. As mentioned, arguing which method is "better" should always be seen in context with the problem at hand. Comparing most non-administrative methods to administrative would usually result in both higher homogeneity and smaller units of measurement simply because of the size, but as shown, even though size matters, it doesn't encompass everything when aiming for homogeneity. Therefore, the discussion should center on usability and the goal of the models. This algorithm is designed to enhance usability and simplicity and at the same time securing small areas of high homogeneity.

### Declaration of conflicting interests

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

### Funding

The author(s) received no financial support for the research, authorship, and/or publication of this article.

### ORCID iD

Rolf Lyneborg Lund https://orcid.org/0000-0003-1013-8620

### References

Andersson E and Malmberg B (2013) *Contexual Effects on Educational Attainment in Individualized Neighborhoods: Differences Across Gender and Social Class* (SPaDE No. 10). Stockholm. Available at: http://www.diva-portal.org/smash/get/diva2:687024/FULLTEXT01.pdf

Åslund O and Skans ON (1985) Will I see you at work? Ethnic workplace segregation in Sweden. *ILR Review* 63(3): 471–493. Available at: http://www.jstor.org/stable/40649713

Barone C (2011) Some things never change: Gender segregation in higher education across eight nations and three decades. *Sociology of Education* 84(2): 157–176.

Bower KM, Thorpe RJ, Rohde C, et al. (2014) The intersection of neighborhood racial segregation, poverty, and urbanicity and its impact on food store availability in the United States. *Preventive Medicine* 58(1): 33–39.

Breetzke GD and Horn AC (2006) Crossing the racial divide: A spatial-ecological perspective of offenders in the City of Tshwane Metropolitan Municipality, South Africa. *GeoJournal* 67(3): 181–194.

Brydsten A, Gustafsson PE, Hammarström A, et al. (2017) Does contextual unemployment matter for health status across the life course? A longitudinal multilevel study exploring the link between neighbourhood unemployment and functional somatic symptoms. *Health and Place* 43: 113–120.

Buck N (2001) Identifying neighbourhood effects on social exclusion. *Urban Studies* 38(12): 2251–2275.

Buttimer A and Seamon D (1980) *The Human Experience of Space and Place*. New York: Taylor & Francis. Available at: https://books.google.dk/books?id=xRHICQAAQBAJ&dq=Homogeneity+in+Residential+Areas&lr=&source=gbs_navlinks_s

Carlsson AC, Li X, Holzmann MJ, et al. (2017) Neighborhood socioeconomic status at the age of 40 years and ischemic stroke before the age of 50 years: A nationwide cohort study from Sweden. *International Journal of Stroke* 12: 815–826.

Charles M and Bradley K (2009) Indulging our gendered selves? Sex segregation by field of study in 44 countries. *American Journal of Sociology* 114(4): 924–976. Available at: http://www.ncbi.nlm.nih.gov/pubmed/19824299

Charles M and Grusky D (1995) Models for describing the underlying structure of sex segregation. *American Journal of Sociology* 100(4): 931–971. Available at: http://www.jstor.org/stable/10.2307/2782156

Clark WAV (2017) Who moves into what kinds of neighbourhoods: Spatial sorting and integration. *Tijdschrift voor Economische en Sociale Geografie*. Epub ahead of print 4 April. DOI: 10.1111/tesg.12264.

Clifford N, French S and Valentine G (2010) *Key Methods in Geography: Geographical Research*. London: SAGE.

Cunha JMP, Da Jimenez MA, Perez JRR, et al. (2009) Social segregation and academic achievement in state-run elementary schools in the municipality of Campinas, Brazil. *Geoforum* 40(5): 873–883.

Damm AP and Schultz-Nielsen ML (2008) *The construction of neighbourhoods and its relevance for the measurement of social and ethnic segregation: Evidence from Denmark. discussion paper series*. 6 October 2008, Working Paper No. 08-17. Denmark: Department of Economics.

Dawkins CJ (2006) Are social networks the ties that bind families to neighborhoods? *Housing Studies* 21(6): 867–881.

Deng Y (2016) Challenges and complications in neighborhood mapping: From neighborhood concept to operationalization. *Journal of Geographical Systems* 18(3): 229–248.

DeSilva S, Pham A and Smith M (2012) Racial and ethnic price differentials in a small urban housing market. *Housing Policy Debate* 22(2): 241–269. Available at: https://search.proquest.com/docview/1020094132?accountid=8144

Feld SL (1981) The focused organization of social ties. *The American Journal of Sociology* 86(5): 1015–1035.

Fiscella K and Fremont AM (2006) Use of geocoding and surname analysis to estimate race and ethnicity. *Health Services Research* 41: 1480–1500.

Fischer CS, Stockmayer G, Stiles J, et al. (2004) Distinguishing the geographic levels and social dimensions of U.S. metropolitan segregation, 1960–2000. *Demography* 41(1): 37–59.

Fone D, Dunstan F, Lloyd K, et al. (2007) Does social cohesion modify the association between area income deprivation and mental health? A multilevel analysis. *International Journal of Epidemiology* 36(2): 338–345.

Freisthler B, Thomas CA, Curry SR, et al. (2016) An alternative to residential neighborhoods: An exploratory study of how activity spaces and perception of neighborhood social processes relate to maladaptive parenting. *Child and Youth Care Forum* 45(2): 259–277.

Galster GC (1989) Residential segregation in American cities: A further response to Clark. *Population Research and Policy Review* 8(2): 181–192. Available at: https://search.proquest.com/docview/60049493?accountid=8144

Galster GC (2010) The mechanism(s) of neighborhood effects theory, evidence, and policy implications. In: *ESRC Seminar*, St. Andrews University, Scotland, UK, 4–5 February 2010, pp. 1–32.

Gans HJ (1961) The balanced community: Homogeneity or heterogeneity in residential areas? *Journal of the American Planning Association* 27(3): 176–184.

Grannis R (1998) The importance of trivial streets: Residential streets and residential segregation. *American Journal of Sociology* 103(6): 1530–1564.

Hwang S (2015) Residential segregation, housing submarkets, and spatial analysis: St. Louis and Cincinnati as a case study. *Housing Policy Debate* 25(1): 91–115. Available at: https://search.proquest.com/docview/1648350535?accountid=8144

Johnelle Sparks P, Sparks CS and Campbell JJA (2013) An application of Bayesian spatial statistical methods to the study of racial and poverty segregation and infant mortality rates in the US. *GeoJournal* 78(2): 389–405.

Johnson JH, Parnell A, Joyner AM, et al. (2004) Racial apartheid in a small North Carolina Town. *Review of Black Political Economy* 31(4): 89–107.

Johnston R, Propper C, Burgess S, et al. (2005) Spatial scale and the neighbourhood effect: Multinomial models of voting at two recent British general elections. *British Journal of Political Science* 35: 487–514.

Jones M and Huh J (2014) Toward a multidimensional understanding of residential neighborhood: A latent profile analysis of Los Angeles neighborhoods and longitudinal adult excess weight. *Health and Place* 27: 134–141.

Jones M and Pebley AR (2014) Redefining neighborhoods using common destinations: Social characteristics of activity spaces and home census tracts compared. *Demography* 51(3): 727–752.

Jost L (2006) Entropy and Diversity. *Oikos* 113: 363–375.

King G, Keohane R and Verba S (1994) *Designing Social Inquiry—Scientific Inference in Qualitative Research*. Princeton, NJ: Princeton University Press. Available at: http://books.google.com/books?hl=en&lr=&id=A7VFF-JR3b8C&oi=fnd&pg=PA248&dq=Designing+Social+Inquiry+-+Scientific+Inference+in+Qualitative+Research&ots=f0iuT7CV5R&sig=zEF7x-p76Qmfs6lQ46CvnvS72Y0

Krieger N, Feldman JM, Waterman PD, et al. (2017a) Local residential segregation matters: Stronger association of census tract compared to conventional city-level measures with fatal and non-fatal assaults (total and firearm related), using the index of concentration at the extremes (ICE) for racial, econ. *Journal of Urban Health* 94(2): 244–258.

Krieger N, Waterman PD, Batra N, et al. (2017b) Measures of local segregation for monitoring health inequities by local health departments. *American Journal of Public Health* 107: 903–906.

Kryvobokov M (2013) Hedonic price model: Defining neighbourhoods with Thiessen polygons. *International Journal of Housing Markets and Analysis* 6(1): 79–97.

Kwan M-P (2013) Beyond space (as we knew it): Toward temporally integrated geographies of segregation, health, and accessibility. *Annals of the Association of American Geographers* 103(5): 1078–1086.

Lagerlund M, Merlo J, Vicente RP, et al. (2015) Does the neighborhood area of residence influence non-attendance in an urban mammography screening program? A multilevel study in a Swedish city. *PLoS ONE* 10(10): e0140244.

Law J, Quick M and Chan PW (2015) Analyzing hotspots of crime using a Bayesian Spatiotemporal modeling approach: A case study of violent crime in the greater Toronto area. *Geographical Analysis* 47(1): 1–19.

Lee BA and Campbell KE (1997) Common ground? Urban neighborhoods as survey respondents see them. *Social Science Quarterly* 78(4): 922–936. Available at: http://www.jstor.org/stable/42863740

Lee BA, Reardon SF, Firebaugh G, et al. (2008) Beyond the census tract: Patterns and determinants of racial segregation at multiple geographic scales. *American Sociological Review* 73(5): 766–791.

Logan JR, Spielman S, Xu H, et al. (2011) Identifying and bounding ethnic neighborhoods. *Urban Geography* 32(3): 334–359.

Macallister I, Johnston RJ, Pattie CJ, et al. (2001) Class dealignment and the neighbourhood effect: Miller revisited. *British Journal of Political Science* 31(1): 41–59.

Massey DS and Denton NA (1988) The dimensions of residential segregation. *Social Forces* 67(2): 281–315.

McIntosh RP (1986) *The Background of Ecology: Concept and Theory*. Cambridge: Cambridge University Press.

Merlo J, Ohlsson H, Chaix B, et al. (2013) Revisiting causal neighborhood effects on individual ischemic heart disease risk: A quasi-experimental multilevel analysis among Swedish siblings. *Social Science and Medicine* 76(1): 39–46.

Ministry for City, Habitation and Rural Districts (2014). *Segregering af de fire største danske byområder*, 17 March, 2014.

Newton PW and Johnston RJ (1976) Residential area characteristics and residential area homogeneity: Further thoughts on extensions to the factorial ecology method. *Environment and Planning A: Economy and Space* 8(5): 543–552.

Östh J, Clark WAV and Malmberg B (2015) Measuring the scale of segregation using k-nearest neighbor aggregates. *Geographical Analysis* 47(1): 34–49.

Östh J, Malmberg B and Andersson EK (2014) Analysing segregation using individualised neighbourhoods. In: Lloyd CD, Shuttleworth I and Wong DW (eds) *Social-Spatial Segregation: Concepts, Processes and Outcomes*. Bristol: Policy Press, pp. 135–162. Available at: https://books.google.dk/books?hl=en&lr=&id=XTNlBAAAQBAJ&oi=fnd&pg=PA135&dq=%22small+area+market+statistics%22+sweden&ots=82wkE3dL-j&sig=G4hkSj-xg5z2wnqKJWxPBwFZG-A&redir_esc=y#v=onepage&q=%22smallareamarketstatistics%22 sweden&f=false

Park RE (1928) Human migration and the marginal man. *American Journal of Sociology* 33(6): 881–893.

Park RE and Burgess EW (2007) *Introduction to the Science of Sociology*. Chicago, IL: University of Chicago Press.

Park RE, Burgess EW and McKenzie R (1967) *The City* (ed. M Janowitz). Chicago, IL: The University of Chicago Press.

Patterson Z and Farber S (2015) Potential path areas and activity spaces in application: A review. *Transport Reviews* 35(6): 679–700.

Pattison P and Robins G (2002) Neighborhood-based models for social networks. *Sociological Methodology* 32: 301–337. Available at: https://search.proquest.com/docview/60445053?accountid=8144

Pickett KE and Pearl M (2001) Multilevel analyses of neighbourhood socioeconomic context and health outcomes: A critical review. *Journal of Epidemiology and Community Health* 55(2): 111–122.

Propper C, Burgess S, Bolster A, et al. (2007) The impact of neighbourhood on the income and mental health of British social renters. *Source: Urban Studies* 44(2): 393–415. Available at: https://www.jstor.org/stable/pdf/43084456.pdf?refreqid=excelsior%3Ad6d56d350d0da6c03b4f5c0f0b1d7ccc

Rao JNK and Thomas DR (1988) The analysis of cross-classified categorical data from complex sample surveys. *Sociological Methodology* 18: 213–269. Available at: https://search.proquest.com/docview/61024092?accountid=8144

Ricotta C and Szeidl L (2006) Towards a unifying approach to diversity measures: Bridging the gap between the Shannon entropy and Rao's quadratic index. *Theoretical Population Biology* 70: 237–243.

Samardzic-Petrovic M, Dragicevic S, Kovacevic M, et al. (2016) Modeling urban land use changes using support vector machines. *Transactions in GIS* 20(5): 718–734.

Sampson RJ (2008) Moving to inequality: Neighborhood effects and experiments meet structure. *American Journal of Sociology* 114(11): 189–231.

Sampson RJ (2012) *Great American City-Chicago and the Enduring Neighborhood Effect*. Chicago, IL: The University of Chicago Press.

Sampson RJ, Morenoff JD and Gannon-Rowley T (2002) Assessing "neighborhood effects": Social processes and new directions in research. *Annual Review of Sociology* 28(1): 443–478.

Söderström M and Uusitalo R (2010) School choice and segregation: Evidence from an admission reform. *Scandinavian Journal of Economics* 112(1): 55–76.

Sundquist J, Sjöstedt C, Winkleby M, et al. (2016) Neighborhood linking social capital as a predictor of drug abuse: A Swedish national cohort study. *Addictive Behaviors* 63: 37–44.

Taylor RB (1997) Social order and disorder of street blocks and neighborhoods: Ecology, microecology, and the systemic model of social disorganization. *Journal of Research in Crime and Delinquency* 34(1): 113–155.

Veenstra G, Luginaah I, Wakefield S, et al. (2005) Who you know, where you live: Social capital, neighbourhood and health. *Social Science and Medicine* 60(12): 2799–2818.

Veldhuizen EM, Stronks K and Kunst AE (2013) Assessing associations between socio-economic environment and self-reported health in Amsterdam using bespoke environments. *PLoS ONE* 8(7): e68790.

Vinikoor LC, Kaufman JS, MacLehose RF, et al. (2008) Effects of racial density and income incongruity on pregnancy outcomes in less segregated communities. *Social Science and Medicine* 66(2): 255–259.

Zingher JN and Thomas MS (2014) The spatial and demographic determinants of racial threat. *Social Science Quarterly* 95(4): 1137–1154.

## Author biography

Rolf Lyneborg Lund is a Ph.D. fellow at the institute for sociology and social work at Aalborg University, Denmark. His main research areas are social inequality in a geographical perspective and computational modelling of geography with register data.

Scandinavian Journal of Educational Research

ISSN: 0031-3831 (Print) 1470-1170 (Online) Journal homepage: https://www.tandfonline.com/loi/csje20

# Moving to Prosperity? The Effect of Prolonged Exposure to Neighborhood Deprivation

Rolf Lyneborg Lund

SCANDINAVIAN JOURNAL OF EDUCATIONAL RESEARCH
https://doi.org/10.1080/00313831.2019.1577753

Routledge
Taylor & Francis Group

Check for updates

# Moving to Prosperity? The Effect of Prolonged Exposure to Neighborhood Deprivation

Rolf Lyneborg Lund

Department of Sociology and Social Work, Aalborg University, Aalborg, Denmark

**ABSTRACT**
Place of living has been associated with a variety of effects but is often considered stationary. Newer research reveals that the accumulation of deprivation conveys many of the effects that were initially thought to be captured by place of birth or current place of living; however, the view of accumulation as a static entity implies that only the length of residency matters. This study uses registry data to follow a cohort ($N$ = 256,345) from birth to age 30 years. It investigates the effects of prolonged exposure to deprived neighborhoods on educational attainment and examines whether these effects are the same for those who accumulate exposure at different times. The study finds that exposure is important for educational attainment but that the effect differs at different life stages.

**ARTICLE HISTORY**
Received 27 August 2018
Accepted 17 January 2019

**KEYWORDS**
Neighborhood effects; deprivation; educational mobility; inverse probability regression adjustment

## 1. Introduction

Research on stratification and sociogeographical deprivation and their effects on later life opportunities is a core topic of sociological theory, and many empirical studies have addressed this issue (Atkinson & Kintrea, 2004; Fone et al., 2007; Hannum, 2005; Sundquist et al., 2016; Wilkinson & Pickett, 2007). Europe and Scandinavia have experienced an increase in concern regarding those who live in highly deprived areas compared to the rest of the population (Breen, Luijkx, Müller, & Pollak, 2009; Hjellbrekke & Korsnes, 2004; Munk & Thomsen, 2018; Riise, Dommermuth, & Lyngstad, 2016; Sharkey, 2013). When political concern revolves around the effects of stratification and deprivation on crime and social problems, efforts are mainly directed toward strengthening educational attainment. For this reason, two of the most extensively studied outcomes in neighborhood effect research are educational attainment and employment.

Abundant empirical research has focused on either the specific effects of living in a specific area and how these affect an individual in the present (Cattell, 2001; Ochieng, 2011) or on being born into deprivation (Carlsson et al., 2017; Ochieng, 2011). The empirical results of neighborhood studies are often mixed; educational attainment in later life is only somewhat linked to the type of neighborhood and is predominantly explained by other social factors (Leventhal & Brooks-Gunn, 2000; Sampson, Morenoff, & Gannon-Rowley, 2002; Sampson & Sharkey, 2008).

Area effects are very difficult to measure due to three overall factors: direct effects, selection and districting. Direct effects involve disentangling the interfamilial, school and general effects at the individual level from the effects of factors associated with the area of residence. This issue is often resolved by Simple statistical control for confounders is often used to resolve this issue, but it cannot address the selection problem because of the nonrandomness of residence. The selection problem is often considered a "placement problem": is placement into the group or the area truly random,

**CONTACT** Rolf Lyneborg Lund rolfll@socsci.aau.dk

which would allow the true effect of the group/area to be isolated, or are individuals selected into these entities? If selection into an area is nonrandom, we cannot disentangle the direct effects that caused an individual to be selected into the area from the effects of area composition in itself (Heckman, 1977; Wooldridge, 1995). In the case of accumulated effects, for which the time an individual lived in specific area types is relevant, selection shifts from being nonrandom only in terms of where one lives to being nonrandom in the sense that where an individual lives in his/her earlier years strongly predict where he or she will live in the future (Hedman, Manley, Van Ham, & ÖSth, n.d.; Næss, Leyland, Smith, & Claussen, 2005; Sampson et al., 2002; Wiborg & Hansen, 2018; Wodtke, Harding, & Elwert, 2011).

The last element is (re)districting. There seems to be a divide in the literature between studies that examine very intricate distinctions between "true" area effects and non-area-based effects (Atkinson & Kintrea, 2001; Bischoff, 2008; Bower, Thorpe, Rohde, & Gaskin, 2014; Gilliland, Olson, & Gauvreau, 2011; McDevitt et al., 1986; Shahzadi, Riaz, Anwar, & Nasreen, 2017) and studies that focus on understanding how much aggregation at the geographical level matters when measuring deprivation (Ferreira, Holan, & Bertolde, 2011; Law, Quick, & Chan, 2015; Malmberg, Andersson, & Östh, 2011; Martinez, 2009; Östh, Clark, & Malmberg, 2015; Östh, Malmberg, & Andersson, 2014; Petrović, Ham, & Manley, 2017; Sampson, 2012; Taylor, Gorard, & Fitz, 2003; Tayman, Schafer, & Carter, 1998). While the first group predominantly uses the smallest available administrative area, i.e., municipalities, parishes or census tracts, as the unit of measurement, the second group uses a multitude of methods to create geographic units that best measure deprivation. This approach leads to the overall problem that when methods that take into account direct effects and the selection problem are used, geography is often overlooked. Additionally, when studies consider geography as the main factor, it is rarely used to locate specific area effects and is used instead to highlight problematic areas and momentary deprivation.

In this paper, I will address not only the methodological problems related to direct effects, selection and districting but also the theoretical problems associated with life course and the accumulation of deprivation, in which an individual's actual life trajectory has a much larger role in later life educational attainment.

## 2. Deprivation and Life Course

The literature on area effects differs widely in scope, methodology and results. As noted above, area deprivation involves a great number of issues, and the existing body of work defines these issues in various ways.

The effects often seem to be reduced to the habitual storage of capital in early years, and most research reduces this predictor to the birth area (Albanese, De Blasio, & Sestito, 2016; Gorard & Siddiqui, 2016; Kelo, 2010; Riise et al., 2016; Warrington, 2005). This reduction to birth area holds some merit since one of the problems in the life course literature is the fact that an individual's early area of residence has a strong correlation with later residence. Methodologically, although focusing on the area of residence during the formative years or at birth theoretically resolves the issue of autocorrelation, this solution is lacking in key aspects. Reducing the effect of area deprivation to smaller sections of a life course and then relying on autocorrelation to "fill in the blanks" between birth, youth, adolescence and then adulthood approximates a life course, but it imposes a very deterministic trajectory. It could be argued that this deterministic trajectory does not take into account the relationship between individuals' accumulated capital and their mobility between different areas or how these factors are constantly evolving and changing.

Research on capital and social mobility often centers around the transmission and accumulation of capital among individuals, generations and areas (Bourdieu, 1986; Gesthuizen, Van Der Meer, & Scheepers, 2008; Moule, Decker, & Pyrooz, 2013; van de Werfhorst & Hofstede, 2007). According to Bourdieu, all aspects of capital involve lifelong accumulation; thus, changes to the environment are expected to have an influence on further capital development (Bourdieu, 1998; Bourdieu & Passeron, 1990). Changes in capital accumulation do not nullify earlier accumulation, but they modify our understanding of the world and change how we make life choices. Considering capital as something that is accumulated implies some sort of determinism; being born into deprivation instills a specific set of capital that lingers throughout one's life, but since capital is accumulated, it seems logical that changes in its composition – as a result of moving, parents' earnings, better education, higher household income, and so on – would result in changes later in life.

The discussion above describes the main reason this paper distinguishes between *accumulation* and *life course*. While accumulation implies that exposure is equal no matter when it happens, the life course approach includes an interaction between accumulation and time; the total exposure is one factor, but when the exposure is applied is another.

Several studies examine the accumulation of deprivation (Crowder & South, 2011; Jackson & Mare, 2007; Kunz, Page, & Solon, 2003; Lindstrom & Massey, 1994; Sampson & Sharkey, 2008; Sharkey, Schwartz, Ellen, & Lacoe, 2014; South & Crowder, 2010), but as noted by Wodtke et al. (2011), most of these studies fail to account for the dynamic selection problem when measuring the effect of neighborhood deprivation. Some studies rely solely on least squares regression when investigating accumulated deprivation (Jackson & Mare, 2007; Pais, South, & Crowder, 2012), while others apply various modifications to standard regression analysis (Lindstrom & Massey, 1994; Sampson & Sharkey, 2008). However, as Wodtke et al. (2011) suggest, these studies either underestimate or overestimate the true effects of neighborhood when dynamic neighborhood selection is not considered.

