HELSINGIN YLIOPISTO
HELSINGFORS UNIVERSITET
UNIVERSITY OF HELSINKI

**Master's thesis**

**Urban Studies and Planning**

# Agent-based models as a tool for exploring complex segregation processes. Simulating scenarios of residential segregation in the Helsinki Metropolitan Area.

Mathew Page

2021

Supervisor:
Mari Vaattovaara

Advisor:
Anssi Joutsiniemi

Master's Programme in Urban Studies and Planning

Faculty of Biological and Environmental Sciences

University of Helsinki

HELSINGIN YLIOPISTO
HELSINGFORS UNIVERSITET
UNIVERSITY OF HELSINKI

| | |
|---|---|
| Koulutusohjelmasta vastaava tiedekunta – Fakulteten ansvarig för programmet – Faculty responsible for the programme<br>Faculty of Science | Tutkinnon myöntävä tiedekunta – Fakulteten som beviljar examen – Faculty granting the degree<br>Faculty of Biological and Environmental Sciences |

Tekijä – Författare – Author

Mathew Page

Tutkielman otsikko – Avhandlingens titel – Title of thesis

Agent-based models as a tool for exploring complex segregation processes. Simulating scenarios of residential segregation in the Helsinki Metropolitan Area.

Koulutusohjelma – Utbildningsprogram – Study programme

Master’s programme in Urban Studies and Planning

| | | |
|---|---|---|
| Tutkielman taso – Avhandlings nivå – Level of the thesis<br>Master’s Thesis | Aika – Datum – Date<br>April 2021 | Sivumäärä – Sidoantal – Number of pages<br>86+1 |

Tiivistelmä – Referat – Abstract

With rising income inequalities and increasing immigration in many European cities, residential segregation remains a key focus for city planners and policy makers. As changes in the socio-spatial configuration of cities result from the residential mobility of its residents, the basis on which this mobility occurs is an important factor in segregation dynamics. There are many macro conditions which can constrain residential choice and facilitate segregation, such as the structure and supply of housing, competition in real estate markets and legal and institutional forms of housing discrimination. However, segregation has also been shown to occur from the bottom-up, through the self-organisation of individual households who make decisions about where to live. Using simple theoretical models, Thomas Schelling demonstrated how individual residential choices can lead to unanticipated and unexpected segregation in a city, even when this is not explicitly desired by any households. Schelling’s models are based upon theories of social homophily, or social distance dynamics, whereby individuals are thought to cluster in social and physical space on the basis of shared social traits. Understanding this process poses challenges for traditional research methods as segregation dynamics exhibit many complex behaviours including interdependency, emergence and nonlinearity. In recent years, simulation has been turned to as one possible method of analysis. Despite this increased interest in simulation as a tool for segregation research, there have been few attempts to operationalise a geospatial model, using empirical data for a real urban area.

This thesis contributes to research on the simulation of social phenomena by developing a geospatial agent-based model (ABM) of residential segregation from empirical population data for the Helsinki Metropolitan Area (HMA). The urban structure, population composition, density and socio-spatial distribution of the HMA is represented within the modelling environment. Whilst the operational parameters of the model remain highly simplified in order to make processes more transparent, it permits exploration of possible system behaviour by placing it in a manipulative form. Specifically, this study uses simulation to test whether individual preferences, based on social homophily, are capable of producing segregation in a theoretical system which is absent of discrimination and other factors which may constrain residential choice. Three different scenarios were conducted, corresponding to different preference structures and demands for co-group neighbours. Each scenario was simulated for three different potential sorting variables derived from the literature; socio-economic status (income), cultural capital (education level) and language groups (mother tongue). Segregation increases in all of the simulations, however there are considerable behavioural differences between the different scenarios and grouping variables. The results broadly support the idea that individual residential choices by households are capable of producing and maintaining segregation under the right theoretical conditions. As a relatively novel approach to segregation research, the components, processes and, parameters of the developed model are described in detail for transparency. Limitations of such an approach are addressed at length, and attention is given to methods of measuring and reporting on the evolution and results of the simulations. The potential and limitations of using simulation in segregation research is highlighted through this work.

Avainsanat – Nyckelord – Keywords
Residential segregation, agent-based model, complex systems, social distance, homophily, residential preference, simulation

Säilytyspaikka – Förvaringställe – Where deposited

University of Helsinki electronic theses library E-thesis/HELDA

Muita tietoja – Övriga uppgifter – Additional information

# Contents

# Abbreviations

| | |
|---|---|
| **ABM** | Agent-Based Model |
| **GAMA** | GIS Agent-based Modelling Architecture |
| **GAML** | GAma Modelling Language |
| **GIS** | Geographic Information System |
| **GUI** | Graphical User Interface |
| **HMA** | Helsinki Metropolitan Area |
| **HSY** | Helsinki Region Environmental Services |
| **KDE** | Kernel Density Estimation |

# 1 Introduction

With rising income inequalities and increasing immigration in European cities, residential segregation remains a key concern for city planners and policy makers. Residential segregation, as a highly visible physical manifestation of social segregation, is commonly described as the degree to which different social groups live apart from one another in urban areas. Whilst residential segregation can be produced through a range of factors, it is ultimately the result of the residential mobility of individual households. Understanding the basis on which residential mobility decisions are made is therefore crucial for the understanding of urban phenomenon such as segregation. Residential mobility decisions may be influenced by a range of push and pull factors, including the residential preferences of individual households. Whilst many factors are likely in play simultaneously, this study will investigate whether preferences alone are sufficient to create segregation. Specifically, preferences for living in neighbourhoods that consist of households who are of a similar social group to oneself will be investigated. This tendency for similarly-disposed individuals to cluster together in virtual (social networks) and physical space, is known as homophily in social-network theory, or social distance dynamics in urban ecological traditions. Many studies have investigated the potential role of social homophily in residential decision-making, particularly with respect to households’ socio-economic position, socio-cultural disposition and ethno-linguistic profile. Several reasons as posited for why such clustering may occur. Benefits may be derived from living with other people with whom one perceives shared common values, traditions or language; and consumption benefits may be gained from living with others who share a similar lifestyle in terms of services and amenities. In some circumstances, however, the resulting segregation can have negative impacts for both individuals and society, particularly when the segregation is not ‘voluntary’ and existing disadvantage is compounded by the effects of spatial segregation. Socio-economic and ethnic segregation remain the focus of research and policy debates, however, in many post-industrial societies there appears to be increasing spatial differentiation and fragmentation also by level of education and other forms of cultural capital (e.g. Boterman et al., 2020; Vaattovaara & Kortteinen, 2003).

The process by which individual household preferences can produce segregation at the level of an urban region is difficult to study, as segregation processes exhibit many of the hallmark characteristics of complex systems, such as interdependence, emergence, self-organisation, nonlinearity and path dependence. Given these properties, it is difficult to predict macro-level patterns by studying individual residential choices alone. Similarly, it is difficult to infer individual behaviour by describing and quantifying segregation patterns at the macro level. Much work on segregation has addressed its potential causes and its impact on the social and urban fabric. Many studies also seek to measure and describe changes in the population composition of neighbourhoods over time. However, our knowledge of the interaction between micro level processes and macro system behaviour remains limited. In order to better understand segregation as a process, new tools and methodologies are required which permit an interrogation of the relationship between the behaviour of individual households and segregation at the level of a city or urban region. This study demonstrates how simulation and agent-based models (ABM) may be used as tool in segregation research.

## 1.1 Purpose of the study

This study contributes to research on the simulation of social phenomena and segregation processes. In particular, this study advances the use of agent-based modelling as a simulation tool for explorative analysis of the complex processes underlying segregation dynamics. Building upon the pioneering work of Thomas Schelling and his early individual-based models, this study seeks to explore household preferences, specifically social homophily, as a potential driver of segregation. Several new features are introduced into the model. Firstly, the model environment is representative of a real urban area, using GIS vector data. Secondly, the composition and distribution of agents is representative of the real population of the Helsinki Metropolitan Area (HMA). Multiple population groups of different sizes are modelled. Finally, the mechanisms of measuring and tracking the evolution of segregation over time are more detailed, introducing spatial relations into the measures and using multiple indices to inform different dimensions of segregation.

A main objective of this study is to develop a geographically-explicit agent-based model capable of simulating the impact that different household residential choice scenarios may have on segregation dynamics. The modelling environment and population structure will be initialised using geospatial data from the HMA. Whilst the focus of this particular study will be on homophily as one potential contributor to residential decision-making, with the introduction of new data sources, the model will permit exploration of any number of decision-making factors by permitting discrete changes in parameters. As ABM remains a rather novel research method in social studies, the development, parametrisation and operational processes of the model will be described in detail to increase its transparency, and permit reflection on the role of simulation in future segregation research.

## 1.2 Research questions

This thesis aims to answer the following questions:

- Is the uncoordinated mobility of individual households, acting entirely upon assumed residential preferences for similar neighbours, capable of producing residential segregation in a realistic urban area?
- How do simulated segregation processes vary when residential preferences are based upon different types of social homophily; namely income, education level, and language group?
- How are these segregation processes affected by increasing intensity of demand for like-neighbours in the neighbourhood and asymmetrical preferences for different groups?

## 1.3 Structure of the study

- The background of the study is introduced in Chapter 2, providing a theoretical framework for the consideration of households' residential choice as a potential factor in residential segregation. The positioning of residential preferences within the broader framework of residential choice will be outlined, and the potential determinants of residential preference explored. Research suggesting social homophily may be an important preference factor in residential sorting in some cities will be presented.
- Chapter 3 describes the complex nature of urban segregation. The means by which individual preferences at the micro-level can produce segregated neighbourhoods at the macro level in such complex systems will be discussed, and the difficulties which this behaviour presents for traditional methods of analysis. Agent-based models are introduced as a tool for exploring the properties and behaviour of such complex systems. Previous applications of agent-based modelling to segregation research are examined.
- The study area is introduced in Chapter 4, including a brief background on residential segregation in the Helsinki Metropolitan Area.
- Chapter 5 describes the data and the specific variables which are used in the analysis.
- The methodology of the study is outlined in Chapter 6. The structure and operation of the agent-based model are explained, and key simulation parameters are defined. The choice of segregation indicators and methods of calculation are discussed.
- The results from the simulations are presented in Chapter 6. Analysis of the initial state of the model is provided to provide a baseline for comparison purposes. The simulations investigate different decision-making scenarios and the spatial segregation implications. The

focus of the experiments is on understanding the processes of segregation, rather than predicting precise outcomes.

- The final chapter turns to a discussion of the results, comparing the findings to previous studies, outlining some of the limitations of simulation as utilised in this paper, and a consideration of the future role of agent-based modelling in studying complex social system and other urban phenomena.

# 2 Background

Segregation is a multi-contextual social phenomenon which can transcend all domains of life, from schools to workplaces, transport, leisure time activities and even in the virtual sphere (Kukk et al., 2019; Y. M. Park & Kwan, 2018; Maarten van Ham & Tammaru, 2016b). Segregation can have many drivers, and is often rooted in systemic inequalities produced through macro structures such as the labour market, economic system and education system (Feitosa et al., 2011, pp. 104–105). Segregation often becomes most visible in residential neighbourhoods, and the pre-eminence of the residential neighbourhood in structuring the daily lives of many individuals makes residential segregation a key focus of concern for policy makers. Residential segregation may be organised and enabled through institution arrangements and policies, it may be economically determined through housing market mechanisms and the nature of the urban structure, it may arise through discriminatory practices, and it may arise through the interplay of individual residential choices (Schelling, 1969, p. 488). This paper will focus on this latter aspect, the capacity for individual residential choices to produce segregation in a city.

Residential segregation is often broadly described as the degree of spatial separation between two or more population groups in a region. Changes in neighbourhood composition arise from the residential mobility of households. As such, it is important to consider on what basis these residential mobility decisions may be made. This section will discuss the various factors that can construct and influence residential preferences, and how these preferences may be differentiated, before turning to the specific type of preference studied in this paper, that of social homophily. Whilst only one specific type of residential preference will be examined in this paper, it is important to place this in the context of residential choice more broadly.

## 2.1 Residential choice

Bourdieu (2005) describes residential choice as being dependent upon two factors; the state of supply of dwellings, and the economic disposition of ‘agents’. An agent’s economic disposition is constituted not only by the economic resources and various forms of capital it possesses, but also by the individual preferences of the agent (Bourdieu, 2005, p. 15). These individual preferences are highly subjective, informed by what Bourdieu (2005, p. 211) terms the agent’s ‘habitus’, a socialised subjectivity constructed through collective and individual histories. Residential preferences are thus shaped, and differentiated, by many determining factors, such as an agent’s economic capital, cultural capital, age, stage of life and family structure (Bourdieu, 2005, p. 25).

The State and other macro-level actors (e.g. building companies, banks) are also active in shaping agents preferences and residential decisions. The State is directly involved in constructing both the state of supply in housing market, as well as shaping the demand. A wide array of instruments employed by the State can contribute to the construction of the economic disposition of the agent. These include housing policy, housing systems, labour market, real estate markets, welfare systems, state policies and investments (Bourdieu, 2005, p. 15). Preferences for home ownership, for instance, can be steered through regulation and financial assistance. An agent's resources may be boosted through welfare systems, incentives and assistance for buying homes. The range of possible housing types and residential milieus available reinforce expectations, and limit choice (Bourdieu, 2005, p. 16). Developers and construction companies can also be instrumental in shaping preferences, by setting standards for particular house styles, or promoting the benefits of urban/suburban lifestyles and neighbourhoods (Bourdieu, 2005). Advertising plays a key role in specifying, reinforcing and shaping expectations in the area of housing (Bourdieu, 2005, p. 89).

There is an important distinction here between household *residential preferences* and *residential choice*. Agent's choices are constrained by at least the state of supply (including the structure, rules and regulation of the housing system), and the agent's economic resources.

## 2.2 Determinants of residential preferences

The many possible determinants of residential preferences are often simplified into three main categories; demographic factors, geographic factors, and socio-economic and cultural factors.

*Demographic determinants* of residential preference largely relate to the household composition and structure, which directly determine the housing needs (Vasanen, 2012, p. 304). A large family, for example, may prefer a dwelling with more bedrooms than a single occupant would require. Demographically determined residential preferences may adapt as household compositions change, and can be impacted by life course events such as having children, marriage and divorce. Whilst preferences of this nature often relate to characteristics of housing units, household demographics may also determine neighbourhood preferences if some types of dwellings are only located in certain areas. For example, a preference for a detached house may preclude certain neighbourhoods.

*Geographic considerations* constitute a key preference component for many households' when making residential decisions, as access is required to certain facilities within and beyond the neighbourhood to fulfil duties in their daily routines (Karsten, 2007, p. 85). Time-geographical literature stresses the importance of time-cost strategies and distance to activity locations in residential decision making (Karsten, 2007, p. 85). Households may have preferences for living close to schools and workplaces for example, as well as other facilities and services required to maintain a desired lifestyle, such as parks, services, etc. Geographic determinants of preference may contribute to segregation if residents of the same social group cluster around certain activity centres. For example, certain types of workplace may only be located in specific areas of an urban region.

*Socio-economic and cultural determinants* of residential preferences may perhaps be most directly related to residential segregation. Many studies have suggested class divisions in residential preferences. Drawing on empirical data from France, Bourdieu (2005) points to a differentiation of preferences based upon various socio-economic factors, including economic capital, cultural capital, and social trajectory. Those working in occupations associated with the highest levels of cultural capital (e.g. teachers, artistic occupations) are more likely to rent, for example, whilst households with hereditary occupations, and more reliant on their *economic heritage* for reproduction, are more likely to invest in owning their dwelling (Bourdieu, 2005, pp. 26–28). For households that value cultural venues and amenities such as specialty shops, cafes and bars, preference for an urban environment and lifestyle may be an important determining factor (Ærø, 2006; Brun & Fagnani, 1994; Karsten, 2007; van Diepen & Musterd, 2009). Urban living and the associated cultural consumption can also serve as a way to demonstrate distinction (Bourdieu, 1984; van Gent et al., 2019). Residential suburbs on the other hand are considered to be main habitat of middle-class families (e.g. Karsten, 2007). In addition to the physical structure and location of residential neighbourhoods, socially and culturally determined household preferences may also relate to the social composition of the neighbourhood in which they live.

