Data Construction of Danish Residential Neighbourhoods for 1986-2016[*]

Af
Anna Piil Damm[**]
Ahmad Hassani[***]
Marie Louise Schultz-Nielsen[****]

Keywords: Residential neighbourhoods, Housing, Clustering
JEL codes: C38, R23, O18

Abstract

Using Danish geo-referenced administrative housing register data, we utilise the Thiessen polygon method to cluster adjacent housing units into 1,961 (macro) Danish residential neighbourhoods with at least 600 households that are delineated by physical barriers, homogenous in terms of housing type and house ownership, compact and have unchanged boundaries over 31 years (1986-2016). Using the same clustering criteria, we construct 8,359 (micro) Danish residential neighbourhoods with at least 150 households such that around 4 micro-neighbourhoods constitute a macro-neighbourhood, which respects the boundaries of the micro-neighbourhoods. The neighbourhoods can be visualised on a map using the grid of hectare cells.

[*] This paper acknowledges the support from and access to Statistics Denmark provided by ECONAU, Department of Economics and Business Economics, Aarhus University. We acknowledge financial support from TrygFonden's Centre for Child Research, Aarhus University, The ROCKWOOL Foundation Research Unit, and VIVE. We thank Jos van Ommeren, Ismir Mulalic, Peter Jensen, Rune Majlund Vejlin, the editor and 2 anonymous reviewers for their helpful comments, Henning Bunzel for his advice in using Danish register-based data, and Drilon Helshani and Geomatic for research assistance. The authors declare that they have no relevant or material financial interests that relate to the research described in this paper.

[**] Department of Economics and Business Economics, Aarhus University, Fuglesangs Allé 4, DK-8210 Aarhus V. Email: apd@econ.au.dk.

[***] Department of Economics and Business Economics, Aarhus University, Fuglesangs Allé 4, DK-8210 Aarhus V. Email: ahass@econ.au.dk

[****] The ROCKWOOL Foundation Research Unit, Ny Kongensgade 6, DK-1472 Copenhagen K. Email: mls@rff.dk

1. Introduction

For many purposes, it can be useful to define residential neighbourhoods in Denmark, which are substantially smaller than municipalities, both in terms of area and population, and reflect social interaction between neighbours. For some purposes, existing administrative divisions of Denmark into postal districts, parishes and school districts may be useful, e.g. for analysis of the effects of merging school districts on the socioeconomic and ethnic mix of pupils in public schools (Bjerre-Nielsen and Gandil, 2020).

By contrast, measurement of residential segregation of socioeconomic and ethnic subgroups of the population and its causes and consequences requires well-defined residential neighbourhoods that are small, relatively homogeneous in terms of population size and area, and reflect social interaction between neighbours. Moreover, analyses of the *evolution* of such types of residential segregation over time require a definition of residential neighbourhoods with unchanged boundaries over time. A related use of residential neighbourhoods for Denmark is for impact evaluations of place-based policies targeted at socially deprived neighbourhoods.[1] In effect evaluations of policies combatting residential segregation of minority groups, it is crucial to have residential neighbourhoods with unchanged boundaries over time that reflect social interactions among neighbours and are relatively homogeneous in terms of population size and area. Furthermore, information regarding residential neighbourhoods is beneficial in a wide range of other studies that require control for local residential conditions.

The purpose of this paper is to describe how such neighbourhoods can be constructed for Denmark using geo-referenced data and administrative registers for the entire period 1986-2016. The current work builds on the approach by Damm and Schultz-Nielsen (2008). That study clustered all 431,233 inhabited hectare cells of Denmark over the period 1985-2004 into 9,404 micro- and 2,296 macro-neighbourhoods (hereafter *the first version)*. In this paper, we describe how we have improved the first version of the clustering approach and updated the neighbourhood measures to cover the entire period 1986-2016, primarily by using the exact location instead of the approximate location (hectare cell) of housing units to divide the inhabited land of Denmark into micro (macro) clusters with at least 150 (600) households. Our use of the exact location instead of the approximate location is also a strength of our algorithm relative to the approach by Bjerre-Nielsen and Gandil (2018). By clustering hectare cells, their algorithm divides the land of Denmark into time-invariant sub-areas (polygons) with at least 100 inhabitants (*k*=100). Another strength of our approach relative to the ap-

[1] For instance, "ghettopakke" policy in Denmark, which is described in more detail in Section 6.

proach by Bjerre-Nielsen and Gandil (2018) is that our approach takes physical barriers to social interaction into account such that neighbourhood boundaries respect physical barriers like major roads and railways.

We construct residential neighbourhoods that satisfy the following criteria. First, a residential neighbourhood should correspond to the geographical area within which an individual socially interacts with other residents. Second, in order to allow for comparisons over time, the boundaries of the neighbourhood should be unaltered over a long period. This criterion rules out the use of administrative divisions such as parishes and school districts. Finally, we should be able to link the residential neighbourhoods thus defined with administrative register data in order to construct annual information about the individual's neighbourhood of residence over a long period of time

Our residential neighbourhoods have unaltered boundaries over the 1986-2016 period. While our clustering criteria are similar to the clustering criteria used in Damm and Schultz-Nielsen (2008) for construction of residential neighbourhoods in Denmark for the 1985-2004-period, our clustering approach is superior to that used in Damm and Schultz-Nielsen (2008). Whereas Damm and Schultz-Nielsen (2008) clustered adjacent hectare cells using the National Square Grid to meet the clustering criteria, we cluster adjacent housing units using the Thiessen polygon method to meet the clustering criteria. This way we obtain residential neighbourhoods with an organic form with boundaries that respect physical barriers like large roads, railways, water and forests. According to our definition of residential neighbourhoods, residents who live in the same hectare cell may belong to two (or even more) different residential neighbourhoods, e.g. if they live on each side of a major road.

The current rules of confidentiality of Statistics Denmark require a minimum of 50 households and a minimum of 100 inhabitants in each neighbourhood. We argue that for analyses of residential segregation over a long time period, it is more useful to construct residential neighbourhoods that have a minimum of 150 households. Furthermore, we argue that for international comparison and analyses of causes and consequences of residential segregation, it is more useful to have a minimum of 600 households. Therefore, we construct micro-neighbourhoods that have at least 150 households and macro-neighbourhoods that have a minimum of 600 households and each comprises around 4 micro-neighbourhoods. In practice, we divide 459,497 inhabited hectare cells of Denmark over the period 1986-2016 into 8,359 micro- and 1,961 macro-neighbourhoods. An average micro- (macro-) neighbourhood includes 291 (1,241) households with standard deviation of 109.3 (394).

Our approach and statistical criteria for clustering housing units into micro-neighbourhoods result in micro-neighbourhoods for Denmark that have a comparable number of inhabitants to the US census blocks, but a lower number of inhabitants to the Small Areas for Market Statistics (SAMS) created by Statistics

Sweden. A strength of our micro-neighbourhoods relative to the US census blocks is that they have unaltered boundaries over time. As for the constructed Danish macro-neighbourhoods, they are comparable with block groups in the US in terms of the number of inhabitants.

The paper is organised as follows. Section 2 describes the first version of Danish residential neighbourhood data constructed by Damm and Schultz-Nielsen (2008), and the value-added of the current version of constructed neighbourhood data relative to the first version. Section 3 provides information about data foundation, whereas Section 4 describes our methodological considerations, in particular the considered criteria for clustering residential properties. Section 5 describes our implementation of the clustering algorithm for construction of the neighbourhood clusters, as well as different characteristics of the final neighbourhood clusters. Section 6 illustrates advantages of the constructed neighbourhood clusters for measurement of residential segregation in Denmark over time. Section 7 concludes.

2. Description of the First Version of Danish Residential Neighbourhoods for 1985-2004

The first version of Danish neighbourhoods was constructed based on geo-referenced and administrative register data, and presented in Damm and Schultz-Nielsen (2008). The geo-referenced data used in the first version was the National Square Grid, which assigns housing addresses in Denmark to squares or 'hectare cells' (100m×100m). In total, there were 431,233 hectare cells with occupied housing addresses on the 1st of January 1985 or 2004, the vast majority (86%) were inhabited in both years. However, very few hectare cells included enough inhabitants to be considered as a neighbourhood

Statistics Denmark has changed their confidentiality requirements between 2006 and today. To meet confidentiality requirements from Statistics Denmark in 2006, each neighbourhood had to include at least 600 households to visualise data for recognisable neighbourhoods, e.g. shown on a map, and at least 150 households to include data for neighbourhoods as a variable in an analysis. Hence, they constructed two levels of neighbourhoods: micro-neighbourhoods comprising at least 150 households and macro-neighbourhoods comprising at least 600 households.[2]

Damm and Schultz-Nielsen (2008) used the following criteria (in order of priority) to cluster hectare cells into micro- (macro-) neighbourhoods: neighbourhoods should be inhabited by at least 150 (600) households, unaltered over time,

[2] Macro-neighbourhoods always respect the boundaries of micro-neighbourhoods.

delineated by physical barriers, comprised a contiguous cluster of cells, compact, homogeneous in terms of type of housing and ownership, relatively small in terms of area, and homogeneous in terms of number of inhabitants. Using these criteria, they clustered hectare cells into 9,404 micro- and 2,296 macro-neighbourhoods. Micro-neighbourhoods include on average 227 and 264 households in respectively 1985 and 2004, while macro-neighbourhoods (same years) include 929 and 1,079 households.

The data on the individual's neighbourhood of residence constructed by Damm and Schultz-Nielsen (2008) has been used in a number of urban economic studies using Danish administrative register data as well as reports written by the Danish Ministry of Cities, Housing and Rural Districts, and has generally functioned very well.[3] Nevertheless, one problem that has emerged is that some addresses in administrative population data cannot be identified in the neighbourhood data. This is mostly a problem in earlier years; in 1986, 8.4 % of the population cannot be assigned to a neighbourhood, in 2004 it is only 0.2 % of the population. Therefore, we have paid special attention to the problem of assigning neighbourhoods to 'historical' addresses in the new construction of neighbourhoods

Besides, using hectare cell information for clustering may result in assigning housing addresses within the same hectare cells but on the opposite sides of a physical barrier (like a river) into one neighbourhood. Since physical barriers decrease the probability of social interaction among residents on the opposite sides of the barriers, implementing the Thiessen polygon method relaxes dependency of clustering approach from administrative definitions like hectare cells, and clustering takes place using the boundaries of each property, which allows neighbourhoods to have organic shapes. Therefore, using Thiessen polygon approach for clustering housing addresses can help us to improve our definition of neighbours through identifying individuals who are living within the same physical barriers and are more likely to have social interactions with each other, which is one of the crucial conditions for construction of neighbourhoods.

Moreover, construction of new roads and railways increase the number of physical barriers over time. As a result, while two residential properties were within the same physical barriers previously, construction of a new highway may split those from each other. To address this problem, the current neighbourhood clusters update the first version through considering the physical barriers in 2015 and including railways as another physical barrier in the clustering procedure.

[3] See e.g. Damm, Schultz-Nielsen and Tranæs (2006), Clark, Westergård-Nielsen and Kristensen (2009), Damm (2014), Knudsen (2014), Hviid (2015), Dustmann, Vasiljeva and Damm (2019), Dustmann and Landersø (2018), Andersen (2015; 2017) and Ministeriet for By, Bolig og Landdistrikter (2014).

3. The Data Foundation

To construct neighbourhood clusters, this study intensively benefits from several Danish high-quality registers. Below follows a brief description of each register that we use to cluster housing units into Danish residential neighbourhoods.

Housing register (BBR): The Housing register provides us with annual information about different characteristics of housing units such as housing types, ownership types, whether a housing unit is occupied (by either owner or renter) or vacant, etc.

Population register (BEF): For the purpose of this study, the Population register contains information about unique housing addresses in the municipalities wherein inhabitants of Denmark live on the 1st of January of each year. Since the Population register records the occupied housing units, using the Population register would help us to exclude vacant housing units from this study. Therefore, each observed housing unit in the Population register resembles a household and we can use housing units in the Population register and households interchangeably. It is also possible to compute the number of inhabitants in each housing unit across years. Besides, we can link the Population register to the Housing register using the unique housing address in the municipality.

The National Square Grid–Denmark: This is a national system of grids, which divides Denmark's land into hectare cells (squares). Hectare cells (100m×100m) are the smallest available grids, and each cell has a unique name. Moreover, these grids are independent from administrative boundaries like postal districts and do not change over time. Since hectare cell information at the building level is available from Statistics Denmark, one can compute different characteristics of residents at the hectare cell level through linking the Population and Housing registers to the hectare cell information. However, to keep the identity of individuals confidential, today Statistics Denmark requires a minimum of 50 households and a minimum of 100 inhabitants in each hectare cell (or a group of hectare cells)

4. Methodological Considerations

A neighbourhood is a geographical area within which neighbours meet and interact socially. Different studies show that the probability of social interactions among the residents of housing units decreases by distance between those housing units (Latané et al., 1995; Wellman, 1996; Butts, 2002), in spite of changes in technology (communication and transport). The optimal size of neighbourhoods is also an important factor. Empirical studies try to find the optimal size of neighbourhoods by conducting surveys and asking participants about the number of neighbours that they know. Results vary by different factors, such as country and socioeconomic characteristics of participants. For instance, Campbell and Lee

(1992) find that participants of their survey in the US know on average 15 neighbours, with variations between 0 and 80. Another Canadian study by Wellman (1979) finds that on average respondents talk to 5 neighbours regularly. It is also important to note that the neighbours known by an individual are not necessarily those within the closest proximity. As an example, Campbell and Lee (1992) show that social networks of low-income individuals are more localised than social networks of high-income individuals. In Denmark, Danckert, Dinesen, and Sønderskov (2017) and Dinesen and Sønderskov (2015) use the number of persons living in the defined radius (starting from 80 meters) of a person's home address to define that person's "immediate neighbourhood".

Our view on the optimal size of neighbourhoods is they should be sufficiently large to take into account that within neighbourhoods, individuals with similar characteristics sort into networks. Moreover, our aim is to construct geographic units which reflect neighbourhoods and have unchanged neighbourhood boundaries over a long time period. Our clusters have to meet the imposed minimum by Statistics Denmark in each year over a long period during which housing units may become vacant and be demolished. With larger clusters, there is room for larger changes within the cluster. The period over which the clusters have to meet the imposed minimum of Statistics Denmark exceeds our current observation period of 1986-2016 because the constructed neighbourhoods should be constructed such that they can be updated to also cover years to come. Finally, we aim to construct measures of neighbourhoods that are similar to neighbourhood measures used in the international literature on residential segregation. These three considerations lead us to use a minimum threshold, which is considerably larger than the imposed minimum by Statistics Denmark of 50 households. For international comparison, our criteria for clustering housing units to form micro- and macro-neighbourhoods are similar to the criteria used by the US Census Bureau to define census blocks and block groups. In particular, we consider seven criteria (listed in their order of priority) that are common between micro- and macro-neighbourhoods, as well as two additional criteria for macro-neighbourhoods.

4.1. Each micro- (macro-) neighbourhood should consist of at least 150 (600) households in all years during the 1986-2016 period

In many countries, census tracts represent the smallest territorial entity for which population data is available. In the US, census tracts are divided into block groups, which constitute of a number of census blocks. The US census tracts have between 1,200 and 8,000 inhabitants and on average 4,000 inhabitants, corresponding to around 2,000 households. Block groups have an optimal size of 1,500 inhabitants (600 households). The US studies of residential segregation tend to use census tracts because census tracts are intended to represent neighbourhoods and have boundaries that change little between censuses (e.g. Cutler, Glaeser and Vigdor, 1999; Massey, Rothwell and Domina, 2009; Iceland

and Weinberg, 2002; Intrator et al., 2016). Alternatively, one can use block groups, which have the advantage of being somewhat smaller (Iceland and Weinberg, 2002). For international comparison, it is useful to construct measures of residential neighbourhoods for Denmark that are similar to block groups in terms of population size and can be aggregated further to resemble census tracts. This is the main reason for which we construct macro-clusters such that they have a minimum of 600 housing units

Moreover, a minimum threshold of 600 households is sufficiently large for most analyses of causes and consequences of residential segregation. In particular, they allow for construction of average socioeconomic characteristics by demographic subgroups within a neighbourhood, allowing researchers to take into account that within the neighbourhood individuals who are similar in terms of demographic and socioeconomic characteristics sort into networks (e.g. Edin et al., 2003; Damm, 2014; Damm and Dustmann, 2014). These are the main reasons for which we choose a minimum of 600 households in all years during our observation period as the threshold for construction of macro neighbourhoods.