Wodtke et al. investigate long-term exposure to concentrated disadvantage and measure how accumulated deprivation affects the chances of attaining a high school degree in a counterfactual setting by using inverse probability weights in a treatment effects setup (Wodtke et al., 2011). This approach allows for a quasi-experimental setup, in which the effect of living in one type of neighborhood is compared to not living there, while selection into neighborhoods is weighted to accommodate unequal correlations between social factors and place of living in a longitudinal sense. By utilizing these doubly robust estimators (see Wooldridge, 2010) in which the primary effect and the selection model represent one method of moments estimator, selection into treatment is addressed. The above study includes tests for when during the life course the deprivation occurs, but it notes that accumulation is by far the most important factor.

What none of the above studies considers is the actual neighborhood: what parameters do we apply to create neighborhoods?

## 3. Conceptual Measurements of Neighborhoods

Even with advancements in GIS and georeferenced data, many studies still use administrative borders (parishes or municipalities) as their smallest unit of reference when investigating area effects (Andersson & Malmberg, 2013; Åslund & Skans, 2010; Cunha, Jimenez, Perez, & Andrade, 2009; Fischer, Stockmayer, Stiles, & Hout, 2004; Söderström & Uusitalo, 2010; Zingher & Steen Thomas, 2014). Even when utilizing smaller areas, such as census tracts in the United States, the usefulness or validity of the areas is rarely questioned (Bower et al., 2014; Krieger, Feldman, et al., 2017; Krieger, Waterman, et al., 2017; Wodtke et al., 2011). As Lee et al. (2008) note, "Most studies implicitly assume that the tract constitutes an appropriate sized spatial unit for capturing segregation". This notion raises some fundamental questions regarding the understanding of place and living: How do we know that the areas we use to contain the social aspects of its inhabitants actually make sense? What is the effect of using only predetermined administrative areas rather than exploring the possibilities of GIS-coded data?

In the case of deprivation, it seems logical that administrative areas help create *a* box, but it may not be the most suitable box if the goal is to isolate deprivation. Selection problems are real and by far one of the most important factors when trying to understand the place of living, but if the question of "selection into what?" is left unaddressed, the selection process will be flawed.

Social behavior is rarely random, and neither is the way we interact and settle. Drawing from the Chicago school of ecological sociology, Sampson (2012) suggests that neighborhoods are more than just areas – they are the collective efficacy of a community. The ways we decide where to live, how to act and how to differentiate among types of capital are rooted in geography, not always in a community sense, but in the sense of belonging (Sampson, 2012). This notion implies the influence of local-level factors and also suggests that social efficacy does not recognize administrative borders. Microclasses, microenclaves and subgroups can easily emerge in otherwise deprived areas, and vice versa. To fully address the deprivation theory, I suggest that deprivation must be the sum of the individuals who have settled in local areas. In this case, this approach implies that homogeneity and, to some extent, concentration play a role. The reason that homogeneity within a confined physical space is extraordinarily important in the case of deprivation, and especially in the case of inverse probability regression adjustment, is that if the treatments are bound by high deviation, treatments that are not composed of the truly disadvantaged and deprived areas one wishes to investigate will result. By utilizing geocoded registry data, it is possible to create more homogenous neighborhoods and to define neighborhoods at the individual level to accommodate the research question, which I address below.

## 4. Methodology

### 4.1. Data

This article utilizes two different types of data: georeferenced data and registries of the Danish population. The first type of data, the georeferenced data, consists of The National Square Grid and a large collection of topographical vector-based object maps that contain roads, streams, lakes, forests and most other place-specific objects found in Denmark. The National Square Grid is a national system of vector grids constructed by the Danish Geodata Agency and Statistics Denmark that measure $100 \times 100$ meters and have unique identifications and spatial references. When considering redistricting, it is very valuable to have access to the smallest units of measurement possible, and the ability to modulate areas in cells that are no larger than $100 \times 100$ meters is ideal for clustering. One could argue that the ideal approach would be to maintain the smallest unit of measurement and not cluster the square grid in any way. However, the use of square grids is impossible because Statistics Denmark operates with very strict confidentiality requirements that require at least 100 individuals per geographical unit before allowing access, and in 2017, less than 1% of the squares were inhabited by more than 100 households. The specific clustering method is presented later.

The other set of data consists of registry data for two different population groups. The first group is the full population, measured yearly from 1980 to 2016, while the second group is a cohort born between 1980 and 1985 and measured yearly between 1980 and 2016. The cohort is the group in which the effect is investigated, while the population sample is used to add background information to the cohort and to calculate deprivation measures within the aforementioned areas. The registries compile individual-level information about education, including full months of total education, including primary school; income, measured as the gross income per year; age; gender; and ethnicity. All these variables are linked among generations and thus also to the cohort. All interval-level data are utilized in the mapping but are categorized into ordinal measures for the entropy measurement. Furthermore, the data consist of other geographical information, such as parish and municipality. All of these data are linked to the square grid after the clustering is performed.

### 4.2. Treatment

#### 4.2.1. Neighborhoods

As described earlier, neighborhoods can easily become simple containers with a heterogeneous inhabitant base. Thus, the main problems in neighborhood studies are as follows: if the administrative areas are too heterogeneous to encompass the social group one wishes to measure, is it even possible to describe the area as a neighborhood? In what areas do we interact with one another, and how do we define the social barriers that create the feeling of "us" and "them"? Many studies have pointed to cohesion within areas and have tried to define what makes a neighborhood (Damm & Schultz-nielsen, 2008; Deng, 2016; Freisthler, Thomas, Curry, & Wolf, 2016; King, Keohane, & Verba, 1994; Östh et al., 2015; Petrović et al., 2017). Feld (1981) points out that even though we live in specific areas, these areas are often divided by specific physical barriers, such as roads, railways and other objects commonly found in both the urban and rural landscape. Thus, by itself, the square grid is as illogical as other administrative area divisions.

Some newer methods of clustering neighborhoods rely heavily on k-means clustering to inductively generate neighborhoods (Andersson & Malmberg, 2013; Dewilde, 2004; Östh et al., 2015; Petrović et al., 2017; Petrović, Ham, & Manley, 2018), and their use has increased since the development of the software Equipop (Östh, 2018). The main advantage of this method is the level of detail at which the data can be processed since the analysis progresses from individual-level data to neighborhood data. The main problem with this methodology is the discretion criteria applied to the data to be analyzed. In most countries where registry data are available, there are some requirements when analyzing individual-level data geographically, and in the case of Denmark, one of the main criteria is that one cannot access individual-level data geographically unless at least 100 people are located in each area. This means that the completely inductive use of k-nearest neighbors is impossible.

With this consideration in mind, I use the square grid to perform clustering to maximize homogeneity in smaller areas. To meet Statistics Denmark's requirement that each geographical area includes at least 100 inhabitants, I use physical barriers[1] to generate the first run of 100 inhabitants, and then I cluster again based on the deprivation measurement described in the next section, with a very important adjustment: clustering occurs only if areas share borders. The two steps are separated in sequence but are recursive in nature. The steps used to arrange the geography into areas that include at least 100 inhabitants to allow for socioeconomic clustering are sparse but very strict. The arguments are as follows: areas must be contained by physical barriers, areas must not be separated by other polygons, areas with the largest percentage of borders are considered first, and the method resulting in the fewest merges is preferred. Furthermore, to meet the goal of obtaining the smallest possible areas with at least 100 inhabitants, the program runs iterations and tests out all possible combinations of areas until all areas are as small as possible include at least 100 inhabitants (Lund, 2018).

The last aspect of the clustering method utilizes the above results when clustering on social factors. Since this is the point at which scale becomes important, steps are taken to keep the areas comparable. For this step, Ward's method (Ward, 1963) is combined with a Lance-Williams recursive algorithm (Lance & Williams, 1967) in which internal homogeneity is preferred to calculate the difference between areas. The recursive argument regarding least possible merges are retained, but instead of securing areas as close to 100 inhabitants as possible, it is changed to secure mean inhabitant counts and ranges that are as equal as possible.

This approach generates areas that are locally anchored and as socioeconomic homogenous as possible while ensuring that the method is reproducible and yields the same areas each time. By using this method, I maximize internal homogeneity and ensure that the area changes only when the deprivation measurement is broken by a barrier. This method results in 7,600 areas with an average of 589 inhabitants per area, a range from 248 to 983 inhabitants and a standard deviation of 212.37. Compared to more commonly used administrative measurements (such as parishes ($N$ = 2,443), which from 25 to 45,187 range in Danish inhabitants with a standard deviation of 3718.69), this method yields smaller, more homogeneous areas with much more comparable inhabitant counts (Lund, 2018).

[1]Roads larger than 5 m across, streams, lakes, green areas larger than 100 $m^2$, fences, larger walls, railroads.

### 4.2.2. Measuring Deprivation

To divide the areas into useful units of deprivation, I use principal component analysis to capture contextual deprivation. Contextual deprivation is measured as a composite index consisting of 5 different yearly measures: area mean income as a percentage compared to mean income in the overall region,[2] percentage of unemployed individuals of working age (18–64 years), the percentage of the population older than 25 years of age with primary education as the highest attained education level, the percentage of welfare recipients and percentage of the population in managerial positions. This method is based on the deprivation measures used by the Danish government to form the ghetto list[3] (Ministry of Transportation, 2018), the British research definition of the Scale of Multiple Deprivation (Payne, 2012), and several studies that aim to conceptualize area deprivation (Atkinson & Kintrea, 2004; Potter, Walker, & Keen, 2012; Sharkey & Elwert, 2011; Wodtke et al., 2011). This index is divided into deciles based on regional levels, thus creating a time-variable measure that changes as the inhabitants change. The first decile contains the least deprived areas, whereas the 10th decile contains the most deprived areas.

To compare, in the most deprived areas, 65% of inhabitants are unemployed, 76% have only a primary education, and 23% receive welfare; additionally, areas of this type have a 30% lower income than the regional mean. In contrast, in the least deprived areas, 7% of the inhabitants have only a primary education, 2% of the population is unemployed, 1.3% of the inhabitants receive welfare, and the mean income is 200% above the regional average.

### 4.2.3. Life Course and Life Trajectories

As Wodtke et al. (2011) show, cumulative exposure to deprivation during childhood has a direct effect on the probability of graduating from high school. This finding points to the fact that time of exposure is irrelevant compared to accumulated time, but it reduces the life course to either moving to or from deprivation and not as a flow of events with different meanings. Life course events can also involve a flow of movement within either the lower or the higher spectrum of neighborhood deprivation and can have an effect on later personal outcomes. To further investigate life course events compared to accumulated deprivation, a sequence analysis is applied to examine two questions regarding the data that could influence the results. First, it is important to understand how stable a life course is with regard to deprivation; general movement patterns, but horizontal movement within the deprivation deciles is unknown. Second, this method determines whether any other type of movement is common; in addition to examining movement from worse to better, and vice versa, it investigates whether moving back and forth between deciles could impact later life outcomes.

The usefulness of sequence analysis is apparent, but the application of this method is far more problematic. The full cohort consists of 358,899 individuals over a span of 30 years, which results in a distance matrix with a size of $3.86^{e+12}$ for a single measure. To accommodate this complexity, the full cohort is split into cohorts by birth year and run individually. As a result, six different distance matrixes are created, and the subsequent clustering is performed individually and then combined into specific types of life sequences.

The sequencing itself is based on yearly data and measures the decile of deprivation each individual inhabits in microareas of his or her life. The full sequence involves living in 1 of 10 types per year over 30 years. The overall form of the data is fairly even, and full data exist for 90% of the population for all 30 years. The 10% without full data include the 2% percent of the cohort that exited due to untimely death and the 8% of people immigrating to or emigrating from Denmark. These last 8% are analyzed, and if more than 50% of the residential data are missing or the outcome year when

[2]Denmark consists of five overall regions with separate taxes and income distributions. By using regions as a reference for the areas they contain, I counteract the fact that less-deprived areas can easily become classified as deprived in a national setting without actually being so.

[3]Each year, the Danish Ministry of Transport publishes a list of areas that fall below a specific point on the scale. This study does not apply the same cutoff points but separates the data into deciles.

the individual was 30 years old is missing, the respondent is dropped from the analysis. If less than 50% of the residential data are missing and the outcome year is present, the respondent receives the first decile value recorded retrospectively. This approach results in a cohort sample of 310,505 respondents.

Regarding distance, research has debated the use of different distance measures before clustering sequences (Abbott & Tsay, 2000; Aisenbrey & Fasang, 2010; Halpin, 2012; Halpin & Chan, 1998). The Hamming distance is chosen for this study because the purpose of this sequence analysis is to identify types of movement between deciles; thus, it was important to align movement events with specific ages and determine not just that an individual moved from one type of area to another *at some point in life* but *the exact point* at which the individual moved. The Hamming distance takes the form of $d(\vec{s}_1, \ \vec{s}_2)$ and is the number of positions by which $s_1$ and $s_2$ differ at a given point in time. This measure is one of the simplest forms of distance measurement in sequence analysis, but it has a few very favorable traits. First, it favors the assumption that events happen at the same time in a sequence; one life event is more like another if it happens to two children when they are both 7 years old than when one is 5 and the other is 16. In other words, in this method, children who move from deprivation to more advantaged neighborhoods are grouped together only if the move happens at roughly the same age. Second, the method is considerably faster than some of the more advanced measures when working with very large datasets.

The overall Hamming distances are clustered using Ward's linkage with a Calinski/Harabasz stopping rule, and since visualization is impossible even when splitting along birth cohorts, data inspection is based on typical sequences in the groups. The Calinski/Harabasz algorithm suggests seven distinct types of sequences, as shown in Table 1. Descriptions of the type of sequence are included in the first column.

The sequences share some very similar traits in that they are based on classical life trajectories in which the most common trend is to live in a middle-class area and not move around much, especially not to significantly better or worse areas. The "unchanged" categories vary, on average, by .9 areas up or down, while the average of the "moving" categories varies by ±2.5 full area types. This result means that individuals in the mobile categories are moving not only a single decile but also more deciles during their adolescent years.

To fully utilize the sequences and to understand whether a turbulent living condition has an effect, a measure of turbulence proposed by Elzinga and Liefbroer (2007) is used:

$$T(x) = log_2\left( Ø(x) \ \frac{s^2_{t,max}(x) + 1}{s^2_t(x) + 1} \right)$$

where $s^2_t(x)$ is the variance of the state durations $t_j \ J = 1, \ \ldots, \ \ell_d(x)$ for sequence $x$, and $s^2_{t,max}(x)$ is the maximum value this variance can take given the total duration $\ell(x) = \sum_j t_j$ of that sequence. This turbulence allows testing for intrasequence turbulence for life courses that seem to be very stable in an overall sequence category but that experience many moves within that sequence or many moves in either direction. This testing also captures little of the effect of moving to either a better or a worse decile, but it helps differentiate whether this trajectory is accomplished in one step or in many steps.

**Table 1.** Sequences.

| Qualitative description of sequence based on pattern | % of cohort in category |
|---|---|
| Average deprivation and unchanged over time | 33.44% |
| Average deprivation and moving to less deprivation | 4.00% |
| Average deprivation and moving to more deprivation | 4.48% |
| Low deprivation and unchanged | 23.63% |
| Low deprivation and moving to more deprivation | 3.74% |
| High deprivation and unchanged | 25.32% |
| High deprivation and moving to less deprivation | 5.40% |

These data are translated into a dummy variable that indicates either very low turbulence (less than two overall changes) or high turbulence (more than two changes), which captures only the degree of moves between deciles and not the actual moves within the same decile. To capture intradecile moves, the total number of moves is counted for each person in the cohort.

### 4.3. Reduction of Neighborhood Confounding

There are two overall model designs in this study, and both are based on counterfactual models that rely on potential outcomes to define the causal effect of time-varying neighborhood exposures on educational attainment at age 30 years.[4] In short, the following models attempt to address time-varying confounders affected by the past area of residence, or what Wodke et al. describe as *dynamic neighborhood selection*. The first problem to consider is the over-control of intermediate variables. Because the past area is a confounder of the area in the following year, the standard procedure is to add a control for year-area, but this also removes the indirect effect of past living areas on future time-varying factors. The second problem to consider is the collider-stratification bias, in which past area functions as a common effect of unobserved factors and prior exposure to a given area. This approach introduces an association between unobserved elements and prior area of residence, and since unobserved factors also affect educational attainment, this bias can easily overinflate the effect of accumulated areas. To address these overall problems, a counterfactual model is introduced, in which the areas

- Are separated by physical barriers,
- Are contained within a single polygon not separated by other polygons,
- Include least 100 households in the years 2000, 2005, 2010 and 2015.

The overall concept of capturing the effects of accumulated deprivation or different types of life courses is based on causal differences among different exposure trajectories, which can be reduced to the counterfactual treatment form of $E(y|x, t) = \mu(x, t, \beta_t)$, where the conditional mean of the treatment effect for the outcome variable $y$, years of full-time education completed, is conditional on covariate $x$ and treatment level $t$, which is years lived in the two most deprived deciles. Since both of the treatments (accumulated deprivation and life course trajectory) are categorical, the treatment assignment model is fitted with multinomial logistic models, while the outcome model is fitted with a linear model for years of education attained. The treatment, then, is not living in the two most deprived deciles, while the effect should be understood as an indicator of how much more or less education the individual would have gained if no deprivation had been accumulated.

The outline of the model is illustrated in Figure 1

For the treatment assignment model, a multinomial logistic model is used, in which the conditional probability of treatment is

$$p(z, t, \gamma) = \frac{\exp(z\gamma_t)}{1 + \sum_{k=1}^{q} exp(z\gamma_k)}$$

where $p(z, t, \gamma)$ is the conditional probability that a person receives treatment $t$ on the condition that covariates $z$ and $\gamma$ are the parameters of the model.

Since there is a problem with dynamic neighborhood selection, stabilized inverse probability weights are included in the calculations, which take the form of

$$sipw_i = \prod_{k=1}^{K} \frac{P(A_k = a_{ki}|\bar{A}_{k-1} = \bar{a}_{(k-1)i,}, L_0 = l_0)}{P(A_k = a_{ki}|\bar{A}_{k-1} = \bar{a}_{(k-1)i,}, \bar{L}_k = \bar{l}_{ki})}$$

[4]See also Holland (1986), Robins, Hernán, & Brumback (2000) and Rubin (1974).

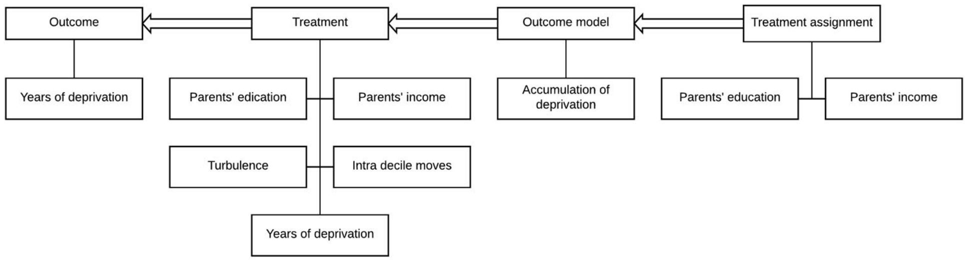



**Figure 1.** Model overview.

The denominator of the weight is the probability that a child is exposed to her actual neighborhood decile at year k conditional on past treatment and confounders. In each wave, inverse probability of treatment (IPT) weighting "balances" treatment assignment across prior confounders by giving more or less weight to children with covariate histories that are underrepresented or overrepresented in the current treatment group. In the weighted pseudopopulation, treatment in each wave is independent of prior confounders; that is, exposure to different neighborhood contexts behaves as if it is sequentially randomized with respect to observed covariates. Conditioning on confounder history is therefore no longer necessary, and an unadjusted model for the observed outcome can be fitted to the weighted observations to estimate the treatment effects of interest.

Since there is very little sample attrition because of either death or migration, survival time censoring is very limited; however, since the outcome is educational attainment, this will be skewed by censoring in the early years. To account for this issue, censoring has been included in the inverse probability weighting (IPW) estimator by introducing a function in which the received treatment, be it accumulated deprivation or life course, $\tau \in \{0 \ldots \tau_n\}$, is calculated by determining which potentially censored outcome is observed, $t = (1 - \tau)\tilde{t}_0 + \tau\tilde{t}_1$, by creating a dummy indicator for treatment censoring where $c$ is 1 for the censored observations: $c = (1 - \tau)(t_0 \geq t_c) + \tau(t_1 \geq t_c)$.

## 5. Analysis

### 5.1. Descriptive Results

Deprivation, as shown in Tables 2 and 3, is easily identified. Considering the area decile at birth, there is a mean difference of more than one year of schooling between children born in the most and least deprived neighborhoods when those individuals are thirty years old. Children born in the least deprived area decile earn approximately 65,000 DKR more per year and have an average of ten fewer days of unemployment a year than children born in the most deprived decile. When moving patterns are examined, a smoothing process seems to emerge; the average centers around average deprivation, but there are almost two whole area deciles of difference from high to low.

**Table 2.** Means of outcomes at age 30 according to decile area of birth.

| Decile area born | Education (years) | Income (Kroner) | Unemployment (days) | Decile area present | *N* |
|---|---|---|---|---|---|
| 1st decile | 13.99 | 317,117 | 25.13 | 4.33 | 28,000 |
| 2nd decile | 14.33 | 331,103 | 20.53 | 4.77 | 28,625 |
| 3rd decile | 14.51 | 342,682 | 19.02 | 5.10 | 28,529 |
| 4th decile | 14.57 | 344,486 | 17.98 | 5.44 | 29,304 |
| 5th decile | 14.66 | 352,566 | 17.18 | 5.71 | 29,045 |
| 6th decile | 14.72 | 351,877 | 17.26 | 5.88 | 28,754 |
| 7th decile | 14.79 | 353,809 | 16.43 | 6.02 | 29,079 |
| 8th decile | 14.87 | 360,589 | 15.33 | 6.22 | 28,765 |
| 9th decile | 14.97 | 371,058 | 14.60 | 6.34 | 29,003 |
| 10th decile | 15.16 | 382,980 | 14.48 | 6.29 | 28,113 |

**Table 3.** Means of outcomes in life course trajectories at age 30.

| Deprivation life courses | Education (years) | Income | Unemployment (days) | Decile area | *N* |
|---|---|---|---|---|---|
| High and unchanged | 14.31 | 331,151 | 21.54 | 4.58 | 65,495 |
| Low and unchanged | 15.10 | 377,020 | 14.06 | 6.36 | 70,154 |
| Average and unchanged | 14.77 | 354,232 | 16.68 | 5.78 | 92,659 |
| High and decreasing | 14.23 | 330,780 | 20.70 | 5.35 | 14,951 |
| Low and increasing | 14.42 | 343,803 | 18.08 | 6.03 | 10,353 |
| Average and decreasing | 14.50 | 345,516 | 18.14 | 5.90 | 12,412 |
| Average and increasing | 14.16 | 326,675 | 20.68 | 5.45 | 11,088 |

The above trend is still noticeable when the birth area is compared to the overall life course of the cohort. Compared with other life course trajectories, being born in a low-deprivation area and staying in the upper deciles results in significantly higher income, more education, less unemployment and less area deprivation later in life.