It is this latter type of preference which will be explored in this study; more specifically, a preference for living in neighbourhoods with residents who are of a similar social group as oneself.

## 2.3 Homophily and social distance dynamics

The tendency for individuals to associate with others who are similar to themselves is known as 'homophily' in social network theory. McPherson et al. (2001, p. 416) define homophily as 'the principle that contact between similar people occurs at a higher rate than among dissimilar people'. In the urban ecological tradition, Robert Park and his Chicago School colleagues, describe homophily in terms of 'social distance' between different population groups. Social distance refers to the degree of dissimilarity between different households in terms of their socio-economic, ethnic or cultural background. All else held equal, households of a similar social status will have low 'social distance' from each other. Ties between individuals with high social distance are thought to dissolve at a higher rate, contributing to the formation of niches within social space (McPherson et al., 2001, p. 415). In a residential context these social distances can become spatial, translated into physical distances, as similarly-disposed households congregate in residential areas creating internally homogenous neighbourhoods (Fossett, 2006, p. 187). As evidence of these social distance dynamics in play, Park (1926, p. 9) points to the Chinatowns and Little Sicilies throughout the world, and their formation on the basis of language, culture and race. Some may find benefits in living amongst similarly-disposed individuals with whom they perceive to share a common set of norms, values, culture, attitudes and ways of life (Clark, 1991; OECD, 2018; Schelling, 1971). Living close to people who share similar preferences, needs, and lifestyles can have the consumption benefits of shared services and facilities, such as specialty store, cultural and religious institutions (OECD, 2018). It may additionally be important in maintaining group identity and coherence for some minority groups (Bakens & Pryce, 2019). In unfavourable circumstances, living in a homogenous area may also be a defensive response for populations who fear attack, and find comfort, acceptance and solidarity in residing in an area that has a relatively homogenous makeup (Fossett, 2006; OECD, 2018).

At the other end of the equation, households with a high social distance from each another are thought to be more likely to display avoidance behaviour, and live separately from each other. Fossett (2006, p. 187) argues that this avoidance behaviour is particularly exercised by higher-status households and households from majority ethnic groups who have an aversion to co-residing in neighbourhoods with lower-status households and minority ethnic groups. Social distances can be mobile, changing over time, particularly through lifecycle events such as changes in occupation, changing socio-economic position etc. (Park, 1926, p. 9).

The mechanism by which these individual preferences can produce homogenous residential neighbourhoods has also been studied by economists such at Thomas Schelling (1969, 1971, 1978) through the use of theoretical models (Fossett, 2006, p. 189). Schelling's simple models of residential segregation were influential in demonstrating how segregated neighbourhoods could emerge from the uncoordinated actions of many individual households who make decisions about where to live based on even a very small preference for similar neighbours.

Mcpherson et al. (2001) contend that homophily exists in multiple dimensions including ethnicity, age, religion, education, occupation, and gender. A mismatch between the social attributes of a household and the composition of the neighbourhood in which they reside has been shown to increase the probability of the household leaving the neighbourhood in some cities (van Gent et al., 2019). Most of the research and modelling of homophily has been in context of ethnic segregation in the US. Whilst ethnic segregation in Western Europe is far less pronounced, recent European studies have reported evidence of clustering on the basis of income, ethnicity and class (Boterman et al., 2020; Musterd et al., 2016; Skifter Andersen, 2015; van Gent et al., 2019).

The importance of social distance dynamics on residential segregation is not without its critics. Particularly in the US, where legal and institutional forms of housing discrimination have a long history, there has been much dialogue between researchers arguing for the pre-eminence of preferences or discrimination as the most important factor in segregation dynamics. Discrimination theorists such as Massey and Denton (1993) and Yinger (1995), for example, submit arguments including an assertion that preference alone is not sufficient to form and sustain segregation, heterogeneity in preferences is not sufficiently represented in modelling, and that minority preferences are integration-promoting. Fossett (2006) addresses these criticism in length, however it is outside of the scope of this paper to delve into this discussion. Preferences and discrimination are not mutually exclusive. Urban ecological theory acknowledges their coexistence, and that one may reinforce and amplify the other in their the impacts, producing segregation (Fossett, 2006, pp. 187–189). This study does not seek to discount the importance of discrimination or other factors that can produce or facilitate segregation. Rather this study is interested in exploring the theoretical question of whether homophilious preferences are capable of producing segregation, in a situation where all other potential causal factors are absent.

## 2.4 Factors in homophilious residential sorting

Socio-economic and ethnic segregation remain the focus of academic and policy debates, however, in many post-industrial societies there appears to be increasing differentiation and fragmentation also by level of education and other forms of cultural capital (Boterman et al., 2020; Vaattovaara &

Kortteinen, 2003; van Gent et al., 2019). Accordingly, this study considers socio-economic status, cultural capital and language as potential factors in homophilious residential sorting.

### ***Socio-economic status***

Due to the legitimate social and health concerns associated with concentrations of poverty and disadvantage, segregation discourse often focuses on clusters of low-income households. However, Marcińczak et al. (2015) report that it is actually the highest social strata which is most segregated in European cities. Gated communities represent an extreme example of the self-segregation of the most privileged in society. Bourdieu (2005, p. 19) sees housing choice as an integral part of class dynamics, describing the house as a material good, 'exposed to the general gaze' on a lasting basis, and therefore playing a large role in revealing household means and taste. This sentiment can be extended to different locations within a city, and the distinction that having an address in certain neighbourhoods can bring. Musterd et al (2016) use individual-level longitudinal data from the Netherlands to examine mismatches between households and their neighbourhood in term of social distance based on income. They found that larger social distances between households and their neighbourhood was associated with an inclination to move, and that the moves often resulted in closing those social distances. This was the case for both those households who were substantially richer than the neighbourhood, and those significantly poorer. Similarly, Van Gent et al. (2019) found that a mismatch of income levels of Dutch couples and the neighbourhood in which they live had a significant effect on the probability of moving, with resettling locations more aligned with their own incomes. In the Finnish context, Vaalavuo et al. (2019) found evidence that upward income mobility is positively related to mobility out of low-income areas, but that this effect of income mobility was weaker amongst households of an immigrant background.

### ***Language and ethnicity***

McPherson et al. (2001, p. 415) contend that homophily in race and ethnicity creates the strongest divides in our personal environments. Particularly for recently-arrived immigrants, living near others who share a culture and language may provide support networks, financial assistance and labour market opportunities (Andersson et al., 2019; Cheshire, 2012). Should the reliance on these ethnic networks be great, however, this can hinder progress in learning the host language, long-term employment prospects and integration (Edin et al., 2003). A significant amount of the research into the role of individual preferences on residential segregation has focussed on ethnic segregation, particularly in the context of the United States (Bruch & Mare, 2006; Clark, 1991; Schelling, 1971). Survey data from the US suggests that whilst Black residential preferences favoured more mixed neighbourhoods than whites, they still did not want to be the minority within the neighbourhood (Bruch & Mare, 2006; Clark, 2009). In Europe, ethnic segregation is generally not as pronounced as in US cities, however in analysis of survey data from six European cities; Górny & Toruńczyk-Ruiz (2014) found a negative association between ethnic diversity in the neighbourhood, and neighbourhood attachment. Similarly, based on analysis of surnames from housing register data in Scotland, Bakens & Pryce (2019) found that the presence of specific ethnic groups in a neighbourhood positively affects the flow of co-ethnic households into these neighbourhoods. Mulder (2007) argues for a greater consideration of the role of family outside the household in residential decision making and ethnic segregation. This is consistent with a Danish study which found that preferences for living close to family and friends are a major pull factor for ethnic minorities for living in homogenous ethnic enclaves (Skifter Andersen, 2015).

In the Helsinki region, Kauppinen and van Ham (2019) report that the intraregional migration of Finnish-origin residents are one of the main forces driving ethnic segregation, whilst the migration patterns of non-Western minority groups serve to reduce ethnic segregation in the region. Survey responses of native residents in Helsinki (N =1,339) suggest that almost 60% of native residents do not desire neighbourhood ethnic mix, and almost 45% of respondents agreed that certain ethnic groups should be avoided when choosing a neighbourhood (Andersson et al., 2017). The same study found that with increasing percentage of immigrants in the neighbourhood, the desire for a lower immigrant proportion increased in a rather linear manner. By contrast, interviews with residents of Somali origin in Helsinki, one of the largest minority ethnic groups in the region, support minority preferences for mixed neighbourhood (Dhalmann, 2013). However, whilst homogenous co-ethnic neighbourhoods were not desired, neither was becoming a minority in a neighbourhood dominated by native Finns due to fears of racial abuse. Whilst new immigrants may find benefits from tapping into established migrant networks in Finnish cities, Huynh (2019) finds that second generation immigrants exhibit mobility patterns in line with native residents.

### *Cultural capital*

Whilst considerably understudied compared to ethnic and income based homophily, there is evidence that socio-cultural homophily may also be an important factor in residential sorting. Van Gent et al (2019) point to sociocultural fractures within and between classes playing out spatially in Dutch cities. In their study of residential sorting amongst couples, they found that with increasing percentages of couples of the same education level and sociocultural group in the neighbourhood, the probability of moving decreased. The effect of socio-cultural homophily was found to be even stronger than income or ethnic background as a predictor of neighbourhood sorting. Hanquinet et al. (2012) report that residents of Brussels with high levels of cultural capital are highly selective in their neighbourhood choices, clustering within specific areas according to either modern or classical orientations. By contrast, no discernible patterns were detected for the ‘culturally disengaged’ (Hanquinet et al., 2012, p. 526). Boterman et al. (2020) found that education was more important than income in neighbourhood sorting in their sample of employed residents aged 24 years or over in Amsterdam. Lower-educated and highly-educated residents were found to be living in very different areas of the city, with university graduates being more segregated than those with the highest incomes. These effects are reinforced and amplified by status inequality and economic competition as households with more resources have advantages when competing for high-quality housing and desirable neighbourhoods (Boterman et al., 2020; Fossett, 2006, p. 188).

Whilst these studies point to evidence suggestive of social homophily being a contributing element in residential mobility, the wide range of factors influencing and constraining residential choice mean that it is often difficult to attribute cause to a single factor. The distinction between stated and revealed preferences also becomes important here (see Hasanzadeh et al., 2019; Vasanen, 2012). Stated preferences may not be acted upon, and residential mobility towards homogenous neighbourhoods may have other causal factors. Different forms of homophily may be of differing importance for different social groups, and some forms of segregation may be subordinate to others. In US, for example, income segregation among Black and Hispanic families has been found to be higher than among white families (Reardon & Bischoff, 2011). Sociocultural differentiation in the Netherlands outweighs income or ethnic background as an explanation for residential mobility and spatial sorting (van Gent et al., 2019, p. 894). Hedman et al. (2011) found evidence that sorting processes on the basis of ethnic, socio-economic and demographic dimensions (including income and education) happen simultaneously in Uppsala, Sweden. Residential preferences and segregation processes are likewise not consistent geographically, and can differ from region to region and from city to city (e.g. Clark, 2009).

# 3 Urban segregation as a complex system

## 3.1 The complex nature of urban segregation

The individual residential preferences of households may seem insignificant when considering residential segregation at the level of a city or urban region, however these individual mobility decisions have been shown to be capable of producing unanticipated segregation at the macro-level (Schelling, 1971). This behaviour is due to urban segregation exhibiting many of the characteristics of a complex system. Similar behaviour can be observed in other complex systems, such as the macro phenomenon of inflations and depressions, which can arise as an unforeseen result of individual decisions regarding purchasing and savings (Schelling, 1971, pp. 145–146). Theories of complex systems emerged in the natural sciences from the 1970s. Central characteristics of complex systems, such as interdependence, emergence, self-organisation, nonlinearity and path dependence can all be present in urban segregation processes.

***Interdependence***

Complex systems are characterised by the interdependence and interconnectivity between the many independent elements that comprise the system, and between the system and its environment.
The effects of the actions or interactions of only one or a few elements of a system will propagate throughout the system, influencing all other related parts, but not in a linear manner. Due to this interdependence, pushing on a complex system "here" often has effects "over there" (Bar-Yam, 2002), which makes problems difficult to solve, as cause and effect may not be obviously related. For example, a household moving from one location to another will affect the composition in both the neighbourhood of departure and arrival. Reactions to this move by other households could prompt further migrations elsewhere in the system if, for instance, friends or family relocate to the new neighbourhood, or new neighbours are unhappy with the change in compositions which this new arrival brings. In the same way, an urban renewal project seeking to address segregation in one neighbourhood can have impacts well beyond its boundaries.

***Self-organisation***

In complex systems, there is no single centralised control mechanism that governs system behaviour. Rather, the global pattern or behaviour is produced through the local interactions of the individual components which comprise the systems. Schelling (1971) demonstrated how unintended and unexpected patterns of segregation can emerge from the uncoordinated actions of individual households who decide where to live, in the absence of any deliberate external control.

### *Emergence*

One of the central concepts in complexity theory is that of ‘emergence’ whereby local interactions between elements at the micro-level give rise to the global properties of the system. Under these conditions, an apparently simple ‘higher-level order’, such as a segregated urban region, can emerge from relatively complex ‘lower-level processes’ (Sawyer, 2005). Feedback between the higher and lower orders occurs in both directions (Figure 1). For example, the local interactions of households can contribute to shaping the emergent higher-level system, but households may equally recognise and react to these emergent patterns of segregation (Feitosa, 2010). The entire system also interacts with the external environment and other macro conditions. As discussed in Chapter 2, there are many interrelated factors that can influence residential choices, which are in turn influenced by and influence upon, many macro conditions such as the real estate market and public policy.

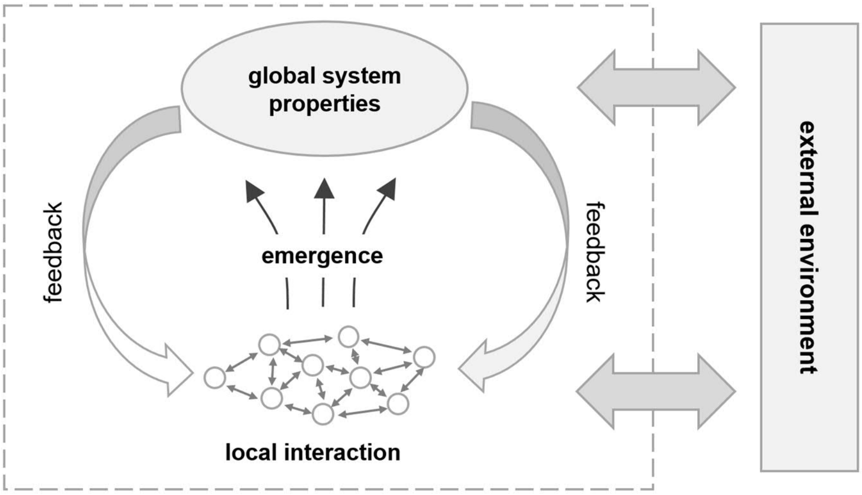


*Figure 1 – Conceptual diagram illustrating the principles of complex systems. Global system properties emerge from interactions and self-organisation at the local level. These properties and structures then feed back into the system, affecting lower-order behaviour. The entire system is simultaneously interacting with the external environment (e.g. macro level policies, labour market, housing market), which influences both the global structure as well as shaping individuals preferences and choices. (Adapted from Adrus, 2005)*

### *Nonlinearity*

This interdependence and feedback between local system components and the emergent global system properties, introduces nonlinearity into the system. Unlike in simple systems, changes in the local properties or rules are not correlated in a linear fashion with outcomes. Small changes at the local context may have surprisingly profound and unexpected impacts on the global behaviour, or may have no discernible influence at all. Whilst the properties of simple systems, even when complicated, can be understood by independently studying each individual component, for complex systems “the whole is more than the sum of the parts”. As such, the overall behaviour of the system and emergent global structure cannot be predicted by having knowledge of the behaviour of the individual elements (e.g. households). As the rate at which segregation occurs is dependent upon many interconnected processes, it can vary greatly depending upon state of the system and the environment. A seemingly stable system state may also be ‘tipped’ by individual actions. For example, Schelling (1972) demonstrated how a previously integrated neighbourhood may be tipped towards segregation through the arrival or departure of an individual household.