In view of the negative association between distance and social interactions, for some types of analyses smaller neighbourhood units may be more appropriate. To give an example, smaller neighbourhood units allow researchers to evaluate how sensitive calculations of residential segregation are to the size of the neighbourhood. The smaller neighbourhoods also have to meet the imposed minimum by Statistics Denmark of 50 households in each year. We choose a minimum threshold of 150 households in each micro-neighbourhood in each year during our observation period so that each micro-neighbourhood is also likely to meet the imposed minimum by Statistics Denmark in years to come, i.e. robust to housing vacancies and demolitions.

Given the negative association between distance and social interactions, imposed minimum by Statistics Denmark, our considerations on the optimal size of neighbourhoods for analyses of residential segregation (sufficiently large to take into account network formation according to the homophily principle, unchanged neighbourhood boundaries over a long time period, and internationally comparable), we set the minimum size of micro- (macro-) neighbourhoods at 150 (600) households, alternatively referred to as housing addresses in the Population Register or occupied housing addresses.[4]

[4] In practice we cluster housing units in the Housing Register and calculate the number of households in each cluster by linking the Housing Register with the Population Register. Since the number of housing units include vacant housing units, in each neighbourhood the number of housing units is at least as high as the number of households.

4.2. Boundaries of neighbourhoods should not change over time

Changes in the boundaries of neighbourhoods result in change in the neighbourhoods wherein each housing unit—and hence residents of that housing unit—is located over time. Since the primary interest of this study is construction of neighbourhoods for measurement of ethnic and socioeconomic residential segregation across Danish residential areas across years, uniqueness of the defined neighbourhoods for each housing unit plays an important role.

4.3. Neighbourhoods are formed by clustering inhabited residential properties based on physical proximity

This criterion is not the first priority because some islands in Denmark have too few households to constitute a neighbourhood that meets the confidentiality requirements of Statistics Denmark of 50 households and our own slightly higher minimum threshold of 150 households for micro-neighbourhoods and 600 inhabitants for macro-neighbourhoods. In such cases, we try to cluster the housing units on the islands with the closest housing units in the mainland.

4.4. Physical barriers are used as boundaries of neighbourhoods and should not cross a neighbourhood

We assume that physical barriers like water (seas, inlets, and lakes), forests, major roads and railways decrease the social interactions among residents on the opposite sides of those barriers. Therefore, we try to cluster housing units within the barriers and use those barriers as boundaries of neighbourhoods. The advantage of using physical barriers as boundaries of neighbourhoods is that physical barriers like lakes do not change across years. As a result, they guarantee that the boundaries of neighbourhoods do not change over time. However, there are special cases in which the number of housing units within those barriers do not comply with our self-imposed minimum of 150 (600) households (locked-in condition), which is the most important criterion for clustering. In those cases, we should cross the barriers and find partner housing units on the other side of the physical barrier. In this regard, we believe that the likelihood of social interactions among residents on the opposite sides of a road is higher than such likelihood for residents of the opposite sides of a big lake. As a result, we need to prioritise physical barriers based on their importance in decreasing social interactions. Hence, we classify the physical barriers by order of priority as follows: (i) large water barriers (seas, inlets, and lakes); (ii) major roads (motorways, major roads of high importance, and other major roads); and (iii) major railways (not local railways).

4.5. Neighbourhood clusters should be formed to be as homogeneous as possible in terms of housing types and ownership of the housing stock

Based on a hypothesis by Becker (1957) and Bailey (1959), the race of the neighbours affects one's preferences for place of residence. Therefore, we expect to see the highest social interactions among individuals who share the same demographic and socioeconomic characteristics. A higher concentration of minorities in specific areas of cities is evidence of households sorting into different neighbourhoods based on similarity in demographic and socioeconomic characteristics. In this regard, one might argue that construction of homogeneous neighbourhoods through clustering housing units of residents with similar demographic and socioeconomic characteristics would guarantee the highest social interaction among the neighbours. However, such a clustering has two disadvantages. First, it would lead to overestimation of residential segregation of demographic and socioeconomic groups. Second, a goal of homogeneity in terms of attributes of residents would contradict our goal of residential neighbourhoods that are unaltered over time. Among others, urban renewals, building repairs, and rent reductions may change the demographic and socioeconomic characteristics of residents without changing the housing stock in the neighbourhood. Consequently, if such housing policies are successful in increasing the mix of demographic and socioeconomic groups, neighbourhoods that were once homogeneous in terms of demographic and socioeconomic characteristics of residents, become heterogeneous after some years, which means that we need to change boundaries of neighbourhoods over time.

Since changing neighbourhoods is in contrast with the criterion of unaltered neighbourhood boundaries, we cluster housing units based on homogeneity in terms of housing types and ownership of the housing stock. This approach has three advantages. Firstly, and most importantly, stock, type, and ownership of housing units do not change too much over time, which guarantees to have unaltered neighbourhood boundaries across years. Secondly, households with similar demographic and socioeconomic characteristics sort themselves into neighbourhoods based on housing and ownership types available in those neighbourhoods. Additionally, different types of housing are distributed across a local housing (labour) market. Such a distribution is exogenous (given) to the households' decisions about sorting themselves to different areas of a local labour market. Thirdly, housing units of the same ownership type located in close proximity to each other are likely to belong to the same association, e.g. landlords' association or public housing association, facilitating social interaction between residents.

For analyses of residential segregation, one may prefer to use neighbourhood units that do not give any weight to homogeneity in terms of housing type and ownership. However, it is not given that pure distance represents social interaction better. Among housing units located within the same distance, our clustering approach gives higher weight to housing units of the same housing type and ownership.

In practice, we divide housing units based on housing types into four groups: (i) farmhouse or semi-detached houses; (ii) townhouses or small-story houses; (iii) dwellings in a multi-story residential building; and (iv) summer houses or other types of housing. We also divide housing units into four categories based on types of ownership: (i) owner-occupied; (ii) private rental; (iii) public housing and publicly rental; and (iv) cooperative housing. The combination of the four groups of housing type and four groups of ownership gives 16 possible *type-ownership* groups or combinations. Consequently, we can assign each housing unit to one out of the 16 possible housing *type-ownership* combinations. For instance, one combination includes all of the public housing and public rental apartments, and another combination includes owner-occupied single-family (farmhouse or semi-detached) houses.

We can find housing and ownership types of each housing unit from the Housing register annually, and assign that housing unit into one out of 16 possible combinations of housing and ownership types. Then, to find similar housing units around each housing unit, we need to give weight to housing type and ownership type. As argued by Damm and Schultz-Nielsen (2008), since our primary interest from construction of neighbourhoods is evaluating ethnic and socioeconomic segregation in Danish neighbourhoods across years, and more importantly socially deprived areas that have attracted lots of attention in recent years, we give a weight of 0.3 for housing type and a weight of 0.7 for ownership type. The reason for giving a higher weight for ownership is that the type of ownership is the main determinant of socially deprived areas in Denmark. While only a small fraction of public housing areas are socially deprived, socially deprived areas tend to be locations with a high concentration of public housing. Therefore, despite the possibility of having variation in housing type in socially deprived areas, ownership is dominated by public housing.

4.6. Neighbourhoods should be homogeneous in terms of number of inhabitants and relatively small in terms of area

To evaluate residential segregations by ethnicity and socioeconomic characteristics, we should calculate the minority group share within each neighbourhood and compare the shares across all of the neighbourhoods. Variation in the minority group share across neighbourhoods in a city stems from two sources: Variation in the number of minority group members and variation in the number of inhabitants. As a result, when neighbourhoods of a city are heterogeneous in terms of number of inhabitants, one should be careful to conclude that relatively high minority group shares result from relatively high concentration of the minority group in question. By contrast, differences in minority group shares across neighbourhoods that are homogeneous in terms of number of inhabitants can without any ambiguity be attributed to differences in the number of minority group members across neighbourhoods. Moreover, a study by Bolster et al. (2006) shows that by increasing the number of inhabitants in a neighbourhood, the estimated neighbourhood effect decreases. Part of the reason may be due to a decrease in social interaction with geographic distance between the individuals. Heterogeneous neighbourhoods in terms of number of inhabitants and area may therefore lead to attenuation bias in the estimation of neighbourhood effects, because the level of residential segregation of ethnic and socioeconomic groups is underestimated in large neighbourhoods

4.7. Neighbourhoods should be compact

Compactness means that other neighbourhood cluster(s) should not split a neighbourhood into two or several parts or surround that neighbourhood. Non-compactness of a neighbourhood decreases social interaction among inhabitants of the separated parts of that neighbourhood. While clustering based on a proximity criterion ensures the compactness of most neighbourhoods, as will be discussed later, we implement an additional stage for clustering housing units to maximise the share of compact neighbourhoods, given the six other criteria.

Compactness means that other neighbourhood cluster(s) should not split a neighbourhood into two or several parts or surround that neighbourhood. Non-compactness of a neighbourhood decreases social interaction among inhabitants of the separated parts of that neighbourhood. While clustering based on a proximity criterion ensures the compactness of most neighbourhoods, as will be discussed later, we implement an additional stage for clustering housing units to maximise the share of compact neighbourhoods, given the six other criteria:

4.7.1. Each macro-neighbourhood should consist of a group of adjacent micro-neighbourhoods

By definition, a group of micro-neighbourhoods should form a macro-neighbourhood. That is, we construct macro-neighbourhoods through aggregating adjacent micro-neighbourhoods

4.7.2. Boundaries of macro-neighbourhoods should respect the boundaries of micro-neighbourhoods

Each micro-neighbourhood should belong to one macro-neighbourhood over time. To meet this condition, macro-neighbourhoods should respect the boundaries of micro-neighbourhoods and should not cross any of the encompassed micro-neighbourhoods.

5. Implementation of the Clustering Algorithm and Final Cluster Characteristics

5.1. Implementation of the Clustering Algorithm

Our clustering algorithm consists of two parts and eight stages, which are presented in more detail below. In the first part, in order to make sure that the number of occupied housing units in the constructed neighbourhoods is higher than our self-imposed minimums (150 and 600 households) in more than one year, and that neighbourhood boundaries do not change, 1986, 2000, and 2015 are selected as pilot years. Then, based on the mentioned criteria and using the Housing register, the clustering algorithm is written, executed, and tested in collaboration with the firm Geomatic. The first part consists of three stages and provides us with neighbourhood clusters in the mentioned three years.

The second part of clustering algorithm starts with linking the observed housing addresses in the Population register during 1986-2016 to the constructed neighbourhoods (hereafter "assigning cluster ID for the housing addresses"). However, due to existence of housing addresses that we cannot observe in the Housing register in those 3 pilot years and some inconsistencies between Housing and Population registers, a group of housing addresses remain without cluster ID in the Population register. In order to increase the number of housing units with cluster ID, stages four and five are implemented. At the end of the fifth stage, we have a group of micro- (macro-) clusters with less than 150 (600) occupied housing units in some years. Those clusters are treated in the sixth stage. After making sure that all the micro- and macro-clusters follow the imposed minimums by Statistics Denmark in each year during 1986-2016, we try to split large neighbourhoods into two or more small neighbourhoods and improve the homogeneity in terms of number of inhabitants in the seventh stage. Finally, in the eighth stage, a group of neighbourhoods that are not compact are identified and addressed. Below, we detail each stage.

Stage 1: Clustering addresses using Thiessen polygons[5]

As mentioned under the 5th criterion in Section 4, each housing unit can take 1 out of 16 combinations of housing *type-ownership* (hereafter "seeds"). One housing unit from each of those 16 seeds is selected in a randomly located area. The following process loops over the currently active seeds (hereafter clusters), and

[5] Thiessen polygons (also known as Voronoi polygons) were introduced by Alfred H. Thiessen (1911) in order to allocate the areas to the nearest weather stations. The idea behind this approach is dividing an area into smaller areas (polygons) based on available values at points in a way that any location within a polygon has the lowest distance to the reference point relative to the other points. The Thiessen polygon approach is applied for analysis of proximity and neighbourhoods, such as catchment areas of businesses.

leads them to select one new address in each round. The address is checked for relevant barriers (water, major roads, major railway). If the recently selected address is outside of the barriers of the cluster, the address is dropped and a new address is selected. Otherwise, the recently selected address is added to the cluster, and jointly with the previous addresses in the cluster the point cluster is shaped as a concave hull.[6] Therefore, each cluster grows as a polygon in an amoeba-like fashion. To ensure that the clusters grow through the entire shape of the polygon rather than the last used point, distances between the new candidate addresses and the polygon are calculated as polygon-to-point. Hence, the candidate point with the lowest distance is selected

The size of clusters (i.e. active seeds) are checked in each loop and set to inactive when they reach the minimum of 150 housing units. Whenever a cluster reaches the minimum number of housing units or cannot find more housing units within the barriers (locked-in condition), a new seed of the same combination is planted at a randomly located address. The script runs until there is no housing unit left to pick.

Figure A1.a illustrates an example of the finished point clusters at the end of the first stage, i.e. for a given location in Denmark. In this figure, addresses (points) that belong to the same cluster are determined by the same colour, and clusters are pure with regard to housing type-ownership. In some cases, housing addresses in the same chain belong to different clusters due to differences in ownership and/or types of those housing units. Figure A1.b presents the drawn polygon for each point, using the Thiessen (Voronoi) polygon function over the points. In particular, it illustrates how the Voronoi polygon function draws a polygon for each point, creating splits halfway between each point combination. In Figure A1.c, the points are dissolved to polygon clusters.

Stage 2: Post-clustering of clusters (optimisation step)

At the end of the first stage, there are micro-clusters that have less than 150 housing units. The solution is clustering these small micro-clusters with the existing clusters. The most important factor in clustering the micro-clusters is proximity. Therefore, it is a prerequisite that all potential micro-clusters should be neighbours of the small micro-cluster (polygon neighbours, sharing a common border). Moreover, those neighbouring micro-clusters should not be separated from the small micro-cluster by physical barriers (within barriers clustering). For each of those neighbouring micro-clusters, a score consisting of the calculated distances in housing type and ownership type is assigned. These distances, as mentioned in Section 4, are weighted with $''Housing\ type \times 0.3 + Housing\ ownership \times 0.7''$ and sorted from small to large. The micro-cluster with the smallest distance is

[6] Using the PostGIS function ST_ConcaveHull.

chosen for clustering (called 'partner') with the small micro-cluster. However, in cases where there are two or more candidate micro-clusters with similar and small distances, the micro-cluster with the smallest number of housing units is selected for clustering. This approach continues until all the micro-clusters reach the limit of 150 housing units. However, some micro-clusters still could not find any partner within the boundaries (locked-in condition), which is treated in the third stage through crossing the barriers.

The bottom-up creation of macro-clusters from micro-clusters works with the same principle as clustering small micro-clusters that are below the size limit: finished micro-clusters find suitable neighbours for clustering until the macro-cluster's minimum size limit (600 households) has been reached. Clustering of micro-clusters to form macro-clusters can also reach a locked-in situation within the physical barriers. This situation is also solved using the cross-barrier method in the third stage.

Stage 3: Cross-barrier clustering of clusters

Some clusters are naturally locked-in by physical barriers and have no chance to reach their size threshold due to an insufficient number of housing units within the barriers. This can be the case for both micro- and macro-clusters, and even more often among macro-clusters because of the higher threshold for the minimum number of households (600). The solution is to let clusters cross the physical barriers to look for potential partners, thereby reaching the size thresholds.

The cross-barrier approach follows a set of rules. First, no other polygon is located between the current small cluster and the partner cluster. That is, the new cluster must be geographically compact. Secondly, for crossing the barriers to find a partner cluster we follow the priority list described under criteria four in Section 4. For instance, a cluster choosing between a potential partner that lies across a water barrier and a partner across a road will choose the partner on the opposite side of the road.