Most of the results in Table 3 correspond with birth area decile, with a few very interesting exceptions. The lowest educational attainment and lowest wage level are not found among respondents living in the most deprived areas with limited mobility but among those born in average areas who experience an increase in deprivation. It also seems that a general change, whether to more or less deprived areas, has a small negative effect on the outcome measurements. In terms of educational attainment, people who lived all of their youth in deprived areas fare better, if only marginally, than children born in deprived areas who move to more prosperous areas. The same tendency can be seen in all non "pure" categories in which there is a change in trajectory; it would seem that staying is better than leaving, regardless of whether leaving involves moving somewhere better.

This finding is very crude, of course, since neither selection nor direct effects are included in these tables, but there is little doubt that area matters.

However, it is hard to disentangle how much of these results are because of area characteristics and how much can be explained simply by interfamilial effects. Area effects and non-area effects are often intertwined, and only by accounting for dynamic selection can we approximate the true accumulated effects of area deprivation.

## *5.2. Continued Exposure and Life Course Trajectory*

Looking purely at the connection between the accumulation of deprivation between the ages of 0 and 18 years and educational attainment at age 30, there is a total effect of .85 years of difference between no accumulated deprivation and maximum accumulation.[5] To attempt to pinpoint the age at which accumulation has the most negative effect, the sample is split into 3 different age groups to measure counts of exposure to decile 10 neighborhoods between the ages of 0 and 5 years, between 6 and 11 years and between 12 and 17 years while controlling for confounders. Not surprisingly, all of the socioeconomic background factors play an important role in educational attainment at age 30, while the number of interdecile moves has a strong negative effect. One could expect this effect to come close to 0 since moves both to and from deprivation are included in this variable, but it seems likely that the effect captured is rooted in smaller moves of one decile up or down and not in continuous moves to or from deprivation. These types of movements also often entail moving a greater distance than moves within the same decile, and it would seem that the more nomadic type of living affects later educational outcomes (Table 4).

The most interesting result in the above table, however, is the difference in the effect of cumulative exposure to deprivation. The most notable effect of accumulation is seen between the ages of 12 and 17 years, followed by the effect observed between the ages of 0 and 5 years. Deprivation has no effect between the ages of 6 and 11 years. This phenomenon could be the result of many factors, but it is most likely due to either a selection problem that is not addressed here or because, beyond the initial

[5]Calculated by simple OLS.

**Table 4.** Fixed effects on education at age 30 (year coef. omitted).

| Covariates | Coef. (SE) |
|---|---|
| Cumulative exposure to decile 10 neighborhood (age 0–5) | 0.0152*** |
| | (0.00554) |
| Cumulative exposure to decile 10 neighborhood (age 6–11) | −0.00403 |
| | (0.00729) |
| Cumulative exposure to decile 10 neighborhood (age 12–17) | 0.0488*** |
| | (0.00578) |
| Gender | 0.665*** |
| | (0.00786) |
| Father's education (7 categories) | 0.422*** |
| | (0.00223) |
| Number of interdecile moves between age 0–17 | −0.207*** |
| | (0.00491) |
| Number of intradecile moves between age 0–17 | 0.0288 |
| | (0.0197) |
| Father's income | 9.45e-08*** |
| | (5.07e-09) |
| Mother's income | 1.85e-08*** |
| | (3.18e-09) |
| Constant | 1.637*** |
| | (0.0142) |
| Observations | 234,299 |
| $R^2$ | 0.182 |

Standard errors in parentheses.
***$p < .01$.
**$p < .05$.
*$p < .1$.

disadvantage of being born in a deprived area, this effect does not manifest before it reaches a critical level in later years of adolescence. Another factor could be schooling. At the age of six years, children have begun their primary education and are, in general, much less likely to move around. That phenomenon could imply that while individuals are still in primary school, much of the effect of deprivation is suppressed, but it becomes a major factor if secondary education is unavailable in the area and the family does not relocate to areas with better opportunities.

Below is the first of two models that account for selection using inverse probability weighted regression adjustments. This model measures the number of years spent living in the two most deprived deciles as a simple count – each year lived in the most deprived area is counted, so that the range is from 0, which is the treatment reference, to 18. For reference, the nondeprived areas, which are coded as a dummy for the two most deprived deciles compared to the rest, are the treatment, while the most deprived areas are treated with nondeprivation. The effects should be read as whether each year of accumulated deprivation is treated with no deprivation.

The tendency is rather unsurprising; accumulating capital from the most deprived areas has a negative effect on educational attainment later in life. When children who, at some point in their childhood, spent a single year in either of the two most deprived neighborhoods are compared to children who never lived in those areas, a difference of 0.32 years of education at age 30 years is observed. This effect continues to increase as accumulation grows and comparing children who lived all their first 18 years in deprivation to children who did not result in a difference of more than 1.4 years of education at age 30.

The overall effect seems to be linear in its relationship to some extent; for each year spent in deprivation, the effect of no deprivation becomes greater. It is interesting that there seems to be no stagnation in the treatment effect. Except for a few isolated years, the effect increases steadily to the full 18 years of accumulated deprivation. This result indicates that even the smallest change in living conditions results in an effect later in life, as long as the change involves moving out of deprivation.

**Table 5.** Multinomial logit IPWRA treatments of area deciles on linear educational attainment.

| | Treatment effect | 95% Conf. interval | |
|---|---|---|---|
| 1 year of deprivation | 0.322***<br>0.071 | 0.181 | 0.463 |
| 2 years of deprivation | 0.408***<br>0.076 | 0.258 | 0.558 |
| 3 years of deprivation | 0.369***<br>0.081 | 0.0209 | 0.530 |
| 4 years of deprivation | 0.516***<br>0.087 | 0.344 | 0.688 |
| 5 years of deprivation | 0.668***<br>0.097 | 0.478 | 0.859 |
| 6 years of deprivation | 0.688***<br>0.097 | 0.497 | 0.879 |
| 7 years of deprivation | 0.855***<br>0.106 | 0.646 | 1.064 |
| 8 years of deprivation | 0.819***<br>0.110 | 0.803 | 1.235 |
| 9 years of deprivation | 0.844***<br>0.119 | 0.609 | 1.078 |
| 10 years of deprivation | 1.087***<br>0.120 | 0.851 | 1.323 |
| 11 years of deprivation | 1.193***<br>.132 | 0.934 | 1.553 |
| 12 years of deprivation | 1.121***<br>0.125 | 0.974 | 1.367 |
| 13 years of deprivation | 1.254***<br>0.127 | 1.005 | 1.403 |
| 14 years of deprivation | 1.211***<br>0.083 | 1.047 | 1.374 |
| 15 years of deprivation | 1.287 ***<br>0.096 | 1.144 | 1.524 |
| 16 years of deprivation | 1.368***<br>0.109 | 1.185 | 1.613 |
| 17 years of deprivation | 1.451***<br>0.122 | 1.080 | 1.660 |
| 18 years of deprivation | 1.458***<br>0.117 | 1.090 | 1.751 |
| PO mean at 0 | 14.523 | 14.49 | 14.54 |

Note: Robust standard errors in parentheses.
***$p < .01$.
**$p < .05$.
*$p < .1$.

The above results help us understand the overall effect of accumulation, but not the change in a life course that could influence how the effect materializes in later educational attainment. Using the clustered sequence analysis results in the aforementioned groups as treatment, and the relationship becomes clearer. In the example below, the comparison group is children born in deprived areas who had very little mobility and who stayed in deprived neighborhoods their whole childhood.

The pure effect of living one's whole childhood in a very deprived neighborhood compared to living in the least deprived neighborhoods yields a difference of more than 1.6 years of education at age 30. This is the result of the most extreme case, in which the effect of constant deprivation is compared to constant wealth. When the rest of the results are compared with the descriptive results, it becomes much clearer that moving away from deprivation has a positive effect regardless of the starting point. Although this finding is not surprising, some results indicate that the effect of moving to a less deprived neighborhood varies with the starting point. When comparing children who lived in constant high-deprivation areas to children who were born in such areas but moved to less deprived areas, the effect is a modest gain of 0.1 years in educational attainment. However, if the starting points were average areas and the trajectories were the same, the effect is a difference of 0.3 years at age 30.

**Table 6.** Multinomial logit IPWRA treatments of life courses on linear educational attainment.

| | Treatment effect | 95% Conf. interval | |
|---|---|---|---|
| Low dep. unchanged vs. high dep. unchanged | −1.683*** (0.0438) | −1.186 | −1.983 |
| Average dep. unchanged vs. high dep. unchanged | −0.610*** (0.0374) | −.378 | −.722 |
| High dep. decreasing vs. high deprivation unchanged | −0.119*** (0.0439) | −.89 | −.187 |
| Low dep. increasing vs. high deprivation unchanged | −0.379*** (0.0463) | −.238 | −.644 |
| Average dep. decreasing vs. high deprivation unchanged | −0.390*** (0.0469) | −.254 | −.587 |
| Average dep. and increasing vs. high deprivation unchanged | 0.0305 (0.0486) | .222 | −.132 |
| PO mean at high dep. unchanged | 14.28 0.03 | .646 | 1.064 |

Note: Robust standard errors in parentheses.
***$p < .01$.
**$p < .05$.
*$p < .1$.

Looking instead at children with a negative trajectory, it seems that it is still better to start off in a rich neighborhood and then experience a decay in capital than it is to start out in an average neighborhood and experience the same decay. The only time that no effect and a slightly negative coefficient are observed is when children with constant deprivation are compared to those that started out in average neighborhoods and moved to more deprived areas. In this case, it seems that the decline in living standards has, at best, the same effect as having spent a full childhood in deprived areas.

Comparing results from the life trajectory treatments to the results regarding simple accumulation reveals a few interesting differences. The accumulation in Table 5 indicates a steady incline in educational gains for less deprived areas compared to the most deprived areas, but the connection might not be as linear when considering the results in Table 6, where the trajectory is included. This result indicates that accumulation matters, as does where one is born, but that the specific trajectory needs to be included to obtain the full picture.

## 6. Discussion and Conclusion

Neighborhood deprivation, ghettoization and the consequences of growing up in the wrong part of town are important aspects of the understanding of social deprivation. It is surprising that even though some research addresses the accumulation of deprivation in some sense (Clampet-Lundquist & Massey, 2008; Crowder & South, 2011; Jackson & Mare, 2007; Sampson & Sharkey, 2008; Wodtke et al., 2011), a relatively small body of work has the specific aim of examining the accumulation of deprivation, and even fewer studies consider the selection problem (Clampet-Lundquist & Massey, 2008; Wodtke et al., 2011) and the districting problem (Altman & Mcdonald, 2011; Damm & Schultz-nielsen, 2008; Dinesen & Sonderskov, 2015; Riva, Gauvin, & Barnett, 2007). This study investigates the connection between exposure to degrees of deprivation during childhood and adolescence and educational outcomes in later life using counterfactual models.

The results are relatively unambiguous: the more deprivation one is exposed to, the less education one can expect to acquire later in life. The difference in accumulation, primarily in the most deprived area compared with the second most deprived area, might differ by only 0.2 years in terms of educational attainment at the age of 30, but when the most deprived neighborhoods are compared to those at the other end of the spectrum, the difference is close to a full year of full-time education. These findings are interesting, but the differences found when comparing life course trajectories are even more noteworthy. These results indicate that the purest life courses of either full deprivation or full wealth result in a difference of more than 1.6 years of educational attainment. In contrast, comparing the pure deprivation group to the group that was born in average neighborhoods and moved to more deprived neighborhoods yields no significant difference in educational attainment later in life.

These results also have interesting theoretical implications. Even though when this topic is approached from a capital accumulation paradigm, the total amount of capital collected accounts for a large portion of the effect, the timing of accumulation is even more important. This finding can easily be understood in a case in which school choice is limited to people who are willing to move, but since a control for distance to school has been added, this should have no effect. These results indicate a more intricate connection between *when* we acquire capital and *how* we internalize the capital.

The overall environment we experience during childhood affects how we perceive the world and how we value education. The constitution of our area of residence has an effect on how able or willing we are to attain education, but it is not enough to consider only the accumulation of capital over a long period of time or only one's birth area. To fully understand educational deprivation from a geo-social perspective, one must consider that even though we understand long-term exposure to chemicals and other physical factors in a purely accumulative sense, it is rare not to consider the age/time of exposure in a life course setting. Some diseases have a higher mortality rate at infancy, while others have a more severe impact as we age. The same applies when we consider areas; if we disregard selection problems, we will estimate the effects poorly. If we disregard districting, we will not isolate specific area effects, and if we disregard exposure timing, we will conclude that the only factor that truly matters is relocation.

Relocation might be a valid solution, but not at any cost – the effects gained by relocating are far smaller than the pure effects of being in the lower end of the deprivation spectrum.

## Disclosure statement

No potential conflict of interest was reported by the author.

## ORCID

*Rolf Lyneborg Lund* http://orcid.org/0000-0003-1013-8620

## References

Abbott, A., & Tsay, A. (2000). Sequence analysis and optimal matching methods in sociology: Review and prospect. *Sociological Methods and Research*, *29*(1), 3–33.

Aisenbrey, S., & Fasang, A. E. (2010). New life for old ideas: The "Second Wave" of sequence analysis bringing the "Course" back into the life course. *Sociological Methods & Research*, *38*(3), 420–462.

Albanese, G., De Blasio, G., & Sestito, P. (2016). My parents taught me. Evidence on the family transmission of values. *Journal of Population Economics*, *29*(2), 571–592.

Altman, M., & Mcdonald, M. P. (2011). BARD : Better Automated Redistricting. *Journal of Statistical Software*, *42*(4), 1–28.

Andersson, E., & Malmberg, B. (2013). *Contextual effects on educational attainment in individualized neighborhoods: Differences across gender and social class (SPaDE No. 10)*. Stockholm.

Åslund, O., & Skans, O. N. (2010). Will i see you at work? Ethnic workplace segregation in Sweden, 1985-2002. *Industrial and Labor Relations Review*, *63*(3), 471–493.

Atkinson, R., & Kintrea, K. (2001). Disentangling area effects: Evidence from deprived and non-deprived neighbourhoods. *Urban Studies*, *38*(12), 2277–2298.

Atkinson, R., & Kintrea, K. (2004). "Opportunities and despair, it's all in there": Practitioner experiences and explanations of area effects and life chances. *Sociology*, *38*(3), 437–455.

Bischoff, K. (2008). School district fragmentation and racial residential segregation: How do boundaries matter? *Urban Affairs Review*, *44*(2), 182–217.

Bourdieu, P. (1986). The forms of capital. In J. Richardson (Ed.), *Handbook of theory and research for the sociology of education* (1st ed., pp. 241–258). Westport, CT: Greenwood.

Bourdieu, P. (1998). *Practical reason: On the theory of action*. Stanford: Stanford University Press.

Bourdieu, P., & Passeron, J.-C. (1990). *Reproduction in education, society and culture*. London: SAGE Publications.

Bower, K. M., Thorpe, R. J., Rohde, C., & Gaskin, D. J. (2014). The intersection of neighborhood racial segregation, poverty, and urbanicity and its impact on food store availability in the United States. *Preventive Medicine*, *58*(1), 33–39.

Breen, R., Luijkx, R., Müller, W., & Pollak, R. (2009). Nonpersistent inequality in educational attainment: Evidence from eight European countries. *American Journal of Sociology*, *114*(5), 1475–1521.

Carlsson, A. C., Li, X., Holzmann, M. J., Ärnlöv, J., Wändell, P., Gasevic, D., … Sundquist, K. (2017). Neighborhood socioeconomic status at the age of 40 years and ischemic stroke before the age of 50 years: A nationwide cohort study from Sweden. *International Journal of Stroke*, *12*(8), 815–826.

Cattell, V. (2001). Poor people, poor places, and poor health: The mediating role of social networks and social capital. *Social Science & Medicine*, *52*(10), 1501–1516.

Clampet-Lundquist, S., & Massey, D. S. (2008). Neighborhood effects on economic self-sufficiency: A reconsideration of the moving to opportunity experiment. *American Journal of Sociology*, *114*(1), 107–143.

Crowder, K., & South, S. J. (2011). Spatial and temporal dimensions of neighborhood effects on high school graduation. *Social Science Research*, *40*(1), 87–106.

Cunha, J. M. P. da, Jimenez, M. A., Perez, J. R. R., & Andrade, C. Y. de. (2009). Social segregation and academic achievement in state-run elementary schools in the municipality of Campinas, Brazil. *Geoforum*, *40*(5), 873–883.

Damm, A. P., & Schultz-nielsen, M. L. (2008). *The construction of neighbourhoods and its relevance for the measurement of social and ethnic segregation : Evidence from Denmark*. Discussion Paper Series.

Deng, Y. (2016). Challenges and complications in neighborhood mapping: from neighborhood concept to operationalization. *Journal of Geographical Systems*, *18*(3), 229–248.

Dewilde, C. (2004). The multidimensional measurement of poverty in Belgium and Britain: A categorical approach. *Social Indicators Research*, *68*(3), 331–369.

Dinesen, P. T., & Sonderskov, K. M. (2015). Ethnic diversity and social trust: Evidence from the micro-context. *American Sociological Review*, *80*(3), 550–573.

Elzinga, C. H., & Liefbroer, A. C. (2007). De-standardization of family-life trajectories of young adults: A cross-national comparison using sequence analysis. *European Journal of Population* (Vol. 23, pp. 225–250).

Feld, S. L. (1981). The focused organization of social ties. *The American Journal of Sociology*, *86*(5), 1015–1035.

Ferreira, M. A. R., Holan, S. H., & Bertolde, A. I. (2011). Dynamic multiscale spatiotemporal models for Gaussian areal data. *Journal of the Royal Statistical Society. Series B, Statistical Methodology*, *73*(5), 663–688.

Fischer, C. S., Stockmayer, G., Stiles, J., & Hout, M. (2004). Distinguishing the geographic levels and social Dimensions of U.S. metropolitan segregation, 1960-2000. *Demography*, *41*(1), 37–59.

Fone, D., Dunstan, F., Lloyd, K., Williams, G., Watkins, J., & Palmer, S. (2007). Does social cohesion modify the association between area income deprivation and mental health? A multilevel analysis. *International Journal of Epidemiology*, *36*(2), 338–345.

Freisthler, B., Thomas, C. A., Curry, S. R., & Wolf, J. P. (2016). An alternative to residential neighborhoods: An exploratory study of how activity spaces and perception of neighborhood social processes relate to maladaptive parenting. *Child and Youth Care Forum*, *45*(2), 259–277.

Gesthuizen, M., Van Der Meer, T., & Scheepers, P. (2008). Education and dimensions of social capital: Do educational effects differ due to educational expansion and social security expenditure? *European Sociological Review*, *24*(5), 617–632.

Gilliland, J. A., Olson, S. H., & Gauvreau, D. (2011). Did segregation increase as the city expanded? The case of Montreal, 1881-1901. *Social Science History*, *35*(4), 465–503.

Gorard, S., & Siddiqui, N. (2016). Grammar schools in England: A new approach to analysing their intakes and outcomes. http://dro.dur.ac.uk

Halpin, B. (2012). *Sequence analysis of life-course data: A comparison of distance measures* (No. WP2012-02). Retrieved from http://www.ul.ie/sociology/pubs/wp2012-02.pdf

Halpin, B., & Chan, T. W. (1998). Class careers as sequences: An optimal matching analysis of work-life histories. *European Sociological Review*, *14*(2), 111–130.

Hannum, E. (2005). Market transition, educational disparities, and family strategies in rural China: New evidence on gender stratification and development. *Demography*, *42*(2), 275–299.

Heckman, J. (1977). Sample selection bias as a specification error (with an application to the estimation of labor supply functions).

Hedman, L., Manley, D., Van Ham, M., & ÖSth, J. (n.d.). Cumulative exposure to disadvantage and the intergenerational transmission of neighbourhood effects.

Hjellbrekke, J., & Korsnes, O. (2004). Educational mobility trajectories and mobility barriers in the Norwegian social space. *International Journal of Contemporary Sociology*, *41*(2), 163–189.

Holland, P. W. (1986). Statistics and causal inference. *Journal of the American Statistical Association*, *81*, 945–960.

Jackson, M. I., & Mare, R. D. (2007). Cross-sectional and longitudinal measurements of neighborhood experience and their effects on children. *Social Science Research*, *36*(2), 590–610.

Kelo, M. (2010). *Beyond 2010: priorities and challenges for higher education in the next decade*. Retrieved from http://www.acasecretariat.be/fileadmin/aca_docs/images/members/2008_Beyond_2010_-_Priorities_and_challenges_for_higher_education_in_the_next_decade.pdf#page=15

King, G., Keohane, R., & Verba, S. (1994). *Designing social inquiry – Scientific inference in qualitative research*.

Krieger, N., Feldman, J. M., Waterman, P. D., Chen, J. T., Coull, B. A., & Hemenway, D. (2017). Local residential segregation matters: Stronger association of census tract compared to conventional city-level measures with fatal and non-fatal assaults (total and firearm related), using the Index of Concentration at the Extremes (ICE) for racial, econ. *Journal of Urban Health*, *94*(2), 244–258.

Krieger, N., Waterman, P. D., Batra, N., Murphy, J. S., Dooley, D. P., & Shah, S. N. (2017). Measures of local segregation for monitoring health inequities by local health departments. *American Journal of Public Health*, *107*(6), 903–906.

Kunz, J., Page, M. E., & Solon, G. (2003). Are point-in-time measures of neighborhood characteristics useful proxies for children's long-run neighborhood environment? *Economics Letters*, *79*(2), 231–237.

Lance, G. N., & Williams, W. T. (1967). A general theory of classificatory sorting strategies: 1. Hierarchical systems. *The Computer Journal*, *9*(4), 373–380.

Law, J., Quick, M., & Chan, P. W. (2015). Analyzing hotspots of crime using a Bayesian spatiotemporal modeling approach: A case study of violent crime in the Greater Toronto area. *Geographical Analysis*, *47*(1), 1–19.

Lee, B. A., Reardon, S. F., Firebaugh, G., Farrell, C. R., Matthews, S. A., & O'Sullivan, D. (2008). Beyond the census tract: Patterns and determinants of racial segregation at multiple geographic scales. *American Sociological Review*, *73*(5), 766–791.

Leventhal, T., & Brooks-Gunn, J. (2000). The neighborhoods they live in: The effects of neighborhood residence on child and adolescent outcomes. *Psychological Bulletin*, *126*(2), 309–337.