***Path dependence***

As a complex system, urban segregation is also characterised by path dependence, as previous interactions and residential mobility decisions shape the system (Feitosa, 2010, p. 37). The current population distribution of a city is the end state of many past residential mobility decisions. In this way, decisions made now, as well as past interactions, can affect future possibilities and outcomes in unknown ways. Residential mobility decisions may be influenced by information, social networks and resources gathered over a lifetime. Previous experiences of households or even shared historical reputations of residential areas may factor into future residential decisions of households.

## 3.2 Simulation as a tool for exploring complex systems

Given these characteristics, complex systems pose significant challenges for researchers seeking to study phenomenon such as segregation. Traditional methods of studying segregation tend to approach segregation at either the macro level, measuring and describing patterns, or the micro level, seeking to identify factors that may influence individual mobility decisions. Due to the behaviour of complex systems, little is known, however, of the nature of the relationship between the two forces (Feitosa, 2010, p. 36). Recognising this need for new methodologies in order to study these micro-macro relations and increase our understanding of complex systems, modelling and computer-based simulations have become increasingly employed for studying complex urban phenomena such as segregation.

***Agent-based models (ABMs)***

A model is a simplification of real-world systems, put into a manipulative form, permitting tests and simulations to gain insights into the properties of the real system (Clarke, 2014). One particular type of model, agent-based models (ABMs), have become increasingly utilised for simulating human mobility and social systems. ABMs, also known as multi-agent models, have a significant advantage over other types of models when it comes to modelling the interaction of behavioural units (the ‘agent’), such as a residential household, and the environment (Clarke, 2014, p. 1218). ABMs simulate the behaviour of a system, as defined by the interactions between agents and the environment. By modifying the simulation parameters, it is possible to test how changes in individual and collective behaviour can impact the system’s aggregate behaviour (Clarke, 2014; Sawyer, 2005). There are many different ways in which ABMs can be used. Simulations may seek to make predictions about what kind of behaviour a system may display in the future, based upon arbitrary or empirical data and rule-sets. For the purposes of the study however, the main interest of ABMs is their use for exploratory purposes; to investigate the properties and dynamics of a system, demonstrate that something is possible, or propose new ideas about a complex situation (Feitosa, 2010, p. 43).

### ***Key components of ABM***

***A***gent-based models are comprised of three main components; decision-making units called *agents*, a set of *rules* which govern the agents behaviour, and an *environment* in which the agents operate, which can both influence and be influenced by the agent (Clarke, 2014).

*Agents* are autonomous units situated in space and time. An agent can perceive its environment, interact with other agents, and undertake actions according to its own rules/goals. Agents are capable of processing information and exchanging information with other agents, in order to carry out their goals (Crooks & Heppenstall, 2012). Agents are heterogeneous. Each agent possesses its own attributes, behaviour and goals; however, similar agents may equally form groups that exhibit common behaviour. Agents can represent any type of autonomous object, such as cars in a traffic simulation model, pedestrians in a crowd management simulation, or households in a residential segregation model.

The model *environment* defines the space in which agents move around and operate. In many ABMs, the environment is abstractly simulated, generally using a two-dimension grid structure, however the integration of geographic information systems (GIS) into the model allows for the inclusion of detailed representations of the real world (Crooks, 2010; Feitosa, 2010). In a model of segregation, for example, the environment can represent an urban region, within which agents can move around, and identify other agents in their geographical vicinity. The environment has its own attributes and can additionally perform actions in interaction with the agents. A residential location may have a price or capacity that increases and decreases based upon interactions with the agents who move in and out of the space for example.

Agents operate autonomously within the environment taking actions according to a *set of rules,* which are defined in the simulation parameters. These rules affect the agent's behaviour and relationship with other agents and the surrounding environment (Crooks & Heppenstall, 2012). Agents can interact directly or indirectly with other agents and the environment, observing their location and behaviour, and can transmit information. In a segregation model, for instance, households can have indirect interactions by detecting the status of households that live in a specific neighbourhood, which may then inform their decisions on whether to stay or leave the neighbourhood (Feitosa, 2010). Rules may be applied to all agents, groups of agents, or to individual agents. Agents may also learn from the interactions and adapt their rules. Rules are generally in the form of 'if-else' statements (e.g. if a chosen space is vacant, move to that space, else, choose a new space). Actions may be triggered in reaction to interaction with other agents, or may be undertaken irrespective of other agents.

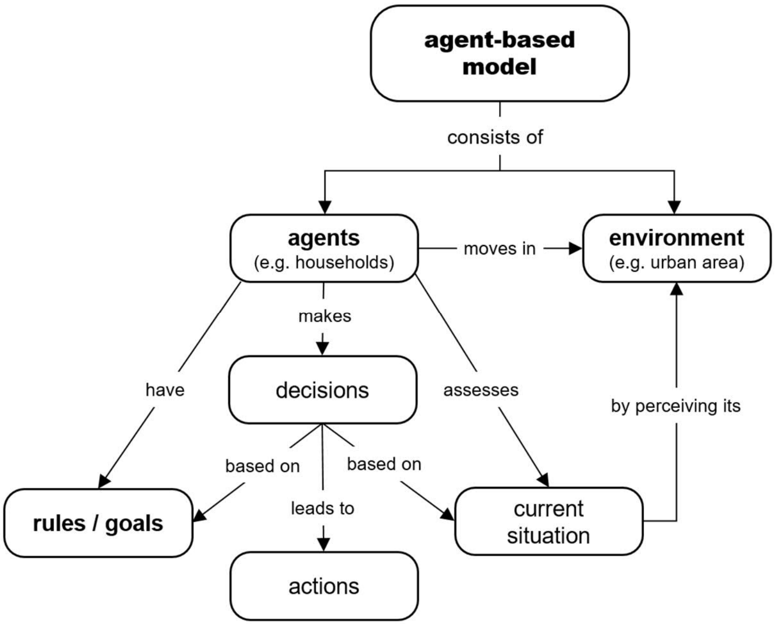


*Figure 2 - Overview of structure an agent-based model and its operation (Adapted from Hall & Virrantaus, 2016)*

Figure 2 demonstrates the interaction of these elements in a simple ABM. Using the example of a residential segregation model, a household can be represented by an agent who acts upon a set of rules that inform their residential mobility and decision-making. The environment may represent an urban area, and residential locations from which the agent can choose. The agent can interact with the environment by moving through space, and also indirectly by detecting the locations and properties of other agents in the environment. The environment changes in response to the migration of agents, and in turn, the state of the environment may affect agents' mobility decisions.

## 3.3 Simulating segregation processes with ABMs

Nobel Prize winning economist Thomas Schelling (1971, 1972, 1978) is credited with developing the first social application of an agent-based model, using simulation as a tool to investigate the dynamics of ethnic segregation in US cities. Schelling's models built the foundation for an individually-based model focused on the actions of agents, representing people, who made decisions about where to live in a simulated world (Ardestani et al., 2018, p. 1). Schelling's models demonstrate how individual household's residential choices can lead to unanticipated and unexpected segregation in a city, even when this was not explicitly desired by any households.

Schelling's 'checkerboard model' (1971) considers the interactions of two different groups of 'agents', representing households, which are equal in all characteristics except for their colour (black or white). The rule-set prescribed that each agent has a preference for living in a neighbourhood that contains at least a certain percentage of like agents of the same colour. The black and white agents were distributed randomly across a checkerboard grid structure, with each square representing a dwelling. The agent's neighbourhood was based on spatial proximity, defined as the surrounding eight squares on the checkerboard. The basic operation of the model can be summarised as follows:

At every iteration, an agent has the possibility to remain in their current location, or move to another vacant space. The agent makes a decision about whether to relocate by assessing the composition of its neighbourhood, and comparing this to its preference threshold. If the percentage of like-agents in the neighbourhood is higher than the defined preference threshold, the agent remains in the location, if not the agent relocates by choosing a vacant space which meets their preference level elsewhere on the board. For example if the agents preference is 50% it requires that at least 4 of its 8 neighbours are of the same colour as itself, if not, it moves to a location which does meet this criteria. Agents are content with a mixed neighbourhood as long as their preference threshold for like-neighbours is satisfied. Each agent takes a turn to decide whether to remain or move until a cycle has been complete.

Schelling conducted many simulations with variations on these parameters, however the result of acting on these preferences for homophily, even when very low, inevitably results in unanticipated segregation of the two groups (Figure 3).

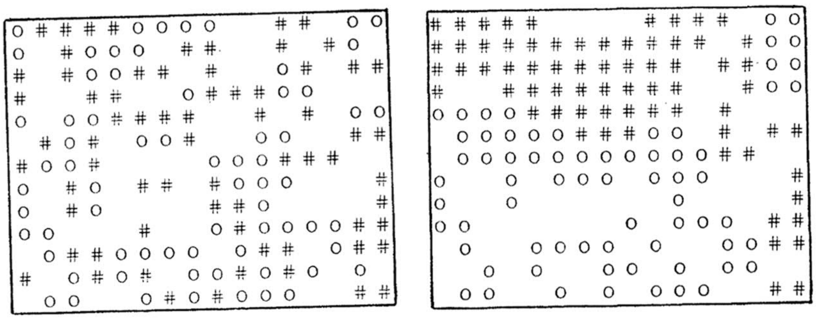


*Figure 3 - An example taken from Schelling (1971) of a random initial distribution (left), and the resulting segregation (right) based on agents of both groups preferring their neighbours to be at least 50% of the same colour as themselves.*

The migration of an individual agent changes the composition in both the neighbourhood of departure and arrival. The neighbourhood they leave becomes less attractive to their own group, and the neighbourhood they move into becomes less attractive to members of the other groups, creating homogenous neighbourhoods over time (Clark & Fossett, 2008). When there are asymmetrical preferences (i.e. one group has a higher demand for like neighbours than the other), the group with the higher preference ended up with more like-neighbours, however not significantly more than the other group. If one of the groups was a minority, but maintained the same preferences as the majority group, then the segregation outcomes were found to be more extreme. Based on the initial random distribution, agents of the minority group are more likely to be unhappy with their current neighbourhood, and so then quickly move towards existing clusters of their own group. The most extreme results were produced when one group was a minority, and also held a higher preference for like-neighbours, even when they are still happy to be a minority in the neighbourhood.

Schelling also found that integrationist preferences (e.g. at least 3 of 8 neighbours from a different group to oneself, but not more than 6) can be insufficient to avoid segregation (Schelling, 1971, p. 165). Schelling notes that particularly when one group is a minority, producing an integrated population requires the conscious and deliberate positioning of the minority group like a mosaic throughout the board, suggesting this is unlikely without a concerted effort.

In all of his simulations, Schelling demonstrated how segregation can occur in the absence of centralised structures or housing discrimination, based on small individual preferences for living in a neighbourhood with neighbours of the same group. Whilst this does not diminish the role that discrimination, and economic factors can have on shaping segregation, it does suggest that individual preferences should not be discounted as having a potential role.

## 3.4 Recent developments in segregation ABMs

Schelling's simple agent-based model of racial segregation dynamics has been used as the basis for many studies of urban segregation in the subsequent decades. These studies have sought to expand the theoretical footings and extend the model in a variety of ways; updating agent preferences to reflect empirical data (Bruch & Mare, 2006; Xie & Zhou, 2012), considering heterogeneous preferences within groups (Xie & Zhou, 2012), adding simple real estate market mechanisms (Zhang, 2004; Crooks, 2010; Feitosa et al., 2011), exploring different neighbourhood definitions (Crooks, 2010), allowing for migration at multiple scales (Ardestani et al., 2018; Jayaprakash et al., 2009), modelling dynamic populations with in-and-out-migration (Ardestani et al., 2018; Benenson et al., 2002; Feitosa et al., 2011), incorporating nested decision making models and multiple categories of sorting (Bruch, 2014; Feitosa et al., 2011), and accounting for geographic features acting as barriers for movement and perception (Crooks, 2010). Consistent with Schelling, the majority of ABM studies have focussed on modelling ethnic segregation processes, however a few have considered segregation on the basis of income (e.g. Benenson et al., 2009; Feitosa et al., 2011) or wealth and status (Benard & Willer, 2007).

### *Incorporating empirical data*

Clark (1991) conducted Shelling-type experiments, implementing empirically-informed preference thresholds drawn from telephone surveys conducted in five US cities, in which the respondents were asked about their preferences for neighbourhoods based on varying proportions of Black/white residents. The results broadly confirmed Schelling's findings and description of preferences; however, Clark notes that preferences of white residents were more likely to drive segregation in the US. Similarly, Bruch & Mare (2006) use survey data about preferred neighbourhood ethnic compositions in Chicago to parametrise agent preferences for co-ethnic neighbours. Departing from Schelling, Bruch & Mare implemented an agent preference function which, rather than a consistent value (e.g. 50%), reflected stepped preference structures and continuous ranges of preferences, depending on the relative proportions of groups in the neighbourhood. Xie & Zhou (2012) used the same survey data as Bruch & Mare (2006), however rather than assuming preferences to be homogenous within agent groups, they interpreted the pattern of responses as a heterogeneous scale of preference (Guttman scale), with different groups of whites having different level of tolerance of Black neighbours. This modification had the effect of reducing the degree of segregation compared to both the models of Schelling (1971) and Bruch & Mare (2006).

### *Multiple sorting criteria*

Fossett (2006) extended Schelling's model to include multiple sorting criteria. In the model, each agent is assigned an ethnicity, a socio-economic status, and preference for housing quality. 'Houses' of differing quality are distributed throughout the grid environment upon initialisation, and agents were randomly distributed, under the condition that the agent has a socio-economic status which is sufficient to afford living in the house. During each cycle, the agent consider their preferences for housing quality, neighbourhood income status, and neighbourhood ethnic composition. If a better alternative is found, the agent can move. During each cycle, a random selection of 25% of the city's households is activated to consider moving, and one out of five of these households (20%) are randomly forced to move to the best available housing unit they find, even if they would prefer to remain in their present location. This parameter ensures that the population remains dynamic and represents potential situations such as relocation due to a job, death of a household, arrival of a new household into the city etc. Fossett finds that both socio-economic and ethnic segregation are capable of forming simultaneously entirely through individual residential decisions.

### *Housing market mechanism*

Zhang (2004) added a simple housing market to the model, whereby spaces had a real estate value that responded to excess demand. Property prices were initially consistent, and whilst black agents had no preference for a particular neighbourhood configuration, white agents preferred white neighbourhoods. Zhang found that in the long run, the neighbourhoods with concentrations of white agents became more expensive, whilst vacancies were much more likely in black neighbourhoods. This was partially due to white agents seeing a greater utility value in paying more to achieve their preferences of living in white neighbourhoods, whilst the black agents saw no added value in paying more, so took the cheaper sites that the white agents had vacated.

### *Dynamic population*

Crooks (2010) shifted the model from a closed system with a static population to a dynamic model allowing for population growth and decline, by adding an 'age' attribute to each agent. Under Schelling's model, the agents were assumed to be 'immortal' and could remain in a space indefinitely should the neighbourhood remain suitable to them (Fossett & Waren, 2005). Crooks (2010) prescribed that agents had one year added to their age each cycle and 'died' once they reached 50. If agents could not find a suitable home, they were assumed to have migrated outside the region and removed from the model. A fixed number of new agents were added in to the model each cycle. Segregation persisted under these conditions; however the dynamic nature of the agents was reflected in geographic variations in the locations of agent clusters, which change over time.