Appendix Figure A1.d presents the micro-neighbourhoods for the same location as in Figures A1.a-A1.c at the end of the third stage where within and cross barriers procedures are implemented over the small micro-neighbourhoods. Appendix Figures A2 and A3 also show examples of the constructed micro- and macro-clusters at the end of the third stage of the clustering algorithm.[7]

Stage 4: Assigning cluster ID to the observed housing addresses in Population registers

At the end of the third stage, we have cluster ID for the observed housing addresses in the Housing register in 1986, 2000, and 2015. Since the aim of this study

[7] We hired the firm Geomatic to construct the algorithm for stages 1-3 such that neighbourhoods were constructed according to our prioritised list of clustering criteria in Section 4.

is to assign cluster ID to all the addresses that we observe in the Population register, we try to link the addresses with cluster ID at the end of the third stage to the observed addresses in the Population register using the unique housing addresses at municipality level.[8] Using all of the available information in Population and Housing registers, we could assign cluster ID for 96.7% of the 75,711,527 housing address-years that we observe during 1986-2016.

Stage 5: Assigning cluster ID to the observed housing addresses using hectare cell information

To assign cluster ID to the last 3.3% of the housing addresses, we can use hectare cell information of housing addresses with cluster ID. However, since we use the Thiessen polygon method to cluster adjacent housing units, there are hectare cells with more than one cluster ID. Therefore, we need to assign a ″dominant″ cluster ID for each hectare cell, defined as the cluster ID to which the highest number of housing units in the hectare cell belong. Appendix B describes our approach in detail.

After finding a dominant cluster ID for each of the observed hectare cells, we assign the dominant cluster ID of these hectare cells to the housing addresses in those hectare cells that do not have a cluster ID. As a result, we assign cluster ID for an additional 2.4% of the housing addresses, while 0.9% remain without cluster ID. Most of these housing addresses are located in 20,676 hectare cells without dominant cluster ID. Therefore, we can link these hectare cells to the adjacent micro-clusters using the minimum distance approach.[9] Having dominant

[8] In fact, we use two variables of KOM and BOPIKOM to identify each housing unit. KOM represents the municipality wherein a housing unit is located, and BOPIKOM is an encrypted housing address, which consists of road number, house number, house letter, floor, and side/door within a municipality. Additionally, OPGIKOM is a shorter version of BOPIKOM, which gives information about road number, house number, and house letter within a municipality. Therefore, housing units in an apartment share a common OPGIKOM. We can observe KOM, BOPIKOM, and OPGIKOM for each housing unit in housing and population registers. Since housing units within an apartment get one cluster ID, using OPGIKOM information of housing units with cluster ID would help us to assign the same cluster ID to other housing units without cluster ID (as be discussed in the fifth stage). We use this approach for small number of housing units within the same building to improve the number of housing units with cluster ID. It is also important to note that there is not any BOPIKOM or OPGIKOM with more than one cluster ID. Moreover, hectare cell information of housing units is available at building (OPGIKOM) level. Hence, we need to use OPGIKOM to assign cluster ID for a group of housing units without cluster ID.

[9] The minimum distance approach is implemented in the following way: in order to assign cluster ID for the 20,676 hectare cells, we identify all of the hectare cells that are within the 3-kilometer distance from each of the hectare cells without dominant cluster ID (hectare cell X). Using dominant cluster ID of all 438,821 hectare cells, we can find micro-clusters that are within the vicinity of hectare cell X. Then, average distances between hectare cell X and all of the hectare cells within each of those adjacent micro-clusters are computed. Later, hectare

cluster ID of these 20,676 hectare cells, we can assign cluster ID to additional 0.6% of the housing addresses in those hectare cells.

Appendix Table A1 shows the observed housing addresses and persons with or without cluster ID in the Population register during 1986-2016. As this table illustrates, we can assign cluster ID to 99.7% of the housing addresses in the Population register. Besides, 98.9% of housing addresses that we observe in the Population register at least once (unique addresses) have cluster ID.

Appendix Figure A4 exhibits the frequency of housing addresses without cluster ID by year, municipality, housing type, and ownership type. According to this figure, the number of housing units without cluster ID has a declining time trend (A4.a). The reason for such a trend is increased precision and compatibility of Housing and Population registers in recent years. Moreover, 30% of housing units without cluster ID are located in large municipalities like Copenhagen, Aarhus, Odense, and Aalborg (A4.b), and, as we expect, more than 50% of those belong to ″other types or not identified″ properties that are not regularly used for residential purposes. Additionally, close to 20% of housing addresses without cluster ID are among the housing types that one can see in rural areas (townhouses, detached single-family houses, and cottages), and about 30% of those have housing types that one can observe in urban areas, i.e. apartments, and row, chain, or double houses (A4.c).

With regard to the type of ownership of housing addresses without cluster ID, housing addresses with unidentified ownership constitute around 45% of those housing units, followed by owner-occupied and private rental housing units (28%), and public housing (15%) (A4.d).[10]

Since all of the micro- (macro-) clusters should meet our self-imposed threshold of minimum 150 (600) households, we need to identify and treat clusters that are too small. In this regard, we define a micro- (macro-) cluster as ″too small″ when that cluster has less than 150 (600) households in one or more years during 1986-2016. Appendix Table A2 shows that 282 micro-clusters and 26 macro-clusters are ″too small″ at the end of stage 5.

Stage 6: Merging ″too small″ clusters in terms of number of households to meet our threshold of a minimum number of households of 150 (600)

In order to solve the problem with ″too small″ micro-clusters (below 150 occupied housing addresses), we identify all the micro-clusters that belong to the macro-cluster wherein the ″too small″ micro-cluster is located. Then, we restrict the selected micro-clusters to those with more than 150 occupied housing addresses

cell X is linked to the micro-cluster with the lowest average distance. Finally, the linked micro-cluster is assigned as the dominant cluster ID for hectare cell X.

[10] Appendix Table A2 shows characteristics of the constructed micro- and macro-clusters in 1986, 2000, and 2015 at the end of this stage.

in years where the ″too small″ micro-cluster has less than 150 occupied housing addresses; such micro-clusters are potential partners for link.[11] Next, we find hectare cells wherein ″too small″ micro-clusters and the potential partners for link are dominant. Finally, as Appendix C describes, we calculate the average distance between all the hectare cells of the ″too small″ micro-cluster and hectare cells of each of the potential partners for link, and link the ″too small″ micro-cluster with a potential partner with the smallest average distance. In fact, this approach helps us meet our criterion of physical proximity of housing addresses, which is our third-most important criterion for clustering.

For ″too small″ macro-clusters, in turn, we identify all of their adjacent macro-clusters that have more than 600 occupied housing addresses. Then, average distances between hectare cells of the ″too small″ macro-cluster and adjacent macro-clusters are calculated. Finally, we link the ″too small″ macro-cluster with the adjacent macro-cluster with the lowest average distance.

Appendix Table A3 shows characteristics of the neighbourhoods after merging ″too small″ clusters with adjacent clusters based on geographical proximity. As expected, the number of micro- (macro-) clusters decreases from 8,241 (1,851) to 7,979 (1,827). Consequently, the average number of housing addresses, inhabitants, and areas of clusters increase.

At the end of this stage, all micro-clusters (macro-clusters) have more than 150 (600) households in each year during 1986-2016. However, 387 micro-clusters and 138 macro-clusters have three times more households than the required minimum, henceforth referred to as ″overly large″.

Stage 7: Splitting ″overly large″ clusters to obtain homogeneous clusters in terms of number of households

Since all clusters have to satisfy our first two criteria of a minimum of 150 (600) households in each micro- (macro-) cluster and unaltered boundaries of clusters over 31 years, the constructed clusters at the end of the sixth stage are larger than our self-imposed thresholds. To obtain clusters that are as homogeneous as possible in terms of the number of households and area, we address the problem of ″overly large″ clusters in a way such that the two first criteria still hold.

[11] Most of the ″too small″ micro-clusters are too small during 1986-1988 and not in the following years. Additionally, because of new constructions and urban renewals, some micro-clusters only exist from 1989. Since we do not observe these micro-clusters during 1986-1988, we cannot link them to the micro-clusters that are too small during 1986-1988. As a result, we restrict the potential partners to those that exist in years where the too small micro-cluster has less than 150 housing addresses.

To address the issue of ″overly large″ micro-clusters, first, we identify all the hectare cells wherein an ″overly large″ micro-cluster is dominant and count the number of housing addresses in the ″overly large″ micro-cluster in each of those hectare cells. Secondly, we calculate the average distance between each of the hectare cells and other hectare cells within the ″overly large″ micro-cluster and sort them based on average distance. Thirdly, the hectare cell with the highest average distance (i.e. the furthest hectare cell) is selected (hectare cell X) and the number of housing addresses in X is calculated. Fourthly, X is linked to its closest hectare cell (Y) to form a new sub-cluster. The number of housing addresses in the new sub-cluster is calculated, and the procedure of adding the closest hectare cells to the new sub-cluster continues (loop) until the number of housing addresses in the new cluster reaches (about) half of the number of housing addresses of the ″overly large″ micro-cluster.

For each ″overly large″ macro-cluster, in turn, we identify all of the included micro-clusters, number of housing addresses in each of the micro-clusters, and hectare cells wherein the micro-clusters are dominant. Then, we calculate the average distance between micro-clusters (i.e. average distance among hectare cells of a micro-cluster and hectare cells of the other micro-clusters within the ″overly large″ macro-cluster) and select the micro-cluster with the highest average distance (micro-cluster Z). Then, Z is linked to its closest micro-cluster (W) to form a new sub-cluster. Later, the number of housing addresses in the new sub-cluster is calculated, and the procedure of adding the closest micro-clusters to the new sub-cluster continues (loop) until the number of housing addresses in the new cluster reaches (around) half of the number of housing addresses in the ″overly large″ macro-cluster.

Appendix Table A4 shows characteristics of the neighbourhoods after splitting the ″overly large" clusters to smaller ones. After the split, the number of micro-clusters (macro-clusters) increases from 7,979 to 8,359 (1,827 to 1,961). Besides, comparison of the clusters at the end of stage 5 (Appendix Table A2) with clusters after splitting shows significant reductions in the mean and standard deviation of the number of occupied housing addresses, inhabitants, and area of each cluster, while medians do not change importantly. The reason is significant changes in the right tails of the distributions of the number of occupied housing addresses and inhabitants and of area (95th percentile). In other words, the procedure of splitting mostly affects clusters that are in less populated areas.

Stage 8: Compactness of clusters

As described in Section 4, ”compactness” means that neighbourhoods should not be split by another neighbourhood into two or several parts or entirely surround another neighbourhood. To check if neighbourhoods are compact, hectare cell information is assigned to each cluster.[12] We then find the nearest hectare cells including housing addresses in cardinal (north, south, east, west) and inter-cardinal (northeast, southeast, southwest, northwest) directions. This information is used to check if each cluster consists of compact hectare cells. A cluster where all hectare cells can be combined through a nearest hectare cell belonging to the same cluster is considered 100% compact. If some of the hectare cells in a cluster are separated from the other part of the cluster, we calculate the share of the cluster that is non-compact. This is done by considering the largest part (measured by average number of housing addresses during 1986-2016) as the compact part of the cluster and all other parts as non-compact. The compactness rate of a cluster in each year is calculated using the number of housing addresses within each cluster and their hectare cell location (formula given in Appendix D).

In line with our expectation, we find that the compactness rate is lower in less populated areas than in cities, because the population density is lower, and the cluster areas are larger and more widespread. However, we also detect a minor number of clusters for which compactness can be improved by moving a smaller share of addresses from one cluster to the adjacent cluster—without violating the minimum of 150 (600) households in each cluster. In all, we move 0.3 % of the housing addresses to the adjacent clusters. By including this adjustment, we finalise the clusters and make a final check of the compactness rate.

Appendix Table A5 reports descriptive statistics on the compactness rate of our micro- and macro-clusters. The average compactness rate is 98.7 among all 8,359 micro-clusters, while the median is 100, which means that outliers in the right tail of distribution decrease the average compactness rate. The same is true for macro-clusters. The average compactness rate of micro-clusters improves slightly from 99.7% in 1986 to 100% at 75th percentile, which is due to the new constructions after 1986 in the more populated cluster areas, where the compactness rate is generally higher. The average compactness of macro-clusters weighted across years is 98.6% and it is stable from 1986 to 2015, while the standard deviation declines from 3.5 to 3.2.

[12] This information includes all hectare cells that are inhabited at least once in the period 1986-2016, but most of the hectare cells are inhabited in all the years.

The completion of this stage leads to our final clusters which we will describe in the next subsection.

5.2. Final Clusters

5.2.1. Descriptive Characteristics of Clusters

In this subsection, we describe the constructed neighbourhoods along various dimensions.

Table 1 summarises the characteristics of the final clusters in 1986, 2000, and 2015. Based on this table, the 459,497 inhabited hectare cells of Denmark are clustered into 8,359 micro- and 1,961 macro-neighbourhoods. Besides, some points are worth taking into consideration. First, there are fewer micro- (macro-) clusters in 1986 and 1987 than in subsequent years. The increase in the number of micro- (macro-) clusters from 8,350 to 8,359 (from 1,960 to 1,961) after 1987 is due to the construction of around 2,200 public housing and cooperative housing units in Odense in 1988. Secondly, differences between mean and median in the number of occupied housing addresses, inhabitants, and area are due to the existence of some clusters that need to find partners (housing addresses) from far distance in order to reach the minimum sizes. Such clusters, which are mostly located in less populated areas, are detectable in the 95th percentile of Table 1. Thirdly, the reason for increase in the number of housing addresses and residents over years is our second criterion for clustering, which states boundaries of neighbourhoods should remain unchanged over time. Consequently, boundaries of neighbourhoods are formed in a way that they include all of the current and future housing units. Therefore, new developments increase the number of housing units, inhabitants, and the number of covered hectare cells by each micro- and macro-cluster.

Table 1. Characteristics of the final clusters in 1986, 2000, and 2015

		Percentiles					All		
		5%	25%	50%	75%	95%	Mean	Std. Dev.	Number
Micro-clusters									
Housing addresses	1986	167	197	246	310	408	260.4	77.2	8,350
	2000	172	214	271	343	468	290.9	101.6	8,359
	2015	172	221	287	368	548	315.2	140.3	8,359
Inhabitants	1986	328	472	571	718	979	606.1	197.8	8,350
	2000	328	472	585	751	1,067	534.0	241.2	8,359
	2015	336	475	605	789	1,218	674.0	322.7	8,359
Area (ha)	1986	2	9	25	58	172	46.5	55.2	8,350
	2000	2	9	26	62	179	48.9	58.0	8,359
	2015	2	9	27	67	181	50.6	59.4	8,359
Too small micro-clusters									0
Overly large micro-clusters									37

Macro-clusters									
Housing addresses	1986	698	870	1,069	1,316	1,662	1,109.6	297.0	1,960
	2000	754	955	1,174	1,476	1,900	1,240.0	375.1	1,961
	2015	771	996	1,248	1,611	2,178	1,343.6	470.1	1,961
Inhabit-ants	1986	1,421.5	1,947	2,447.5	3,100	4,189	2,582.2	837.7	1,960
	2000	1,454	2,005	2,505	3,262	4,532	2,702.8	959.7	1,961
	2015	1,503	2,058	2,636	3,472	4,929	2,876.9	1,133.2	1,961
Area (ha)	1986	10	46	113	322	601	197.6	197.7	1,960
	2000	10	43	121	336	626	207.9	208.0	1,961
	2015	11	44	124	351	635	215.2	213.2	1,961
Too small macro-clusters									0
Overly large macro-clusters									8

Notes: This table shows characteristics of the constructed micro- and macro-clusters in terms of the number of housing addresses in the Population Register, inhabitants and area at the end of stage 8 (i.e. final stage) in 1986, 2000, and 2015. The sources of this table are own calculations based on data from Housing register, Population register, hectare cell information of housing units, the constructed neighbourhood clusters, and information from Geomatic. "Percentile" columns show the characteristics at 5-95 percentile rank. "All" columns use all of the micro- and macro-clusters to compute average and standard deviation of each of the characteristics. Area (ha) means the number of hectare cells in which a micro-cluster (macro-cluster) is dominant. Too small micro-clusters (macro-clusters) are micro-clusters (macro-clusters) with less than 150 (600) housing addresses in at least one year during 1986-2016. Overly large micro-clusters (macro-clusters) present the number of micro-clusters (macro-clusters) with more than 450 (1,800) housing addresses in all of the observed years during 1986-2016.