Lindstrom, D. P., & Massey, D. S. (1994). Selective emigration, cohort quality, and models of immigrant assimilation. *Social Science Research*, *23*(4), 315–349.

Lund, R. L. (2018). From the dark end of the street to the bright side of the road: Automated redistricting of areas using physical barriers as dividers of social space. *Methodological Innovations*, *11*(3), 1–15.

Malmberg, B., Andersson, E., & Östh, J. (2011). How does segregation vary between different urban areas? Using a k-nearest neighbor approach to segregation measurement. In ENHR Conference 2011 (p. 16).

Martinez, J. (2009). The use of GIS and indicators to monitor intra-urban inequalities. A case study in Rosario, Argentina. *Habitat International*, *33*(4), 387–396.

McDevitt, T. M., Hawley, A. H., Udry, J. R., Gadalla, S., Leoprapai, B., & Cardona, R. (1986). Migration plans of the rural populations of third world countries: A probit analysis of micro-level data from Asia, Africa, and Latin America. *Journal of Developing Areas*, *20*(4), 473–490.

Ministry of Transportation. (2018). Liste over ghettoområder pr. 1. December 2017 – Transportministeriet. Retrieved from https://www.trm.dk/da/publikationer/2017/liste-over-ghettoomraader-1-dec-2017

Moule, R. K., Decker, S. H., & Pyrooz, D. C. (2013). Social capital, the life-course, and gangs. In *Handbook of life-course criminology: Emerging trends and directions for future research* (pp. 143–158). New York, NY: Springer New York.

Munk, M. D., & Thomsen, J. P. (2018). Horizontal stratification in access to Danish university programmes. *Acta Sociologica (United Kingdom)*, *61*(1), 50–78.

Næss, Ø., Leyland, A. H., Smith, G. D., & Claussen, B. (2005). Contextual effect on mortality of neighbourhood level education explained by earlier life deprivation. *Journal of Epidemiology and Community Health*, *59*(12), 1058–1059.

Ochieng, B. M. N. (2011). The effect of kin, social network and neighbourhood support on individual well-being. *Health & Social Care in the Community*, *19*(4), 429–437.

Östh, J. (2018). Equipop. Retrieved from http://equipop.kultgeog.uu.se/index.html

Östh, J., Clark, W. A. V., & Malmberg, B. (2015). Measuring the scale of segregation using k-nearest neighbor aggregates. *Geographical Analysis*, *47*(1), 34–49.

Östh, J., Malmberg, B., & Andersson, E. K. (2014). Analysing segregation using individualised neighbourhoods. In D. Llyod Christipher, Ian Shuttleworth, & David W. Wong (Eds.), *Social-spatial segregation: Concepts, processes and outcomes* (pp. 135–162). Clifton: Policy Press.

Pais, J., South, S. J., & Crowder, K. (2012). Metropolitan heterogeneity and minority neighborhood attainment: Spatial assimilation or place stratification? *Social Problems*, *59*(2), 258–281.

Payne, R. A. (2012). UK indices of multiple deprivation – A way to make comparisons across constituent countries easier. *Health* (Vol. 4). Scientific Research Pub.

Petrović, A., Ham, M. van, & Manley, D. (2017). *Size matters? Exploring scale for neighbourhood effects*. In AGILE conference 2017. Springer.

Petrović, A., Ham, M. van, & Manley, D. (2018). Multiscale measures of population: Within- and between-city variation in exposure to the sociospatial context. *Annals of the American Association of Geographers*, *108*(4), 1057–1074.

Potter, C., Walker, G., & Keen, B. (2012). Engaging fathers from disadvantaged areas in children's early educational transitions: A UK perspective. *Journal of Early Childhood Research*, *10*(2), 209–225.

Riise, B. S., Dommermuth, L., & Lyngstad, T. H. (2016). Intergenerational transmission of age at first birth in Norway. *European Societies*, *18*(1), 47–69.

Riva, M., Gauvin, L., & Barnett, T. A. (2007). Toward the next generation of research into small area effects on health: A synthesis of multilevel investigations published since July 1998. *Journal of Epidemiology and Community Health*, *61*(10), 853–861.
Robins, J. M., Hernán, M. A., & Brumback, B. (2000). Marginal structural models and causal inference in epidemiology. *Epidemiology 11*, 550–560.
Rubin, D. B. (1974). Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of Educational Psychology 66*, 688–701.
Sampson, R. J. (2012). *Great American City – Chicago and the enduring neighborhood effect (1.)*. Chicago, IL: The University of Chicago Press.
Sampson, R. J., Morenoff, J. D., & Gannon-Rowley, T. (2002). Assessing "Neighborhood Effects": Social processes and new directions in research. *Annual Review of Sociology*, *28*(2002), 443–478.
Sampson, R. J., & Sharkey, P. (2008). Neighborhood selection and the social inequality. *Demography*, *45*(1), 1–29.
Shahzadi, M., Riaz, M. F., Anwar, S., & Nasreen, S. (2017). How unequal is the size of middle class in the rural urban areas of Punjab province. *International Journal of Social Economics*, *44*(2), 253–266.
Sharkey, P. (2013). *Stuck in place : Urban neighborhoods and the end of progress toward racial equality*. Chicago: The University of Chicago Press.
Sharkey, P., & Elwert, F. (2011). The legacy of disadvantage: Multigenerational neighborhood effects on cognitive ability. *American Journal of Sociology*, *116*(6), 1934–1981.
Sharkey, P., Schwartz, A. E., Ellen, I. G., & Lacoe, J. (2014). High stakes in the classroom, high stakes on the street: The effects of community violence on student's standardized test performance. *Sociological Science*, *1*(May), 199–220.
Söderström, M., & Uusitalo, R. (2010). School choice and segregation: Evidence from an admission reform. *Scandinavian Journal of Economics*, *112*(1), 55–76.
South, S. J., & Crowder, K. (2010). Neighborhood poverty and nonmarital fertility: Spatial and temporal dimensions. *Journal of Marriage and Family*, *72*(1), 89–104.
Sundquist, J., Sjöstedt, C., Winkleby, M., Li, X., Kendler, K. S., & Sundquist, K. (2016). Neighborhood linking social capital as a predictor of drug abuse: A Swedish national cohort study. *Addictive Behaviors*, *63*, 37–44.
Taylor, C., Gorard, S., & Fitz, J. (2003). The modifiable areal unit problem: Segregation between schools and levels of analysis. *International Journal of Social Research Methodology*, *6*(1), 41–60.
Tayman, J., Schafer, E., & Carter, L. (1998). The role of population size in the determination and prediction of population forecast errors: An evaluation using confidence intervals for subcounty areas. *Population Research and Policy Review*, *17*(1), 1–20.
van de Werfhorst, H. G., & Hofstede, S. (2007). Cultural capital or relative risk aversion? Two mechanisms for educational inequality compared. *The British Journal of Sociology*, *58*(3), 391–415.
Ward, J. H. (1963). *Hierarchical grouping to optimize an objective function. Journal of the American Statistical Association*, *58*(301), 236–244.
Warrington, M. (2005). Mirage in the desert? Access to educational opportunities in an area of social exclusion. *Antipode*, *37*(4), 796–816.
Wiborg, Ø. N., & Hansen, M. N. (2018). The Scandinavian model during increasing inequality: Recent trends in educational attainment, earnings and wealth among Norwegian siblings. *Research in Social Stratification and Mobility*, *56*, 53–63.
Wilkinson, R. G., & Pickett, K. E. (2007). The problems of relative deprivation: Why some societies do better than others. *Social Science and Medicine*, *65*(9), 1965–1978.
Wodtke, G. T., Harding, D. J., & Elwert, F. (2011). Neighborhood effects in temporal perspective: The impact of long-term exposure to concentrated disadvantage on high school graduation. *American Sociological Review*, *76*(5), 713–736.
Wooldridge, J. M. (1995). Selection corrections for panel data models under conditional mean independence assumptions. *Journal of Econometrics*, *68*(1), 115–132.
Wooldridge, J. M. (2010). *Econometric analysis of cross section and panel data* (2nd ed., Vol. 58). Cambridge, MA: MIT Press.
Zingher, J. N., & Steen Thomas, M. (2014). The spatial and demographic determinants of racial threat. *Social Science Quarterly*, *95*(4), 1137–1154.

# I Like the Way You Move

- The effect of moving to and from different types of areas during adolescence on later socioeconomic outcomes

Rolf Lyneborg Lund
Aalborg University

# 1 Introduction

The rural/urban divide is both a geopolitical and social issue in most of the world. Unequal geographical distribution of wealth, education and work opportunities is evident in almost all European countries. One of the most prominent goals of the European budget is to redistribute funds to agriculture through the CAP (Common Agricultural Policy) and, since 2000, to strengthen the rural areas of Europe with the introduction of the RDP (Rural Development Policy) (European Commission, 2013). In the European context, there is concern with the increased focus on both inter and intracountry social cohesion following the inclusion of newer eastern European countries in the EU; concern also stems from the increase in nonwestern immigrants seeking asylum (European Commission, 2017).

With the increasing tendency toward depopulation in rural areas and the massive influx of people into urban areas, the focus on geographically defined deprivation has splintered into discussions of deprivation at different levels in rural settings. Whereas urban settings often are described as ghettos or slums, rural ones are described as backwoods towns or hinterlands (Davidson, 1996).

With regard to decision making about support for different environments, the political agenda in the EU is differentiated in the sense that urban and rural areas are seen to have different symptoms. Where urban ghettos are seen as struggling with crime and poor health (Law, Quick, & Chan, 2015; Rotolo & Tittle, 2006; Sampson, Morenoff, & Gannon-Rowley, 2002), rural areas are seen to be challenged by low educational attainment and low intergenerational mobility (Bibby, 2013; Lee, Árnason, Nightingale, & Shucksmith, 2005; Milbourne, 2007; Mulugeta, 2004), but research has so far not been interested in analyzing these perceptions. The problem with comparing different deprived areas and disregarding the urban/rural spectrum is that such an approach overlooks a key aspect of the different social conditions that exist in different places. Consequently, deprivation is often either viewed as a non-setting-specific problem, where deprivation is deprivation without regard to urbanization (Atkinson & Kintrea, 2004; Friedrichs, Galster, & Musterd, 2003; Galster, 2010; Garner & Raudenbush, 1991; Sharkey & Elwert, 2011; Sridharan, Tunstall, Lawder, & Mitchell, 2007; Wodtke, Harding, & Elwert, 2011); or, rurality is seen, in itself, as a sort of deprivation by definition, wherein there are only individual exceptions to the rule (Bower, Thorpe, Rohde, & Gaskin, 2014; Crenshaw, 1992; Ocana-Riola, Sanchez-Cantalejo, & Fernandez-Ajuria, 2006; Voss, Long, Hammer, & Friedman, 2006). Although the literature on neighborhood effects is growing and we are learning more about how to measure the effects of spatial deprivation and why such deprivation develops (Sampson, 2012), we are somewhat limited in regard to the differentiated effects that exist in different deprived settings. This paper will examine deprivation in different settings of urban and rural life to understand how the effects of growing up in deprivation might differ between urban and rural areas when measuring adult income levels, educational attainment and labor market affiliation.

## 2 The concept of deprivation and space

As mentioned earlier, studies of deprivation primarily focus on deprivation in a dichotomous or gradient form (Carlsson et al., 2017; Potter, Walker, & Keen, 2012; Sampson, 2012). In other words, scholars analyze either degrees of deprivation or deprivation as a yes/no question, and studies often focus on somewhat arbitrarily defined spaces.

The first overall problem with studying deprivation in a geographically detached sense is that there is an undefined underlying assumption that deprivation is homogenous across geographical space. While most quantitative studies do report some measure of variance between deprived areas, this measure is often mentioned to justify the scope of the analysis and is not further explored (Atkinson & Kintrea, 2004; Garner & Raudenbush, 1991; Potter et al., 2012). Deprived areas are seen as a homogenous group to be analyzed without consideration of the actual setting of deprivation. While it could be the case that the effects of deprivation are uniform over space, it is problematic that this assumption is left untested, especially because policies to relieve or help with local problems often differ in both type and scope when targeting problems in urban settings compared to rural settings (European Comission, 2007; European Commission, 2017; Ministry of Transportation, 2018). Newer studies that are interested in scale point to the fact that the type of problem being investigated must be considered when choosing the scale of research (Lund, 2018; Petrovi, Ham, & Manley, 2018; Petrović, Ham, & Manley, 2017). Human phenomena can be homogenous across vast distances but can also be heterogeneous even across a single street in a small town (Feld, 1981). The problem with external heterogeneity cannot be solved theoretically; it must be solved by carefully considering the phenomena in question and reacting to the way the human population disperses over geographical space.

The second problem when disregarding geographical placement is the scale of deprivation. While specific placement is a problem of between variance, scale is a problem of variance within areas of deprivation. One of the major problems when trying to understand deprivation in a geographical setting is that deprivation in itself can be an elusive entity to isolate. Even at the city level, deprivation is often found in specific segments or neighborhoods, and labeling a city, parish or municipality as a deprived area is only obscuring the bigger, or in this case smaller, picture.

There are studies that focus on an even smaller level of aggregation, namely the local neighborhood level (Bower et al., 2014; Jones & Pebley, 2014; Logan, Spielman, Xu, & Klein, 2011; Malmberg, Andersson, & Östh, 2011; Wodtke et al., 2011), and the argument for a very small aggregation level is to isolate one's particular research aim as much as possible to exclude as much "noise" as possible. Where some studies use smaller administrative areas such as census tracts consisting of either block-level or street-level data, as in some American studies (Bower et al., 2014; Gage, Dyke, & Maccluer, 1986; Krieger et al., 2017), or smaller statistical units of measurement such as the Small Areas for Market Statistics (SAMS) used in Sweden (Lagerlund, Merlo, Vicente, & Zackrisson, 2015; Merlo et al., 2013) others use more inductive clustering techniques such as k-means clustering or Bayesian methods (Ferreira, Holan, & Bertolde, 2011;

Johnelle Sparks, Sparks, & Campbell, 2013; Malmberg et al., 2011; Östh, Clark, & Malmberg, 2015; Petrović et al., 2017). Although studies that utilize smaller sets of administrative data are more precise in isolating the local area, they still fail to account for the actual distribution inside the neighborhoods and do not account for homogeneity. In the end, it is impossible to know if the lower internal heterogeneity occurs because of simple data smoothing or because the administrative areas capture the local area better. On the other end of the spectrum, when using the k-means and other clustering techniques, a problem of both statistical discretion criteria and fuzziness of borders arises. Although these methods capture the local from a bottom-up perspective, they rely on individual-level data access on a geographical scale, which in a Danish context is impossible due to rules of anonymity on registers. Another problem with the nearest neighbor approach is that the geography of habitation is disregarded. That approach relies solely on clustering individuals and does not consider how the surrounding area indicates neighborhood formation[1].

This paper will explore a new method of automated redistricting that takes into account both segments of the local area: size and geography.

[1] For a more thorough discussion of the methodologies see (Lund, 2018).

# 3 Understanding differences in space

Although the rural/urban divide is used in many different settings, most countries operate with a clear classification of rurality and urbanity. The European Union introduced the NUTS-system (European Comission, 2007) to classify area aggregation levels and to classify rurality inside different European countries. Many countries use some variation of these units to further classify areas according to some degree of rurality vs. urbanity. The Department of Environment, Food and Rural Affairs in the UK applies 6 different types of residential areas ranging from major conurbations to dispersed areas (Bibby, 2013), while Denmark uses a 4-category system ranging from outer municipality to urban municipality (Ministeriet for Fødevarer Landbrug og Fiskeri, 2011).

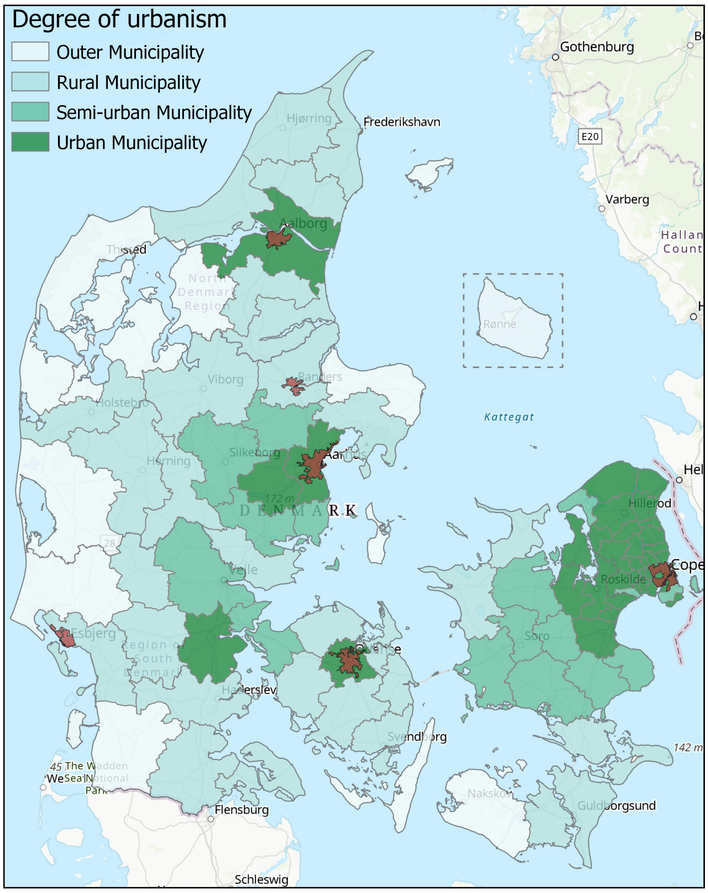
In most cases, these categories follow some very specific overall statistical traits. In Denmark, the 4 categories are defined as follows (Ministeriet for Fødevarer Landbrug og Fiskeri, 2011):

1. Urban Municipality - cities or surrounding areas where at least half of the citizens live in cities with more than 3000 inhabitants and within a 30-minute drive from one of the larger cities in Denmark (Copenhagen, Aarhus, Odense or Aalborg).
2. Semiurban municipality - city areas where at least half of the citizens live in cities with more than 3000 inhabitants but more than a 30-minutes drive from one of the larger cities in Denmark.
3. Rural Municipality - rural areas where less than half of the inhabitants live in cities with more than 3000 inhabitants but are within a 30-minute drive of one of the larger cities in Denmark.

4. Outer Municipality - rural areas where more than half of the inhabitants live in towns smaller than 3000 inhabitants and are farther than a 30-minute drive from one of the larger cities in Denmark.

As seen in figure 1, the only municipalities that are categorized as urban are those that center on larger cities. Copenhagen, Aarhus, Odense and Aalborg are all large enough to account for half of the population inside the municipality who also live in large cities, while the 5$^{th}$ largest city, Esbjerg, and the 6$^{th}$, Randers, fall into the rural category because of how many people in those municipalities who live outside the larger city and instead in smaller, more rural cities.

**Figure 1: Municipality types and the six largest cities (red)**



This situation means that the placement of outer municipalities creates what the Danish media has dubbed: “The rotten banana”. The main reason for this name is the lower levels of educational attainment in these municipalities and the fact that the poorer municipalities are located in the half-circle ranging from the upper west coast to the lower southeastern part of the country. This description, however, does not tell us if these areas are all problematic, in what sense they are problematic and if they yield any significantly different life course results for people growing up in these parts of the country. Like all other parts of Denmark, these parts contain not only impoverished neighborhoods but also rich ones. The real question here is whether the deprived neighborhoods in the outer and more rural parts of Denmark have a doubly negative effect or if the effect is contained to the actual type of neighborhood in which one grows up.

Deprivation is often defined as a specific issue that is almost completely separated from levels of rurality. When considering the very specific measurements often applied to deprivation studies, this definition means that not only will the more rural areas be more represented because of their often-lower levels of educational attainment but also the effect of deprivation will be skewed by a very nongeographical approach to a very geographical topic.

In this paper, I will explore different effects of deprivation by considering deprivation as a very local and geographical phenomenon. Does growing up in deprivation in an urban setting change our life chances compared to growing up in more rural deprivation?

# 4 Methodology

## 4.1 Data

The two main data sources utilized in this paper are georeferenced data and register data for the total Danish population. The georeferenced data consist of the national square grid that divides Denmark into vectors of 100 by 100 meter cells and topographical maps that contain information about buildings, roads, rivers, railroads and so on. The georeferenced data are linked to the registers, but because Statistics Denmark has very strict discretion criteria for anonymity, the data must be clustered to at least 100 inhabitants per measurable geographical unit before further linking them to individual-level data. The methodology behind this is described thoroughly below.

The second set of data is a variety of Danish registers with two overall population types: first, a full population of all persons living in Denmark measured in yearly intervals between 1980 and 2016; and second, a cohort of children born between 1980 and 1985. Both the full population and the cohort have yearly measured individual-level information about educational attainment, labor force affiliation and income as well as gender, age, ethnicity and other demographic characteristics. The cohort is linked to family background and overall aggregated area characteristics as well as to a proximity measurement for all other areas in the country to further investigate overall moving patterns and physical mobility.

## 4.2 Measuring the local

As mentioned earlier, isolating deprived neighborhoods can be difficult when using administrative areas. This study reverses the concept of neighborhoods as something that can be defined by somewhat arbitrary borders (like municipalities, cities or parishes) and instead introduces an automated redistricting based on a more inductive, recursive algorithm to isolate very small socioeconomic clusters (Lund, 2018). Since the early days of the Chicago school, space and place have together been understood as a social function of the inhabitants located within (Waterman, Park, Burgess, & Boyd, 2006). The way we create social cohesion and efficacy inside neighborhoods might be hard to describe and measure, but the way we shape the landscape to allow this cohesion is not (Entwisle, Rindfuss, Walsh, Evans, & Curran, 1997; Feld, 1981; Lund, 2018; White, Kim, & Glick, 2005). We shape our cities, communities and housing after principles of closeness, and we separate these bonds with physical barriers (Feld, 1981). Although we often expect these barriers to be in the form of walls, this is, for the most part, not the case. Instead, we use more subtle separators of space – either actively or subconsciously – in the form of roads, railways, rivers, lakes, forests or other objects that might not be built as separators but nevertheless act as such (Feld, 1981; Lund, 2018). Using this logic, the areas are built opposite to administrative areas by looking at the way we decide to cluster in an already existing geography.

The methodology relies on two distinct steps: first, define the rules for overall geographical subdivision and ensure that enough inhabitants are located in each geographical entity. Danish register data have very specific discretion rules in regard to geographical clustering and require at least 100 inhabitants per geographical unit before an actual merge between geography and individual data can be performed. A simple, programmable reduction of arguments is as follows:

1. Apply separators on the total of Danish geography.
2. Merge with the square grid and form borders to follow the grid.
3. Count inhabitants in newly formed areas and flag areas that have fewer than 100 inhabitants.