### *Neighbourhood definition*

In a separate experiment, Crooks (2010) investigated the impact of neighbourhood ‘vision’ – the agents definition of what their neighbourhood was - using increasing Euclidean buffers. Larger neighbourhood definitions, were found to take longer to stabilize and produced more extreme patterns of segregation (see also, Fossett & Waren, 2005; O’Sullivan et al., 2003). Laurie & Jaggi (2003) found similar results using increasing degrees of rook contiguity neighbourhoods, however the effect was inversed when a lower preference threshold (30% in their case) was used. With a smaller neighbourhood vision, micro clusters are clearly visible, however, as vision increased, more stable and integrated neighbourhoods were formed.

### *Geographic Barriers*

Many ABM applications generally adopt a symmetrical grid to represent the environment. Under these assumptions, there are no geographical constraints that may influence agents’ mobility behaviour or conception of their neighbourhood. Using a vector-based model to break away from the uniform checkerboard, Crooks (2010) investigated the potential impacts of geometric barriers on agents’ perceptions of neighbourhood. Crooks attempts to replicate the impact of physical features that may be present in an urban environment, such as rivers, railways and highways. The model remained conceptual, unrepresentative of any real urban area. Others studies (e.g. Yin, 2009) have used the geographical features of real urban areas, but have not incorporated empirical population data, randomly generating and distributing agent groups throughout the environment.

### *Geographically-explicit models based on empirical data*

Despite these advances, there have been limited attempts to operationalise a geographically-explicit model using empirical data for a realistic urban area. The limited progress in advancing realistic models is due in part to historical issues concerning data availability, technical capabilities, and computational requirements. Many advances have been made on this front in recent years. Detailed geospatial data is now available in many cities, and several software programs, many of which are free, have simplified the technical requirements for constructing ABMs. Computing capacity continues to increase, and it is now feasible to operationalise an agent-based model using GIS data on a personal computer. At the time of writing, a very limited number of studies can be found which have attempted to model realistic cities and geographical regions using empirical data (Anderson et al., 2020; Ardestani et al., 2018; Benenson et al., 2002; Bruch, 2014; Feitosa et al., 2011; Perez et al., 2019; Yin, 2009). The key parameters of these studies will be discussed here.

Benenson et al (2002) were at the forefront of advancing realistic ABM, with their Schelling-type model simulating ethnic residential dynamics in the Yaffo district of Tel Aviv. Using time-series building-based registry data, Benenson et al. modelled household residential mobility patterns between 1955 and 1995 using a geographically-explicit model of the district. Voronoi polygons were generated to represent each building in the district, and the neighbourhood definitions respected the physical street patterns of the city. The agents were dynamic in that they accounted for steady in-migration, and probable outmigration was accounted for if agents could not find a suitable home.  When deciding where to relocate, agents ranked the potential residential opportunities by their utilities prior to making a choice, which Benenson et al (2009) find was important to making the model robust in its outcome.

Bruch (2014) used empirical register data and an accurate GIS map of city blocks for the cities of Atlanta, Chicago, and Los Angeles, to examine the role of income in racial residential segregation, constructing a multiple attribute choice model. A 20% sample of known households were generated as representative agents and allocated to census blocks which were capable of holding multiple agents. Rather than all agents being activated each cycle, a randomly chosen 0.01% sample of households is given an opportunity to migrate to a new location each iteration. Each agent evaluates the neighbourhood in which it lives (census block group) as well as a random 10% sample of vacant housing units elsewhere in the metro area. This rule-set served to reduce computation requirements, and was justified on the basis that not all households actively consider changing residence every year, and when they do, they are likely to only view a fraction of the available properties.

Ardestani et al (2018) used census data for the New Zealand city of Auckland to model the population dynamics of four major ethnic groups (Europeans, Asians, Pacific people and Maori). The entire metropolitan region was modelled as distinct districts, and a multi-scalar approach was adopted, with agents having the possibility to relocate across different scales based on empirically informed population mobility rates and income levels of each ethnic group. Whilst agents were created for the entire population of the metropolitan area (847,113 individuals) based on the census area-level data, only a fraction of these were stochastically activated for potential migration each cycle. This was justified as being representative of the actual moving patterns, with the number of relocating agents informed by census mobility values for each ethnic group, but also serves to reduce the computational costs of running the model. Besides the different scales of migration, the model also included a much broader set of determinants of agent behaviour, including the empirically-informed economic situation of each ethnic group as a factor in their migration range, as well as empirical vacancy rates, replicating the real estate market and housing policy measures. The decision-making parameters were calibrated using the 1991 and 2006 census-based benchmarks, before forecasting into the future for a further 15 years. Ardestani et al (2018) argue that the combination of these factors, as well as the accounting of in-migration and out-migration permit their model to simulate changes both in macro factors (immigration, population growth, housing development and policy) as well as micro-level factors (e.g. agent income, agent preferences for neighbourhoods etc.).

Perez et al (2019) use an ABM to simulate the dynamics of new immigrant populations arriving in Montreal, Canada. 12,000 representative agents were generated in total, based on census data. The neighbourhood definition varied for different components of the decision-making module. Agents considered the proportion of co-ethnic neighbours (African, American, Asian, and European) for a neighbourhood defined by a 1 km Euclidean radius, whilst the proportion of co-language neighbours (speaking predominantly one of the official language of English or French) considered a 3 km neighbourhood. Distance to public transport, amenities and green areas were also included in the decision-making module and had different distance buffers again. The population remained static, with no population growth/decline modelled, and native-born residents were not considered. Three scenarios were run which activated different elements of the decision-making module, and 8 different thresholds for requirements of co-ethnic neighbours and public transport stops were tested. These 24 different simulations run from the 2011 census baseline for 4 cycles (‘years’) and compared to the 2016 census data of population proportions. The model was largely able to reproduce the real movements of migrants moving to Montreal.

Anderson et al (2020) extended the model of Perez et al (2019) to Toronto and Vancouver, additionally investigating the effects of disaggregating ethnicity groups. Expanding residents of ‘Asian’ origin into Chinese, South Asian and Other Asian sub-groups they find that both the spatial patterns and the degree of separation between different ethnic groups is influenced by the degree of aggregation. 20,000 representative agents were initiated and assigned ethnicities and starting locations based on the census data.

Feitosa et al (2011) present perhaps the most highly-specified model published. Modelling the entire population of São José dos Campos in Brazil (106,951 households), a series of simulations are run, designed to investigate different segregation scenarios arising from individual agent preferences, labour market conditions, real estate markets and responses to urban policies. Under the proposed MASUS (multi-agent simulator for urban segregation) model, household agents were assigned variables including tenure status, age, income, education, household size, and presence of children. The neighbourhood spaces included variables such as land value, dwelling offers, infrastructure, type of settlement, distance to the CBD, distance to the agent's original neighbourhood, and proportions of social groups. Urban growth and sprawl are modelled, governed by dedicated sub-modules, as is the real estate market. Other sub-modules deal with the creation of new households and the removal of existing agents from the model to replicate population growth and decline according to control variables. The decision-making module utilises a nested logit framework, and relocation decisions depend upon calculated and ranked utility values and a Monte Carlo simulation. Agents perceptions take into consider both environmental conditions, as well as population compositions within their area with different distances specified for different variables. Reciprocally, agent mobility decisions affect the environmental landscape such as land value. Using this comprehensive model, Feitosa et al (2011) run a number of tests, comparing the simulation outcomes to empirical data using the period 1991-2000 (9 cycles), exploring the segregation effects of increasing/decreasing inequality, and testing the efficacy of an anti-segregation policy (dispersing poverty through the use of housing vouchers).

The different approaches adopted in each of the models described above reflect variances in data availability, computational considerations, and most importantly, the particular research question or process which they wish to investigate. Whilst some studies have sought to calibrate their models using time-series data to try an explain historic patterns of migration or make projections about future states, this comes with many risks. Even if a model can replicate past movement patterns based on a prescribed rule-set, it remains unknown whether those parameters were significant in the historic mobility of residents in the real system. Other undetected mediating or confounding factors may be involved. Over-prescribing attributes to a model risks making assumptions which cannot be validated, and may lead to incorrect causal conclusions being made. One major limitations of ABM is the inability to validate such projections. The purpose of simulation in this thesis is to explore one particular mechanism which may have a role in segregation dynamics. For this reason, the model will possess a simpler operational rule-set in order to make the relationship between individuals actions and macro outcomes more transparent.

# 4 Study area and its characteristics

The Helsinki Metropolitan Area (HMA) is the most populated urban area in Finland. In 2019, just over 1.1 million people were residing within the capital region, representing over a fifth of the total population of Finland. The HMA is comprised of four neighbouring municipalities; Helsinki (632,907 residents), Espoo (278,538 residents), Vantaa (223,536 residents) and Kauniainen (10,153 residents) (Statistics Finland, 2019b).

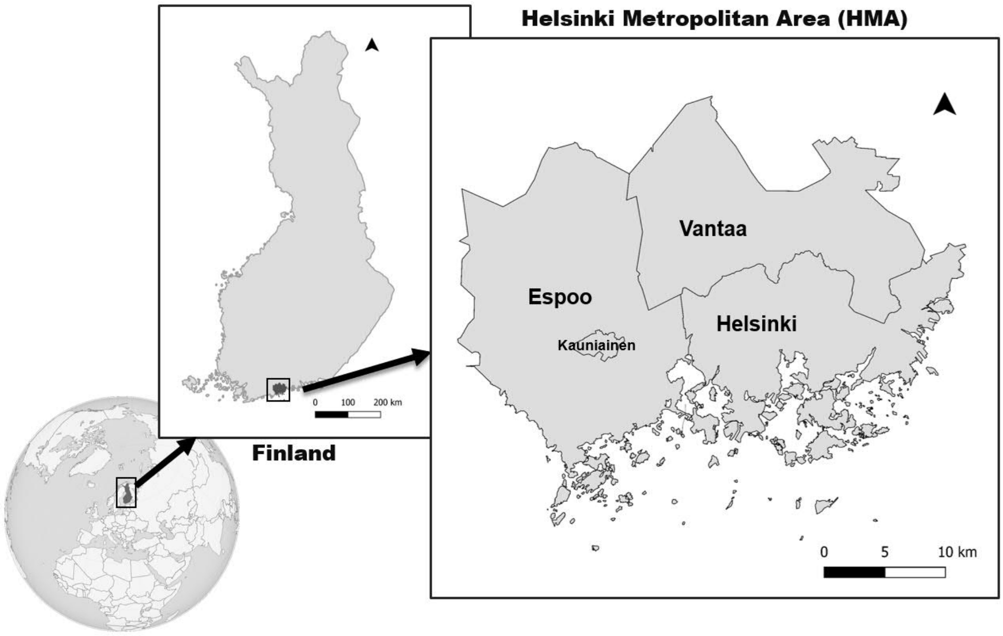


*Figure 4 – Location map with municipal subdivisions of the Helsinki Metropolitan Area (HMA).*

The HMA is one of the few urban regions in Finland that is forecast to increase its population in coming years. The Greater Helsinki Region, which includes a ring of 10 commuter municipalities, is expected to reach two million resident by 2056, growth of around 25% from current levels (City of Helsinki, 2020). Within the HMA, the inter-municipal dynamics are complex. Helsinki has been losing 3,000–3,500 residents per year on average to other municipalities in the region since 2000, and this rate is forecast to continue (City of Helsinki, 2020).

## 4.1 Residential segregation in the Helsinki Metropolitan Area

Finland, with its Welfare State ethos, is reputed for its low levels of inequalities, and broadly speaking, segregation within the HMA is very low compared to other European capitals. However, this has been changing due to growing wealth inequalities, a restructuring of the labour market favouring those possessing tertiary degrees, and significant growth is forecast in the percentage of foreign residents in the region. As the capital, and most populated region in Finland, the HMA will experience these challenges most acutely.

### *Socio-economic segregation*

Socio-economic segregation in a city is often a symptom of wider wealth inequalities in a society. Finland is well positioned on this front, boasting one of the lowest GINI coefficients in Europe and Helsinki is regarded as one of the least segregated capitals in Europe (Maarten van Ham & Tammaru, 2016a), a title the City of Helsinki explicitly aims to defend (City of Helsinki, 2017).

Socio-economic segregation is often highly related to the tenure structure and housing systems of a city or region. Low-income households are often concentrated in social rental housing, and thus their role and positioning within the urban structure can accentuate or moderate concentrations of poverty. Many large housing estates were built in the HMA in the 1960s and 1970s, however an explicit socio-spatial tenure mixing policy has resulting in the stock of municipal rental housing being rather evenly distributed throughout the city (Vaattovaara et al., 2018). This policy began already in Helsinki in the 1970’s, and was quickly adopted by other municipalities in the metropolitan area, including Espoo and Vantaa, Despite this, in recent years, there has been signs that socio-economic segregation in growing, particularly in the east and north-east of the region.

Helsinki has the highest share of low-income households (Table 1) of the four municipalities. The least socio-economically advantaged areas are mainly located in eastern and north-eastern Helsinki (Figure 5), whilst the wealthiest residents have traditionally been settling on the western shores (Vaattovaara & Kortteinen, 2003). Espoo and Kauniainen have significantly larger shares of high-income households compared to Helsinki and Vantaa.

*Table 1 – Household income categories and median income for the Helsinki Metropolitan Area (2017).* (Source: Statistics Finland, 2018, 2019b)

| | **Low-income** | | **Medium-income** | | **High-income** | | **Median household income** |
|---|---|---|---|---|---|---|---|
| | Number | % | Number | % | Number | % | |
| **Espoo** | 20,158 | *17%* | 58,987 | *49%* | 40,313 | *34%* | 41,978 € |
| **Helsinki** | 69,253 | *21%* | 174,503 | *53%* | 85,588 | *26%* | 33,263 € |
| **Vantaa** | 18,369 | *18%* | 60,991 | *60%* | 22,457 | *22%* | 36,044 € |
| **Kauniainen** | 503 | *12%* | 1,648 | *38%* | 2,158 | *50%* | 57,363 € |
| **HMA average** | 108,283 | *20%* | 296,129 | *53%* | 150,516 | *27%* | |

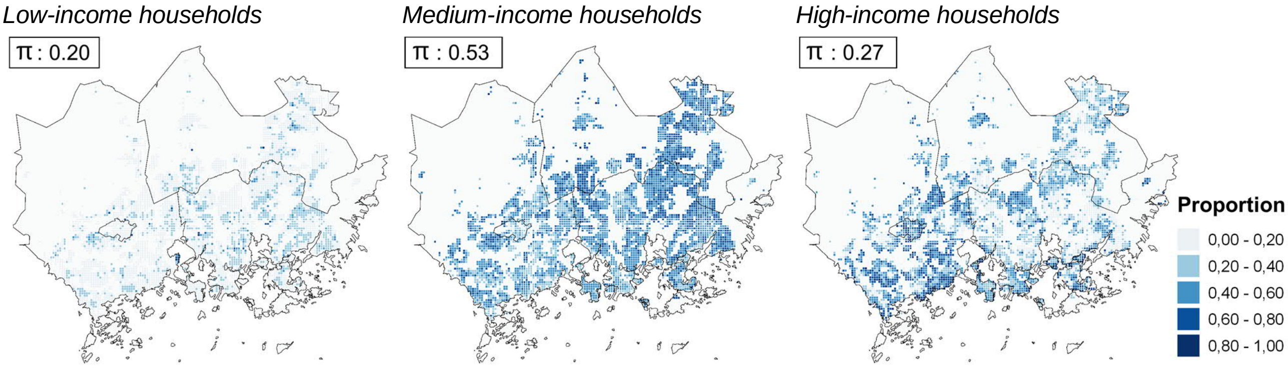


*Figure 5 – Distribution of income groups in the HMA. Local proportions of each group (%) within each 250m grid cell are displayed on the maps, whilst the group's absolute proportion in the HMA region (π) is indicated above each map. Source: Statistics Finland* (2019b)

The structure of socio-economic segregation in the Helsinki region has been strongly influenced by the economic recession in the 1990's, and the subsequent shift in international labour market conditions (Vaattovaara et al., 2018). Areas within the region which had higher proportions of unskilled workers became indebted with high unemployment, whilst those with a tertiary education became better placed to take advantage of the subsequent boom in the ICT sector in the mid-90s, fuelled by telecommunication company Nokia. Accordingly, areas which had a higher proportion of the population with high-level education were able to rebound more quickly (Vaattovaara & Kortteinen, 2003)

As each municipality in Finland has independent control of the planning and housing systems within their administrative boundaries, inter-municipal competition for working and high-income household who contribute most to their tax base has been a contributing factor for differentiation between municipalities. A large portion of the municipalities tax revenues come from the residents income taxes, and as the municipalities are responsible for social welfare and other public expenses, the benefits of attracting high-income workers is two-fold (Vaattovaara et al., 2018). Helsinki has a higher density structure compared the other municipalities. Whilst over 85% of Helsinki's housing units are in apartment buildings, this ratio is much lower in Espoo (60%), Vantaa (65%) and Kauniainen (51%). In Espoo and Vantaa, around one in four housing units is a detached home, and one in three for Kauniainen. (Statistics Finland, 2020) The commuting ring of municipalities, feature even more detached housing. The movement of wealthy households out of the housing estates and into the detached periphery has resulted in a differentiation between the housing estates and the rest of the region (Vaattovaara et al., 2018)

### *Socio-cultural segregation (level of education)*

Finland has been at the top of international rankings for educational outcomes and educational equalities for many years, regularly appearing at the top of the OECD PISA rankings (OECD 2019). There is limited evidence that the residential neighbourhood one grows up in is a deciding a factor in educational outcomes in Helsinki (Kauppinen, 2007, p. 440). Despite this, school selection strategies by parents, have been shown to amplify segregation both within the classroom and in the residential domain through the medium of school catchment zoning (Bernelius & Vaattovaara, 2016; Bernelius & Vilkama, 2019). Highly-educated families are more likely to undertake such school selection strategies (Bernelius & Vaattovaara, 2016), providing a mechanism by which highly-educated households may cluster on the basis of shared values.