Table 2 shows average characteristics of the micro- and macro-clusters during 1986-2016. An average micro-cluster consists of 291 occupied housing addresses (median 270), and 637 inhabitants (median 585). Additionally, an average macro-cluster contains 1,241 occupied housing addresses (median 1,173) and 2,717 inhabitants (median 2,520). Since the standard deviation of the number of occupied housing addresses and inhabitants at both levels of neighbourhood clusters are below 50% of the mean, and the mean and median are close to each other, we can say that the constructed neighbourhoods are homogeneous in terms of the numbers of occupied housing addresses and inhabitants. Our later comparison of characteristics of neighbourhood clusters, municipalities, and postal districts provide further documentation for this claim

Table 2. Average Characteristics of the final clusters in 1986-2016

	Percentiles					**All**				
	5%	**25%**	**50%**	**75%**	**95%**	**Mean**	**Std. dev.**	**Min.**	**Max**	**Number**
Micro-clusters										
Housing addresses	171	213	270	324	474	291.3	109.3	150	2,690	8,359
Inhabitants	326	427	585	752	1,068	637.4	256.9	160	6,356	8,359
Area (ha)	<10	<10	26	63	179	48.9	57.7	<10	443	8,359
Macro-clusters										
Housing addresses	742	945	1,173	1,474	1,930	1,241.6	394.0	606	4,922	1,961
Inhabitants	1,440	1,994	2,520	3,278	4,564	2,717.2	992.4	868	11,590	1,961
Area (ha)	10	43	121	338	624	207.9	207.2	<10	998	1,961

Notes: This table shows average characteristics of the constructed micro- and macro-clusters in terms of the number of housing addresses in the Population Register, inhabitants and area at the end of stage 8 (i.e. final stage) during 1986-2016. The sources of this table are own calculations based on data from Housing register, Population register, hectare cell information of housing units, the constructed neighbourhood clusters, and information from Geomatic. "Percentile" columns show the characteristics at 5-95 percentile rank. "All" columns use all of the micro- and macro-clusters to compute average and standard deviation of each of the characteristics. Area (ha) means the number of hectare cells in which a micro-cluster (macro-cluster) is dominant

However, an average micro-cluster covers 49 hectare cells, which is around twice as large as the median of 26 hectare cells. The same story holds for macro-clusters, where the mean and median are 208 and 121 hectare cells, respectively. The reason for large differences between the mean and median is that we try to divide all of the inhabited land of Denmark into neighbourhoods by respecting our self-imposed minimum number of households. Since the densities of inhabited land are nontrivially lower in rural than urban areas, neighbourhoods in rural areas cover more hectare cells

As reported in Table 3, an average macro-cluster consists of 4.26 micro-clusters (median 4) with standard deviation of 1.27 (i.e. one third of the mean) suggesting that the macro-clusters are homogeneous in terms of the number of micro-clusters that they cover.

Table 3. Number of micro-clusters in each macro-cluster for the final clusters

Year	Percentiles					All				
	5%	25%	50%	75%	95%	Mean	Std. dev.	Min.	Max.	Number
1986	3	4	4	5	7	4.26	1.27	2	8	1,960
2000	3	4	4	5	7	4.26	1.28	2	11	1,961
2015	3	4	4	5	7	4.26	1.28	2	11	1,961

Notes: This table shows the number of micro-clusters in each macro-cluster in 1986, 2000, and 2015. The sources of this table are own calculations based on data from Housing register, Population register, hectare cell information of housing units, the constructed neighbourhood clusters, and information from Geomatic. "Percentile" columns show the number of micro-clusters in each macro-cluster at 5-95 percentile rank. "All" columns use all of the macro-clusters to compute average and standard deviation of the number of micro-clusters in each macro-cluster.

In addition, Appendix Table A6 shows that the vast majority of hectare cells are occupied by only one micro-cluster. On average, each hectare cell belongs to 1.09 micro-cluster (median 1), and the hectare cells in the 95th percentile are occupied by two micro-clusters. Moreover, Appendix Table A7 reports the number of hectare cells belonging to one or more than one micro-cluster in 1986, 2000, and 2015. Among 422,049 inhabited hectare cells in 2015, for instance, more than 91% belong to one micro-cluster, and extra 7% cover two micro-clusters. As a result, less than 2% of hectare cells are divided between three to six micro-clusters. 91% of the 459,497 inhabited hectare cells during 1986-2016 are occupied by only one micro-cluster.

5.2.2. Comparison of constructed neighbourhoods with municipalities and postal districts

A relevant question is what are the advantages of the constructed neighbourhoods relative to other definitions of neighbourhoods such as municipalities and postal districts? To answer this question, first note that, to obtain neighbourhoods that reflect social interactions between residents, we have constructed the neighbourhoods based on proximity of housing units, physical barriers, and homogeneity in terms of housing and ownership types. By contrast, municipalities and postal districts are defined for administrative purposes and vary substantially in terms of the mentioned criteria. Secondly, Appendix Table A8 compares the number of occupied housing addresses and inhabitants in municipalities, postal districts, and the constructed micro- and macro-clusters in 1986, 2000, and 2015. For brevity, we compare characteristics of these neighbourhoods in 2015; however, the same story holds for other years. On average, there were 26,632 occupied housing addresses and 57,168 inhabitants in each municipality (medians of 17,885 and 42,601, respectively) with standard deviations that are (about) equal to the respective mean. The standard deviation of the number of housing addresses (inhabitants) at micro- and macro-cluster levels is below 50% of the mean at that level. The difference between the minimum and maximum is modest; micro- (macro-) clusters have between 150 and 2,432 (614 and 4,888) occupied housing addresses. Besides, important features like geographic distances between housing units, physical barriers, as well as housing and ownership types, which can decrease interactions among inhabitants of a neighbourhood, are systematically considered in our procedure of clustering housing units to neighbourhoods.

By contrast, comparison of the minimum and maximum across municipalities reveals very substantial variation in the number of occupied housing addresses and inhabitants, which documents that municipalities are heterogeneous in terms of the number of households and inhabitants. Postal districts are even more heterogeneous on those dimensions. Comparison of the standard deviation with the respective mean reveals that the standard deviation of the number of housing addresses (inhabitants) is twice as large as the mean at postal district level. Besides, the minimum and maximum number of housing addresses (inhabitants) are very different. In 2015, at least one postal district had fewer than 10 housing addresses and inhabitants. In the same year, at least one postal district had 60,665 housing addresses and 92,537 inhabitants.

5.2.3. Comparison with the first version of clusters by Damm and Schultz-Nielsen (2008)

Our current neighbourhood clusters have at least four advantages relative to the first version of neighbourhoods constructed by Damm and Schultz-Nielsen (2008). First, instead of using hectare cells, current neighbourhoods benefit from the Thiessen polygon approach and consideration of additional physical barriers for clustering, in particular major railways. Such an approach helps us put any kind of administrative definition aside and consider physical barriers as boundaries of neighbourhoods. Secondly, we use all the occupied housing addresses during 31 years (1986-2016) rather than two years of 1985 and 2004 to make sure that all clusters respect our self-imposed minimums (that exceed the imposed minimums by Statistics Denmark) across several years, and boundaries of neighbourhood are well defined for such a long period. In other words, we assign cluster ID for each housing unit that we observe in the Population register (i.e. occupied housing unit) at least once during 1986-2016. It is also possible for us to assign cluster ID to new residential properties in the future by using hectare cell information of those housing units. Thirdly, we identify the neighbourhood for almost 100% of the housing units in all calendar years over the 1986-2016 period, whereas Damm and Schultz-Nielsen (2008) have a problem of identifying the neighbourhood for historical addresses. Finally, the current neighbourhood clusters update the first version through considering the physical barriers in 2015 and including major railways as another physical barrier in the clustering procedure.

Appendix Table A9 compares the first and current versions of micro- and macro-neighbourhoods in terms of the number of occupied housing addresses, inhabitants, and area in 1986[13] and 2004. To begin with, the current version clusters 459,497 inhabited hectare cells of Denmark into 8,359 micro- and 1,961 macro-clusters, while the first version clustered 431,233 inhabited hectare cells into 9,086 micro- and 2,295 macro-clusters. Because of decrease in the number of neighbourhood clusters, we expect the current neighbourhoods to be larger than the first version. For example, on average, there were 296 occupied housing addresses and 642 inhabitants in each current micro-clusters in 2004, while those numbers were 272 and 555 in the first version of micro-clusters, respectively. Besides, the average number of hectare cells in the current version of micro-clusters is marginally larger than that average in the first version of micro-clusters (48.9 vs. 47.5). However, comparison of the ratio of the standard deviation relative to the mean reveals lower variation regarding the number of housing addresses, inhabitants, and area of the current micro-clusters compared to the first version. The same story holds for macro-neighbourhoods in the second panel of Appendix Table A9. Therefore, the current neighbourhood clusters are marginally more homogeneous than the first version of neighbourhoods on the aforementioned dimensions.

Finally, according to the last two rows of Appendix Table A9, the number of micro-clusters in each macro-cluster is slightly higher in the current version relative to the first version of neighbourhood clusters (4.2 vs. 4.0). Since the reduction in the number of macro-clusters (from 2,295 to 1,961) exceeds the reduction in the number of micro-clusters (from 9,086 to 8,359), the higher number of micro-clusters per macro-cluster in the current version is as expected.

5.2.4. Comparison with small neighbourhoods in the US[14] and Sweden

As mentioned in Section 4, the criteria that we use to construct neighbourhoods through clustering housing units are similar to those used by the US Census Bureau to define census blocks and block groups.

Census blocks are the smallest statistical areas for which decennial census data is collected and tabulated. Boundaries of census blocks are formed by visible features such as roads, streets, railroads, and streams, and by nonvisible features like

[13] Since the start year of clustering in the first version was 1985, we compare characteristics of the first version in 1985 with the characteristics of the current version in 1986.

[14] All of the definitions are from Chapter 11 of ”Geographical Areas Reference Manual” section of the United States Census Bureau website (https://www.census.gov/) which was updated on May 16, 2018, and ”Geographic Terms and Concepts” in Appendix A of ”2010 Census Summary File 1” of the United States Census Bureau.

selected property lines, school districts, and county limits. While census blocks in cities are bounded on all sides by streets, census blocks in rural and remote areas can be large and bounded by a variety of features. There are also census blocks without any inhabitant. Besides, to expand the geographic coverage of census blocks, the US Census Bureau has expanded the Census Blocks Statistics Program each decade since 1940, which means that the boundaries of census blocks have been altered over time.

Block groups, in turn, are the next level of statistical divisions above the census blocks. In other words, a block group is a combination of census blocks, and generally defined to contain between 600-3,000 inhabitants, with an optimal population size of 1,500 inhabitants (600 housing units). The block group is the smallest geographic area for which the decennial census data is tabulated and published.

Among Scandinavian countries, Statistics Sweden created Small Areas for Market Statistics (SAMS) from dividing municipalities (for large municipalities) and election districts (for small municipalities) to 9,200 small geographic areas in 1994. With exception of minor adjustments in 2003, boundaries of SAMS division have remained constant across years (Statistics Sweden, 2005; Åslund et al., 2017). SAMS areas are the smallest geographic entities that are homogeneous in terms of housing and ownership types, and encompass (on average) 1,000 inhabitants (Åslund et al., 2011).

The constructed micro-neighbourhoods in Denmark are comparable in the number of inhabitants to census blocks in the US and SAMS areas in Sweden. Nevertheless, in comparison to the census blocks, the boundaries of the constructed micro-neighbourhoods are unaltered over time, and are relatively small in comparison with SAMS areas. Moreover, in terms of the number of inhabitants, the constructed Danish macro-neighbourhoods are comparable with block groups in the US. By clustering 2-3 adjacent macro-neighbourhoods into a larger cluster, one can obtain Danish neighbourhoods that are similar to US census tracts in terms of population size.

6. Relevance of our Neighbourhood Data for Measurement of the Effects of Interventions into Socially Deprived Neighbourhoods

According to the Public Housing Act, the Ministry of Transport, Building, and Housing has published annual lists of ‘ghettos’ (their term for socially deprived neighbourhoods) since 2010. In order to define an area as a ghetto, the Ministry considers different ethnic and socioeconomic characteristics of areas with at least 1,000 adjacent public housing units. These characteristics and their thresholds are subject to change over time, and in 2018 the Danish Government introduced a definition of socially deprived neighbourhoods and started to reserve the term

″ghetto area″ for socially deprived neighbourhoods that have at least 50% non-Western immigrants and descendants (see Appendix E for its definition of socially deprived neighbourhoods and ghetto areas).

Public housing associations on the Government's lists of ghetto areas and socially deprived neighbourhoods are eligible for large amounts of funds from ″Landsbyggefonden″ for social programmes for their tenants and physical improvements of the neighbourhoods in order to improve the socioeconomic and ethnic mix in the neighbourhoods. In the latest ″ghettopakke″ proposed by the Danish Government in the spring 2018, DKK 12 billion (about USD 1.8 billion) will be allocated to initiatives that improve the socioeconomic mix in the socially deprived areas such as demolition of public housing units and construction (conversion of public housing) of (to) owner-occupied properties during the 2019-2026 period (Regeringen, 2018). Moreover, public housing associations are required to decrease the share of public housing units for families in ″hard ghetto areas″ to 40% before 2030 (see Appendix E for the Danish Government's definition of ″hard ghetto areas″).[15]

Policies to improve the socioeconomic mix in socially deprived neighbourhoods include construction (conversion of public housing) of (to) owner-occupied properties. However, since the Ministry computes the socioeconomic mix in areas with adjacent public housing projects using data for residents in only public housing units and excluding residents in other ownership types in the neighbourhood, the effects of such a policy are not detectable. Similarly, the approach used by the Ministry is not able to capture spillover effects of policies to improve the socioeconomic mix in the socially deprived neighbourhoods through demolition and sales of existing blocks of public housing; residents in such blocks have to move to another neighbourhood. In other words, perhaps the policy of dispersing tenants of public housing units into other neighbourhoods just displaces the deprivation from one neighbourhood to other neighbourhoods that are not under consideration of the Ministry. To evaluate the degree of success of such policies, one should use well-defined residential neighbourhoods such as ours rather than the intervention areas. To distinguish between intervention and control areas, those neighbourhoods should be relatively homogeneous in terms of housing and ownership types. Moreover, neighbourhood units should be homogeneous in size (area and the number of inhabitants) to avoid attenuation bias. Additionally, to make comparisons of neighbourhoods across years possible, boundaries of those neighbourhoods should not change over time.

Appendix Table A10 presents characteristics of the 16 socially deprived neighbourhoods (among those, 7 neighbourhoods were on the list of hard ghetto areas)

[15] Trafikstyrelsen 2018; Transport, Bygnings- og Boligministeriet 2018; Law number 1322 dated 27/11/2018: Lov om ændring af lov om almene boliger m.v., lov om leje af almene boliger og lov om leje, Section 168 a.

in four municipalities of Copenhagen, Aarhus, Odense, and Høje-Taastrup based on the latest definition by the Ministry in December 2018. Besides, Appendix Figures A5 to A8 project those socially deprived neighbourhoods (in each of the four municipalities) and our constructed micro-neighbourhoods on maps. For instance, Appendix Figure A5 shows the location of the 7 socially deprived neighbourhoods in Copenhagen in 2018. As this figure illustrates, our micro-neighbourhoods cover not only all those neighbourhoods, but also the rest of the residential areas of Copenhagen. The same argument holds for all the inhabited land of Denmark. As the constructed neighbourhood clusters are linked to the Population register in a way that residential neighbourhoods for 99.53% of residents of Denmark during 1986-2016 are identifiable (Appendix Table A1), they allow for computation of the share of socioeconomic and ethnic groups across micro- and macro-neighbourhood over a period of 31 years. While each of the considered neighbourhoods by the Ministry (adjacent blocks of public housing with at least 1000 residents) are large in terms of area and number of residents,[16] our constructed neighbourhoods are smaller and more homogeneous in terms of number of inhabitants, but less homogeneous with respect to housing type and ownership. Each of the socially deprived neighbourhoods consists of two to six micro-neighbourhoods. In order to have consistent measures, one can calculate the share of socioeconomic and ethnic groups across our neighbourhoods using the same definitions as the Ministry uses to define socially deprived neighbourhoods (Appendix E).