To satisfy the discretion criteria, it is important to secure at least 100 inhabitants per area before actual data can be applied to the geography. This requires further steps that are optimized to secure four separate criteria: have at least 100 inhabitants per area, merge areas so that as few merges as possible take place, merge areas so that the areas are as physically small as possible, and merge so that merges are as close to the 100-rule is possible. These criteria are made to secure areas that are small both physically and in number of inhabitants, but the overall advantage of this optimization is that merge solutions can be objectively evaluated and the best version can be chosen.

The overall computational problem with a recursive model is one of permutations. With 21.384 raw areas sharing an average of 5.4 borders and evaluating four overall cost-functions, choosing the optimal starting point becomes impossible because each model choice restarts all other merges. This would result in a theoretical lowest number of calculations of septendecilliards of permutations and thus is not efficient to run in either loops or recursions. To circumvent this, prioritizations are introduced in the form of percentage shared borders and a more linear form of merges. A reduced form is as follows (see appendix A for pseudocode):

1. Flagged and unflagged areas are treated equally.
2. Percentage borders shared by flagged areas and the surrounding areas are calculated as well, as follows:
   a. Inhabitant size after merge
   b. Geographical size after merge
   c. A counter to keep track of the number of merges
3. A gradient descent cost function that is gradually becoming higher with an increase in all values defined above.
4. Starting with areas with the highest percentage of shared borders, calculate the cost of the merge on number of inhabitants[2]; geographical size[3] is calculated by a recursive function and the number of merges that is required to get to the specific point.
5. Recursion continues until the cost-function ceases to decrease in 10 consecutive runs.

The main advantage of this method is that it yields the same result every time and that the areas are small in both inhabitant size and geographical size (Lund, 2018).

[2] This cost function is based on an infinitely large cost at inhabitant size below 100, 0 at 100 and then gradually increasing from 100.

[3] Measured as $1m^2=1$ and then incrementally larger for each square meter increase after merge.

## 4.3 Measuring deprivation

To classify areas in regard to deprivation, a composite index is introduced comprising standardized measurements for average area income as a percentage of the mean income of the region it belongs to, percentage of welfare recipients, percentage of unemployed individuals and percentage of the area population in managerial positions. This is a reduced form of the method used by the Ministry of Transportation to form the ghetto list, which declares specific areas to be vulnerable areas or even ghettos. The main point of deviance from the complete method is by omitting measurements of educational attainment and percentage of nonwestern immigrants. Because Denmark only houses four major universities that are all located inside the major cities, this could cause a somewhat skewed picture compared to more labor market-specific measurements, and confusion would also be introduced by movers and stayers at different stages of the life course. Educational attainment is still measured as an outcome to further our understanding of educational differences in different local settings but is not used to classify deprivation directly.

Because this study is interested in measuring deprivation and its effects over a 30-year time period, the index is averaged over the first 18 years of the cohorts' life course so that areas that are affected by potential outliers and fluctuating deprivation levels are excluded and then only considered to be deprived if the average level of deprivation is constantly considered to be in the lowest $10^{th}$ decile.

## 4.4 Neighborhood confounding and dynamic selection

In addition to the aforementioned problems with between- and within-variation when investigating area effects, we also have problems with both neighborhood confounding and overall selection on both the level of rurality and the level of deprivation (Wodtke et al., 2011). Where the normal selection problem with neighborhoods primarily consists of nonrandom assignment to neighborhoods, the temporal problem with neighborhood selection becomes even more complex because of autocorrelation: where one lives one year is a strong predictor for the type of area one lives the following year. Wodtke et al. (2011) describe this problem as dynamic selection and raise concerns with using life course or time data in neighborhood studies without resolving the problem with time-varying confounders. Even in a normal regression design, this selection becomes impossible to take into account because a simple control for a lagged neighborhood type would cause a massive overestimation of neighborhood effects because of autocorrelation.

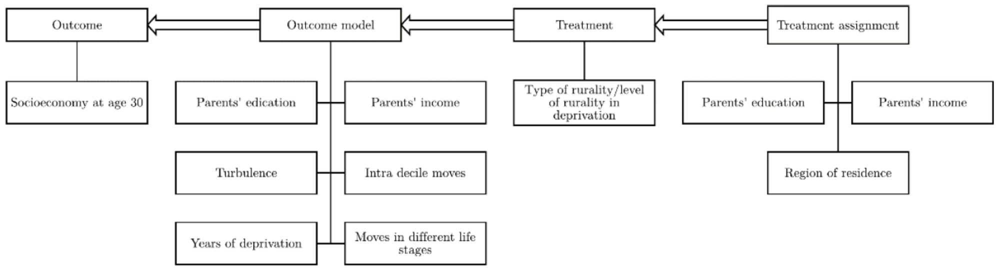
Because randomizing housing is somewhat implausible, a counterfactual framework is introduced with the use of treatment effects and inverse probability weighted regression adjustments (Wooldridge, 1995, 2010). These doubly robust methods combine an outcome regression with a model for the exposure or, in this case, the selection, where the main concern is selection into the treatment (Wooldridge, 2010). This means that there are four overall parts of the model. First, an outcome, which in this paper is either yearly number of days of unemployment at age 30, yearly income at age 30 or months of fulltime education completed.

Second, estimators are used to predict socioeconomic measurements such as parents' educational attainment, parents' income, intra and inter-area movement when growing up, years living in deprived neighborhoods and when the moves to and from deprivation occurred in the life course. By adding control for life course events and accumulation of deprivation in a temporal perspective, it becomes possible to isolate area effects away from the potential problem of comparing respondents who only briefly lived in deprivation to those who have lived in deprivation their entire lives.

The third part is the treatment. Three different types of treatment are used in this paper, where the first treatment is the effect of growing up in either outer, rural, semi-urban or urban municipalities and moving to an urban setting after turning 18 years of age. The second treatment is the same setting but instead reduces to growing up in any of the four municipality types in deprived neighborhoods and moving to urban settings. The third treatment is the effect of growing up in deprivation and staying in the municipality type after the age of 18. This last treatment is the pure effect of the four municipality types because it consists of the stayers - those who grew up and never left.

The last part of the model is the regression adjustment, where the probability of belonging to any category in the treatments listed above is calculated and adjusted for when estimating the treatment effects in the first steps of the model. In reduced form, the adjustment must contain measurements that could cause selection into the treatments in part three of the model. Because most of the treatments are dependent on parental socioeconomic status during childhood and adolescence, the main adjustment is through parents' income and educational attainment, with control for parents gaining more education and advancing in the labor force later in the child's life. Another potential confounder is the region of growing up. Although the proportion of children growing up in deprivation is somewhat equal between regions, it is not clear whether the socioeconomic parameters predict deprivation differently among regions. Thus, a time-varying measurement of region is included (Lund, 2019).

An overview of the methodological design can be seen in figure 2 below.




In this paper, the treatments are in most cases restricted to only concern the 10th decile of deprived areas but differ in categories of either municipality type or different moving patterns between municipality types. All models do, however, contain a categorical treatment variable, which means that the treatments are modeled in a multinomial logistic regression with the form of

$$p(z,t,\gamma) = \frac{\exp(z\gamma_t)}{1 + \sum_{k=1}^{q} \exp(z\gamma_k)}$$

where $p(z,t,\gamma)$ is the conditional probability that a person receives treatment $t$ on the condition that covariates $z$ and $\gamma$ are the parameters of the model.

$$sipw_i = \prod_{k=1}^{K} \frac{P\left(A_k = a_{ki} \middle| \bar{A}_{k-1} = \bar{a}_{(k-1)i,}, L_0 = l_0\right)}{P\left(A_k = a_{ki} \middle| \bar{A}_{k-1} = \bar{a}_{(k-1)i,}, \bar{L}_k = \bar{l}_{ki}\right)}$$

The denominator of the weight is the probability that a child is exposed to a specific neighborhood decile at year k conditional on past treatment and confounders. In each wave, inverse probability of treatment (IPT) weighting "balances" treatment assignment across prior confounders by giving more or less weight to children with covariate histories that are underrepresented or overrepresented in the current treatment group (Lund, 2019; Wodtke et al., 2011). In the weighted pseudopopulation, treatment in each wave is independent of prior confounders; that is, exposure to different neighborhood contexts behaves as if it is sequentially randomized with respect to observed covariates. Conditioning on confounder history is therefore no longer necessary, and an unadjusted model for the observed outcome can be fitted to the weighted observations to estimate the treatment effects of interest (Lund, 2019; Wodtke et al., 2011).

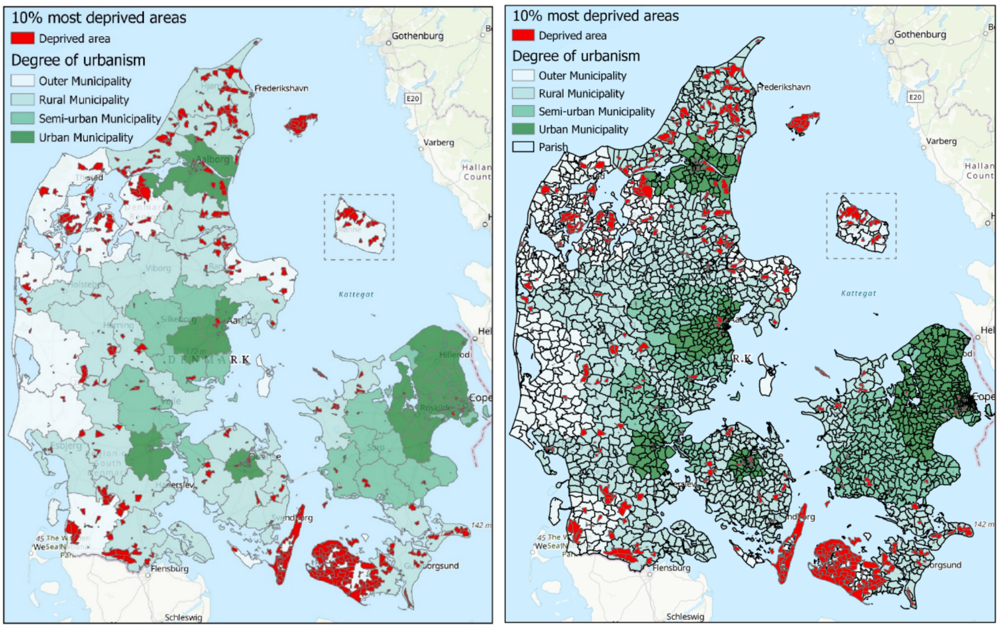
Because there is very little sample attrition due to death or migration, survival time censoring is very limited; however, because the outcome is educational attainment, this will be skewed by censoring in the early years. To account for this issue, censoring has been included in the inverse probability weighting (IPW) estimator by introducing a function in which the received treatment, be it accumulated deprivation or life course, $\tau \in \{0 \dots \tau_n\}$, is calculated by determining which potentially censored outcome is observed, $t = (1-\tau)\tilde{t}_0 + \tau\tilde{t}_1$, by creating a dummy indicator for treatment censoring where $c$ is 1 for the censored observations: $c = (1-\tau)(t_0 \geq t_c) + \tau(t_1 \geq t_c)$ (Aisenbrey & Fasang, 2010; Lund, 2019).

# 5 Analysis

## 5.1 Descriptive results

Looking at the general trends depicted in figure 3 in the traditional division of municipality types (left) and the same with added borders for parishes (right), it becomes clear that both parish and municipality are bad indicators of homogenous deprivation.

**Figure 3: Deprived areas divided into municipality types (left) and with added borders for parish (right)**




Morans I = 0,29, Z = $Morans\ I = 0.29^{***}, Z = 151{,}91$

Although a Global Morans I of 0.29[4] indicates strong evidence for clustering, it is interesting that the clustering, even at the local parish level, is less evident. Parts of Denmark such as Lolland, which is the large cluster of deprived areas in the southeastern part of Denmark, have a very high

[4] Both Global and Local Morans I calculated with the full index of deprivation, with all areas having values standardized between 0 and 1, from the least amount of deprivation to the most.

concentration, and by looking closer at the capitol Copenhagen, there seem to be parts that are clustered but also satellite areas of deprivation (see appendix B). Across the rest of Denmark are smaller clusters of deprivation, but usually only part of a parish is affected. This finding indicates that even though some municipalities and parishes with deprived areas cluster together to form a disproportionately large deprived area, the deprivation is local and does not affect the whole of any of the administrative areas except for the case of Lolland.

Table 1 depicts the percentage distribution of deprived areas compared to nondeprived areas in the four municipality types, and there is a massive overrepresentation of deprived areas in the outer municipality category.

Table 1: % deprived areas by municipality type in 2016

| Municipality type | Nondeprived areas | 10% most deprived areas |
|---|---|---|
| Outer municipality | 80.25 | 19.75 |
| Rural municipality | 89.68 | 10.32 |
| Semiurban municipality | 94.39 | 5.61 |
| Urban municipality | 93.98 | 6.02 |

Most outer-municipality types of areas have a much lower population density than in the city, so even though the raw count of deprived areas in the outer municipalities is almost three times as high as in an urban setting, the population count is about the same with a slight overrepresentation in urban areas. The mean inhabitant count of other municipality areas is 121 inhabitants, while the mean for urban areas is 451.

Looking at the cohort, 309,152[5] children were born between 1980 and 1986 in all area and municipality types. Even though the land mass covered by outer and rural municipalities as seen in figure 1 covers more than 70% of Denmark, less than 30% of the cohort is born there, while almost 58% of the cohort is born in urban municipalities. The same is the case for the full population, where the overall percentages are only slightly more leveled out due to an aging population living in the more outer and rural municipalities.

Looking at the overall trends with regard to mobility behavior across physical mobility patterns, it becomes clear that children and adolescents from outer and rural municipalities move farther than their semiurban and urban counterparts, while they move only slightly more often than the urban group. It is clear that the main change in moving behavior occurs after the age of 18, when the adolescents move from their family home and in many cases in the direction of educational institutions.

Table 2: Overall mean differences between municipality types (cohort)

| | Outer Municipality | | Rural Municipality | | Semiurban Municipality | | Urban Municipality | |
|---|---|---|---|---|---|---|---|---|
| Variable | Mean | SD | Mean | SD | Mean | SD | Mean | SD |
| Total distance moved | 322.62 | 296.29 | 309.85 | 279.00 | 239.78 | 217.82 | 264.34 | 231.31 |
| Distance moved age 0-7 | 23.40 | 60.13 | 23.64 | 60.57 | 19.64 | 48.23 | 19.56 | 51.05 |
| Distance moved age 8-12 | 25.42 | 67.06 | 24.51 | 63.06 | 20.38 | 50.92 | 19.22 | 52.04 |
| Distance moved age 13-17 | 27.46 | 72.44 | 24.76 | 64.82 | 21.01 | 52.83 | 18.97 | 53.82 |
| Distance moved age 18-24 | 110.47 | 131.81 | 104.59 | 126.86 | 79.59 | 100.47 | 89.55 | 108.10 |
| Distance moved age 25-30 | 111.11 | 129.38 | 108.98 | 129.41 | 82.18 | 97.57 | 94.62 | 116.94 |
| Mean N of moves | 1.18 | 1.58 | 1.16 | 1.48 | 1.18 | 1.44 | 1.01 | 1.26 |
| Mean education level | 2.42 | 1.47 | 2.67 | 1.64 | 2.70 | 1.66 | 3.90 | 2.17 |
| Mean % of unemployment | 0.07 | 0.18 | 0.06 | 0.18 | 0.06 | 0.18 | 0.07 | 0.18 |
| N | 19235 | | 71901 | | 38964 | | 179052 | |

Note: Distances measured in kilometers, education in months and unemployment in days.

[5] With about 4% missing due to untimely death or migration, which are not present in the sample.

Not surprisingly, educational attainment is much higher for individuals who grew up in the city because of the location of universities in Denmark. Without living in an urban area, disregarding those few who are willing to commute hundreds of kilometers each day as students, it is almost impossible to attain a master's degree. Thus, education is much higher for those living in urban areas. Surprisingly, both unemployment and income levels are very similar.

The above table mainly focuses on an overview of municipality types and not on the dynamics of mobility and deprivation. Table 3 shows the average differences between different types of life trajectories and motilities, where each individual in the cohort is divided into four overall sequences. The individual is placed in a starting municipality by place of birth and by living the majority of their childhood and adolescence in that municipality type. After the age of 18, they chose to either stay in their main childhood municipality (stayers) or move to either of the other three. This approach is used for the full cohort and for the reduced cohort that lived in deprivation.

Considering the differences between income at age 30 and the different trajectories, there are a few interesting findings.

**Table 3: Mean difference between life trajectories (cohort)**

| Type of move | Income | | Education | | Unemployment | | N | |
|---|---|---|---|---|---|---|---|---|
| | µ all | µ dep. | µ all | µ dep. | µ all | µ dep. | N all | N dep |
| Outer no moves | 299906.2 | 290442.1 | 163.8 | 159.3 | 21.2 | 26.4 | 12032 | 1886 |
| Outer to rural | 304347.6 | 288340.2 | 170.1 | 164.9 | 21.7 | 29.6 | 6062 | 847 |
| Outer to semi | 315912.9 | 296114.9 | 173.4 | 169.2 | 20.2 | 32.2 | 1465 | 172 |
| Outer to urban | 329673.8 | 305442 | 186.7 | 183.4 | 20.9 | 24.6 | 11430 | 1176 |
| Rural to outer | 292010.3 | 269452.3 | 163.0 | 157.4 | 24.9 | 31.6 | 3869 | 577 |
| Rural no moves | 301844 | 279760 | 166.0 | 160.8 | 22.8 | 28.8 | 48196 | 6794 |
| Rural to semi | 324918.9 | 285904.1 | 170.1 | 165.5 | 18.7 | 23.3 | 6371 | 858 |

| | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| Rural to urban | 322847.8 | 299162 | 186.1 | 180.2 | 22.2 | 28.4 | 34382 | 4012 |
| Semi to outer | 267205.8 | 255109.6 | 157.5 | 156.1 | 32.3 | 30.6 | 695 | 140 |
| Semi to rural | 294354.7 | 281229.6 | 164.0 | 160.9 | 22.8 | 29.8 | 5081 | 633 |
| Semi no moves | 306638.1 | 275728.8 | 164.8 | 159.1 | 22.7 | 27.9 | 20755 | 2746 |
| Semi to urban | 319935.1 | 295844.1 | 183.9 | 179.7 | 22.1 | 26.5 | 16372 | 1347 |
| Urban to outer | 264723.7 | 241448.9 | 155.6 | 151.0 | 30.0 | 43.5 | 1538 | 413 |
| Urban to rural | 285580.4 | 259715.8 | 162.9 | 155.4 | 26.0 | 32.2 | 7876 | 1100 |
| Urban to semi | 297385.7 | 269033.7 | 163.7 | 157.2 | 23.3 | 34.3 | 7404 | 784 |
| Urban no moves | 310210.6 | 280576.3 | 175.3 | 169.1 | 23.9 | 30.5 | 101702 | 1886 |

First and foremost, when looking at the full cohort, it is clear that the higher income groups are the ones that move towards the city or stay in the city. In general, moving to any more urbanized municipality type than the childhood municipality seems to have a uniformly positive effect. It is, however, interesting that the groups with the highest spikes in income at age 30 are the individuals who move from either an outer or a rural municipality to an urban municipality. The same trend can be seen in educational attainment and unemployment, where moving towards urban areas from more stratified municipalities results in significantly higher education and lower unemployment. Some of these findings can be explained by the almost solely urban locations of Danish universities, but some selection must be involved in the group moving from outer and rural municipalities because educational attainment is much higher for this group. Moving from semi-urban or urban municipalities to outer or rural municipalities does, however, have some very negative effects on all parameters, but because N is very small, it is hard to determine these effects precisely.

When looking at the deprived life trajectories, a different story emerges. All municipalities gain higher income, higher education and lower unemployment when moving to the city, but considering the stayers, it is by far the people who live and stay in the outer municipalities who have the highest income and the lowest amount of unemployment. This finding does not take into account possible confounders or underlying tendencies that can explain this trend, a discussion of which will follow below.

## 5.2 Do you like the way you move?

Here, subgroups of the cohort are compared. Table 4 refers to everyone in the cohort who was born and grew up in a given municipality. This excludes people who moved across municipality types during childhood and adolescence and those who stayed behind from birth to the age of 30. The main goal of the first analysis is to understand what it means to grow up in one type of municipality and what the effect of moving to the city after the age of 17 has on unemployment, income and educational attainment disregarding deprivation. In other words, what does a move to the city do to a given person's later life trajectory?

The reference group is people who move from an outer municipality to the city after the age of 17. All effects are treatment effects where the probability of treatment is equal, as described earlier.

Looking at unemployment, the potential outcome of the reference group is 19 days of unemployment during a year. All others who move to the city experience an increase in unemployment, but even though it is significant, the practical implication of 4 more days of unemployment than the reference group is nominal.

More interesting is the difference in income. By using the log of income, the interpretation of income becomes a percentage difference in income levels. While there are no differences between moving to the city from outer or rural municipalities, the effect becomes highly significant when comparing outer municipality movers to those who move from semi urban municipalities to the city and those who stay in the city. On average, the difference in income levels between outer municipality movers and semi-urban movers is 4.3%-points in income, while the difference between outer municipality movers and those staying in the city is more than 8%-points.

Table 4: IPWRA[6] on different move sequences on socioeconomic factors at age 30 for all types of neighborhoods[7]

| VARIABLES | Unemployment (days) | Income (ln) | Education (months) |
|---|---|---|---|
| From outer to city (ref.) | - | - | - |
| | - | - | - |
| From rural to city vs. ref | 1.640* | -0.0114 | -0.864** |
| | (0.846) | (0.0100) | (0.435) |
| From semi to city vs. ref | 2.592*** | -0.0432*** | -3.677*** |
| | (0.904) | (0.0109) | (0.387) |
| Stay in city vs. ref | 4.566*** | -0.0888*** | -11.03*** |
| | (0.761) | (0.00959) | (0.332) |
| PO mean at outer | 19.11*** | 12.58*** | 188.7*** |
| | (0.719) | (0.00908) | (0.317) |
| Observations | 132,567 | 133,642 | 133,645 |

Robust standard errors in parentheses
*** p<0.01, ** p<0.05, * p<0.1

The same trend can be seen in educational attainment. While the difference between outer and rural movers is minuscule, the differences become much larger between other categories. The difference between outer movers and city stayers is 11 months of completed education more for the outer movers, and even though the inhabitants who stay in the city have easier access to education, the difference could be explained by a desire to seek

[6] Inverse probability weighted regression adjustment

[7] The treatment assignment probability is fitted with multinomial logistic regression, while the outcome model is fitted with a linear model. This is the case for this model and all subsequent models, and thus the coefficient shown is β

education, and thus the move to the city is to gain access. In that sense, the city stayers are a less selected group than the people who move to the city.

Although this difference can be explained, it is harder to make the same assumption about the difference between outer- and rural movers. The move is about the same in distance, but compared to the rural movers, the effect is almost four months of education more if one moves from an outer municipality to the city. It is clear, however, that there is a difference in later life chances when comparing where one grows up.