Within the Helsinki metropolitan region, there is evidence of differentiation in education levels of households across the metropolitan area (Table 2). The share of residents with only basic education is higher in Vantaa (26%) than in Espoo (19%) and Helsinki (20%). Vantaa also has a considerably lower share of tertiary-educated residents than the other municipalities.

*Table 2 – Educational attainment statistics for HMA* (Source: Statistics Finland, 2019b)

| | **Basic education** | | **Secondary education** | | **Tertiary education** | |
|---|---|---|---|---|---|---|
| | Number | *%* | Number | *%* | Number | *%* |
| **Espoo** | 40,330 | *19%* | 86,194 | *41%* | 85,131 | *40%* |
| **Helsinki** | 108,641 | *21%* | 218,807 | *42%* | 197,673 | *38%* |
| **Vantaa** | 46,307 | *26%* | 86,531 | *50%* | 41,942 | *24%* |
| **Kauniainen** | 1,020 | *13%* | 2,915 | *37%* | 3,975 | *50%* |
| **HMA** | 196,298 | *21%* | 394,447 | *43%* | 328,721 | *36%* |

Following the ICT boom in the mid-90's, Vaattovaara & Kortteinen (2003), refer to a growing spatial division of the highly-educated, migrating into specialised neighbourhoods. This selective migration of highly-educated households was concentrated to the West of the region, producing an East-West divide, visible when comparing maps of education attainment in the HMA (Figure 6).

Kauniainen and several areas within Espoo are characterised by clusters of highly educated individuals, with those with at least an undergraduate university degree representing more than half of the residents.

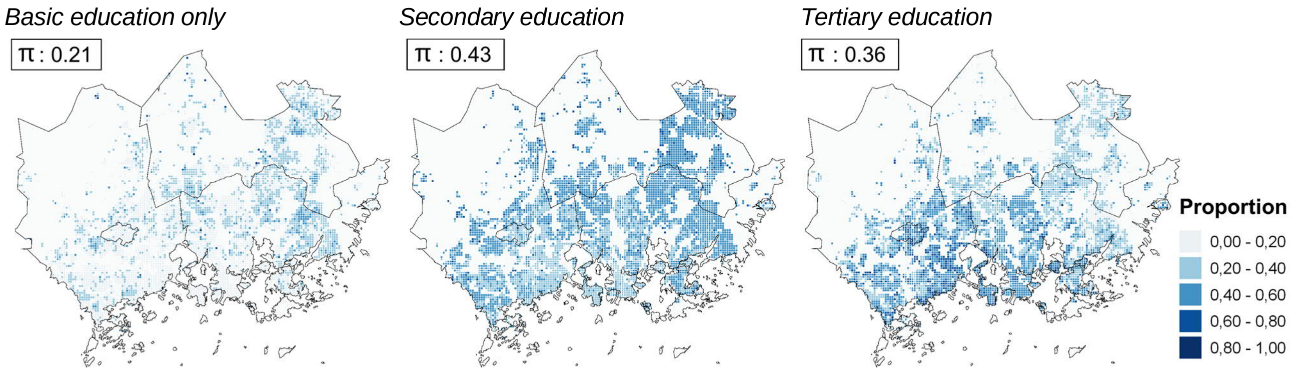


*Figure 6 – Differentiation by educational attainment in the HMA. Local proportions of each group (%) within each 250m grid cell are displayed on the maps, whilst the group's absolute proportion in the HMA region (π) is indicated above each map. Source: Statistics Finland* (2019b)

Within Helsinki, the city centre to the south, as well as the western coastline are similarly overrepresented with highly educated households. Whilst the majority of grid cells in Espoo have at least 25% of university-educated residents, in Helsinki and Vantaa there is more differentiation between areas. Several areas in the east of Helsinki have less than one in four households with a tertiary qualification. Vaattovaara & Kortteinen (2003) posit that one potential reason for the migration of the highly educated to the west may reflect differing preferences on dwelling type for those with higher economic and cultural capital, with Espoo and Kauniainen offering detached, owner-occupier homes**.**

### *Ethno-linguistic segregation*

Finland has two official languages, Finnish and Swedish. The Finnish-speaking population represented 76.6% of the HMA population in 2019 (compared to 87.3% nationally), whilst Swedish-speakers accounted for 5.5% (5.2% nationally) (Table 3) (HSY, 2020; OSF, 2019).

*Table 3 - Mother-tongue by municipality Source: HSY* (2020)

| | **Finnish** | | **Swedish** | | **Other** | |
|---|---|---|---|---|---|---|
| | Number | % | Number | % | Number | % |
| **Espoo** | 217,084 | *75%* | 20,017 | *7%* | 52,634 | *18%* |
| **Helsinki** | 511,997 | *77%* | 36,704 | *6%* | 112,278 | *17%* |
| **Vantaa** | 180,788 | *77%* | 5,567 | *2%* | 47,763 | *20%* |
| **Kauniainen** | 5,778 | *59%* | 3,180 | *32%* | 851 | *9%* |
| **HMA** | 915,647 | *77%* | 65,468 | *5%* | 213,526 | *18%* |

Those residents with a foreign mother-tongue (languages other than Finnish or Swedish) represent a diverse range of nationalities, with former Soviet Union countries, Estonia, Somalia and Iraq being the main countries of origin (City of Helsinki, 2019c). The share of foreign-language speakers in

HMA is 17.9%, well above the national average of 7.5%. Russian, Estonian, Somali and Arabic are the most commonly spoken foreign languages in HMA, with large communities also of English, Chinese, Kurdish, Persian, Albanian and Vietnamese, language groups (OSF, 2020). A full breakdown of the major language groups in the HMA is provided in Appendix 1.

Immigration has historically been low in Finland, and only really began to increase from the 1990s. In 2019, almost half of all residents with a foreign background in Finland lived in the Helsinki Metropolitan Area (City of Helsinki, 2019d). Around 210,000 (18%) HMA residents were born outside Finland (OSF, 2019). Whilst this share of the population with a foreign background is still significantly lower than that of the capital cities of its Nordic neighbours (in Oslo and Stockholm, the share of residents with a foreign background is closer to 25%), the number, and the proportion, of residents of foreign background is growing rapidly (Nordregio, 2017). By 2035, the number of foreign language speakers in the capital region is expected to more than double from current levels to around 400,000 (28% of the total population), of which 23% would be born in Finland (City of Helsinki, 2019b). The differentiation between HMA municipalities will continue, with Helsinki set to increase its share of foreign-language speakers to 26%, Espoo to 30% and Vantaa to 34% (City of Helsinki, 2019a). The largest foreign language groups in Helsinki and Espoo are expected to be Middle Eastern, North African and Far Eastern languages, whilst Russian and other former Soviet Union languages will be most prevalent in Vantaa. (City of Helsinki, 2019a)

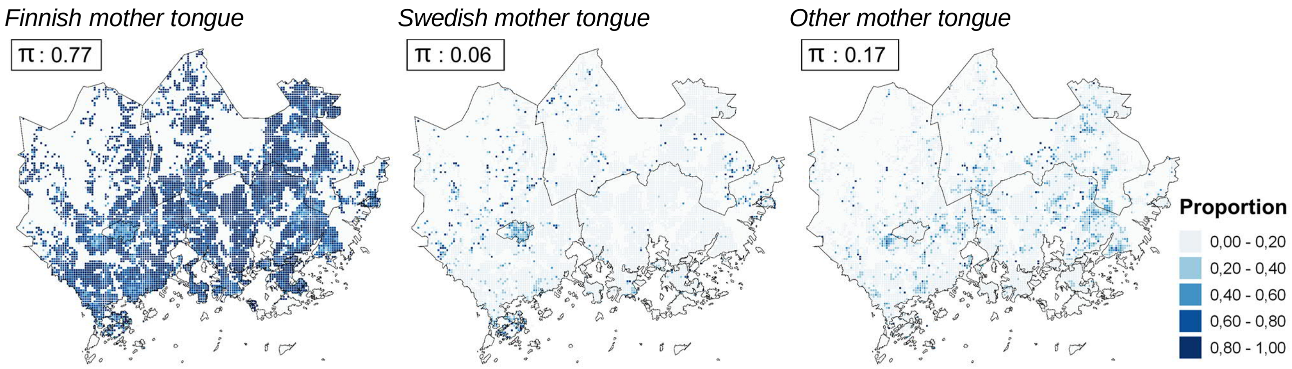


*Figure 7 – Distribution of language groups in the HMA. Local proportions of each group (%) within each 250m grid cell are displayed on the maps, whilst the group's absolute proportion in the HMA region ($\pi$) is indicated above each map. Source: HSY* (2020)

Large proportions of foreign-language speakers are visible in the east of Helsinki and Vantaa, and along the northern railway line (Figure 7). In Espoo, foreign-language speakers are overrepresented in Espoo central, and the university area of Otaniemi. The small municipality of Kauniainen is home to a large Swedish-speaking community.

Historically, residents with a language other than Finnish or Swedish as their first language were overrepresented in socially rented housing estates (Vaattovaara et al., 2018), and socio-spatial mixing policies thus held a role in preventing segregation. Since 2006, this reliance by on state-subsidised housing has however been decreasing, with private rental becoming more common for foreign-born residents (City of Helsinki 2020c).

Whilst Kauppinen & van Ham (2019) identify intraregional migration by Finnish-origin residents as the main process driving ethnic segregation in the Helsinki region, this cannot be conflated with language groups. Kauppinen & van Ham (2019) consider residents having at least one foreign-born parent as being of immigrant-origin, however it is possible that Finnish or Swedish may be their first language. Additionally, Huynh (2019) reports that home-leaving second-generation immigrants in Finnish cities display similar mobility patterns to their native-born peers.

# 5 Data

## 5.1 Data description

The availability of population information with rather high geographic precision makes the HMA a good case study for an ABM. In order to create the agents and model environment representing the HMA, two different types of geospatial register data were combined. Statistics Finland maintains a database with population information, aggregated to a 250 metre x 250 metre grid structure (Statistics Finland, 2019b). Variables for income and education were extracted from the Statistics Finland 2019 grid database; however, there are no variables that can be used to approximate language groups or ethnicity. For this, the Helsinki Region Environmental Services Authority (HSY) maintains a building-based register for all municipalities in the HMA. The HSY SeutuData dataset includes a variable for mother tongue, recording residents as either Finnish-speaking, Swedish-speaking or ‘other languages’ (non-Finnish or Swedish speakers). Whilst mother tongue is an imprecise representation of ethnicity, sharing a common language can be a key factor in social homophily. The variables extracted from these two databases is presented in Table 4 and described in detail below. Both datasets are in geospatial vector format; however, the point-based HSY register data was appended to the same 250-metre grid used by Statistics Finland to allow consistent treatment within the agent-based model. Any points that did not correctly intersect the grid were removed from analysis.

## 5.2 Variables

To maintain consistency in the simulations, all variables have been subdivided into three groups (e.g. high-income, medium-income, low-income). The absolute proportions of each group in the model is representative of the actual ratios in the HMA.

*Table 4 - List of geospatial variables and attributes used for modelling*

| Variables | Attributes | Source |
|---|---|---|
| **Language** | Finnish mother tongue<br>Swedish mother tongue<br>Other mother tongue (not Finnish or Swedish) | HSY SeutuData 2019 |
| **Socio-economic** | Low-income household (income deciles 1-2)<br>Medium-income household (income deciles 3-8)<br>High-income Household (income deciles 9-10) | Statistics Finland Grid Database, 2019 |
| **Cultural Capital** | Basic Education only<br>Matriculation Exam / Vocational Training<br>University Education (Bachelor level or higher) | Statistics Finland Grid Database, 2019 |

### *Household income groups*

Statistics Finland reports for each grid cell, the number of households within three pre-designated income groups; high, medium and low-income households (Statistics Finland, 2019a). The income categories are formed on deciles, created by listing all households in order of their equivalent disposable monetary income and then creating 10 groups with equal number of households. The equivalent disposable monetary income is calculated by dividing disposable household income by the number of consumption units in the household, in order to account for shared consumption benefits of larger households. The OECD's adjusted consumption unit scale is used, whereby the first adult of the household receives the weight 1, other household members over the age of 13 receive the weight of 0.5 and any children under 13 receive the weight of 0.3. The low-income group is comprised of households who fall into the lowest two deciles of equivalent disposable monetary income, earning at most 16,979 euro per year. The middle-income group consists of those households falling into deciles 3-8, earning between 16,980 and 35,297 euro per year. Households in deciles 9 and 10 are categorised as high-income households, and earn in excess of 35,297 per year. Due to reporting lags, the household income values in the 2019 dataset are accurate for the year ended 31 December 2017. For data protection, income categories for grid cells with less than 10 households have been made confidential and not reported.

Whilst income can be considered a good indicator of a household’s financial position, it is not a perfect indicator of wealth, and in particular accumulated intergenerational housing wealth, which may reproduce spatial patterns (van Ham et al., 2014)

### *Educational attainment*

Educational attainment, along with occupation and other composite measures, has been used in previous studies considering the role of cultural capital in homophilious residential sorting (e.g. Boterman et al., 2020; Bourdieu, 2005; van Gent et al., 2019). In Finland, basic education is compulsory until 15 years of age (year 9). Of those students electing to continue with their education in Finland, approximately half continue to vocational upper secondary education (VET) whilst the other half  complete general upper secondary education, leading to a matriculation examination (Finnish National Agency for Education, 2020). Educational attainment has been re-classified into three groups of agents as follows; individuals having completed basic level education only, individuals having completed upper secondary education only (completed a matriculation examination or vocational diploma), and finally, individuals who have completed tertiary education (bachelor degree level or higher). Educational attainment, as defined in the Statistics Finland dataset only considers those residents old enough to have completed upper secondary education, thus only residents aged 18 years or over are included. For grid cells with less than 10 residents over 18 years, educational data has been made confidential and not reported. The data is accurate as at 31 December 2018.

### ***Language groups***

An individual’s mother tongue has been used in this study in order to examine language groups as a potential source of social homophily. Ethno-linguistic homophilious preferences have been explored in other studies, including the bilingual context of Quebec (Perez et al., 2019).