Finally, since the dominant neighbourhood for each of the inhabited hectare cells of Denmark is available, it is possible to assign neighbourhood ID to the newly constructed residential properties using the hectare cell information of those new constructions to update the current version of the neighbourhood clusters to cover years after 2016.[17]

7. Conclusions

By clustering housing units, we have constructed data to measure residential neighbourhoods in Denmark with unaltered boundaries over the 1986-2016 period at two levels: 8,359 micro- and 1,961 macro-clusters. In terms of number of inhabitants, our macro-clusters compare to the US block groups, while our micro-

[16] As an example, the number of residents in socially deprived neighbourhoods of Copenhagen vary between 1,034 and 6,526 persons.

[17] The neighbourhood clusters as well as the codes and datasets for updating the neighbourhood clusters to cover years after 2016 will be available to other researchers as soon as the authors have finished their study for which they have constructed the neighbourhoods. For more information, please contact the authors.

clusters are comparable to census blocks in the US and Small Areas for Market Statistics (SAMS) in Sweden. Our macro-clusters are homogeneous in terms of the number of micro-clusters that they comprise; on average, each macro-cluster consists of 4 micro-clusters. Moreover, both micro- and macro-clusters are homogeneous in terms of number of inhabitants and households, albeit less so in terms of area. They are substantially smaller and more homogeneous in terms of number of inhabitants and households than existing administrative geographical units like municipalities and postal districts and delineated by natural borders and major roads and railways.

Our constructed residential neighbourhoods are superior to the residential neighbourhoods constructed by Damm and Schultz-Nielsen (2008) for the 1985-2004 period, most importantly because our residential neighbourhoods are unaltered over the 1986-2016 period, but also because they are more homogeneous in terms of number of inhabitants and households and allow us to assign neighbourhood of residence ID to almost 100% of the housing addresses in the Population register. However, the longer period of coverage comes at the cost of being slightly larger.

There were in total 459,497 inhabited hectare cells during 1986-2016, 91% of which belong entirely to one micro-cluster. Our micro-clusters are appropriate for measurement of residential segregation over a 31 years period and the effects of public policies that aim at improving the social and ethnic mix in socially deprived neighbourhoods. Finally, our residential neighbourhood clusters can be updated to cover a longer period.

References

Andersen, H. S. 2015. ″Indvandring, integration og etnisk segregation: udviklingen i indvandrernes bosætning siden 1985″ [Immigration, integration and ethnic segregation: development in settlement patterns since 1985]. SBI Forlag, vol. 2015:01.

Andersen, H. S. 2017. ″Selective moving behavior in ethnic neighbourhoods: white flight, white avoidance, ethnic attraction or ethnic retention?″ *Housing Studies,* 32 (3): 296-318.

Åslund, O., P. A. Edin, P. Fredriksson, and H. Grönqvist. 2011. ″Peers, neighborhoods, and immigrant student achievement: evidence from a placement policy″. *American Economic Journal: Applied Economics,* 3: 67-95.

Åslund, O., I. Blind, and M. Dahlberg. 2017. ″All abroad? Commuter train access and labor market outcomes″. *Regional Science and Urban Economics,* 67: 90-107.

Bailey, M. 1959. ″Note on the economics of residential zoning and urban renewal″. *Land Economy,* 35: 288-290.

Becker, G. 1957. *The Economics of Discrimination.* Chicago: University of Chicago.

Bjerre-Nielsen, A., and M.H. Gandil. 2018 ″Privacy in spatial data with high resolution and time invariance″. https://github.com/abjer/privacy_spatial/raw/master/paper/privacy_spatial.pdf

Bjerre-Nielsen, A. and M.H. Gandil. 2020 ″Defying Attendance Boundary Policies and the Limits to Combating School Segregation″, manuscript.

Bolster, A., S. Burgess, R. Johnston, K. Jones, C. Propper. and R. Saker. 2006. ″Neighbourhoods, Households and Income Dynamics: A Semi-Parametric Investigation of Neighbourhood Effects″. *Journal of Economic Geography*, 7(1): 1-38.

Butts, C. 2002. *Spatial Models of Large-Scale Interpersonal Networks*. Doctoral Dissertation, Department of Social and Decision Sciences, Carnegie Mellon University.

Campbell, K. E., and B. A. Lee. 1992. ″Sources of Personal Neighbor Networks: Social Integration, Need, or Time?″ *Social Forces* 70(4): 1077-1100.

Clark, A. E., N. Westergård-Nielsen, and N. Kristensen. 2009. ″Economic Satisfaction and Income Rank in Small Neighbourhoods″. *Journal of the European Economic Association*, 7(2/3): 519-527.

Cutler, D. M., E. L. Glaeser, and J. L. Vigdor. 1999. ″The rise and decline of the American ghetto″, *Journal of Political Economy*, 107(3): 455-506.

Damm, A. P. 2014. ″Neighborhood Quality and Labor Market Outcomes: Evidence from Quasi-Random Neighborhood Assignment of Immigrants″. *Journal of Urban Economics*, 79: 139-166.

Damm, Anna P. and Christian Dustmann, "Does Growing Up in a High Crime Neighborhood affect Youth Criminal Behavior?", *American Economic Review* 2014, 104(6): 1806-1832.

Damm, A.P., M.L. Schultz-Nielsen, and T. Tranæs. 2006. *En befolkning deler sig op?* [Segregation of a Population?] Gyldendal.

Damm, A.P. and M.L. Schultz-Nielsen. 2008. "Danish Neighbourhoods: Construction and Relevance for Measurement of Residential Segregation", *Danish Journal of Economics (Nationaløkonomisk Tidsskrift)*, 146(3): 241-262.

Danckert, B., P.T. Dinesen, and K.M. Sønderskov. 2017. "Reacting to Neighbourhood Cues? Political Sophistication Moderates the Effect of Exposure to immigrants", *Public Opinion Quarterly*, 81(1): 37-56.

Dinesen, P.T. and K.M Sønderskov. 2015. "Ethnic Diversity and Social Trust: Evidence from the Micro-Context", *American Sociological Review*, 80(3): 550-573.

Dustmann, C., and R. Landersø. 2018. "Child's Gender, Young Fathers' Crime, and Spillover Effects in Criminal Behavior". ROCKWOOL Foundation Research Unit Study Paper 127.

Dustmann, C., K. Vasiljeva. and A.P. Damm. 2019. "Refugee Migration and Electoral Outcomes". *The Review of Economic Studies,* 86: 2035-2091.

Edin, P. P. Fredriksson and O. Åslund. 2003. "Ethnic Enclaves and the Economic Success of Immigrants – Evidence from a Natural Experiment". *Quarterly Journal of Economics,* vol. 118: 329-357.

Hviid, S.J. 2015. *Dynamic Models of the Housing Market*. Aarhus University, Department of Economics and Business Economics, PhD dissertation 2015-21.

Iceland, J. and D.H. Weinberg. 2002. "Racial and ethnic residential segregation in the United States: 1980-2000, Census 2000 Special Reports". Washington, DC: U.S. Bureau of the Census.

Intrator, J., J. Tannen and D.S. Massey. 2016. "Segregation by race and income in the United States 1970-2010". *Social Science Research*, 60: 45-60.

Knudsen, L.B. 2014. Mapping og registerdata. In: Jørgensen, A. and H.L. Jensen (eds.) *Introduktion til mapping-metoder: Metodeserie for social- og sundhedsvidenskaberne.* [Introduction to mapping methods: Series on methods in social and health sciences] Vol. 5. Syddansk Universitetsforlag, Ch. 4: 56-68.

Latané B., J. Liu, A. Nowak. and L. Zhenget. 1995. "Distance matters: Physical space and social impact". *Personality and Social Psychology Bulletin,* 21(8): 795-805.

Law number 1322 dated 27/11/2018. Lov om ændring af af lov om almene boliger m.v., lov om leje af almene boliger og lov om leje (https://www.retsinformation.dk/Forms/R0710.aspx?id=205151, accessed on the 20th of Jan. 2021).

Massey, D.S., J. Rothwell and T. Domina. 2009. "The Changing Bases of Segregation in the United States". *The ANNALS of the American Academy of Political and Social Science*, 626: 74-90.

Ministeriet for By, Bolig og Landdistrikter. 2014. ″Analyse – segregering i de fire største danske byområder″ [Analysis – segregation in the four largest urban areas in Denmark]. Report published on the 17th of March 2014.

Regeringen. 2018. ″Ét Danmark uden parallelsamfund – ingen ghettoer i 2030″ [A united Denmark without parallel societies – no ghettos in 2030]. 1st of March 2018, https://www.regeringen.dk/publikationer-og-aftaletekster/%C3%A9t-danmark-uden-parallelsamfund/

Statistics Sweden. 2005. Geografin i statistiken – regionala indelningar i Sverige. Meddelanden i samordningsfrågor för Sveriges officiella statistik 2005:2 [Geography in statistics – regional divisions in Sweden. Reports on Statistical Co-ordination for the Official Statistics of Sweden 2005:2]. Örebro.

Thiessen, A. J., and J. C. Alter. 1911. ″Precipitation Averages for Large Areas″. *Monthly Weather Review*, 39: 1082-84.

Trafikstyrelsen. 2018. Aftale mellem regeringen (Venstre, Liberal Alliance og Det Konservative Folkeparti) og Socialdemokratiet, Dansk Folkeparti og Socialistisk Folkeparti om: Initiativer på boligområdet, der modvirker parallelsamfund, 9th of May 2018 [Agreement between the Government and the Social Democratic Party, the Danish People's Party and the Socialistic People's Party about: Housing initiatives that counteract parallel societies] (www.trafikstyrelsen.dk/~/media/Dokumenter/10%20Bolig/Bolig/Almene %20boliger/Boligaftaler/Boligaftale%202018.pdf)

Transport-, Bygnings- og Boligministeriet. 2018. ″Liste over hårde ghettoområder pr. 1. dec. 2018″ [List of hard ghetto areas on the 1st of Dec. 2018]. Note, 1st of Dec. 2018.

United States Census Bureau website. 2018. ″Geographical Areas Reference Manual″. https://www.census.gov/ updated on May 16, 2018).

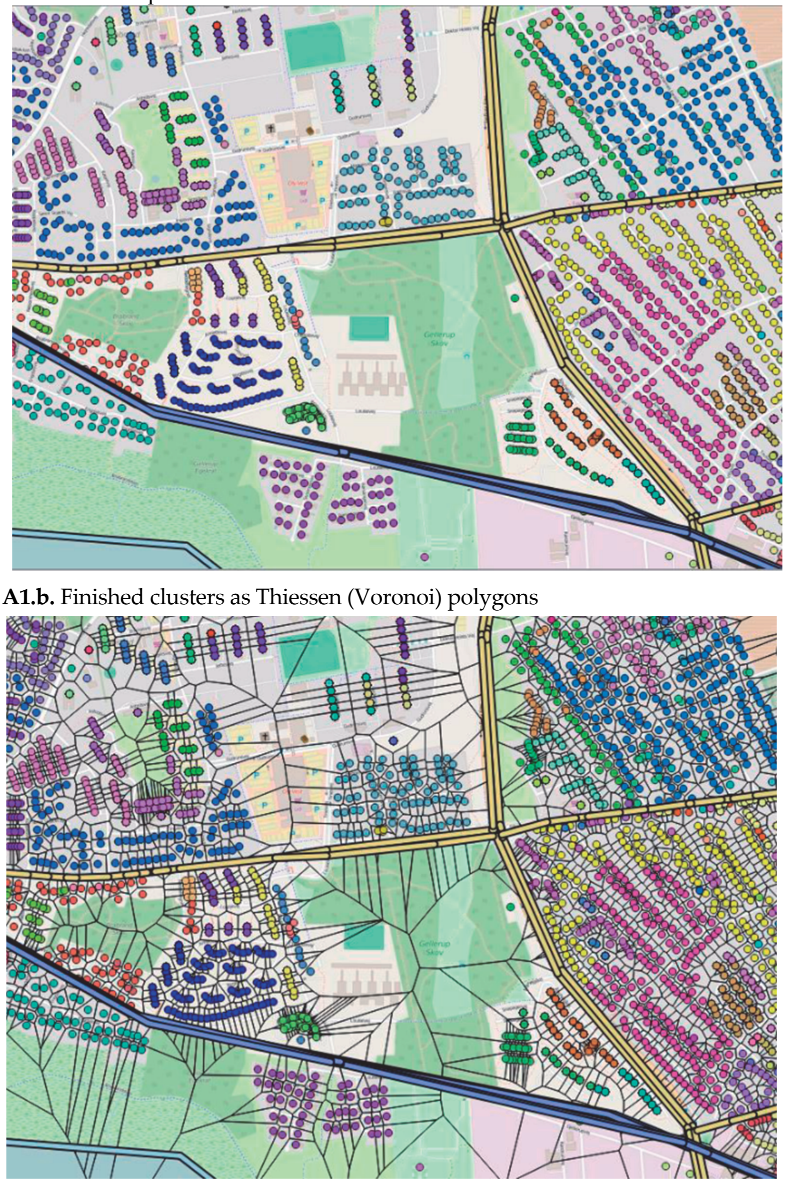
United States Census Bureau. 2010. ″2010 Census Summary File 1″ Appendix A-Geographic Terms and Concepts, A-10 & A-11.

Wellman, B. 1979. ″The Community Question: The Intimate Networks of East Yorkers″. *American Journal of Sociology*, 84(5): 1201-1231.

Wellmann, B. 1996. ″Determinants of Recent Immigrants' Locational Choices″. Federal Reserve Bank of Atlanta, *Working Paper* 98-3

Appendix Figures

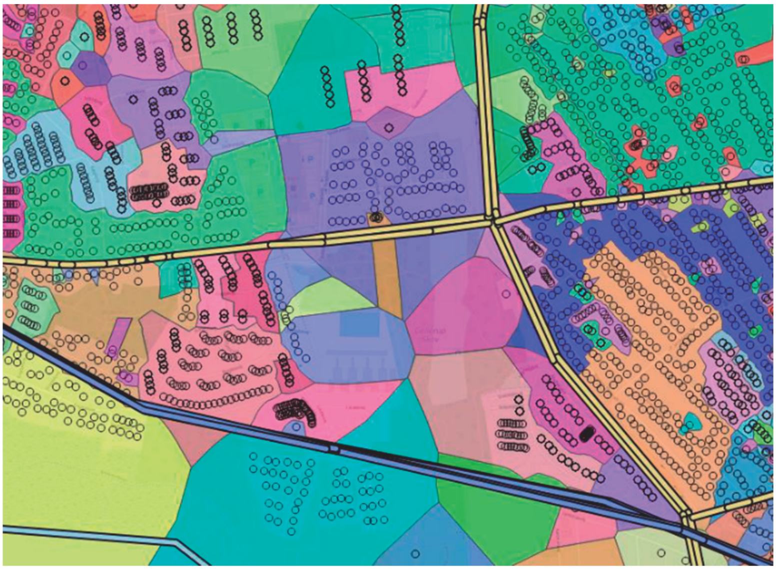
Figure A1. Graphical illustration of the constructed micro-neighbourhoods during stages one to three



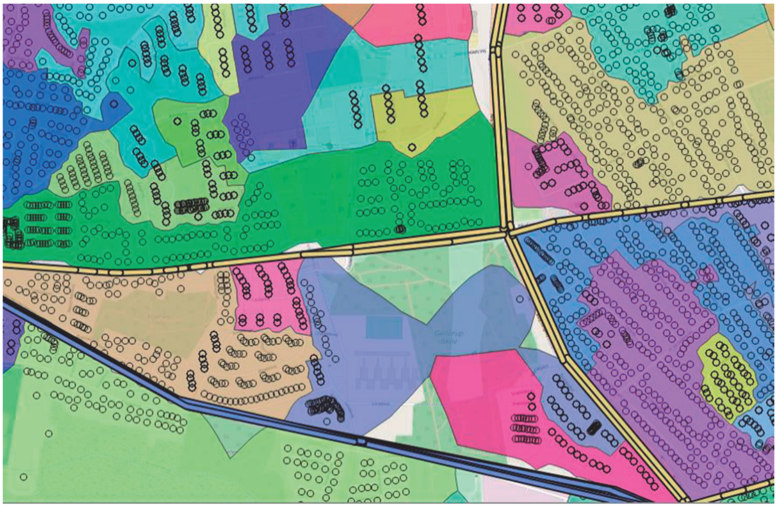
A1.a. Finished point clusters

A1.b. Finished clusters as Thiessen (Voronoi) polygons

A1. c. Dissolved Thiessens (Voronoi) polygons



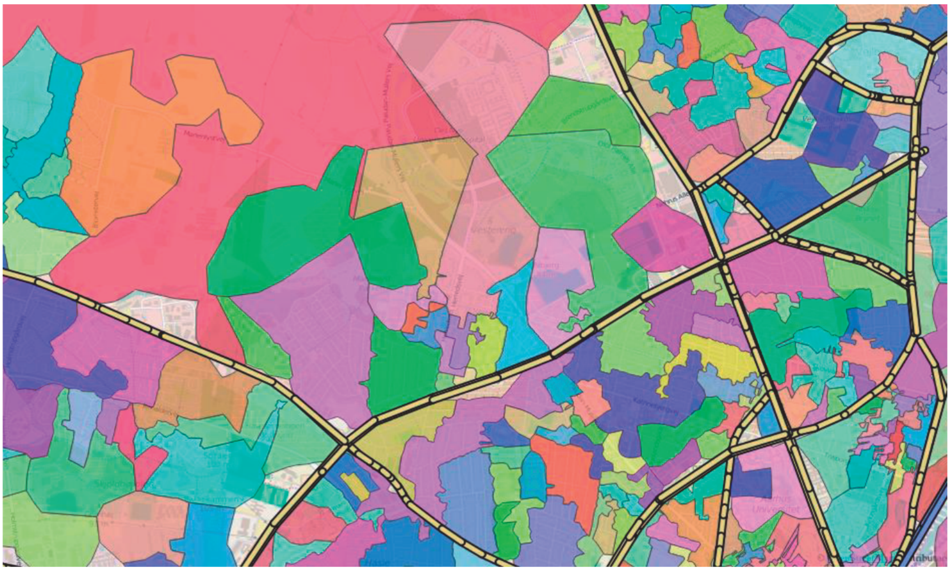
A1.d. Finished micro neighbourhoods



Notes: This figure graphically illustrates the construction of micro-neighbourhoods during stages one to three. Finished point clusters in stage one are shown in (a). The finished clusters are illustrated as Thiessen (Voronoi) polygons in (b). Dissolved Thiessen (Voronoi) polygons are shown in (c), and the finished micro-neighbourhoods in stage three are presented in (d).