Looking more directly at deprivation, the same comparison is made for movers, isolating those who grew up in deprived neighborhoods in their respective municipalities. One issue to consider is the effect of growing up in an outer municipality compared to other starting points and moving to the city; another is living in deprived neighborhoods in these municipalities. Do these people experience the same effects as the full cohort or is it worse to grow up in deprivation in the outskirts than in the city?

Table 5: IPWRA on different move sequences on socioeconomic factors at age 30 for individuals growing up in deprived neighborhoods

| VARIABLES | Unemployment (days) | Income (ln) | Education (months) |
|---|---|---|---|
| From outer to city (ref.) | - | - | - |
| | - | - | - |
| From rural to city vs. ref | 2.115 | -0.0226 | -3.905*** |
| | (2.840) | (0.0230) | (1.049) |
| From semi to city vs. ref | 1.175 | -0.0578* | -3.707*** |
| | (3.299) | (0.0308) | (1.257) |
| Stay in city vs. ref | 4.383 | -0.102*** | -12.06*** |
| | (2.690) | (0.0218) | (1.001) |
| PO mean at outer | 24.88*** | 12.52*** | 184.9*** |
| | (2.523) | (0.0201) | (0.938) |
| Observations | 12,106 | 12,214 | 12,198 |

Robust standard errors in parentheses
*** p<0.01, ** p<0.05, * p<0.1

Table 5 mirrors the previous table but only for those who grew up in deprivation in any type of neighborhood but moved to the city. Overall, the effects are comparable, but there are a few interesting differences. The potential outcome mean of unemployment is 6 days more for the outer movers compared to all outer movers from the previous table, and although the trend is the same, there are no significant differences. This finding is mainly due to the lowering of N.

More interesting is educational attainment and income. The effect of moving from deprivation in outer municipalities compared to all other types of moves is better than for the overall cohort, with an almost 6%-point higher income than semi-movers and 10%-points higher income than city stayers. Outer movers acquire approximately 4 months more education than both rural and semi movers and more than a full year more than city stayers. Comparing the potential outcome means of outer movers from the previous model, none of the change can be ascribed to a lowering of the income levels or educational attainment of the full cohort, which implies that the effects are more prominent for outer movers in general.

The most logical explanation is drive. While the difference between city stayers and outer movers can be explained by selection in terms of the desire to move to the city to gain education, the difference between outer and rural/semi is less logical. All of these people move to the city, but they experience a clear, significant difference in outcomes. It could be that young people moving from outer municipalities who decide to move to the city must make a much more informed decision compared to other movers because the change of place is by far the biggest in distance and in concept. Those who chose to make that move are more driven by a specific goal, while those much closer to the urban areas might be drawn to the city for numerous reasons.

### 5.3 Stuck in the middle (with you)

One life trajectory is to move from the known childhood home to new places, while another is to stay in place. The following analysis is focused on those who stayed behind – not in deprivation but in the same type of area where they grew up.

Table 6 is based on the same treatment effects methodology as earlier but now comparing those who stayed in outer municipality types both as children and as adults and never moved away to similar counterparts in different municipalities. In other words, what happens to those who stay behind?

Table 6: IPWRA on people staying in birth municipality on socioeconomic factors at age 30 for all types of neighborhoods

| VARIABLES | Unemployment (days) | Income (ln) | Education (months) |
|---|---|---|---|
| Outer - staying (ref.) | - | - | - |
| | - | - | - |
| Rural - staying vs. ref | 1.454* | -0.0427* | 2.332*** |
| | (0.760) | (0.0255) | (0.430) |
| Semi - staying vs. ref | 2.483*** | -0.0266** | 1.330*** |
| | (0.815) | (0.0129) | (0.416) |
| City - staying vs. ref | 0.882 | -0.065*** | 10.82*** |
| | (2.264) | (0.203) | (1.345) |
| PO mean at outer | 18.33*** | 12.56*** | 166.4*** |
| | (0.639) | (0.00847) | (0.313) |
| Observations | 147,265 | 147,760 | 147,762 |

Robust standard errors in parentheses
*** p<0.01, ** p<0.05, * p<0.1

Considering first the full cohort in table 6, the differences from the movers presented in the earlier models are subtle. Unemployment is slightly higher among rural and semi-urban stayers compared to the outer stayers, while the urban stayers have an insignificantly higher number of days unemployed. However, it does appear that even those who stay behind in outer municipalities are prospering significantly more in regard to income. The income level for stayers in urban settings is 6.5% lower compared to the outer stayers. The differences become less evident the closer to the outer municipalities the cohort lives.

Educational attainment is less surprising. Because of the placement of universities, there is no access to master's-level education in the outer municipalities, and thus, there is no real possibility to gain education higher than lower tertiary. Staying in rural areas results in a potential outcome mean of 166 months of education, while the city counterpart has almost 11 months more of fulltime education. The other parts have less of an advantage but still a higher level of education.

The following analysis is based on those who grew up in deprived neighborhoods and stayed in the same municipality type from birth to the age of 30, meaning that they could potentially leave deprivation behind but stay in the same type of municipality. More than 75% of the cohort that grew up in deprivation and stayed in the same municipality type also stayed in the highest decile of deprivation.

Comparing the potential outcome means between the full cohort and the deprived cohort, the difference in unemployment is five days, with more than 23 days of unemployment on average per year for the cohort that grew up in a deprived outer municipality area. Although the difference between the full group and the deprived group is large, the number of days of unemployment for urban residents in deprivation is the largest effect measured of all groups with more than 7 days.

Table 7: IPWRA on people staying in birth municipality on socioeconomic factors at age 30 for individuals growing up in deprived neighborhoods

| VARIABLES | Unemployment (days) | Income (ln) | Education (months) |
|---|---|---|---|
| Outer - staying (ref.) | - | - | - |
| | - | - | - |
| Rural - staying vs. ref | 3.479** | -0.162** | 6.445*** |
| | (1.938) | (0.0768) | (2.087) |
| Semi - staying vs. ref | 4.102** | -0.180** | 6.099*** |
| | (0.0106) | (0.0774) | (2.111) |
| City - staying vs. ref | 7.042*** | -0.220*** | 14.62*** |
| | (0.0102) | (0.0769) | (2.070) |
| PO mean at outer | 23.588*** | 12.63*** | 157.4*** |
| | (0.0100) | (0.0766) | (2.058) |
| Observations | 27,691 | 27,805 | 27,746 |

Robust standard errors in parentheses
*** p<0.01, ** p<0.05, * p<0.1

The income difference between urban and outer stayers has the same tendency as the unemployment differences, where the outer stayers experience an increase in income of 22% compared to the urban stayers. This indicates that staying in deprivation has a very different effect depending on where one grows up. Although the educational attainment is equally skewed to where the urban stayers have more than 1 year and 3 months more fulltime education than their outer counterparts, this does not pay off in employment and income.

# 6 Discussion & conclusion

Much of the 21st century literature on mobility and place of living is occupied with movers compared to stayers in a way that suggests a higher level of mobility is a requirement of existence in the modern world (Clampet-Lundquist & Massey, 2008; Clark, 2017; Savage & Egerton, 1997; Woelfel & Murero, 2004); as Zygmunt Bauman put it in more dramatic terms, “The new elite is not defined by any locality: it is truly and fully exterritorial” (Bauman, 2001). Although Baumans’s point is less tangible than simple moving patterns and physical mobility, the point is still interesting when comparing different groups of society, specifically where they grow up and in what setting. When the elite is considered exterritorial, what then happens to those who are less mobile? What happens, when you are born and raised in deprivation but the deprivation is dispersed by geography and different spatial settings?

This analysis has revolved around two subjects: deprivation and location. When one is left out, the other loses its nuances and becomes a matter of rural as a yes/no question or deprived as a yes/no question. Considering a cohort’s place of birth and later life outcomes is not a new approach to understanding the effects of neighborhood deprivation (Erola, Jalonen, & Lehti, 2016; Hedman, Manley, van Ham, & Östh, 2015; Jackson & Mare, 2007; Tanner, Davies, & O’Grady, 1999), and neither is the use of different scales of measurement when considering *where* the deprivation is located (Merlo et al., 2013; Östh et al., 2015; Reibel, 2007; Sridharan et al., 2007). However, most scholars never consider the fact that deprivation is something rooted in a geographical space.

This rootedness almost becomes a mantra in this setting because we need to ask how we understand rootedness and what it means to be rooted in a specific geographical space. The first part of the analysis focuses on those who are not rooted: those who move, either in general or move far away from deprivation. The effect of moving from outer and rural areas to more urban areas is much larger than for those who decide to stay. This, as discussed, is most likely due to this group being a selected group of individuals, but the fact that the outer movers experience the largest effects is interesting. There is no doubt that the movers from deprived neighborhoods have a much more disadvantaged starting point. All potential outcome means indicate a contrast with the full group. However, in almost all cases, it is the outer municipality group that experiences the largest effects compared to the other groups, with regard to both those who move and those who stay behind.

The differences when comparing the stayers to the movers are interesting. Not only do the full cohort and the deprived movers from the outer and rural municipalities benefit more than their urban and semi-urban counterparts, the outer stayers fare much better when considering both income and unemployment rate and especially when compared to all other groups that grew up in deprivation. This finding indicates two distinct results. First, it indicates that deprivation is not equal even when considering a rather broad spectrum such as degree of urbanization, and second, it indicates that later life outcomes are reliant on where one experienced deprivation.

This conclusion raises an interesting question: is education always necessary? The possibilities of the city and perhaps of the new urban world are often portrayed as centered on knowledge and education, but moving to the city might not always be the best way to change the life chances of those who do not choose to pursue education.

# 7 Literature


Aisenbrey, S., & Fasang, A. E. (2010). New Life for Old Ideas: The "Second Wave" of Sequence Analysis Bringing the "Course" Back Into the Life Course. *Sociological Methods & Research*, *38*(3), 420–462.

Atkinson, R., & Kintrea, K. (2004). "Opportunities and despair, it's all in there": Practitioner experiences and explanations of area effects and life chances. *Sociology*, *38*(3), 437–455.

Bauman, Z. (2001). *Community - Seeking Safety in an Insecure World*. Cambridge: Polity Press.

Bibby, P. (2013). *Urban and Rural Area Definitions for Policy Purposes in England and Wales: Methodology (v1.0)*.

Bower, K. M., Thorpe, R. J., Rohde, C., & Gaskin, D. J. (2014). The intersection of neighborhood racial segregation, poverty, and urbanicity and its impact on food store availability in the United States. *Preventive Medicine*, *58*(1), 33–39.

Carlsson, A. C., Li, X., Holzmann, M. J., Ärnlöv, J., Wändell, P., Gasevic, D., … Sundquist, K. (2017). Neighborhood socioeconomic status at the age of 40 years and ischemic stroke before the age of 50 years: A nationwide cohort study from Sweden. *International Journal of Stroke*, *12*(8), 815–826.

Clampet-Lundquist, S., & Massey, D. S. (2008). Neighborhood Effects on Economic Self-Sufficiency: A Reconsideration of the Moving to Opportunity Experiment. *American Journal of Sociology*, *114*(1), 107–143.

Clark, W. A. V. (2017). Who Moves Into what Kinds of Neighbourhoods: Spatial Sorting and Integration. *Tijdschrift Voor Economische En Sociale Geografie*.

Crenshaw, E. (1992). Cross-National Determinants of Income Inequality: A Replication and Extension Using Ecological-Evolutionary Theory. *Social Forces*, *71*(2), 339–363.

Davidson, O. G. (1996). Broken heartland: The rise of America's rural ghetto, 220.

Entwisle, B., Rindfuss, R. R., Walsh, S. J., Evans, T. P., & Curran, S. R. (1997). Geographic information systems, spatial network analysis, and contraceptive choice. *Demography*, *34*(2), 171–187.

Erola, J., Jalonen, S., & Lehti, H. (2016). Parental education, class and income over early life course and children's achievement. *Research in Social Stratification and Mobility*, *44*, 33–43.

European Comission. (2007). *Regions in the European Union. Statistics*.

European Commission. (2013). *Agricultural Policy Perspectives Brief N°5*.

European Commission. (2017). *My Region , My Europe , Our Future*.

Feld, S. L. (1981). The Focused Organization of Social Ties. *The American Journal of Sociology*, *86*(5), 1015–1035.

Ferreira, M. A. R., Holan, S. H., & Bertolde, A. I. (2011). Dynamic multiscale spatiotemporal models for Gaussian areal data. *Journal of the Royal Statistical Society. Series B, Statistical Methodology*, *73*(5), 663–688.

Friedrichs, J., Galster, G., & Musterd, S. (2003). Neighbourhood effects on social opportunities: the European and American research and policy context. *Housing Studies*, *18*(6), 797–806.

Gage, T. B., Dyke, B., & Maccluer, J. W. (1986). Estimating Mortality Level for Small Populations: An Evaluation of a Pair of Two-Census Methods. *Population Studies*, *40*(2), 263–273.

Galster, G. C. (2010). The Mechanism (s) of Neighborhood Effects Theory, Evidence, and Policy Implications. *ESRC Seminar*, 1–32.

Garner, C. L., & Raudenbush, S. W. (1991). Neighborhood Effects on Educational Attainment: A Multilevel Analysis. *Source: Sociology of Education*, *64*(4), 251–262.

Hedman, L., Manley, D., van Ham, M., & Östh, J. (2015). Cumulative exposure to disadvantage and the intergenerational transmission of neighbourhood effects. *Journal of Economic Geography*, *15*(1), 195–215.

Jackson, M. I., & Mare, R. D. (2007). Cross-sectional and longitudinal measurements of neighborhood experience and their effects on children. *Social Science Research*, *36*(2), 590–610.

Johnelle Sparks, P., Sparks, C. S., & Campbell, J. J. A. (2013). An application of Bayesian spatial statistical methods to the study of racial and poverty segregation and infant mortality rates in the US. *GeoJournal*, *78*(2), 389–405.

Jones, M., & Pebley, A. R. (2014). Redefining Neighborhoods Using Common Destinations: Social Characteristics of Activity Spaces and Home Census Tracts Compared. *Demography*, *51*(3), 727–752.

Krieger, N., Feldman, J. M., Waterman, P. D., Chen, J. T., Coull, B. A.,

& Hemenway, D. (2017). Local Residential Segregation Matters: Stronger Association of Census Tract Compared to Conventional City-Level Measures with Fatal and Non-Fatal Assaults (Total and Firearm Related), Using the Index of Concentration at the Extremes (ICE) for Racial, Econ. *Journal of Urban Health*, *94*(2), 244–258.

Lagerlund, M., Merlo, J., Vicente, R. P., & Zackrisson, S. (2015). Does the neighborhood area of residence influence non-attendance in an urban mammography screening program? A multilevel study in a Swedish city. *PLoS ONE*, *10*(10).

Law, J., Quick, M., & Chan, P. W. (2015). Analyzing Hotspots of Crime Using a Bayesian Spatiotemporal Modeling Approach: A Case Study of Violent Crime in the Greater Toronto Area. *Geographical Analysis*, *47*(1), 1–19.

Lee, J., Árnason, A., Nightingale, A., & Shucksmith, M. (2005). Networking: Social capital and identities in European rural development. *Sociologia Ruralis*, *45*(4), 269–283.

Logan, J. R., Spielman, S., Xu, H., & Klein, P. N. (2011). Identifying and Bounding Ethnic Neighborhoods. *Urban Geography*, *32*(3), 334–359.

Lund, R. L. (2018). From the dark end of the street to the bright side of the road – automated redistricting of areas using physical barriers as dividers of social space. *Methodological Innovations*, *11*(3), 1–15.

Lund, R. L. (2019). Moving to Prosperity? The Effect of Prolonged Exposure to Neighborhood Deprivation. *Scandinavian Journal of Educational Research*, *0*(0), 1–17.

Malmberg, B., Andersson, E., & Östh, J. (2011). How does segregation vary between different urban areas? Using a k-nearest neighbor approach to segregation measurement. In *ENHR Conference 2011* (p. 16).

Merlo, J., Ohlsson, H., Chaix, B., Lichtenstein, P., Kawachi, I., & Subramanian, S. V. (2013). Revisiting causal neighborhood effects on individual ischemic heart disease risk: A quasi-experimental multilevel analysis among Swedish siblings. *Social Science and Medicine*, *76*(1), 39–46.

Milbourne, P. (2007). Re-populating rural studies: Migrations, movements and mobilities. *Journal of Rural Studies*, *23*(3), 381–386.

Ministeriet for Fødevarer Landbrug og Fiskeri. (2011). *Landdistriktredegørelse 2011*.

Ministry of Transportation. (2018). Liste over ghettoområder pr. 1. december 2017 - Transportministeriet.

Mulugeta, E. (2004). Swimming against the Tide: Educational Problems and Coping Strategies of Rural Female Students in Ethiopia. *Eastern Africa Social Science Research Review*, *20*(2), 71–97.

Ocana-Riola, R., Sanchez-Cantalejo, C., & Fernandez-Ajuria, A. (2006). Rural Habitat and Risk of Death in Small Areas of Southern Spain. *Social Science & Medicine*, *63*(5), 1352–1362.

Östh, J., Clark, W. A. V., & Malmberg, B. (2015). Measuring the scale of segregation using k-nearest neighbor aggregates. *Geographical Analysis*, *47*(1), 34–49.

Petrovi, A., Ham, M. Van, & Manley, D. (2018). Multiscale measures of population: Within- and between-city variation in exposure to the sociospatial context. *Annals of the American Association of Geographers*, *108*(4), 1057–1074.

Petrović, A., Ham, M. van, & Manley, D. (2017). Size matters? Exploring scale for neighbourhood effects. In *AGILE conference 2017*. Springer. Springer.

Potter, C., Walker, G., & Keen, B. (2012). Engaging fathers from disadvantaged areas in children's early educational transitions: A UK perspective. *Journal of Early Childhood Research*, *10*(2), 209–225.

Reibel, M. (2007). Geographic information systems and spatial data processing in demography: A review. *Population Research and Policy Review*, *26*(5–6), 601–618.

Rotolo, T., & Tittle, C. R. (2006). Population Size, Change, and Crime in U.S. Cities. *Journal of Quantitative Criminology*, *22*(4), 341–367.

Sampson, R. J. (2012). *Great American City - Chicago and the Enduring Neighborhood Effect* (1.). Chicago: The University of Chicago Press.

Sampson, R. J., Morenoff, J. D., & Gannon-Rowley, T. (2002). Assessing "Neighborhood Effects": Social Processes and New Directions in Research. *Annual Review of Sociology*, *28*(2002), 443–478.

Savage, M., & Egerton, M. (1997). Social Mobility, Individual Ability and the Inheritance of Class Inequality. *Sociology*, *31*(4), 645–672.

Sharkey, P., & Elwert, F. (2011). The Legacy of Disadvantage: Multigenerational Neighborhood Effects on Cognitive Ability. *American Journal of Sociology*, *116*(6), 1934–1981.

Sridharan, S., Tunstall, H., Lawder, R., & Mitchell, R. (2007). An exploratory spatial data analysis approach to understanding the relationship between deprivation and mortality in Scotland. *Social Science & Medicine*, *65*(9), 1942–1952.

Tanner, J., Davies, S., & O'Grady, B. (1999). Whatever happened to yesterday's rebels? Longitudinal effects of youth delinquency on education and employment. *Social Problems*, *46*(2), 250–274.

Voss, P. R., Long, D. D., Hammer, R. B., & Friedman, S. (2006). County child poverty rates in the US: a spatial regression approach. *Population Research and Policy Review*, *25*(4), 369–391.

Waterman, W. C., Park, R. E., Burgess, E. W., & Boyd, W. K. (2006). *The City.The Story of Durham.* (M. Janowitz, Ed.), *Social Forces* (Vol. 5). Chicago: The University of Chicago Press.

White, M. J., Kim, A. H., & Glick, J. E. (2005). Mapping Social Distance: Ethnic Residential Segregation in a Multiethnic Metro. *Sociological Methods & Research*, *34*(2), 173–203.

Wodtke, G. T., Harding, D. J., & Elwert, F. (2011). Neighborhood effects in temporal perspective: The impact of long-term exposure to concentrated disadvantage on high school graduation. *American Sociological Review*, *76*(5), 713–736.

Woelfel, J., & Murero, M. (2004). Spaces and Networks: Concepts for Social Stratification. *Research in Social Stratification and Mobility*, *22*(04), 57–71.

Wooldridge, J. M. (1995). Selection corrections for panel data models under conditional mean independence assumptions. *Journal of Econometrics*, *68*(1), 115–132.

Wooldridge, J. M. (2010). *Econometric Analysis of Cross Section and Panel Data*. *MIT Press* (2nd ed., Vol. 58).