Finland is a bilingual country, with Finnish and Swedish as the official languages. The HSY building register contains information on the mother tongue of residents permanently residing in each building as at 31 December 2019. The total number of residents who have Swedish or ‘other languages’, defined as Non-Finnish or Swedish speakers, are recorded as separate variables in the register. The number of Finnish-speaking individuals was calculated by deducting these two groups from the total number of residents in each building. As discussed previously, the non-Finnish or Swedish speaking residents are a highly heterogeneous group in terms of language, country of origin etc. An individual may additionally have Finnish or Swedish as their mother tongue, however be a second-generation immigrant with attachments to other linguistic communities for example. Despite this limitation, language group has been included in this study as it also serves to demonstrate another potential factor in the development of segregation, that of group size. The Finnish-speaking residents represent over three-quarters of all residents in the area, whilst the Swedish-speaking population represents only 5% and all other language groups combined represent 17%. Compared to the income and education groups, language groups represent a clear majority/minority dynamic which is often underexplored in Schelling-type models which tend to use equal number of agents in each group. The relative size of each group can influence the potential for each group to achieving their prescribed preference thresholds in unexpected ways.

In order to maintain a consistent input format for the model, the HSY building-based point data was aggregated to the same 250m x 250m grid used by Statistics Finland from which the other two variables are derived. Several points did not intersect the grid directly, and were removed from analysis. This resulted in approximately 3% of the population being excluded.

### ***Number of agents***

Whilst it is possible to generate the entire population as agents, the computation costs of doing so are significant, particularly when multiple simulations are conducted for each scenario. As all simulations in this study were conducted on a standard personal laptop computer, the number of agents has been reduced to a representative population of 100,000 agents (Table 5). The reductions were calculated at each grid cell to ensure that the local proportions remained consistent with the actual population. As these calculations were made at the grid cell level, rounding on thousands of cells produces a slight deviation in the final number of agents for each sorting variable. Very few segregation ABMs have run more than 100,000 agents. Those few which have run a larger number of agents generally activate only a subset of the total agents each cycle.

*Table 5 - Total agents created by sorting variable, compared to HMA population. Deviances from 100,000 agents are due to rounding on thousands of grid cells.*

| | HMA Population | | Agents Created | |
|---|---|---|---|---|
| | Number | % | Number | % |
| **Education groups** | | | | |
| Basic education | 196,298 | *21%* | 21,313 | *21%* |
| Secondary education | 394,447 | *43%* | 42,967 | *43%* |
| Tertiary education | 328,721 | *36%* | 35,791 | *36%* |
| *Individuals aged over 18** | *919,466* | | *100,071* | |
| **Income groups** | | | | |
| Low-income | 108,283 | *20%* | 19,491 | *19%* |
| Medium-income | 296,129 | *53%* | 53,429 | *53%* |
| High-income | 150,516 | *27%* | 27,190 | *27%* |
| *Total households** | *554,928* | | *100,110* | |
| **Language groups** | | | | |
| Finnish | 915,647 | *77%* | 77,222 | *78%* |
| Swedish | 65,468 | *5%* | 5,281 | *5%* |
| Other | 213,526 | *18%* | 16,916 | *17%* |
| *Total individuals** | *1,194,641* | | *99,419* | |

** Income groups are calculated at a household level, whilst education and language variables are individual measures. Educational attainment is recorded for all individuals over 18 years. Mother tongue is recorded for all individuals in the HMA.*

# 6 Methodology

The main method of analysis is a geospatial agent-based model which incorporates representations of selected features of the urban structure and population of the HMA. Within the model, various simulation experiments are conducted by modifying specific parameters concerning agent's residential mobility behaviour and testing for impact on the system. The resultant patterns of segregation are compared to a baseline situation, and the process tracked through various measures.

The agent-based model used in this study was built using the GAMA (GIS Agent-based Modelling Architecture) simulation platform. First developed by the Vietnamese-French research team MSI (Taillandier et al., 2019), GAMA is an open-source project which provides a complete modelling and simulation development environment. GAMA accepts vector data as an input and treats agents as spatialised objects as in GIS (Taillandier et al., 2019). This allowed the agents and model environment features to be created from geospatial vector data for the Helsinki region. Some basic Schelling-type models already exist within the model library. The GAML (Gama Modelling Language) scripts from these models were adapted in the process of building this model.

## 6.1 Data processing

All data processing was completed using the R software package and programming language. For each sorting variable, two shapefiles were created, one for the agents, and another for the environment. The agent shapefile records the starting location of the agents as point geometries, representing the centroid of the grid cell in which they reside. As well as a location, agents have a Group ID and an Agent ID, which can be tracked over time throughout the simulations. The environment shapefile contains polygon geometries for all occupied grid cells, and a corresponding Grid ID. The total number of agents located within each grid cell also saved to the environment shapefile.

## 6.2 Simulation parameters

Three significant parameters affecting the operation of the model in this study are the neighbourhood definition, preference thresholds for like-neighbours for each agent group, and the vacancy rate. All of these parameters have been made modifiable parameters in the graphical user interface (GUI) version of the ABM. For the purpose of this study, neighbourhood definition and vacancy rate have been held constant. Only the agent preference for like-neighbours, as a percentage of the neighbourhood, will be modified to test for its effect.

### *Neighbourhood definition*

Schelling-type agent-based models consider that an agent’s likelihood of moving residence is affected by the agent’s assessment of the composition of its ‘neighbourhood’. The neighbourhood definition is very important in agent-based modelling, as in any spatial analysis. The network structure of agents and neighbourhoods can have a large impact on the behaviour of the system, with different networks potentially inducing very different system-wide behaviour. Neighbourhood definitions are often simplified to binary ‘in/out’ administrative subdivisions that are deemed to be commonly experienced and understood by residents. From a segregation perspective, bounded neighbourhood definitions of this nature are particularly relevant if one wishes to study segregation with relation to school catchment zones, which have clearly defined boundaries for example. However, what constitutes a neighbourhood, and how residents may perceive a neighbourhood when deciding to relocate or remain, can be considerably more complex and subjective. A definition of what constitutes a neighbourhood may be highly personal, concerning a resident’s activity spaces or actual visited locations. It is equally possible to interpret neighbourhoods in many different ways within the modelling environment.

As per Schelling (1971), the neighbourhood definition adopted for this study considers a neighbourhood to be personalised from the agent’s perspective based on spatial proximity, irrespective of all administrative boundaries. ‘Neighbourhood distance’ is coded as a modifiable parameter that calculates a Euclidean buffer from the agent’s location. A neighbourhood distance buffer of 500m has been held as a constant for all experiments, sufficiently large to include at least the adjacent grid cells. An agent is deemed to have perfect knowledge about the composition of the population residing within their current neighbourhood.

### *Vacancy*

In order for agents to migrate to new locations, there must be vacant ‘dwellings’ for them to move into. An 8% vacancy rate has been prescribed in the model, consistent with the empirical HMA vacancy rate in 2019  for empty and not regularly inhabited homes (Statistics Finland, 2019c). The vacancy rate is specified by increasing the capacity of each grid cell (the number of agents which can be housed within) by 8% above the empirically informed initial distribution. I.e. if a grid cell contains 100 households, 8 additional places are allocated in this cell. As the number of agents and initial distribution varies based on the grouping variable being testing, the number and distribution of vacancies also varies under each variable. Additionally, due to rounding on thousands of cells, the effective vacancy rate of is slightly less than the 8% prescribed.

***Preference threshold***

A number of assumptions have been made concerning co-group preferences. Firstly, preferences have been assumed to be consistent within each group**.** That is, all agents of any given group share the same preference value for co-group agents within their neighbourhood. Heterogeneity of preferences has not been modelled (see Xie & Zhou, 2012). Secondly, whilst agents are concerned about the *ratio* of agents within the neighbourhood, they are indifferent to the *configuration* of agents within their neighbourhood.

The intensity of co-group preferences is the main parameter of analysis in this study. Three different scenarios were run, each with different preference levels. The preference thresholds were chosen with respect to previous studies, the number of groups in the simulation, the relative proportions of groups in the HMA. All other variables are held constant.

The scenarios are as follows:

- ***Scenario A***: 30% preference for co-group neighbours for all groups.
  Agents are willing to live in mixed neighbourhoods as long as 30% of their neighbours are of the same group as themselves. With three groups in each simulation, if an agent lives in a neighbourhood with less than 30% of its own group, it will necessarily be a minority within the neighbourhood.
- ***Scenario B***: 50% preference for co-group neighbours for all groups.
  Under this simulation, agents are happy to live in mixed neighbourhoods, however require that they are the majority group in the neighbourhood (at least 50%).
- ***Scenario C:*** 70% preference for one group, and 30% preference for both other groups.
  This scenario aims to test asymmetric preference thresholds, and the potential for one group with a higher same-group preference to influence the overall segregation levels in a city. Based upon the literature suggesting the increased importance of these groups in shaping segregation processes in some cities, the following agent groups were assigned the higher preference; high-income agents (high economic capital), agents with tertiary education (high cultural capital), and agents with Finnish as a mother tongue (large majority group).

The different scenarios are summarised in Table 6.

*Table 6 – Preference thresholds for co-group neighbours under each simulation scenario*

| | **Percentage of neighbours** | | |
|---|---|---|---|
| | Scenario A | Scenario B | Scenario C |
| **Education groups** | | | |
| Basic education | 30% | 50% | 30% |
| Secondary education | 30% | 50% | 30% |
| Tertiary education | 30% | 50% | 70% |
| **Income groups** | | | |
| Low income | 30% | 50% | 30% |
| Medium income | 30% | 50% | 30% |
| High income | 30% | 50% | 70% |
| **Language groups** | | | |
| Finnish | 30% | 50% | 70% |
| Swedish | 30% | 50% | 30% |
| *Other* | 30% | 50% | 30% |

***Rules of interaction***

A set of rules is prescribed for the interaction between agents and between agents and the environment. The dominant rule governing agent actions is summarised as follows:

- Each cycle agents will move, when possible, from areas where the population composition of their neighbourhood does not meet the prescribed preference threshold for similar (co-group) neighbours to a new location.
- Agents have perfect information about the composition of their current neighbourhood, but no information about areas outside their neighbourhood. New locations are chosen at random, and the agent will assess the neighbourhood composition only on the next cycle.
- If a chosen location does not have remaining capacity, the agent chooses a new location.

## 6.3 Simulation process

*Initialisation of the model*

When a simulation is initiated (Figure 8), the initial state of the system is constructed from two separate shapefiles containing the geospatial data representing the agents and the environment:

- Spaces representing the modelling environment are created using the polygon geometries of the grid cells from the environment shapefile. As well as importing the grid_ID, the initial number of starting agents within each cell is multiplied by a vacancy rate to set the maximum capacity of each cell.

- The agents are created using the point geometries from the agent shapefile. The group_ID and agent_ID are read from the shapefile. The agents' locations are recorded by intersecting the agents' geometries with the environment polygons. Each agent is specified to be located at the centroid of the grid cell, so the neighbourhood definition of each agent within a grid cell is shared. As each agent is allocated to a space, the capacity of that space reduces by one.
- Global attributes are stored for each agent and space. These attributes are the key parameters which may be modified each simulation. The prescribed neighbourhood distance, and the specified preference threshold corresponding to each individual agent's group_ID is stored. Each of the spaces take the prescribed vacancy rate and multiply it by the number of agents they contain to calculate their maximum capacity.
- All calculated fields are updated. Agents calculate their total neighbours and the number of neighbours which are of the same group_ID as themselves, using the neighbourhood distance parameter to determine the extent of their neighbourhood. The environment grid cells record the number of agents of each group located within their perimeter, and the number of agents from each group who are meeting their preference threshold.

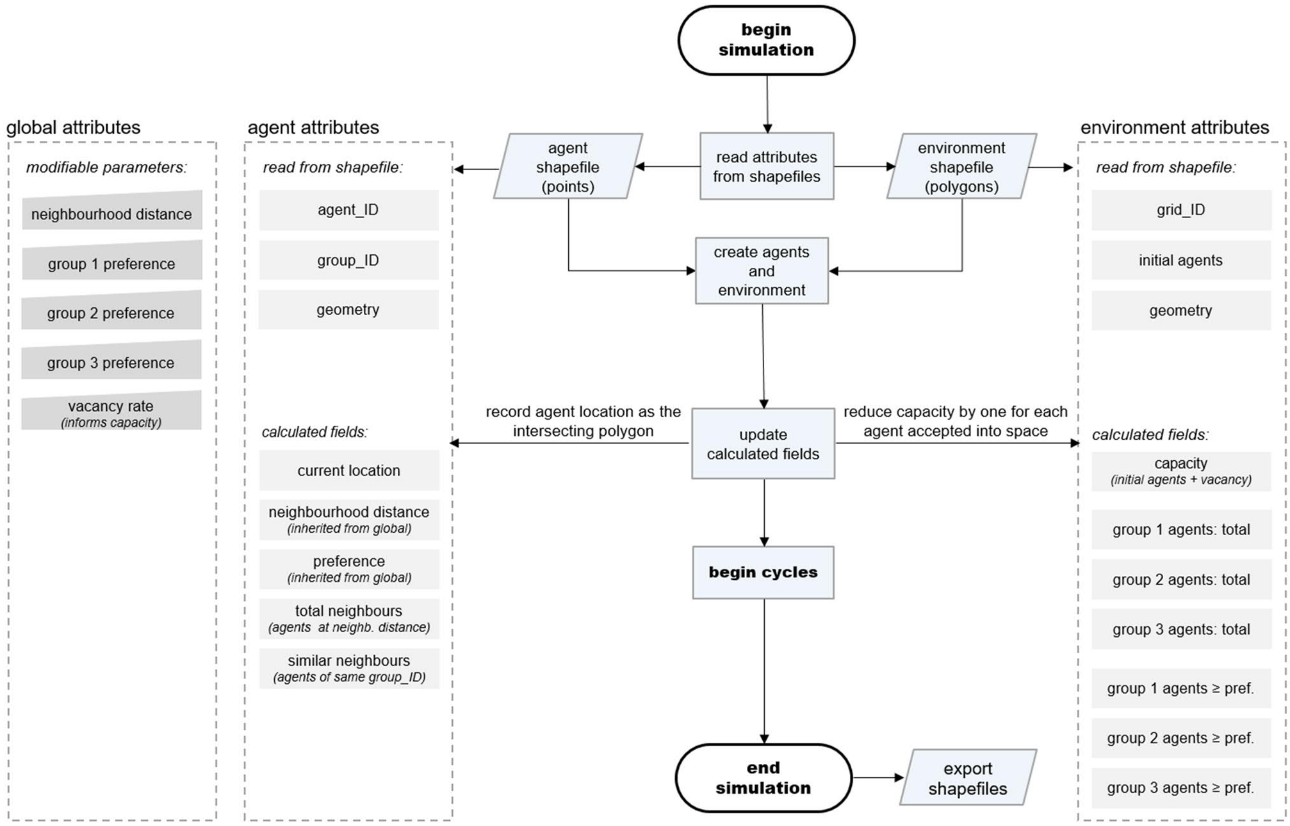


*Figure 8 – Model initialisation and sequence diagram. Global attributes have been made modifiable parameters in the GUI experiment, with their values inherited by the agents and environment. For the simulations in this paper, only the preference values for each groups are adjusted according to the scenarios.*

*Simulation cycles*

Once initialisation of the model is complete, the first cycle of the simulation begins (Figure 9):

- During each cycle, each agent has the opportunity to decide whether they remain in their current location, or whether they migrate to a new location.
    - An agent is activated in the order in which they were created.
    - The neighbourhood of the agent is calculated using the specified Euclidean distance buffer (500m in this study), and a list of all the other agents residing within the neighbourhood is compiled.
    - The agent assesses the neighbourhood composition and compares this to its specified preference threshold. If the preference threshold is not achieved, the agent will vacate its current location and migrate to a new location.
    - If migrating, the agent chooses a new vacant cell at random. If the space has capacity, the agent is accepted, otherwise another space is chosen at random until a space with sufficient capacity is found. The capacity of the space of departure increases by 1, and the new space decreases its capacity by one.
    - The next agent is then activated, and this process continues until all agents have been activated.
- At the end of each cycle:
    - Statistics are updated if running the GUI experiment (statistics, maps, and graphs).
    - A shapefile with the current state of the simulation is exported for later analysis.
    - A new cycle begins.