Source: Geomatic

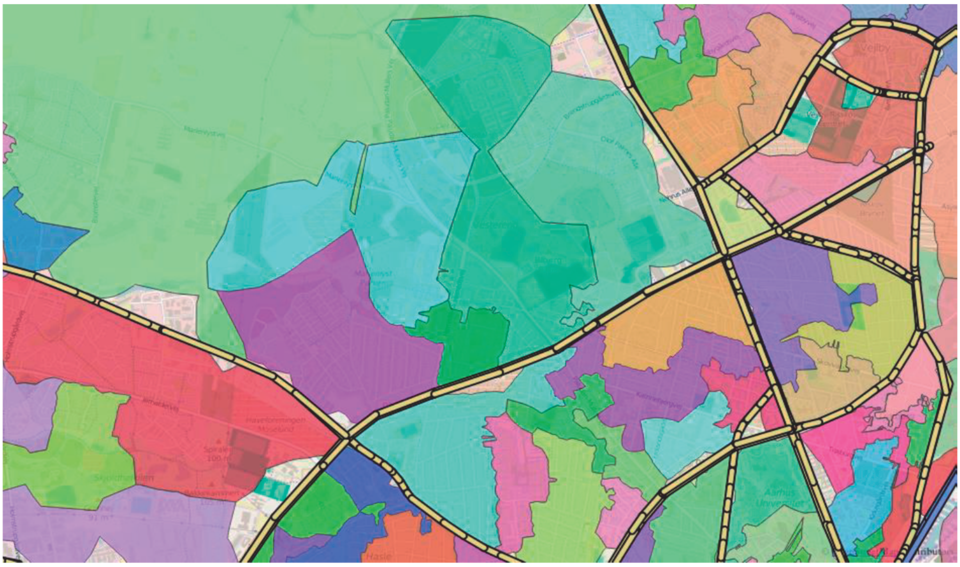
Figure A2. Constructed micro-neighbourhoods at the end of the third stage



Notes: This figure shows the constructed micro-clusters at the end of the third stage. Each colour shows one micro-cluster.

Source: Geomatic.

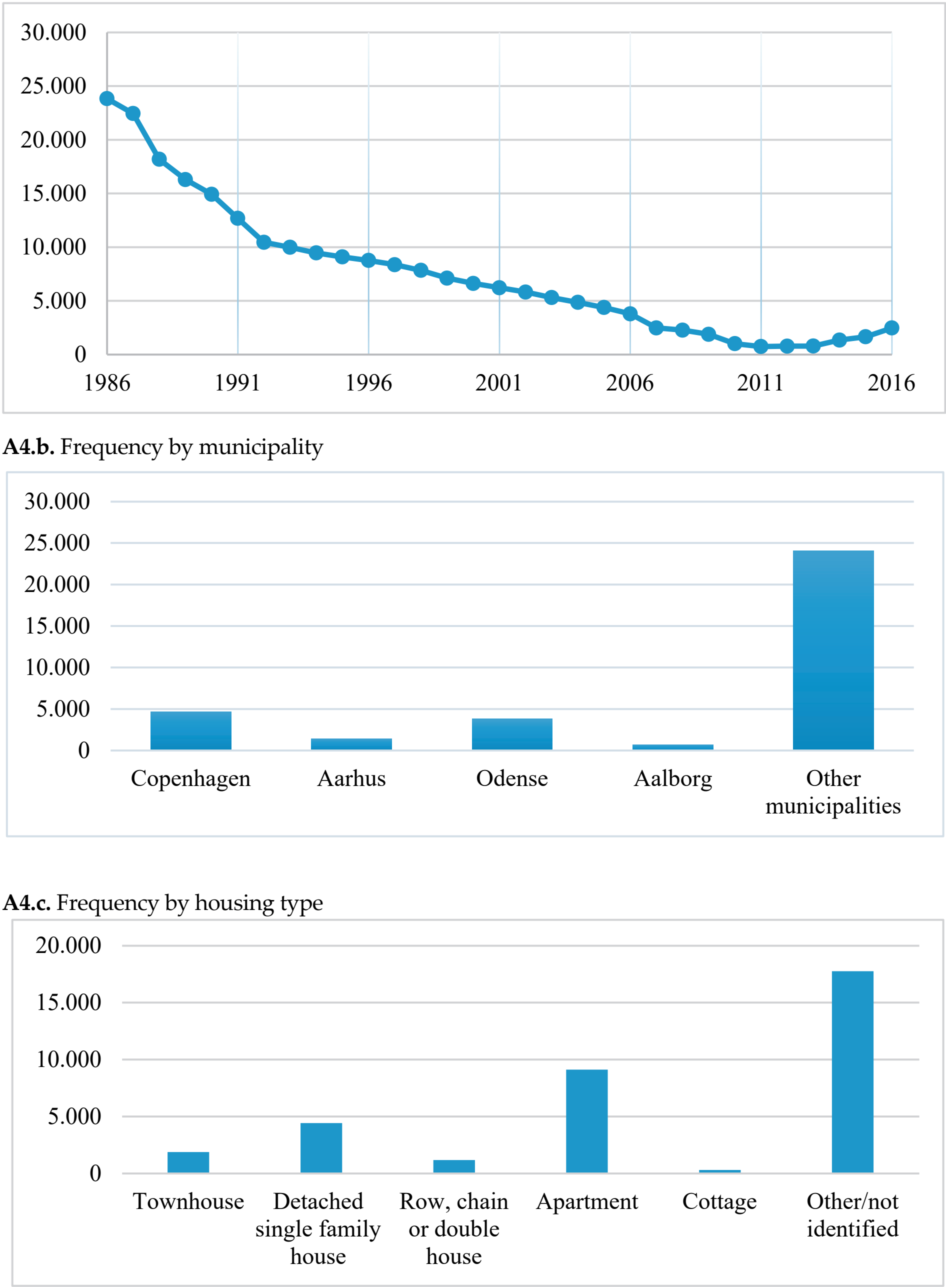
Figure A3. Constructed macro-neighbourhoods at the end of the third stage



Notes: This figure shows the constructed macro-clusters at the end of the third stage. Each colour shows one macro-cluster.

Source: Geomatic.

Figure A4. Characteristics of housing addresses without cluster ID by year

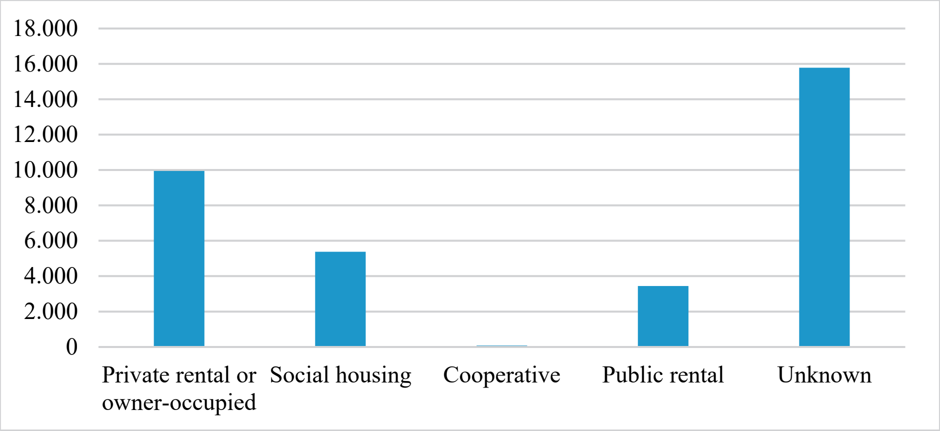


A4a. Frequency by year

A4.b. Frequency by municipality

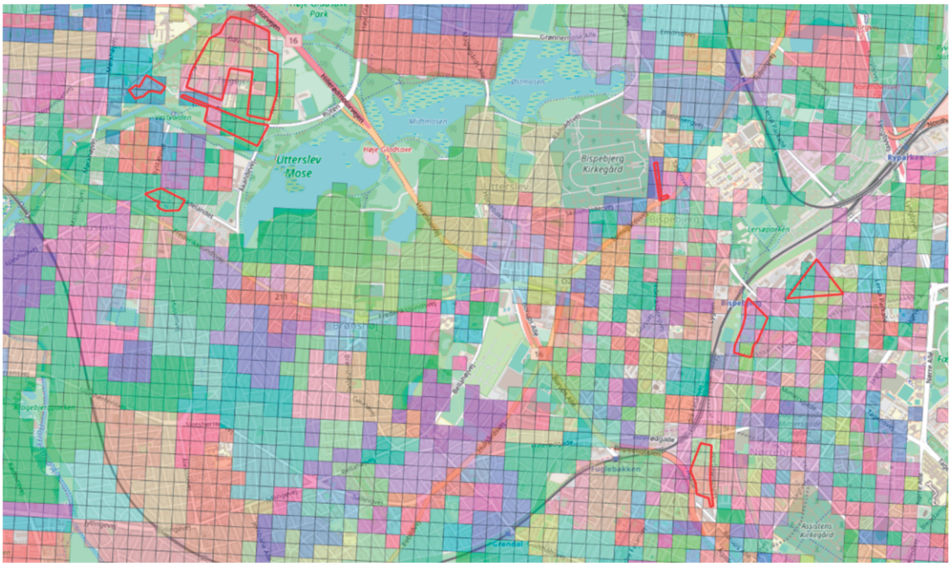
A4.c. Frequency by housing type

A4.d. Frequency by ownership type



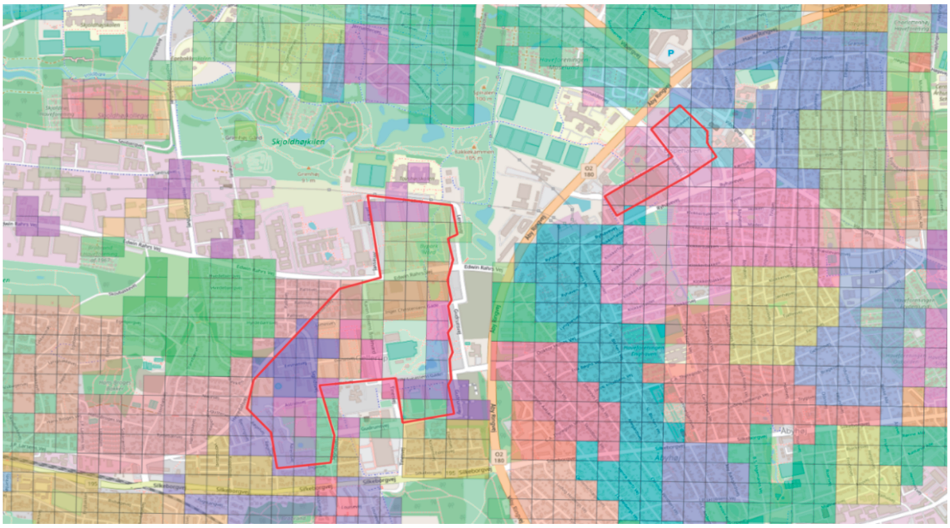
Notes: This figure shows the characteristics of the housing addresses that are observed in the Population register during 1986-2016 but do not have cluster ID. The considered characteristics are: frequency by year (a), frequency by municipality (b), frequency by housing type (c), and frequency by ownership type. Sources are own calculations based on data from Statistics Denmark and Geomatic

Figure A5. Socially deprived neighbourhoods in Copenhagen in 2018



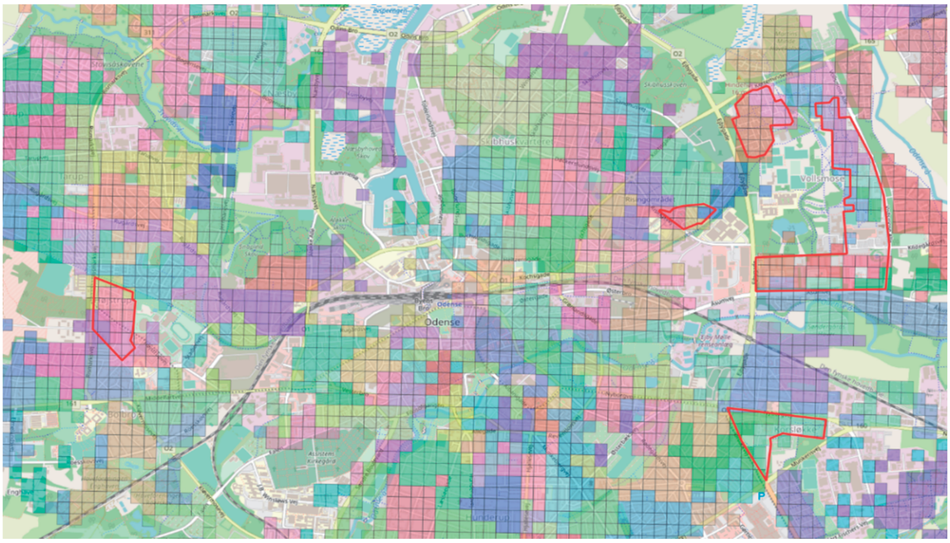
Notes: This figure shows socially deprived neighbourhoods (ghetto areas) in Copenhagen based on the Ministry of Transport, Building, and Housing's report in December 2018 (Appendix Table A10), as well as the dominant micro-neighbourhood of each hectare cell. Boundaries of socially deprived neighbourhoods are determined by red colour. Each colour shows clusters of hectare cells in which a micro-cluster is dominant

Figure A6. Socially deprived neighbourhoods in Aarhus in 2018



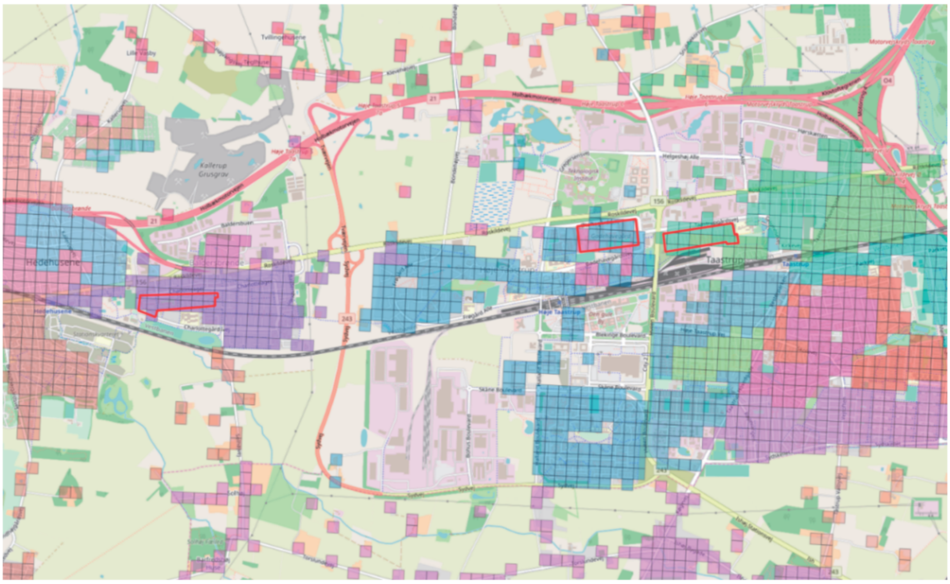
Notes: This figure shows socially deprived neighbourhoods (ghetto areas) in Aarhus based on the Ministry of Transport, Building, and Housing's report in December 2018 (Appendix Table A10), as well as the dominant micro-neighbourhood of each hectare cell. Boundaries of socially deprived neighbourhoods are determined by red colour. Each colour shows clusters of hectare cells in which a micro-cluster is dominant

Figure A7. Socially deprived neighbourhoods in Odense in 2018



Notes: This figure shows socially deprived neighbourhoods (ghetto areas) in Odense based on the Ministry of Transport, Building, and Housing's report in December 2018 (Appendix Table A10), as well as the dominant micro-neighbourhood of each hectare cell. Boundaries of socially deprived neighbourhoods are determined by red colour. Each colour shows clusters of hectare cells in which a micro-cluster is dominant.