# 8 Appendices

## 8.1 Appendix A

Pseudocode:

### Code

**Data:** J=initial shape-file type data with rough defined areas (j),
SG=Square Grid containing inhabitants per 100m2 (s)
**Result:** Shapefile with smaller areas with inhabitant count set to X¡
and borders moved to fit square grid type

```
/* Code begins */
for s in j:
    if squares are primarily within defined areas:
        assign square to area=j in file J;
    else:
        disregard square in dictionary
if sqare j = area j:
    update data J and move borders to fit square;
    count = number of inhabitants for square to area j in file J;
else:
    pass;
for j in J:
    if count of residents < x:
        proc=0;
    else:
        proc=1;
if proc==0:
    define(borderproc)
       percentage border shared with neighboring areas;
else:
    pass;

define(costfunction);
```

The cost function is defined by

$$y = borderproc * inhabitantcount * geosize$$

Where each type is set in a specific way to prioritize small, very close neighborhoods. These are set as:

borderproc:

$$0 = \infty, 100 = 1$$

inhabitantcount:

$$< 100 = \infty, 100 = 1, 100 + x_1$$

where $x_1$ = each additional inhabitant after potential merge

geosize:

$$0 = 1, 0 + x_2$$

where $x_1$ = each additional $\mathrm{m}^2$ after potential merge

while the gradient descent parameters are set as

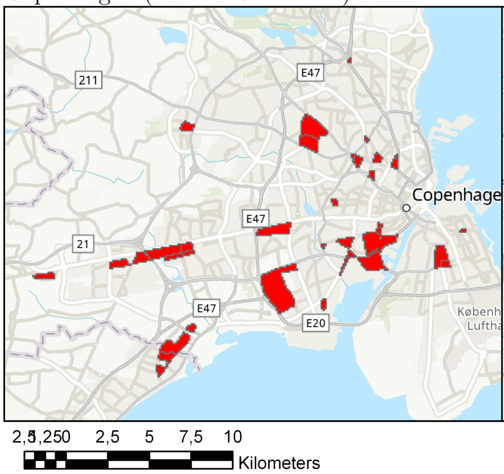
$$\theta = 1$$

$$\alpha = .1$$

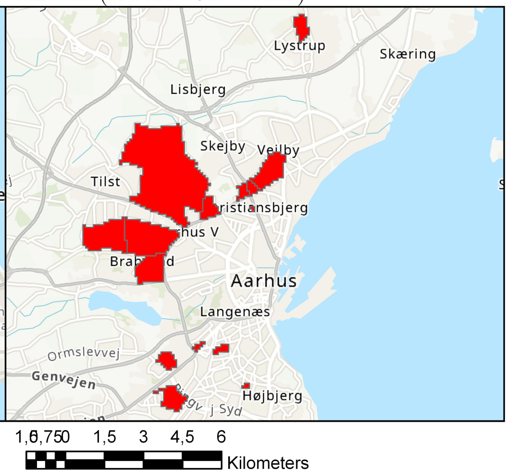
```
define(runlist)
  while proc==0:
    for j in J:
      write merge order list to largest shared borders first for each j;
      write inhabitant size for all possible merges;
      count number of merges for all possible merges
    if proc==1:
      pass;

/* recursion starts here                                  */
with(costfunction)
  for j in runlist:
    run recursion to minimize cost of merges;
    end when cost function stalls
end
```

## 8.2 Appendix B

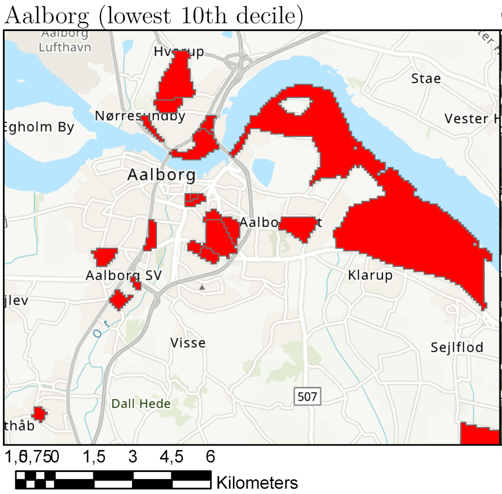
Copenhagen (lowest 10th decile)



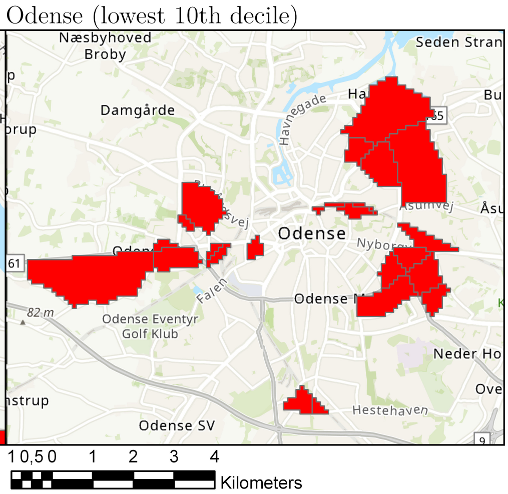

Aarhus (lowest 10th decile)



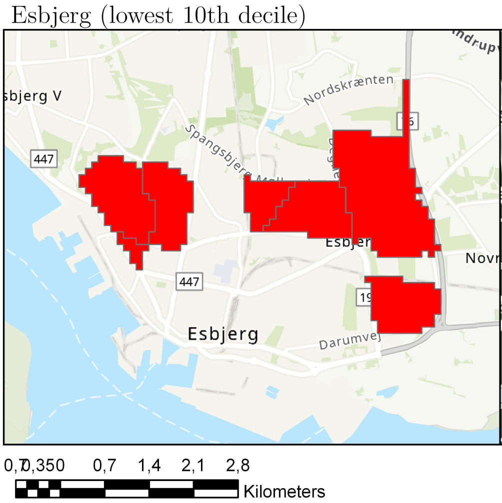

Aalborg (lowest 10th decile)



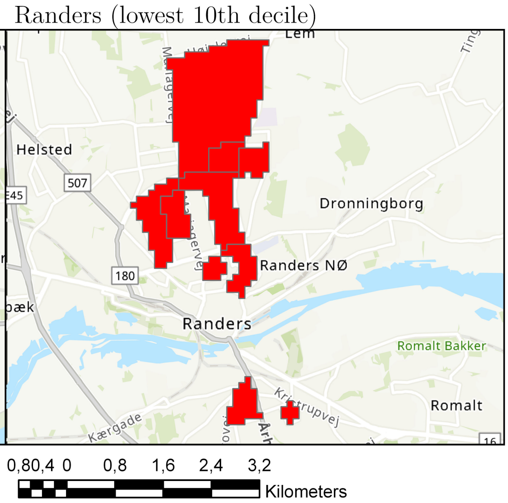

Odense (lowest 10th decile)




Esbjerg (lowest 10th decile)




Randers (lowest 10th decile)



NORDIC JOURNAL
OF RELIGION AND SOCIETY
RESEARCH PUBLICATION

UNIVERSITETSFORLAGET

Volume 32, No. 1-2019, p. 55–70
ISSN online: 1890-7008
https://doi.org/10.18261/issn.1890-7008-2019-01-04

# Social Geographical Patterns in Membership of the Established Church in Denmark

**Rolf Lyneborg Lund**
Aalborg University, Denmark
rolfll@socsci.aau.dk

**Anja Jørgensen**
Aalborg University, Denmark
anjaj@socsci.aau.dk

**Ole Preben Riis**
University of Agder, Norway
olepriis@gmail.com

## ABSTRACT

The Church of Denmark was established in 1849 and is regarded as a pillar of traditional national identity. This status is being challenged by a steady decline in membership in recent decades. The Capital area is especially prone to low membership rates, and this regional pattern remains when the analysis controls for income and education. Furthermore, the local membership rate is also related to affiliation to the neighbourhood. Our detailed analysis is based on public register data on the individual level combined with geographical mapping information. Denmark is thereby divided into micro-aggregated areas in order to locate varying church membership rates. While some local variation can be explained by socioeconomic status, some also depend on residential belonging to different types of local communities. Our analysis points to the sense of attachment to place of residence as a major factor in explaining membership of the established church.

Keywords:
Geography of religion, secularization, church membership rates, spatial inequality

## INTRODUCTION

The present study investigates local variations in church membership in Denmark. The analysis is restricted to the established church for both practical and theoretical reasons: It is the only church of which local membership data are available, and many specific factors influence the membership level of each religious minority community. In order to eliminate the possible influence of migration on church membership, the analysis is based on Danish citizens only.

The Danish established church is an institution with a dual character. On the one hand, it is a religious institution based on the Lutheran confession of Augsburg. One the other, it is a social institution, a framework for collecting a broad membership with divergent personal beliefs. The social dimension is recognized by both church officials and common members. It explains why commitment to the church remains relatively high, despite reservations about its dogma and a general tendency towards a privatized form of religiosity (Andersen and Riis 2002). Survey studies even indicate a slight increase in public trust in the church's ability to give a response to the spiritual needs of its members (Andersen and Lüchau 2011).

Membership of the established church often expresses a civil religious stance rather than a confessional Lutheran one (Sundback 2000; Riis 2000, 2008). Grace Davie has characterized the typical Danish stance to the established church as "belonging without believing" in contrast to the British stance of "believing without belonging" (Davie 1999).

Denmark is often described as a homogeneous society. However, in-depth studies reveal marked local variations (Andersen 2005; Damm, Schultz-Nielsen and Tranæs 2006)[1] and local patterns of community attachment, which relate to church membership. To this extent, it seems relevant to investigate how this local variation affects different aspects of human life and especially the church membership rates in a Danish context.

### Church membership

One of the historically integrating institutions is the established church, the Evangelical-Lutheran church, which is described by the constitution of 1849 as "the church of the people of Denmark," and the state is obliged to support it. The constitution simultaneously secures personal religious freedom. A clear majority remains as church tax-paying members, despite infrequent attendance at services and a widespread scepticism concerning its confession (Riis 2011; Andersen and Riis 2002).

The question about membership of the established church was addressed in the 1990 survey on Religious and Moral Pluralism (Gustafssson and Petterson 2000). The most common reasons given referred to the church ritual services. Furthermore, many referred to the civil religious function of the established church: to protect the cultural heritage, because it historically integrates the people, and because it is associated with Danish citizenship, and because it is open to all, even to people with a weak faith. Most of the members reject the position that there is only one true religion (Gustafssson and Petterson 2000, 209). Therefore, membership is mostly based on the ritual services and the civil religious function of the church.

1. We prefer to refer to the original sources rather than to secondary presentations in English.

Nearly all children used to be formerly baptised in the established church. However, the rate of baptism has declined markedly since the turn of the new millennium. This trend could be explained sociologically by a general change in values towards autonomy and authenticity in late modern societies (Inglehart 1990; Taylor 2007). For Denmark, the European Values Study indicates that independence and personal choice are stressed as values (Gundelach and Riis 1992; Gundelach 2002, 2011).

Empirical studies of the church statistics have pointed to urbanization as an explanatory factor (Riis 2012). Copenhagen is most of all characterized by its heterogeneity and internal contrasts. It contains districts where the most affluent and the most deprived Danes live. It is a patchwork of local communities. Some parts of Copenhagen are characterized by a high rate of immigrants while others are characterized by a high rate of people registered as ethnically Danish. In some neighbourhoods, the inhabitants are highly cosmopolitan, having regular connections to regions outside Copenhagen, while the links to other parts of the capital are weaker. Few inhabitants from the affluent neighbourhoods regularly visit the poorer districts.

Neighbourhoods in the non-capital areas of Denmark are also characterized by social differences. Some have declining populations, some attract new residents, and some vacillate between a small population at winter and a large one in the summer. Some rural communities have a high rate of mobility and thereby a low degree of local rootedness.

The declining membership rates cannot be explained by economic or educational factors alone. The type of community also has an effect. People with a similar socio-economic status may react differently, according to the type of local community. The affluent districts in the capital area form an enclosed elitist culture that stresses cosmopolitanism and autonomy, including the personal choice of one's religion (Andersen and Riis 2002). People with a high socio-economic status in areas outside the capital are more embedded in the local tradition.

The decline of membership rates that seems non-uniform between local communities and not constant in regards to socioeconomic background leaves a gap in how we understand membership rates. We further investigate what factors are relevant when trying to explain both the geographical and the socioeconomic trend in church membership rates.

## The meaning of local community

Regional differences in religious and moral orientations have been pointed out in several former studies. The urban sociologist, Jens Tonboe (2001; Thygesen, Knudsen, and Keiding 2005), distinguished between four regions characterized by different norms according to the European Values Survey: 1) The Capital Area, 2) The Urbanized provincial part of the country, 3) The Close Periphery and, 4) The Distant Periphery (Tonboe 2001). Moral strictness was maintained in the Periphery, while the Capital was characterized by moral laissez-faire. His analysis was based on relatively large units of analysis and thereby omitted the basic question of the sociological meaning of the local communities. This can better be addressed by analysing geographical units of analysis that are sufficiently small to allow local patterns of belonging, embedding and neighbourliness to emerge.

The sociological variable focused on in the following analysis is local belonging, or how places form a framework for social relatedness to place, whether these are family ties, friendships or neighbour relations (Jørgensen et al. 2016); this framework is ascribed with an inherent value by the residents. This dimension corresponds with a so-called "people- or relational-centered perspective" opposed to a "power, position or capital-centered" perspective on social space by Jørgensen (2017). Neighbourhood is a sociologically meaningful entity, even though its definition and boundaries can be perceived in many ways, even among individuals living in the same physical setting. (Sampson 2011, 230). Neighbourhoods in late modern societies sometimes constitute a community in the sense of forming tight-knit bonds and sharing values; however, in many cases they do not (Sampson 2011, 229). Neighbourhood is a sociologically meaningful entity, even though its definition and boundaries can be perceived in many ways, even among individuals living in the same physical setting (Sampson 2011, 230).

Face-to-face interactions among residents of a neighbourhood may stimulate social ties among residents that support collective goals, such as public order or control of crime. These shared expectations and mutual trust among neighbourhood residents promote a sense of cohesion or belonging, which Sampson (2011) calls Collective Efficacy. This aspect of face-to-face interaction is inherently better understood in small units where people recognize others than in large, anonymous units. According to Sampson, the root of the collective efficacy to a neighbourhood is the intersection of practices and social meanings with a spatial context (Sampson 2011, 230). Robert Sampson is critical to the way Coleman defines Social Capital as primarily a resource that is realized through social relationships (Coleman 1988) and argues that: Social networks foster the conditions under which collective efficacy may flourish, but they are not sufficient for the exercise of social cohesion and social control (Sampson 2011). Networks have to be activated in order to be meaningful and in this sense collective efficacy can be defined as a link between mutual trust, shared expectations among residents and a willingness to intervene and interact (Sampson, Raudenbush, and Earls 1997).

The quantity, quality and diversity of institutions and organizations have an impact on a neighbourhood (Sampson 2011, 233). It is important to be aware of the organizational density and the levels of participation in relation to these organizations, as organizational density is not an equivalent to coordinated action for local interests. Sampson, Raudenbush and Earls (1997) have in this way constructed a measure of collective efficacy combining informant ratings of the capacity for informal social control with social cohesion defining neighbourhood "as a variably interacting population of people and institutions in a common place." This means that network-density, attachment to place, civic participation, disorder, organizational density, identity and capacity for collective action is variable and analytically separable from structural variables and possible consequences.

By introducing belonging as related to church membership, we suggest that residential relatedness to place is important for understanding the driving forces beside the variables of age, income, education, stability and density. Belonging can be characterized as social relatedness to place, which refers to all types of local social relations such as family, friends, neighbours – the associational life that tie people to place.

Figure 1 and 2 illustrates the association between geographical place of residence and both local activities measured in local election participation and participation in local activities (figure 1), and number of social associations measured in associations per 1000 inhabitants (Jakobsen, Sørensen and Johansen, 2014). In the figures below, red indicates low election participation and local activities while green indicates high values. The overall tendencies that emerge from the maps are uniform and display some of the difference in localized social differentiation. A high number of social associations is especially characteristic of Western Jutland. It is only in the Capital that local activity seems especially low. This could indicate an active/passive relationship in the local environment. When population density becomes sufficiently high, there is no need for local volunteering and participation since organized services are widely available. When the population density becomes low, the local community needs to commit themselves in order to obtain options similar to those in the Capital.

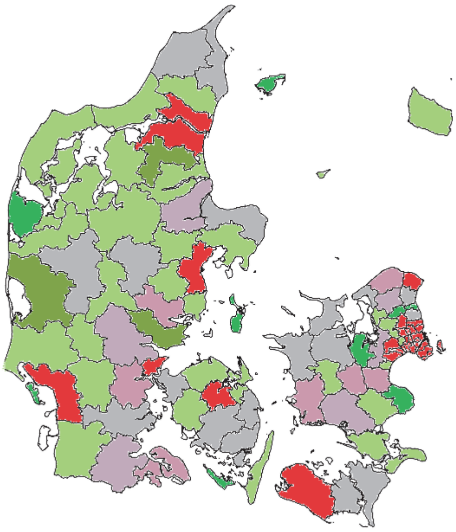


**Figure 1.** Map of locally active residents

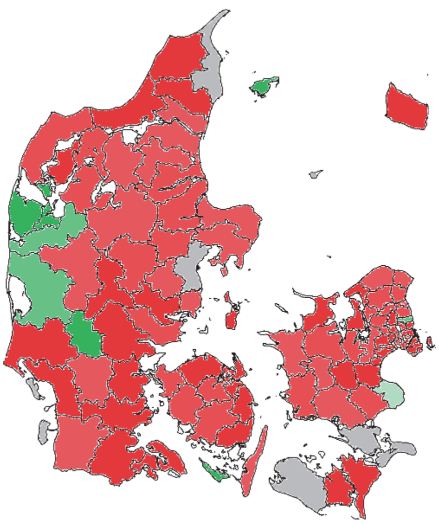


**Figure 2.** Map of social associations

Several studies in Denmark have indicated the relevance of studying social relatedness to place at the neighbourhood level (Jørgensen et al. 2016). For instance, the local distribution of education and economic capital depends on social space (Andersen 2005; Damm, Schultz-Nielsen and Tranæs 2006). Furthermore, in-depth local studies have demonstrated the importance of neighbour support in poor or wealthy settings for economic inequality (Munk and Andrade 2014).

In this study, we therefore pursue a dual perspective on social space as an integrative framework: On the one hand, a vertical perspective that indicates patterns of power, position or capital. This directs attention to demographic and socio-economic processes of change in and to variations of these processes across different neighbourhoods. On the other hand, a horizontal perspective that tries to illuminate the local collective efficacy.

## IDENTIFYING NEIGHBOURHOODS

Ecological studies of neighbourhoods are typically based on formal, administrative local units, such as parishes. However, such units do not represent sociologically meaningful communities, where people interact in daily life and share a sense of belonging. Our analysis aims to identify local units, confined by physical barriers where the inhabitants may interact in daily life. We argue that people who live in such a neighbourhood have a good chance to share values and beliefs. Our study thus follows up on former studies of social neighbourhoods (Damm and Schultz-Nielsen 2008; Feld 1981; Freisthler et al. 2016; Jones and Huh 2014; Jones and Pebley 2014; Kwan 2013; Patterson and Farber 2015). However, our study utilizes more detailed localities that are sociologically meaningful. The identification of the localities is not determined theoretically, but pointed out empirically by an algorithm based on barriers that hinder interaction, such as roads, streams or forests (Damm and Schultz-Nielsen 2008; Deng 2016; Freisthler et al.2016; King, Keohane, and Verba 1994). One of the official local units is the parish. The shape of the parishes goes back to the Reformation era, and it has hardly been regulated with the urbanization of the last two centuries. A parish is centred on a specific church, but its membership does not always identify as a social group.

## DATA

The analysis is based on two different types of data. The first segment of data originates from The National Square Grid, a national system of vector grids constructed by The Danish Geodata Agency and Statistics Denmark. The National Square Grid is linked to each person in the Danish public registers. It is possible to place each person living in Denmark inside a square by 100x100 meters. However, our units of the present analysis are larger, both due to considerations on confidentiality and on the borderlines of local identity.

The second set of data consists of register data for the total of the Danish population over 18 years of age in the years 2000, 2005, 2010 and 2015. The registers compile information on education, income, age, gender, and ethnicity for individuals.[2] Membership of the established church is identified by payment of church tax. As for the rate of infant membership, births are assigned to a mother living in that area.

## METHODOLOGY

### Creating areas

To capture the idea of the physical space as a determinant factor for social space, we utilize a spatial clustering algorithm. It first considers physical barriers, i.e. roads, railways and creeks and holds all polygons as undividable by these barriers, simultaneously applying the square grid to these areas to calculate the number of people living inside each confined area.

The algorithm works in two steps; the first is to apply the barriers in question. Physical barriers are thought to create also a social barrier which separates "us" from "the people on the other side of the road" (Feld 1981). There is no preconception about how many inhabitants one polygon may contain (Damm and Schultz-Nielsen 2008). Thereby the population is distributed into the first array of areas. The total amount of new areas is 20,940 of which only 28 per cent meet the minimum of 150 households required by Statistics Denmark.

In order to fulfil the discretion demand stated by Statistics Denmark, a second clustering is performed and starts out by making all areas applicable for clustering while only considering adjacent areas that share the longest borders. This leads to the next set of arguments, which identifies the least possible number of mergers while retaining the smallest total number of inhabitants over 150 persons in each area. The main point in the above criteria is to make the algorithm work in a way that results in the fewest possible number of area mergers. The problem in selecting a specific point to start the selection process is that the final merge would vary extremely and would be different each time a different starting polygon as selected. By setting these criteria, the algorithm consequently creates the same mergers if the process was to be repeated. This results in a map of Denmark consisting of 8043 base areas.

### Measuring church membership

The analysis of the relationship between church membership, the type of residential area and the socioeconomic factors is based on six steps. The first step is to perform a simple linear regression on an individual level. This will be the baseline model in order to under-

2. Income is measured as the net total yearly tax income per year of each individual including surplus of a partner. Education is measured as full months completed education and includes mandatory primary education. Area stability is a measure by the ratio of people living in the same area in both 2010 and 2015 in proportion to those present in only one of these years. Population density is measured as the number of people per square kilometer. The rate of senior citizens is measured as the percentage of inhabitants over the age of 67 compared to the rest of the area population. This cut-off point is the typical age of retirement from employment. This information is then linked to the square grid after the clustering has taken place.

stand the effect when time and place is not considered. The next step controls for time in a fixed effects model in order to investigate how much of the observed effect is due to the passing of time. Step 3 fixes the effect of parish and time while step 4 fixes the effect of micro-area and time. These models allow us to see how much of the effect is due to parishes versus local settings; whether non-membership can be explained by the functioning of the church or by the social character of neighbourhoods. The time and area fixed effects are fitted as:

$$Y_{ia} - \bar{Y}_a = \beta_0 + \beta_i(X_{ia} - \bar{X}_a) + (\varepsilon_{ia} - \varepsilon_{ia}) + (u_{ia} - \bar{u}_a)$$

It determines the difference between $Y_{ita}$ I – church membership – for the $i$'$^{th}$ person in $a$'$^{th}$ area for the $t^{th}$ time, and the mean of the area at $t$, $\bar{Y}_{ta}$. Similar calculations are made for the socioeconomic factors – income and education – where $X_{ita}$ refers to data for each individual inside an area at $t$ and this measure is subtracted by $\bar{X}t_a$ – the mean of the area at $t$.

Step 5 and 6 omits the individual level, as they are based on aggregated area data, in order to include measurements of stability (the proportion of the inhabitants people that moves in or out of an area) and population density (inhabitants per square kilometer). The analysis utilizes a time fixed regression analysis of the effect on the church membership rate within an area ($Y_{it}$) determined by socioeconomic indicators for each area. This is done for both parishes and for micro areas.

$$Y_{it} = \beta_0 + \beta_{it}X_{it} + \varepsilon$$

This allows us to compare the influence of factors that influence church membership at both the individual level and at the level of the specific areas.

## ANALYSIS

### Demographic characteristics

Former sociological studies of individual commitment to the church have pointed to gender as a major factor (i.e. Gundelach and Riis 1992, 32). Church life is mainly supported by women. However, the local variation in gender composition is too small to be usable in our analysis. It is also a well-known fact that commitment to the church is stronger among the older membership (Gundelach and Riis 1992, 17).

The age composition has an influence on the social and cultural life of a neighbourhood. Religion can be a resource for collective memories, as pointed out by Hervieu-Léger (1993). We may add that the nurturing of local traditions implies a local diversification. A neighbourhood with many retired people seems to halt the trend towards leaving the church among the younger inhabitants. However, our findings demonstrate only a weak relationship between the age composition of local areas and church membership. Furthermore, our effort to elaborate the analysis by focusing on the proportion of retired people did not indicate a simple and clear relationship. Only the last decile indicates a slight increase in church membership rates. So, despite a sound theoretical backing, our hypothesis about age as a factor determining church membership could not be confirmed at this level of analysis.