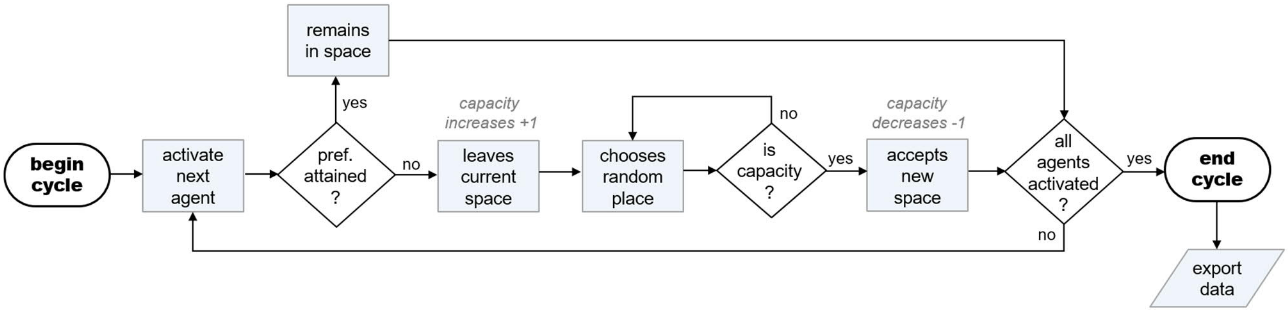


*Figure 9 – Process diagram for each simulation cycle.*

*End of simulation*

The simulation ends once the specified number of cycles has been reached, or the operator stops the model. The final state of the simulation is exported as a shapefile.

***Number of cycles***

All simulations in this study are terminated after 20 cycles have elapsed. Cycles do not correspond directly with any specific time frames in the real world, however, these can be considered as yearly intervals. In practice, this would mean all households consider once every year whether they would like to move a new location or stay in their existing residence for another year.

## 6.4 Measuring segregation

In order to assess the impact of the agents' behaviour on segregation patterns within the study area, several indices have been calculated at the end of each cycle. As segregation can have many dimensions, there is a need for multiple measures to sufficiently describe these different aspects of segregation (Massey & Denton, 1988; Reardon & O'Sullivan, 2004). In an extensive literature review and analysis of 20 different segregation indices used in segregation research, Massey & Denton (1988) outline 5 conceptually distinct dimensions of segregation. These dimensions are evenness, exposure, clustering, centralisation and concentration. Each of these dimensions entail different social and behavioural implications. Reardon and O'Sullivan (2004) propose that if the spatial relationships of the areal units is taken into consideration in the calculations, Massey & Denton's five dimensions can be reclassified into just two measures; spatial evenness (of which clustering is the opposite) and spatial exposure (of which isolation is the opposite) (Figure 10).

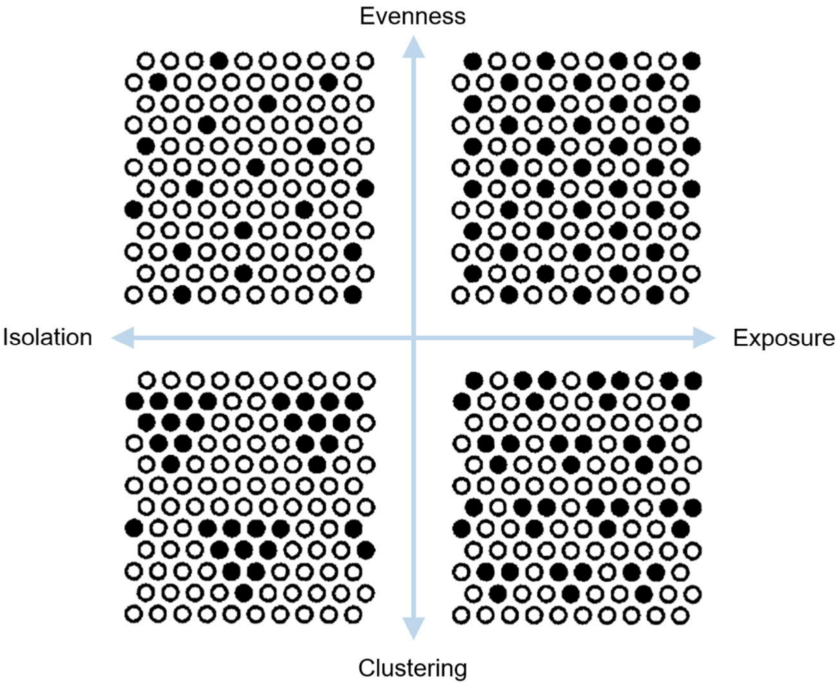


*Figure 10 – Dimensions of spatial segregation* (from Reardon & O'Sullivan, 2004)

*Spatial evenness* refers to the distribution of different social groups across areal units in a metropolitan area. When different population groups are evenly distributed throughout the city, and all areal units contain population distributions that reflect the city as a whole, segregation is low. Conversely, if certain groups are overrepresented or underrepresented in some areas, or in the extreme, share no common areal units, segregation is high. An uneven distribution can result in spatial clustering, with groups living in more segregated contiguous enclaves.

*Spatial exposure* considers opportunities for contact between groups, and the extent to which members of different groups have a common residence (i.e. share an areal unit). The opposite measure of exposure is isolation, calculated as the likelihood that a member of a certain group comes into contact with another member of the same group. Low exposure to other groups, or high isolation, indicates that a group is segregated. Isolation, as with exposure, depends in part on the overall composition of the metropolitan region. For this reason, the overall size of the group's proportion at the global level will influence the calculation, limiting comparison of isolation index values between groups.  Reardon & O'Sullivan (2004) consider that centralisation (clustering towards urban core) and concentration (the physical amount space which different groups occupy) become subcategories of the spatial evenness/clustering dimension under this interpretation.

### *Calculation of spatial segregation measures*

In line with Reardon and O'Sullivan (2004) this study has adopted measures for spatial evenness and spatial isolation. Specifically, spatial versions of the Information Theory Index ($H$), a measure of evenness, and the Isolation index ($Q_m$) are calculated for each group. Both of these indices are traditionally aspatial, in that they consider each spatial unit as being an independent entity with no relation to others, ignoring the spatial arrangement of the metropolitan region (see Reardon & O'Sullivan, 2004). Despite the fact that population data is regularly reported in spatial units such as census zones, the relative locations of these units to one another is usually disregarded in many studies of segregation (O'Sullivan, 2009). As such, these aspatial measures suffer from known issues such as 'checkerboard problem' (See Reardon & O'Sullivan, 2004; White, 1983)

Spatial elements are factored into the calculations of $H$ and $Q$ presented in this paper using kernel density estimation (KDE). In KDE, a smoothly curved surface is fitted over each data point (grid centroid), based on a kernel density function which defines the shape and size of the kernel (Figure 11a). The volume under the surface is equal to the value of the relevant population field for each point. The surface value of the kernel is highest at the location of the point and diminishes with increasing distance from the point (a "distance-decay" effect) (Reardon & O'Sullivan, 2004). The surface value reaches zero at the specified bandwidth distance from the point. By using a kernel function with a bandwidth sufficiently large to cross the boundary of the areal units, nearby populations are considered to be closely related to the population of these areal units (ESRI, 2020). The local population intensity of each point is calculated by adding together the values of all the intersecting kernel surfaces at the point (Figure 11b).

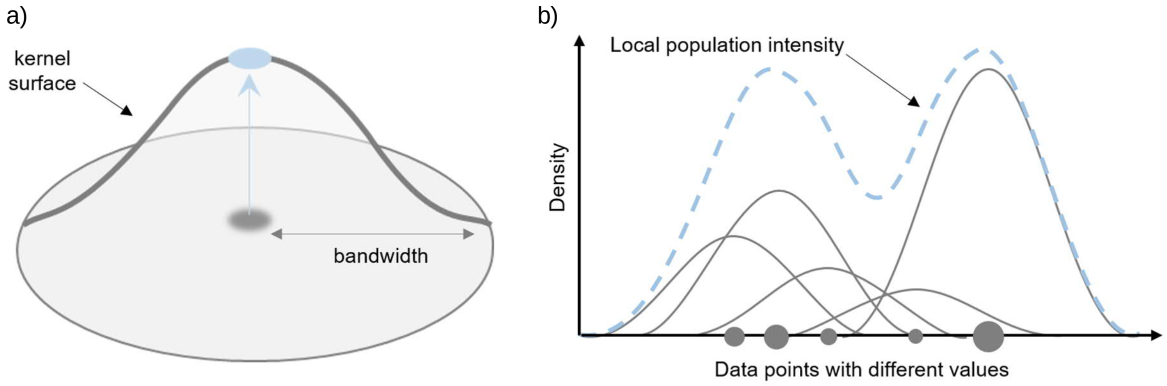


*Figure 11- Conceptual illustration of kernel density estimation for a single population data point (a) and calculation of local population intensity from the accumulation of intersecting kernels (b). The shape of the kernel surface is determined by the choice of kernel function. In this study, a quartic kernel with 500m bandwidth has been use in all spatial measures. Adapted from Lee et al. (2020) and Odland (2020).*

After testing several different kernel types and bandwidths, a quartic (bi-variate) kernel function with a 500m bandwidth was chosen. For any grid cell, this considers several surrounding cells, with locations in closer proximity contributing more to the calculation of the 'local population intensity' than distant locations. Spatial measures using KDE are indicated in the paper with a superimposed tilde (~) above the symbol.

Whilst the global indices for evenness and isolation express the degree to which the city is segregated overall, as a single entity, they do not provide any detail of local differentiation between smaller subdivisions of the urban region. For this reason, in addition to the global indices, local decompositions of the segregation measures have also been calculated (Feitosa et al., 2007; Iceland, 2004; White, 1986). These local measures display how much each locality contributes to the global segregation measure of a city. By mapping the local measures, differentiation between different areas can be visualised, and the most critical areas identified.

*Spatial information theory index (H)*

The information theory index (*H*), also known as the 'multi-group entropy index', or the multigroup version of Theil's H (Theil, 1972), is a measure of evenness, describing how evenly the groups are distributed through the entire city. *H* measures the average difference between a unit's group proportions and that of the system as a whole (Iceland, 2004; Theil, 1972). *H* has a value between 0 and 1 with a lower value representing a more integrated, and less segregated metropolitan region. The value will be 0 if all areas have the same composition as the entire metropolitan area and will be equal to 1 if all areas contain one group only, indicating maximum segregation. All indices have their advantages and disadvantages, however for this study, *H* is deemed more suitable than other comparable measures such as the diversity index (*D*) which is also frequently used. When taken at a local level, *D* can reveal the areal units which are more/less similar than the population composition of the whole, however *H* provides a more accurate definition. The local values of *H* range from positive to negative, providing details about which local areas have higher or lower diversity than the global index (Barros & Feitosa, 2018). This provides detail that is hidden in the global index and permits a more informative visualisation when mapping the results by areal units.

Whilst *H* is more sensitive to the grouping systems (i.e. number of groups included in the analysis) than *D* (Barros & Feitosa, 2018), the classing system does not change between simulations in this study, and therefore allows comparison with the baseline measure without concern.

The local entropy measures (diversity scores) from which *H* is calculated are influenced by the relative size of the various groups in a metropolitan area, however, this is adjusted for in the formulation of *H*, which reports how evenly groups are distributed across metropolitan area neighbourhoods, regardless of the size of each of the groups (Iceland, 2004). *H*, like most other measures, is however, affected by the scale of the areal units from which it is calculated. A coarser subdivision usually produces a lower index value than a finer subdivision.

The spatial information theory index ($\tilde{H}$) is the weighted average deviation of each localities entropy from the metropolitan-wide entropy, expressed as a fraction of the metropolitan area's total entropy:

$$\tilde{\mathrm{H}} = \sum_{j=1}^{J} \frac{t_j\,(E - \tilde{\mathrm{E}}_j)}{T\,E}$$

Where $E$ is the entropy in the metropolitan region $R$, and $\tilde{\mathrm{E}}j$ is the entropy in locality $j$, defined as follows:

$$E = -\sum_{m=1}^{M} \pi_m \ln(\pi_m)$$

$$\tilde{E}j = -\sum_{m=1}^{M} \tilde{\pi}_{jm} \ln(\tilde{\pi}_{jm})$$

Where:

- $T$ is the total population of the metropolitan region $R$
- $t_j$ is the total population of tract $j$
- $M$ is the number of groups in the metropolitan region $R$
- $J$ is the total number of areal units in the metropolitan region $R$
- $\pi_m$ is the proportion of group m in the metropolitan region $R$
- $\tilde{\pi}_{jm}$ is the proportion of group m in locality $j$ (calculated using KDE)
- *ln* is the natural logarithm function

Following Iceland (2004), when the proportion of group m in is 0, then the log is set to 0, ensuring that the absence of a particular group does not increase the total score of *H*.

*Spatial isolation index ($\tilde{Q}_m$)*

Isolation and exposure measures attempt to quantify the *experience* of segregation, that is, who are residents likely to physically encounter and come into contact with, in their neighbourhood of residence (Massey & Denton, 1988). As the relative size of each group is taken into consideration in the calculation of the isolation index a group which represents a larger proportion of the region's population will have a higher isolation index than a smaller group. As the group proportions are held constant in this study, comparison is possible against the baseline measure for each group, however, the results cannot be compared between groups without acknowledging this property. The isolation index $\tilde{Q}_m$ is computed as the average of the proportion of group *m* in each areal unit, weighted by the proportion of group *m* in the overall population *R*. The isolation index is a value between 0 and 1, where 1 signifies maximum isolation. This may be interpreted as the probability that a randomly chosen member of group *m* shares the same areal unit with another member of the same group.

The global Isolation index $\tilde{Q}_m$ is:

$$\tilde{Q}_m = \sum_{j=1}^{J} \left(\frac{N_{jm}}{N_m}\right)\left(\frac{\tilde{L}_{jm}}{\tilde{L}_j}\right)$$

Where:

- $J$ is the total number of area units in the metropolitan region $R$
- $N_{jm}$ is the population of group m in in the area unit $j$
- $N_m$ is the population of group m in the metropolitan region $R$
- $\tilde{\text{L}}_{\text{jm}}$ is the local population intensity of group m in locality $j$ (calculated using KDE)
- $\tilde{\text{L}}_j$ is the local population intensity of locality $j$ (calculated using KDE)

and the local index $\tilde{Q}_{jm}$ is:

$$\tilde{Q}_{jm} = \left(\frac{N_{jm}}{N_m}\right)\left(\frac{\tilde{\text{L}}_{jm}}{\tilde{\text{L}}_{\text{j}}}\right)$$

All segregation indices were calculated using shapefiles exported directly from the GAMA model using the SEGREG Plugin in QGIS (Sousa, 2020). The results were then aggregated for the 10 simulation runs to present a single index value for each scenario and for each sorting variable.

*Additional Calculations from GAMA*

In addition to the traditional segregation measures, data concerning the state of the agents is also exported from the model each cycle. Specifically, the proportion of agents which are attaining their preferred neighbourhood composition has been reported for each group. This statistic provides additional information about the processes of migration of the agents, and can help explain why some simulation result in higher segregation than others.