Figure A8. Socially deprived neighbourhoods in Høje-Taastrup in 2018



Notes: This figure shows socially deprived neighbourhoods (ghetto areas) in Høje-Taastrup based on the Ministry of Transport, Building, and Housing's report in December 2018 (Appendix Table A10), as well as the dominant micro-neighbourhood of each hectare cell. Boundaries of socially deprived neighbourhoods are determined by red colour. Each colour shows clusters of hectare cells in which a micro-cluster is dominant.

Appendix Tables

Table A1. Individuals and addresses with and without cluster ID in the Population register

	Housing addresses		Inhabitants	
	Number	**Unique Number**	**Number**	**Unique Number**
Total	75,711,527	3,168,283	165,949,732	7,970,834
With cluster ID	75,479,083 Success: 99.69%	3,133,661 Success: 98.91%	165,177,102 Success: 99.53%	7,956,967 Success: 99.82%
Without cluster ID	232,444	34,622	772,630	283,377

Notes: This table shows the observed housing addresses and persons in the Population register during 1986-2016 with and without cluster ID. The sources of this table are own calculations based on data from Housing register, Population register, hectare cell information of housing units, the constructed neighbourhood clusters, and information from Geomatic. ″Number″ represents each housing unit-year or person-year in the Population register. ″Unique Number″ reports each housing address or person in the Population register once rather than several times over 1986-2016. These observations are divided into two categories. First, a group with cluster ID (i.e. identified neighbourhood). Second, a group without cluster ID (unidentified neighbourhood).

Table A2. Characteristics of the clusters after stage 5

		Percentiles					**All**		
		5%	**25%**	**50%**	**75%**	**95%**	**Mean**	**Std. dev.**	**Number**
Micro-clusters									
Housing addresses	1986	164	191	241	313	454	264.2	98.7	8,230
	2000	169	209	266	348	514.5	295.1	123.6	8,240
	2015	168	215	283	375	597.5	319.7	159.3	8,240
Inhabitants	1986	307	465	568	734	1,044	614.9	234.3	8,230
	2000	313	464	581	764	1,149	643.2	279.6	8,240
	2015	315.5	464	601	808	1,303	684.6	359.2	8,240
Area (ha)	1986	2	8	26	58	175	47.1	56.7	8,230
	2000	2	9	27	63	182	49.6	59.8	8,240
	2015	2	9	27	68	185	51.3	61.51	8,240
Too small micro-clusters									282
Overly large micro-clusters									353
Macro-clusters									
Housing addresses	1986	687	865	1,095	1,403	1,959	1,174	409.5	1,851
	2000	739	952	1,204	1,583	2,254	1,313	491.2	1,851
	2015	758	993	1,284	1,730	2,539	1,423	588.6	1,851
Inhabitants	1986	1,353	1,941	2,478	3,313	4,857	2,734	1,108.0	1,851
	2000	1,389	1,997	2,566	3,459	5,242	2,863.4	1,220.3	1,851
	2015	1,442	2,048	2,707	3,693	5,776	3,047.9	1,404.4	1,851
Area (ha)	1986	10	42	117	329	655	209	223.2	1,851
	2000	10	41	123	345	698	220.3	235.7	1,851
	2015	11	43	127	360	712	228.0	241.9	1,851
Too small macro-clusters									26
Overly large macro-clusters									135

Notes: This table shows characteristics of the constructed micro- and macro-clusters in terms of the number of housing addresses in the Population Register, inhabitants and area at the end of stage 5 in 1986, 2000, and 2015. The sources of this table are own calculations based on data from Housing register, Population register, hectare cell information of housing units, the constructed neighbourhood clusters, and information from Geomatic. "Percentile" columns show the characteristics at 5-95 percentile rank. "All" columns use all of the micro- and macro-clusters to compute average and standard deviation of each of the characteristics. Area (ha) means the number of hectare cells in which a micro-cluster (macro-cluster) is dominant. Too small micro-clusters (macro-clusters) are micro-clusters (macro-clusters) with less than 150 (600) housing addresses in at least one year during 1986-2016. Overly large micro-clusters (macro-clusters) present the number of micro-clusters (macro-clusters) with more than 450 (1,800) housing addresses in all of the observed years during 1986-2016.

Table A3. Characteristics of the clusters after merging the small clusters

		Percentiles					All		
		5%	**25%**	**50%**	**75%**	**95%**	**Mean**	**Std. dev.**	**Number**
Micro-clusters									
Housing addresses	1986	167	197	248	321	469	272.8	101.1	7,970
	2000	172	214	275	359	535	304.7	127.6	7,979
	2015	172	220	292	388	622	330.2	164.9	7,979
Inhabitants	1986	333	480	583	754	1,078	635.0	236.2	7,970
	2000	331	478	601	791	1,186	664.2	284.5	7,979
	2015	336	480	619	830	1,341	707.0	367.0	7,979
Area (ha)	1986	2	9	27	60	180	48.7	57.9	7,970
	2000	2	10	28	66	186	51.2	61.0	7,979
	2015	2	10	28	70	188	53.0	62.6	7,979
Too small micro-clusters									0
Overly large micro-clusters									387
Macro-clusters									
Housing addresses	1986	699	878	1,110	1,418	1,975	1,190.4	407.0	1,827
	2000	755	965	1,217	1,604	2,267	1,330.2	494.5	1,828
	2015	773	1,003.5	1,297	1,744	2,565	1,441.4	594.3	1,828
Inhabitants	1986	1,430	1,970	2,515	3,345	4,885	2,770.2	1,100.0	1,827
	2000	1,454	2,028.5	2,602.5	3,514.5	5,266	2,899.4	2,602.5	1,828
	2015	1,499	2,083.5	2,731	3,748	5,787	3,086.2	2,731	1,828
Area (ha)	1986	11	43	119	336	669	212.0	224.8	1,827
	2000	11	43	125	353	708	223.1	237.4	1,828
	2015	11	45	128	370	718	230.8	243.7	1,828
Too small macro-clusters									0
Overly large macro-clusters									138

Notes: This table shows characteristics of the constructed micro- and macro-clusters in terms of the number of housing addresses in the Population Register, inhabitants and area at the end of stage 6 (i.e. merging small clusters with adjacent clusters) in 1986, 2000, and 2015. The sources of this table are own calculations based on data from Housing register, Population register, hectare cell information of housing units, the constructed neighbourhood clusters, and information from Geomatic. пPercentile” columns show the characteristics at 5-95 percentile rank ”All” columns use all of the micro- and macro-clusters to compute average and standard deviation of each of the characteristics. Area (ha) means the number of hectare cells in which a micro-cluster (macro-cluster) is dominant. Too small micro-clusters (macro-clusters) are micro-clusters (macro-clusters) with less than 150 (600) housing addresses in at least one year during 1986-2016. Overly large micro-clusters (macro-clusters) present the number of micro-clusters (macro-clusters) with more than 450 (1,800) housing addresses in all of the observed years during 1986-2016.

Table A4. Characteristics of the clusters after splitting the large clusters

		Percentiles					All		
		5%	25%	50%	75%	95%	Mean	Std. dev.	Number
Micro-clusters									
Housing addresses	1986	167	197	246	310	407	260.4	76.9	8,350
	2000	172	215	271	344	467	290.9	101.3	8,359
	2015	173	222	287	367	548	315.2	140.2	8,359
Inhabitants	1986	329	472	571	717	978	606.1	197.2	8,350
	2000	329	475	585	751	1,067	634.0	240.6	8,359
	2015	336	476	605	789	1,217	674.9	322.7	8,359
Area (ha)	1986	2	9	26	58	172	46.5	55.0	8,350
	2000	2	9	26	63	179	48.9	57.8	8,359
	2015	2	9	27	67.5	181	50.6	59.3	8,359
Too small micro-clusters									0
Overly large micro-clusters									27
Macro-clusters									
Housing addresses	1986	698	869	1,068	1,315.5	1,662.5	1,109.6	297.3	1.960
	2000	754	955	1,174	1,477	1,899	1,240.0	375.2	1,961
	2015	771	995	1,248	1,611	2,178	1,343.6	470.2	1,961
Inhabitants	1986	1,418.5	1,949.5	2,448	3,103.5	4,198	2,582.2	838.1	1.960
	2000	1,448	2,003	2,507	3,259	4,530	2,702.8	959.7	1,961
	2015	1,503	2,058	2,640	3,464	4,929	2,876.9	1,133.5	1,961
Area (ha)	1986	10	42.5	112.5	322	601	197.6	197.6	1.960
	2000	10	43	120	336	627	207.9	207.9	1,961
	2015	11	44	125	351	636	215.2	213.1	1,961
Too small macro-clusters									0
Overly large macro-clusters									8

Notes: This table shows characteristics of the constructed micro- and macro-clusters in terms of the number of housing addresses in the Population Register, inhabitants and area at the end of stage 7 (i.e. splitting large clusters to smaller clusters) in 1986, 2000, and 2015. The sources of this table are own calculations based on data from Housing register, Population register, hectare cell information of housing units, the constructed neighbourhood clusters, and information from Geomatic. "Percentile" columns show the characteristics at 5-95 percentile rank. "All" columns use all of the micro- and macro-clusters to compute average and standard deviation of each of the characteristics. Area (ha) means the number of hectare cells in which a micro-cluster (macro-cluster) is dominant. Too small micro-clusters (macro-clusters) are micro-clusters (macro-clusters) with less than 150 (600) housing addresses in at least one year during 1986-2016. Overly large micro-clusters (macro-clusters) present the number of micro-clusters (macro-clusters) with more than 450 (1,800) housing addresses in all of the observed years during 1986-2016.

Table A5. Compactness of final clusters

		Percentiles					All		
		5%	25%	50%	75%	95%	Mean	Std. dev.	Num ber
Micro-clusters									
% Compactness **(Annual)**	1986	100	100	100	99.7	95.6	98.6	6.3	8,350
	2000	100	100	100	100	95.3	98.6	6.0	8,359
	2015	100	100	100	100	95.2	98.7	5.7	8,359
% Compactness **(Weighted across years)**		100	100	100	99.8	95.5	98.7	5.7	8,359
Macro-clusters									
% Compactness **(Annual)**	1986	100	100	99.9	99.2	91.7	98.6	3.5	1,960
	2000	100	100	99.9	99.2	91.5	98.6	3.4	1,961
	2015	100	100	100	99.3	91.4	98.6	3.2	1,961
% Compactness **(Weighted across years)**		100	100	99.9	99.2	91.5	98.6	3.2	1,961

Notes: This table shows characteristics of the constructed micro- and macro-clusters in terms of compactness at the end of stage 8 (i.e. final stage) in 1986, 2000, and 2015. The sources of this table are own calculations based on data from Housing register in the Population Register, Population register, hectare cell information of housing units, the constructed neighbourhood clusters, and information from Geomatic. "Percentile" columns show the compactness at 5-95 percentile rank "All" columns use all of the micro- and macro-clusters to compute average and standard deviation of each of the characteristics. Compactness (annual) is calculated using formula (1), and compactness (weighted across years) is calculated using formula (2).

Table A6. Distribution of the number of micro-clusters in each hectare cell after final stage

Year	Percentiles					All				
	5%	25%	50%	75%	95%	Mean	Std. dev.	Min.	Max.	Number
1986	1	1	1	1	2	1.08	0.311	1	6	387,448
2000	1	1	1	1	2	1.09	0.315	1	6	407,852
2015	1	1	1	1	2	1.09	0.316	1	6	422,049

Notes: This table shows the number of observed micro-clusters in each of the inhabited hectare cells (alternatively, the number of micro-clusters that share an inhabited hectare cell with each other). The sources of this table are own calculations based on data from Housing register, Population register, hectare cell information of housing units, the constructed neighbourhood clusters, information from Geomatic. ″Percentile″ columns show the number of micro-clusters in each hectare cell at 5-95 percentile rank. ″All″ columns use all of the clusters to compute average, standard deviation, minimum, and maximum of the number of micro-clusters in each hectare cell.

Table A7. Number of micro-clusters in each hectare cell after final stage

	1	2	3	4	5 and 6	Total
1986	355,350	29,595	2,290	194	19	387,448
2000	373,180	31,991	2,456	205	20	407,852
2015	385,797	33,476	2,551	205	20	422,049
Average **(1986 – 2016)**	422,031	34,606	2,627	213	20	459,497

Notes: This table shows the number of the inhabited hectare cells with 1 to 6 observed micro-cluster(s) in 1986, 2000, and 2015 (alternatively, the number hectare cells that belong to 1 to 6 micro-cluster(s)). The sources of this table are own calculations based on data from Housing register, Population register, hectare cell information of housing units, the constructed neighbourhood clusters, information from Geomatic. ″Average (1986-2016)″ shows average number of the inhabited hectare cells with 1 to 6 micro-cluster(s) during 1986-2016.