**Table 1.** Indices of socioeconomic indicators within strata of church member rates (2015)

| Decentile | Income | Education | Area stability | Pop. density | Ratio of old age pensonists |
|---|---|---|---|---|---|
| 1st decentile | 312705,1 | 4,09 | 0,63 | 4,62 | 0,21 |
| 2nd decentile | 288146 | 3,69 | 0,68 | 1,92 | 0,22 |
| 3rd decentile | 274583,4 | 3,55 | 0,67 | 1,69 | 0,22 |
| 4th decentile | 270438,2 | 3,47 | 0,69 | 1,26 | 0,23 |
| 5th decentile | 267117,7 | 3,41 | 0,70 | 1,20 | 0,22 |
| 6th decentile | 268022,5 | 3,37 | 0,71 | 0,86 | 0,23 |
| 7th decentile | 267603,5 | 3,32 | 0,73 | 0,78 | 0,23 |
| 8th decentile | 268105,2 | 3,29 | 0,73 | 0,61 | 0,23 |
| 9th decentile | 265977,9 | 3,24 | 0,75 | 0,58 | 0,23 |
| 10th decentile | 261772,9 | 3,15 | 0,75 | 0,47 | 0,24 |

One of the classic themes in sociology is the impact of urbanization, characterized by population density (Simmel 1903; Wirth 1938) and by population turnover (Park, Burgess, and McKenzie [1967] 1925). Urbanization is typically seen as a force that leads to superficial and unstable social relations. Sociology of religion has also pointed to urbanization as the main factor behind the decline in church membership in the Capital area (i.e. Riis 2015). It is therefore relevant to focus on this issue here.

Our data show a clear relationship between urbanization – indicated by population stability and population density – and membership of the established church as well as the descriptive map shown below. Neighbourhoods with high population density and low degree of stability are often characterized by low membership rates. Population density and high turnover can be seen as carrying a special culture or way of life which is less locally embedded and less committed to the traditional social institutions – such as the church.

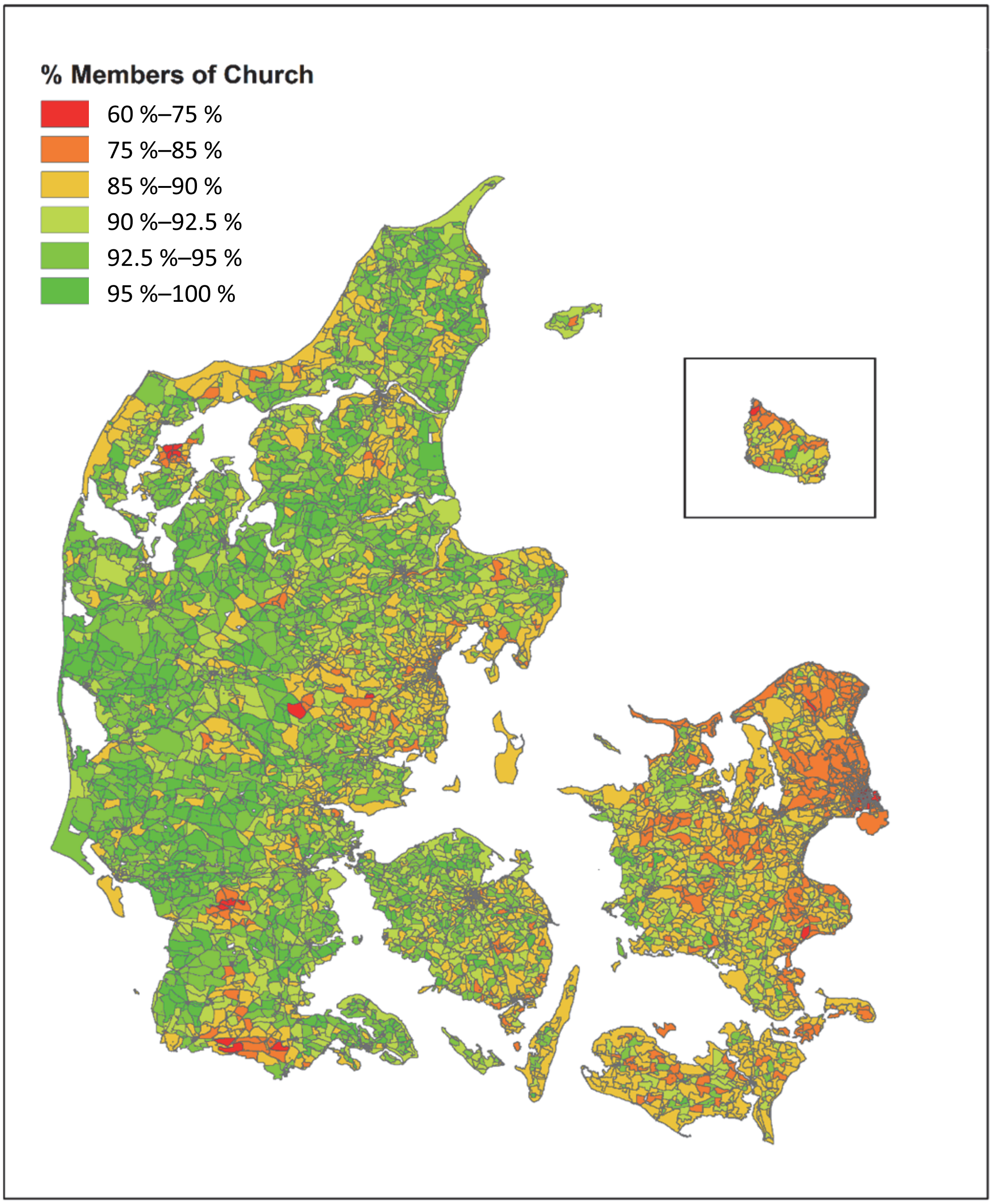



**Figure 3.** Percentage membership of the established Church in Denmark

### Economic and educational status

Because of the civil religious character of the established church of Denmark, church membership should be a common feature, and therefore not related to social status, measured by income and education. In a society where church membership is a selected marker of identity, such as the USA, we should expect religious commitment to be related to social status. It is therefore surprising to find that in Denmark, higher social status, measured by both income and education, is related to lower membership rates. The difference in mean income between the lowest decile membership areas compared with the highest when looking at area average income levels is roughly 50,000 kroners a year.

This pattern is surprising, as the high-status elite was formerly regarded as the main supporters of the traditional institutions in society. Our data indicate the formation of a new elite, which is withdrawing itself from the established church. The socio-economic elite in the Capital is not identical with the elite not living in the Capital area. We focus here on the 1 per cent richest of the population in Zealand versus Jutland. Thereby, a clear pattern is visible: The church membership rate is much lower in the affluent districts closer to the Capital, whereas it remains close to the national average of 88 per cent in the affluent districts of Jutland. The elite located closer to the capital overall fits the trend of high economic status resulting in lower membership; but this is only partially true in Jutland, where the membership rates in this group comes much closer to the national average of 88 per cent. This indicates that something else other than the mere socio-economic status influences decisions about leaving or remaining in the church.

The overall correlations between the aggregated level indicators and church membership differ. The measure that correlates most strongly with membership is educational attainment.

As for area stability, the association is positive, indicating higher membership rates in the most stable neighbourhoods. In general, the lowest church membership rates are to be found in areas in the larger cities characterized by a high income, a high educational attainment, living neighbourhoods with a frequently shifting residential base. This becomes even more evident in table 2, where we compare effects at the individual level with effects over time and time/space.

**Table 2.** Regression models on church membership

| | Indiv. effects b/(se) | Year FE b/(se) | Parish FE b/(se) | Area FE b/(se) | Parish Avg. b/(se) | Area Avg. b/(se) |
|---|---|---|---|---|---|---|
| Age | 0.000*** (0.00) | 0.001*** (0.00) | 0.000*** (0.00) | 0.000*** (0.00) | 0.000 (0.00) | 0.000* (0.00) |
| Gender | 0.037*** (0.00) | 0.037*** (0.00) | 0.040*** (0.00) | 0.039*** (0.00) | 0.031** (0.01) | 0.027*** (0.01) |
| Education in months | -0.001*** (0.00) | -0.001*** (0.00) | -0.001*** (0.00) | -0.000*** (0.00) | -0.002*** (0.00) | -0.001*** (0.00) |
| Ethnicity | -0.741*** (0.00) | -0.740*** (0.00) | -0.734*** (0.00) | -0.706*** (0.00) | -0.927*** (0.01) | -0.837*** (0.01) |
| Income | -0.000*** (0.00) | -0.000*** (0.00) | -0.000*** (0.00) | -0.000*** (0.00) | 0.000 (0.00) | -0.000 (0.00) |
| Year=2000 | | 0.000 (.) | 0.000 (.) | 0.000 (.) | 0.000 (.) | 0.000 (.) |
| Year=2005 | | -0.007*** (0.00) | -0.004*** (0.00) | -0.007*** (0.00) | -0.002*** (0.00) | -0.004*** (0.00) |
| Year=2010 | | -0.016*** (0.00) | -0.009*** (0.00) | -0.017*** (0.00) | -0.006*** (0.00) | -0.006*** (0.00) |
| Year=2015 | | -0.029*** (0.00) | -0.022*** (0.00) | -0.030*** (0.00) | -0.012*** (0.00) | -0.016*** (0.00) |
| Pop. density | | | | | -0.001*** (.000) | -0.003*** (.000) |
| Area stability | | | | | 0.0321*** (0.004) | 0.0503*** (0.004) |
| Constant | 0.924*** (0.00) | 0.926*** (0.00) | 0.916*** (0.00) | 0.904*** (0.00) | 1.099*** (0.03) | 0.990*** (0.01) |
| R-sqr<br>Dfres<br>N | 0.281<br>15729053<br>15729059.0 | 0.282<br>15729050<br>15729059.0 | 0.294<br>11925298<br>11927537.0 | 0.301<br>15721008<br>15729059.0 | 0.945<br>4176<br>6415.0 | 0.975<br>16079<br>24129.0 |

* p<0.05, ** p<0.01, *** p<0.001

The individual, pooled regression in column 1 shows that church membership tends to increase slightly with age. This is consistent with the former observation that church membership declines through new generations due to declining rates of baptism. Likewise, education and income have a negative impact on church membership, which implies that more educated and wealthier individuals are less likely to be members of the church. Also, males and immigrants are less likely to be members. Adjusting for time effects, the same overall tendency holds for all the variables and it becomes clear that the tendency is a decline in membership rates. Looking at the parish and area fixed effects, the analysis shows a larger effect on the decline in the local areas than in the parishes. This indicates that the decline depends more on the local community than on the local church parish. Membership of the church depends on the type of local community rather than on the local church and its vicar. While the general decline in membership is due to lower rates of baptism, much of the variation in membership depends on the character of the local area.

The final models investigate the aggregated effects of parishes and areas. It shows the same patterns as the previous models – that income no longer holds a significant effect. It shows that the clustering effect of the areas levels out the pure effect of income as area averages are used for comparison. The high-income areas in Jutland are evening out the relative accumulation of wealth in the Capital region. This indicates that the effect of social status is not uniform in its geographical distribution.

One explanation as to why income no longer holds any significant effects on church membership and why education increases its effect could be that the cultural and economic elites differ not only in place but also in life choices. The cultural elite are much less likely to be members of the church, and even though this tendency is more prominent in the Capital, this seems to be the most universal effect we found.

## CONCLUSION

One of the traditional pillars of a national Danish identity is the established Evangelical-Lutheran church. Since the turn of the new millennium, church membership has declined markedly, indicating a basic change in its function, from a civil religious institution, for all baptised Danes, to a confessional one, for believing Protestants. This decline is especially marked in the Capital area. Our findings reconfirm the contrast between the Capital area and the rest of the country. However, by aggregating data at this level, local variations are evened out. Our more detailed studies show that these two parts are both heterogeneous. Denmark should rather be seen as a jigsaw puzzle of many small local units with their own social composition and local identity. One of the factors behind church membership is the local associational life.

Those local areas characterized by a low membership in the established church are also generally characterized by a weak participation in local activities. This pattern is more prevalent in the Capital than in the rest of Denmark, and deviations can be found in both. Church membership can thus be understood as an habitual part of the local associational life.

Our findings show that the church membership rate is especially low among the socio--economic and educational elites, and especially in the Capital area. This could be interpreted as 'secularization from above'. However, the image of the trend-setting elite in the Capital versus the laggards in the rest of Denmark is too simplified. Several elites are competing for social status, including challengers from the more rural parts of Denmark. Our detailed local analysis demonstrates that high-status districts have many common characteristics across the county.

Regarding church membership, the rate remains quite high in high-status districts in the non-capital areas of Denmark. Correspondingly, Edin Tabak (2015) suggests that there is no contradiction between being attached to place and being more oriented towards the cosmopolitan. He sees this dichotomy as a typical modernist narrative of unification and stabilization of space (Tabak 2015). In his ANT projection a local community is just one of many possible heterogeneous networks to which individuals could be attached (Tabak 2015). Place is not "what holds people together" it is "what is held together" in a certain place that is interesting. This perspective enables us to see that church membership is not just a question of belonging to a certain place or neighbourhood, but also a matter of individuals downloading the culture that is distributed in specific places.

The major sociological factor that emerges from this analysis is the social efficacy, or the sense of attachment to the place of residence. In some local units, the residents carry and nurture their own, local culture, formed through their everyday community life. Everyday interaction in a neighbourhood is the basis for membership of collective institutions, such as the established church. Theoretically, our study points to a focus on local micro-sociological processes that form the link between structural patterns and individual preferences. Methodologically, our study points to a focus on local variations that are seen as residuals in studies focusing only on the major trends. Local residuals are not statistical noise; they form the key to a further level of sociological investigation. Our analysis points to two branches of further studies. The first branch consists of comparative local case studies of communities that are similar in socio-economic composition but differ in membership of the established church. This allows identifying those elusive socio-cultural factors that trigger disaffiliation. The other branch aims to investigate the social processes that sustain or hinder a local, social efficacy.

## REFERENCES

Andersen, Hans Thor. 2005. *Storbyens ændrede socialgeografi* [The changed social geography of the big city]. Kulturgeografiske skrifter, bind 15. Copenhagen: A. Reitzels Forlag.

Andersen, Peter B. and Peter Lüchau. 2004. "Tro og religiøst tilhørsforhold i Europa" [Faith and religious affiliation in Europe]. In *Danskernes Særpræg*, edited by Peter Gundelach, 245–68. Copenhagen: Hans Reitzels Forlag.

______. 2011 "Islam and Labor Market Integration in Denmark." In *Religion and Social Problems*, edited by Titus Hjelm, 46–64. New York: Routledge.

Coleman, James S. 1988. "Social capital in the creation of human capital." *American Journal of Sociology* 94: 95–120.

Damm, Anna Piil and Marie Louise Schultz-Nielsen. 2008. *The Construction of Neighbourhoods and its Relevance for the Measurement of Social and Ethnic Segregation: Evidence from Denmark*. Discussion Paper No. 3719. Bonn: The Institute for the Study of Labor.

Damm, Anna Piil, Marie Louise Schultz-Nielsen, and Torben Tranæs. 2006. *En befolkning deler sig op?* [A population divides]. Copenhagen: Gyldendal.

Davie, Grace. 1999. "Believing without Belonging. Is this the future of religion in Britain?" *Social Compass* 37: 456–69. doi: 10.1177/003776890037004004.

Deng, Yeng. 2016. "Challenges and complications in neighborhood mapping: from neighborhood concept to operationalization." *Journal of Geographical Systems* 18 (3): 229–48. doi: 10.1007/s10109-016-0232-z.

Feld, Scott. L. 1981. "The Focused Organization of Social Ties." *The American Journal of Sociology* 86 (5): 1015–35.

Freisthler, Bridget, Thomas, Crystal A., Susanna Curry, and Jennifer Price Wolf. 2016. "An Alternative to Residential Neighborhoods: An Exploratory Study of How Activity Spaces and

Perception of Neighborhood Social Processes Relate to Maladaptive Parenting." *Child and Youth Care Forum* 45 (2): 259–77. doi: 10.1007/s10566-015-9329-7.

Gundelach, Peter and Ole Riis. 1992. *Danskernes Værdier* [The values of the Danes]. Copenhagen: Forlaget Sociologi.

Gundelach, Peter. 2002. *Danskernes værdier 1981–1999* [The values of the Danes 1981–1999]. Copenhagen: Hans Reitzels Forlag.

______. 2011. *Stabilitet og forandringer* [Stability and change]. In *Små og store Forandringer,* edited by Peter Gundelach, 11–29. Copenhagen: Hans Reitzels Forlag.

Gustafsson, Göran and Thorleif Pettersson. 2000. *Folkkyrkor och religiös pluralism – den nordiska religiösa modellen* [The national church and religious pluralism – the Nordic model of religion]. Stockholm: Verbum.

Hervieu-Léger, Danièle. 1993. *Le religion pour mémoire* [Religion as a Chain of Memory]. Paris: CERF.

Inglehart, Ronald. 1990. *Cultural Shifts in Advanced Industrial Society.* Princeton: Princeton University Press.

Jakobsen, Leif Henrik, Stig Yding Sørensen, and Kasper Damgaard Johansen. 2014. *Virketrang i Danmark* [The urge to effectualism in Denmark]. Teknologisk Institut, RegLab.

Jones, Mark and Jimi Huh. 2014. "Toward a multidimensional understanding of residential neighborhoods: A latent profile analysis of Los Angeles neighborhoods and longitudinal adult excess weight." *Health and Place* 27: 134–41. doi: 10.1016/j.healthplace.2014.01.011.

Jones, Malia, and Anne R. Pebley. 2014. "Redefining Neighborhoods Using Common Destinations: Social Characteristics of Activity Spaces and Home Census Tracts Compared." *Demography* 51 (3): 727–52. doi: 10.1007/s13524-014-0283-z.

Jørgensen, Anja, Lisbeth B. Knudsen, Mia Arp, and Henrik Skov. 2016. "Zones of Belonging – experiences from a Danish study on belonging, local community and mobility." *Geoforum Perspective* 15 (29). doi: https://doi.org/10.5278/ojs.perspektiv.v15i29.1450.

Jørgensen, Anja. 2017. "Fra By over rum til territorialitet-og tilbage igen?! (From city through space to territoriality – and back again!?)". In *Sociologi,* edited by Carsten B Laustsen, and Anders Blok, 55–70. Copenhagen: Hans Reitzels Forlag.

King, Gary, Robert O. Keohane, and Sidney Verba. 1994. *Designing Social Inquiry: Scientific Inference in Qualitative Research.* Princeton, NJ: Princeton University Press.

Kwan, Mei-Po. 2013. "Beyond space (as we knew it): Toward temporally integrated geographies of segregation, health, and accessibility: Space–time integration in geography and GIScience." *Annals of the Association of American Geographers* 103 (5): 1078–86.

Munk, Martin D., and Stefan Bastholm Andrade. 2014. "Do Neighbors Matter When Analyzing Income Inequality?: An analysis of the consequences and reproduction of poor and wealthy neighborhoods." Paper presented at Dansk Sociolog Kongres.

Park, Robert E., Ernest W. Burgess, and Roderick D. McKenzie. (1925) 1967. *The City*. Chicago: University of Chicago Press.

Patterson, Zachary and Steven Farber. 2015. “Potential Path Areas and Activity Spaces in Application: A Review.” *Transport Reviews* 35 (6): 679–700. doi: 10.1080/01441647.2015.1042944.

Riis, Ole. 2000. Den europæiske integrations udfordring til danskernes religiøse identitet. In *Kirken og Europa,* edited by Jakob Balling, 158–98. Aarhus: Aarhus Universitetsforlag.

______. “Le dilemme nordique” [The Nordic dilemma]. In *Religion(s) et identité(s) en Europe*, edited by Antonela Cappelle-Pogacean, Patrick Michel and Enzo Pace, 73–98. Paris: Sciences Po.

______. 2011. “Religion.” In *Sociologi – en grundbog til et fag*, edited by Heine Andersen, 361–80. Copenhagen: Hans Reitzels Forlag.

______. 2012. “Kirkestatistisk status for den danske folkekirke” [Church statistical status for the Danish National Church.]. In *Religionsstatistikk og medlemsforståelse*, edited by Ingunn Breistein and Ida Marie Høeg, 193–212. Oslo: Akademia Forlag.

______. 2015. “Sekulariseringens teorier” [Theories of secularization]. In *Religionsforskning før og nu*, edited by Armin Geertz and Tim Jensen, 7–35 Copenhagen: Gyldendal.

Sampson, Robert J., Stephen W. Raudenbush, and Felton Earls. 1997. “Neighborhoods and violent crime: A multilevel study of collective efficacy.” *Science* 277 (5328): 918–24. doi: 10.1126/science.277.5328.918.

Sampson, Robert J. 2011. “Neighborhood effects, causal mechanisms and the social structure of the city.” In *Analytical Sociology and Social Mechanisms,* edited by Pierre Demeulenaere, 227–49. New York: Cambridge University Press.

Simmel, Georg. 1903. “Soziologie der Konkurrenz” [Sociology of competition]. *Neue Deutsche Rundschau* 14 (4): 1009–23.

Sundback, Susan. 2000. “Medlemsskapet i de nordiska kyrkorna” [The membership in the nordic churches]. In *Folkkyrkor och religiös plualism – den nordiska modellen,* edited by Göran Gustafsson and Thorleif Pettersson, 34–73. Stockholm: Verbum.

Tabak, Edin. 2015. “Information Cosmopolitics: A model of information practices.” In *Information Cosmopolitics: An Actor-Network Theory Approach to Information Practices,* 121–37. Waltham, MA: Chandos Publishing.

Taylor, Charles. 2007. *A Secular Age.* Cambridge, Mass: The Bellknap Press of Harvard University Press.

Thygesen, Lau Caspar, Lisbeth B. Knudsen, and Niels Keiding. 2005. “Modelling Regional Variations of First-Time births in Denmark 1980–1994 by an Age-Period-Cohort Model.” *Demographic Research* 13 (23): 573–96. doi: 10.4054/DemRes.2005.13.23.

Tonboe, Jens C. 2001. *Moralens geografi* [The geography of morality]. Sociological working paper. Aalborg.

Wirth, Louis. 1938. “Urbanism as a Way of Life.” *American Journal of Sociology* 44 (1): 1–24.

# SUMMARY

In this thesis, I point to the importance of asking not only questions about the neighborhood effects on the inhabitants that lives there but also asking more fundamental question about what a neighborhood is, how me measure it and what scale means to the way we process the effects. By using a combination of selection models and automated redistricting, I show that scale is very important when investigating neighborhood deprivation. Using administrative borders to isolate deprived areas are inadequate to reveal the intricate and often small clusters that are truly deprived. Furthermore, I show that deprivation is not one thing; deprivation in different geographical settings has a variety of different effects on later life outcomes of the residents. Thus, I argue that place is diverse and complex and that neighborhood research must account for the geographical difference between neighborhoods to fully understand the underlying mechanisms.

SSN (online): 2246-1256
ISBN (online): 978-87-7210-472-0

AALBORG UNIVERSITY PRESS