Table A8. Municipalities, Postal Districts, and the final Neighbourhood clusters

	Year	Type of neighbourhood	Percentiles					All				
			5%	25%	50%	75%	95%	Mean	Std. dev	Min.	Max.	Number
Housing addresses	1986	Municipality	7,725	10,906	15,742	24,069	48,359	22,206	29,921	46	270,927	99
		Postal districts	9	78	473	1,734	9,068	1,970	4,315	<10	44,550	1,030
		Micro-clusters	167	197	246	310	408	260	77	150	853	8,350
		Macro-clusters	698	870	1,069	1,316	1,662	1,109	297	606	2,067	1,960
	2000	Municipality	3,770	12,406	17,885	26,340	52,133	24,627	31,257	44	274,981	99
		Postal districts	10	86	543	2,153	10,927	2,355	5,125	<10	52,287	1,022
		Micro-clusters	172	214	271	343	468	290	101	152	1,595	8,359
		Macro-clusters	754	955	1,174	1,476	1,900	1,240	375	621	4,208	1,961
	2015	Municipality	3,459	12,406	17,885	26,340	55,060	26,632	33,754	39	289,609	99
		Postal districts	13	92	589	2,435	11,716	2,549	5,447	<10	60,665	1,034
		Micro-clusters	172	221	287	368	548	315	140	150	2,432	8,359
		Macro-clusters	771	996	1,248	1,611	2,178	1,343	470	614	4,888	1,961
Inhabitants	1986	Municipality	8,354	27,377	39,604	56,587	111,858	61,649	55,473	121	472,138	99
		Postal districts	16	149	1,117	4,671	21,144	4,705	9,313	<10	80,318	1,030
		Micro-clusters	328	472	571	718	979	606	197	150	1,560	8,359
		Macro-clusters	1,421	1,947	2,447	3,100	4,189	2,582	837	614	5,331	1,961
	2000	Municipality	7,394	28,258	41,846	59,474	115,117	53,781	58,853	99	494,221	99
		Postal districts	20	161	1,267	5,095	23,929	5,151	10,307	<10	92,537	1,022
		Micro-clusters	328	472	585	751	1,067	634	241	169	3,356	8,359
		Macro-clusters	1,454	2,005	2,505	3,262	4,532	2,702	959	907	10,697	1,961
	2015	Municipality	6,276	29,030	42,601	59,285	115,446	57,168	68,250	91	580,184	99
		Postal districts	27	195	1,309	5,602	25,010	5,461	11,091	<10	120,080	1,034
		Micro-clusters	336	475	605	789	1,218	674	322	176	5,673	8,359
		Macro-clusters	1,503	2,058	2,636	3,472	4,929	2,876	1,133	868	11,440	1,961

Notes: This table compares municipalities, postal districts, micro-clusters, and macro-clusters in terms of the number of housing units and the number of inhabitants in 1986, 2000, and 2015. The sources of this table are own calculations based on data from Housing register, Population register, hectare cell information of housing units, the constructed neighbourhood clusters, information from Geomatic. "Percentile" columns show the characteristics at 5-95 percentile rank. "All" columns use all of the neighbourhoods (municipality, postal district, micro-, or macro-neighbourhood) to compute average, standard deviation, minimum, and maximum of the number of micro-clusters in each hectare cell.

Table A9. Comparing the first and the current version of neighbourhoods in 1985 and 2004

	Year	Version	Percentiles					All		
			5%	25%	50%	75%	95%	Mean	Std. Dev.	Number
Micro-clusters										
Housing addres-ses	1985	First	158	180	217	272	366	234.9	73.9	9,086
	1986[1]	Current	167	197	246	310	408	260.4	77.2	8,350
	2004	First	163	196	245	313	474	272.7	115.0	9,086
		Current	172	217	276	349	485	296.3	108.3	8,359
Inhabi-tants	1985	First	289	409	513	653	949	555.6	220.1	9,086
	1986	Current	328	472	571	718	979	606.1	197.8	8,350
	2004	First	289	411	526	700	1,114	592.2	285.5	9,086
		Current	329	473	590	759	1,097	642.3	254.7	8,359
Area (in hectare)		First	2	6	22	58	190	47.5	64.6	9,086
		Current	2	9	26	63	179	48.9	57.7	8,359
Macro-clusters										
Housing addres-ses	1985	First	631	712	859	1,072	1,460	929.9	293.4	2,295
	1986	Current	698	870	1,069	1,316	1,662	1,109.6	297.0	1,960
	2004	First	653	798	985	1,237	1,831	1,079.7	396.2	2,295
		Current	759	964	1,188	1,511	1,954	1,263.2	394.1	1,961
Inhabi-tants	1985	First	1,160	1,583	1,994	2,608	3,878	2,200.0	904.2	2,295
	1986	Current	1,421.5	1,947	2,447.5	3,100	4,189	2,582.2	837.7	1,960
	2004	First	1,180	1,618	2,090	2,807	4,310	2,344.5	1,039.0	2,295
		Current	1,454	2,012	2,534	3,299	4,607	2,737.8	994.8	1,961
Area (in hectare)		First	7	27	88	268	668	187.9	236.6	2,295
		Current	10	43	121	338	624	207.9	207.2	1,961
Micro-clusters in each macro-cluster		First	3	3	4	5	6	4,0	1.3	2,295
		Current	3	4	4	5	7	4.26	1.28	1,961

Notes: This table compares characteristics of the current neighbourhood clusters with the characteristics of the constructed neighbourhoods by Damm and Schultz-Nielsen (2008)—called "First" version—in 1985/1986, and 2004. Data sources for the current version are from own calculations based on data from Housing register, Population register, hectare cell information of housing units, the constructed neighbourhood clusters, information from Geomatic. The data source for the "First" version is from Tables 2 and 3 of Damm and Schultz-Nielsen (2008). The current version of clusters covers 31 years from 1986 to 2016, while the first version includes two years of 1985 and 2004. Therefore, we compare neighbourhood characteristics of the current version in 1986 with neighbourhood characteristics of the first version in 1985. However, we can compare characteristics of two versions of neighbourhoods in 2004. "Percentile" columns show the characteristics at 5-95 percentile rank. "All" columns use all of the neighbourhood clusters to compute average and standard deviation. "Housing addresses" refer to housing addresses in the Population Register.

[1] Current version of clusters covers 31 years from 1986 to 2016, while the first version includes two years of 1985 and 2004. Therefore, we compare neighborhood characteristics of current version in 1986 with neighborhood characteristics of the first version in 1985.

Table A10. Characteristics of the socially deprived areas in 4 Danish municipalities based on the Ministry of Transport, Building, and Housing report in 2018

Municipality	Area	Number of residents	Unrelated to labour market (%)	Non-Western immig./ Desc. (%)	Criminals (%)	Primary education (%)	Gross income (%)
		>=1000	> 40%	>50%	≥1.98%	>60%	<55%
		1 Jan 2018	2016-2017	1 Jan 2018	2016-2017	1 Jan 2018	2017
Copenhagen	Lundtoftegade	1,541	35.0	55.5	2.73	62.3	56.5
	Aldersrogade	2,231	34.3	71.2	1.6	66.0	53.5
	Mjølnerparken	1,694	41.9	82.6	2.16	77.4	49.9
	Gadelandet/ Husumgård	1,034	30.9	68.4	1.88	71.8	54.1
	Tingbjerg/ Utterslevhuse	6,526	27.6	73.1	1.83	76.3	51.4
	Bispeparken	1,567	31.4	59.1	2.91	62.2	53.5
	Hørgården	1,591	33.4	51.8	2.28	61.6	57.6
Aarhus	Bispehaven	2,215	48.1	67.6	2.18	74.5	56.1
	Skovgårdsparken	1,442	41.9	69.7	1.74	65.4	60.3
	Gellerupparken/ Toveshøj	5,614	52.5	79.4	2.93	83.6	53.5
Odense	Solbakken mv.	1,324	42.5	52.4	2.25	65.3	56.3
	Korsløkkeparken Øst	1,899	46.5	62.6	1.66	66.4	60.9
	Vollsmose	7,763	52.2	76.0	2.74	78.8	53.0
Høje-Taastrup	Tåstrupgård	2,448	28.7	64.5	1.78	83.5	53.0
	Charlotteager	1,753	35.6	53.6	1.99	70.2	57.1
	Gadehavegård	2,186	39.4	56.5	1.86	72.5	54.8

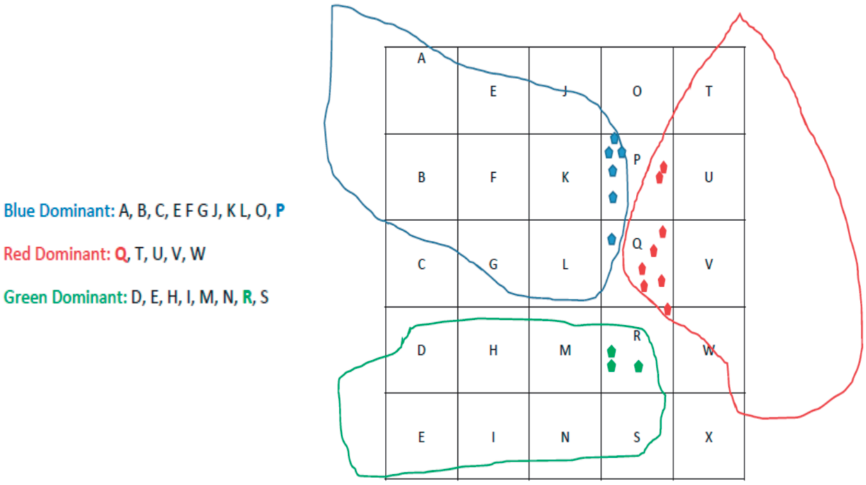
Notes: This table is extracted from the Ministry of Transport, Building, and Housing's report on 1st December 2018, and shows different characteristics of "ghetto areas" in four municipalities of Copenhagen, Aarhus, Odense, and Høje-Taastrup. Besides, "hard ghetto areas" are highlighted.

Appendix B. Assigning dominant cluster ID for hectare cells

Since we use the Thiessen polygon method to cluster adjacent housing units, there are hectare cells with more than one cluster ID. In fact, there are 438,821 hectare cells with cluster ID at the end of the fourth stage. Among those hectare cells, 400,774 hectare cells have one cluster ID, while 38,047 hectare cells have more than one cluster ID.[18]

As Figure B1 graphically illustrates, for each of those 400,774 hectare cells that are occupied by only one micro-cluster, we assign that micro-cluster as dominant cluster ID of the hectare cell. However, for each of the 38,047 hectare cells that are occupied by more than 1 micro-cluster, we count the number of housing addresses from each micro-cluster in that hectare cell, and assign the micro-cluster with the highest number of housing addresses in the hectare cell as dominant cluster ID of that hectare cell. For instance, in Figure B1, hectare cell P is occupied by two micro-clusters (i.e. Blue and Red). Since there are 5 housing addresses from Blue micro-cluster and 2 housing addrespses from Red micro-cluster in hectare cell P, we assign Blue micro-cluster as dominant cluster ID for hectare cell P.

Figure B1. Thematic explanation of assigning dominant cluster ID to hectare cells

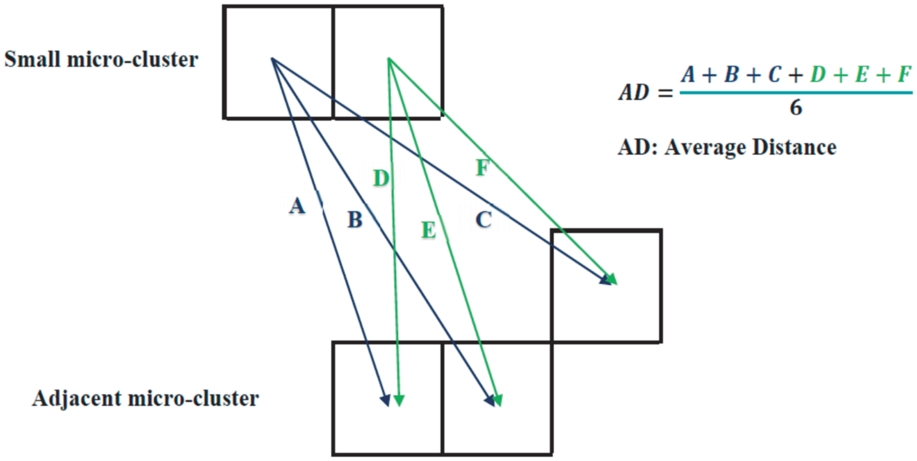

Blue Dominant: A, B, C, E F G J, K L, O, P
Red Dominant: Q, T, U, V, W
Green Dominant: D, E, H, I, M, N, R, S
A
E
J
O
T
B
F
K
P
U
C
G
L
Q
V
D
H
M
R
W
E
I
N
S
X

Notes: This figure graphically demonstrates the procedure of assigning dominant cluster ID for each hectare cell. Each letter shows a hectare cell; each of the polygons resembles a micro-cluster, and each of the points show a housing unit. Housing units in a micro-cluster has the same colour as the micro-cluster (polygon) has.

[18] More than 35,000 of those hectare cells have 2 cluster IDs.

Appendix C. Calculating the average distance between all the hectare cells of a ”too small” micro-cluster and hectare cells of the potential partners

Appendix Figure C1 thematically shows the way of calculating average distances between hectare cells of small micro-cluster and the hectare cells of a potential micro-clusters.



Notes: This figure graphically illustrates the approach of calculating average distance (AD) between small and each of the adjacent micro-clusters. Each of the arrows A-C show distance from one of the hectare cells of small micro-cluster to all of the hectare cells of the adjacent micro-clusters (the same for arrows D-F). And AD shows the formula for calculating average distance from a small micro-cluster to each of the adjacent micro-clusters.

In practice, however, there are macro-clusters with more than one ”too small” micro-cluster. In such cases, we calculate average distances between each of the ”too small” micro-clusters and all of the potential partners for link. Then, we link one of the ”too small” micro-clusters with one of the potential partners with which the average distance is minimum and form a new micro-cluster that is no longer too small. Later, we repeat the mentioned procedure for other ”too small” micro-clusters until all of the ”too small” micro-clusters within a macro-cluster are merged with one of the adjacent micro-clusters.

Appendix D. Calculating compactness of neighbourhoods

The compactness rate of a cluster in each year is calculated using the number of housing addresses within each cluster and their hectare cell location. Practically, we calculate two different measures of compactness at micro- and macro-neighbourhood levels. The following formula shows our first measure of the compactness ratio of cluster i (micro or macro) in year t (1986-2016), CR_{it}:

$$CR_{it} = \frac{BOPI_{it} - \sum_j^J BOPIPROB_{ijt}}{BOPI_{it}} \tag{1}$$

where $BOPI_{it}$ is the total number of housing addresses in that cluster, and $BOPIPROB_{ijt}$ is the number of housing addresses in that cluster, which are located in non-compact hectare cell j.

Our second measure of compactness of each neighbourhood during 1986-2016 is the weighted average of the compactness ratio of that cluster i across all years t, CR_i:

$$CR_i = \sum_t^T \frac{BOPI_{it}}{BOPITOT_i} \times CR_{it} \tag{2}$$

where $BOPITOT_i$ is the total number of observed housing addresses in that cluster during 1986-2016. Hence, the overall compactness ratio for each cluster is calculated through weighting the yearly compactness ratios by the size of the cluster (number of housing addresses in each year).

These two measures of the compactness ratio help us check whether the constructed neighbourhoods respect the compactness criterion or not. Additionally, future users of these neighbourhoods can evaluate the robustness of their estimates by excluding (including) neighbourhoods that have compactness ratios below (above) a certain threshold

Appendix E. Definition of socially deprived neighbourhoods and ghetto areas by the Ministry of Transport, Building, and Housing

Since 1st December 2018, the Ministry of Transport, Building, and Housing defines socially deprived neighbourhoods and 'ghetto areas' on the basis of the following five criteria that apply only to physically adjacent housing units owned by public housing associations, which have at least 1,000 residents:[19]

1. The proportion of residents aged 18-64 years, who are neither attached to the labour market nor enrolled in education, exceeds 40 per cent, calculated as an average across the latest two years.
2. The share of persons convicted for violations of the Penal Code, the Weapons Act or the Drugs Act constitutes at least three times the national average, calculated as the average across the latest two years.
3. The proportion of residents aged 30-59, who only have primary education, exceeds 60 per cent.[20]
4. The average gross income of taxable residents between the ages of 15 and 64, excluding residents enrolled in education, constitutes less than 55 per cent of the average gross income for the same group in the region.
5. The proportion of immigrants and descendants from non-Western countries exceeds 50 per cent.

A ″socially deprived area″ meets at least two of criteria 1-4. A ″ghetto area″ meets at least two of criteria 1-4 and additionally criteria 5. Finally, a ″hard ghetto area″ is an area which has been a ″ghetto area″ for the latest 4 years (exception for the years 2018-2020: for the latest 5 years). According to these definitions, there were 43 ″socially deprived areas″ in 2018 (Ministry of Transport, Building, and Housing, ″Liste over udsatte boligområder pr. 1. december 2018″, Notat 2018). Of these, 29 were ″ghetto areas″; of the 29 ″ghetto areas″, 15 were ″hard ghetto areas″ (Ministry of Transport, Building, and Housing, ″Liste over ghettoområder pr. 1. december 2018″, Notat 2018).

[19] Law number 1322 dated 27/11/2018: Lov om ændring af lov om almene boliger m.v., lov om leje af almene boliger og lov om leje, Section 61a.

[20] Since 2018, only education taken or approved in Denmark is included. The information on educational attainment can therefore be drawn without measurement error from the Education administrative register. Presumably due to disregard of education obtained abroad since 2018, the Ministry has increased the threshold from 50 to 60 per cent compared to the definition for 2017.