## 6.5 Model validation

Validation of the results of agent-based models is difficult, especially when used for exploratory simulations such as in this thesis. Validation of the model in this case came through an iterative process of incremental adaptation when writing the code, in order to verify that the agent’s behaviour and interactions were operating as specified before adding in additional parameters. Rather than relying on error codes for validation, the model was checked for anomalies at every stage as more parameters and complexity were built in. Data on the initial state of the model was exported from GAMA, and compared to the empirical data to ensure no errors occurred during importation.  Due to the stochastic nature of the models, particularly the random choice of new location when migrating, it is possible, and indeed likely that there will be different outcomes from any number of simulations. To account for potential fluctuations in the processes and outcomes, ten separate simulations were conducted with identical parameters for each of the scenarios for each grouping variable. Presenting the results as box and whisker plots allows visual identification of the variances between different runs. All segregation indices and measurements are calculated on the aggregate mean results of the 10 simulations

## 6.6 Visualising segregation

Most segregation ABMs use simple checkerboard environments, with one agent occupying one space. This makes visualisation simple and permits visual assessment of any clustering. Indeed, some studies rely entirely on visual assessment to assess their modelling results. As this model considers a realistic urban structure, with more than two groups and with many agents per space, visualising the results in an easily interpretable way becomes more difficult. A number of visualisation option were tested throughout the development of this thesis, and the reasons for dismissing these options are outlined briefly here for reference.

Presenting the location of all agents for the three population groups on one map requires that only one group is visualised per grid cell. A multi-group map was achieved by highlighting the majority group in each grid cell. The absolute proportion held by this majority group was portrayed through an increasing colour gradient (Figure 12).

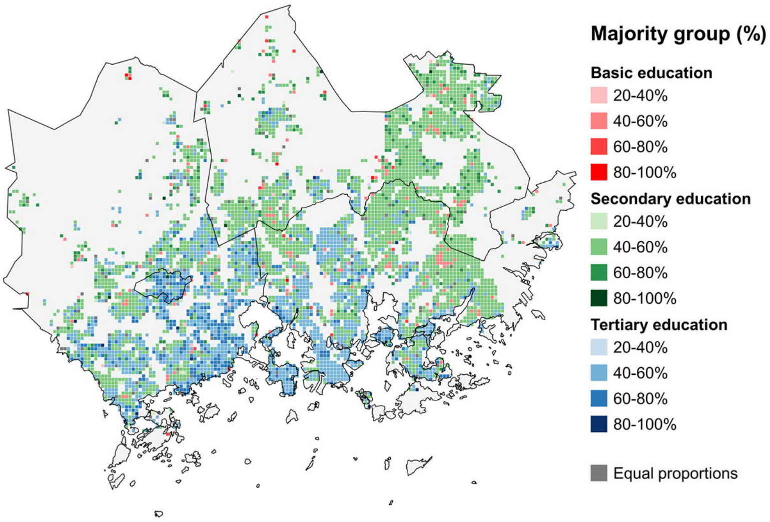


*Figure 12 - Multi-group map indicating the majority group in each grid cell for education groups at baseline (cycle =0)*

Whilst providing an attractive map, the simple majority/non-majority rule demands the loss of too much detail to be acceptable. With three groups, over 65% of the grid cells population may be hidden if one group obtains a 34% share. Particularly when a groups’ absolute size in the region is smaller, this was deemed an unacceptable loss of detail. Additionally, density is not reflected under this classification. Creating additional subcategories and adjusting transparencies to reflect the population density and urban structure was tested, however the addition of further classifications only served decrease legibility.

Considering the need for individual maps for each population group, kernel density heat maps were explored to reflect both the location and the density of the each population group (Figure 13). These maps are easy to interpret without losing important details, presenting a min-max scale which is reclassified for each group. However, KDE already forms the basis of the employed spatial isolation measures, and visualising the local $\tilde{Q}_m$ values achieves largely the same visual outcome.

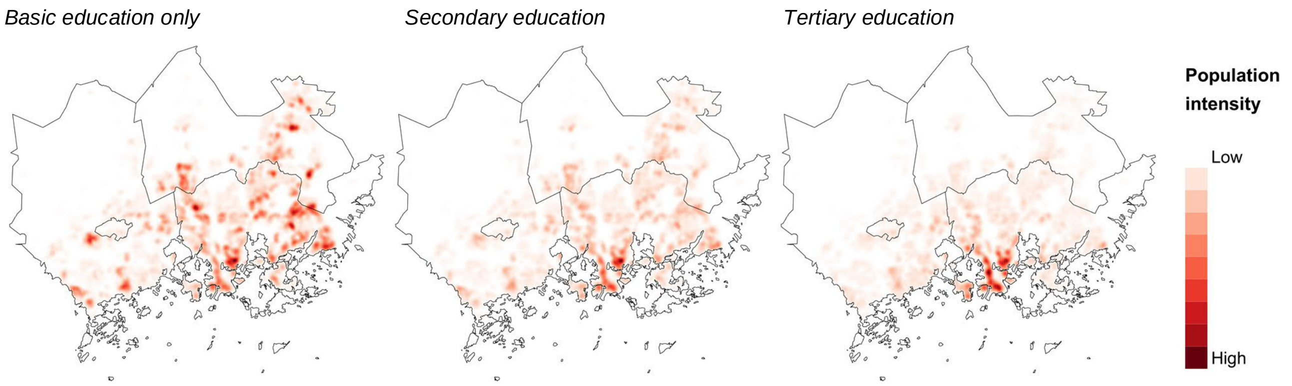


*Figure 13 - Kernel density heat map of education groups at baseline (cycle =0). The intensity of colours are relative to each group, so clusters of minority groups will appear as intense as majority groups, even though their absolute number is much lower.*

This visualisation of local $\tilde{Q}_m$ values was ultimately chosen as it allows the identification of geographic areas that are contributing more or less to the isolation of a group, whilst directly referring back to the global isolation indices (Figure 14). Classifications are consistent across groups, which permits visual comparison across groups for the local values.

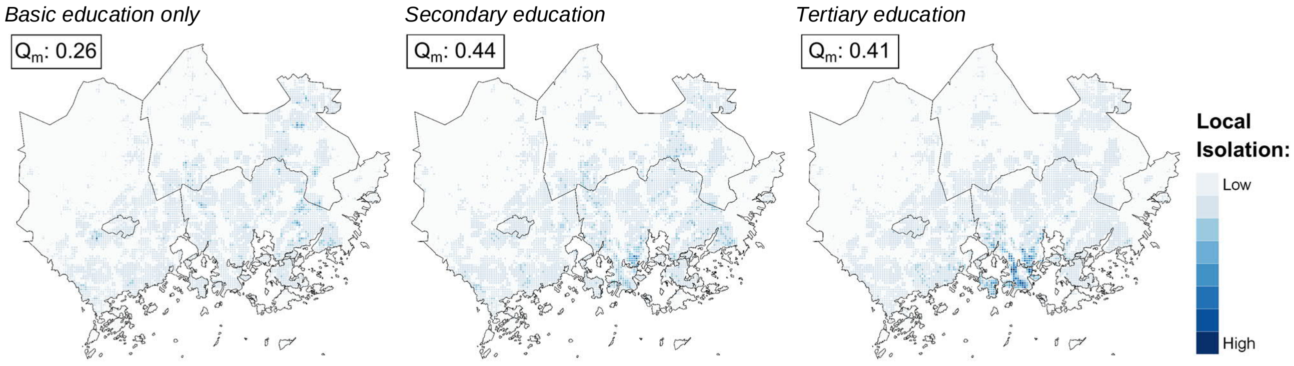
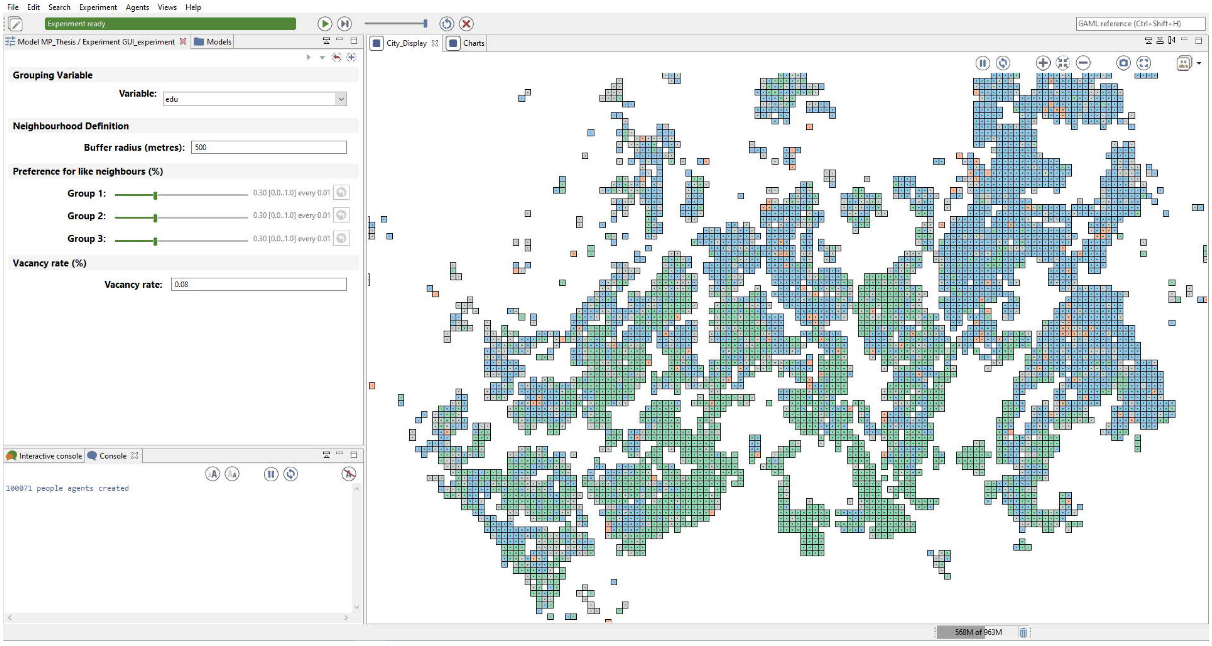


*Figure 14 - Isolation index ($\tilde{Q}_m$) of the baseline state for education groups. The sum of the local $\tilde{Q}_m$ measures as mapped are equal to the global $\tilde{Q}_m$ figure indicated above each map. Darker colours are contributing more to the global index. The classification of local values is consistent across all groups.*

## 6.7 Graphical user interface

The simulations for this study were carried out using batch processing, running multiple simulations simultaneously in order to reduce the computational load and the duration of each cycle. A graphical user interface (GUI) experiment was also constructed to demonstrate the accessibility of ABMs for non-experts. In a GUI experiment, users not familiar with the GAML language can test different scenarios using a simplified user interface (Figure 15). When coding, specific parameters such as neighbourhood distance and preference thresholds can be passed to the GUI to allow easy testing of different scenarios using dialogue boxes. A dynamic map of the environment is presented, along with charts that track the progress of the simulations over time. The attributes of the individual agents and spaces can be interrogated easily by the user at any time during the simulations.



Grouping Variable

Variable: edu

Neighbourhood Definition

Buffer radius (metres): 500

Preference for like neighbours (%)

Group 1: 0.30 [0.0..1.0] every 0.01

Group 2: 0.30 [0.0..1.0] every 0.01

Group 3: 0.30 [0.0..1.0] every 0.01

Vacancy rate (%)

Vacancy rate: 0.08

```
100071 people agents created
```

*Figure 15* - *Graphical user interface in GAMA. Key parameters can be explored without needing to modify the script. A live map and charts (on a separate tab) allow the user to track the progress throughout the simulation.*

# 7 Results

## 7.1 Initial state of simulation

The initial state of each simulation will briefly be described here with an explanation of how to interpret the maps and different segregation indices. As the 100,000 agents are representative of the regional proportion of each group and the population density of the HMA, the initial (baseline) state of the simulation is largely consistent with the HMA figures as outlined earlier when describing the characteristics of the study area. Maps detailing the local variations in the segregation indices have been provided as a visual support for understanding the global measures, however it is not the purpose of this thesis to make precise geographic predications about future states of segregation in the HMA. The results can only speak to the behaviour of the system as prescribed within the model. Rather, the maps can serve to help explain changes against the baseline measures and the impact of density and the urban structure on the results.

### *Income groups – Initial state (baseline scenario)*

With a global $\tilde{H}$-value of 0.06, the three income groups are very evenly distributed throughout the urban fabric of the HMA (Figure 16). Local areas which are less diverse than the overall region, in blue, contribute a positive score to the $\tilde{H}$-value. The local values listed in the legend are multiplied by 1000 as the values are very small, so the contribution of most grid cells to the global $\tilde{H}$-value is less than 0.0001. Negative local $\tilde{H}$-values, in orange/red, indicate that the local populations are more integrated than the regional proportions. Precisely, this means that the three groups are closer to being equally represented in the local area than their respective proportions at the regional level.

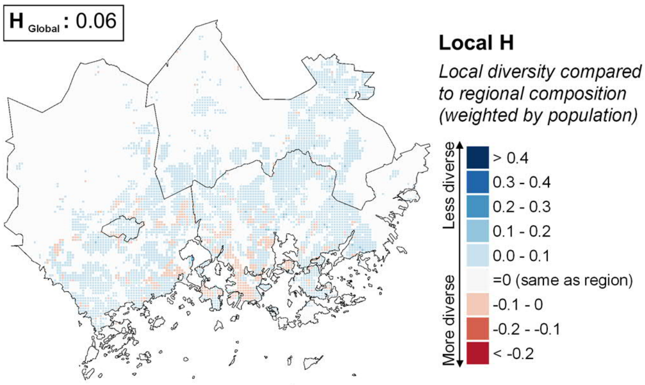


*Figure 16 - Information Theory Index ($\tilde{H}$) for the baseline state of income groups. The sum of the local $\tilde{H}$ values as mapped are equal to the global $\tilde{H}$ figure indicated above the map. Local values take both the regional population composition and density into account. Local values in the legend are multiplied by 1000 for legibility.*

The medium-income group has a higher isolation index ($\tilde{Q}_m$) value (Figure 17), due in a large part to its greater overall representation in the region at 53%, whilst high-income (27%) and low-income groups (20%) have smaller regional proportions. Darker colours indicate that these areas contribute more to the global isolation values. The local units' $\tilde{Q}_m$ values take both the regional population composition and population density into account, so concentrations of groups in denser areas will be more apparent, as is the case with high-income households in the city centre of Helsinki.

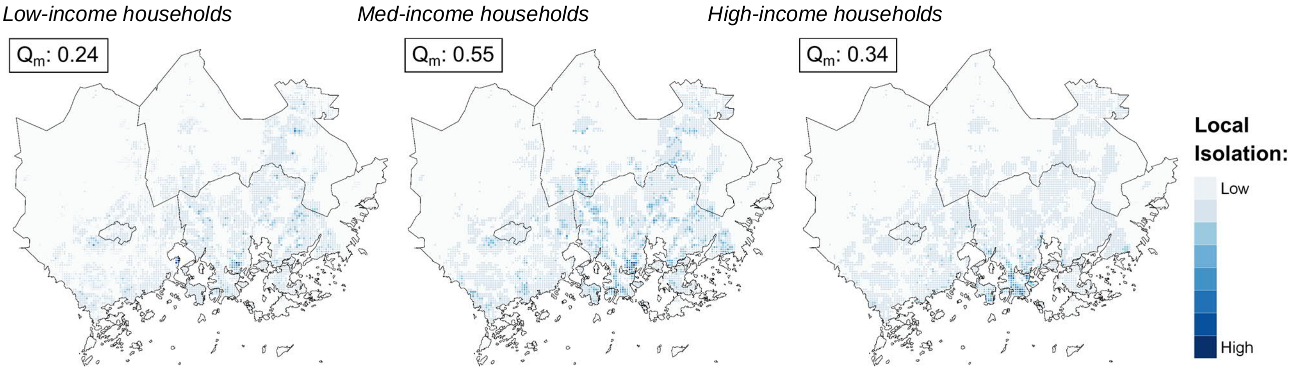


*Figure 17 - Isolation index ($\tilde{Q}_m$) values for the baseline state of income groups. The sum of the local $\tilde{Q}_m$ measures as mapped are equal to the global $\tilde{Q}_m$ figure indicated above each map. As $\tilde{Q}_m$ is affected by the regional proportions of each group, the global values cannot be directly compared across groups.*

### *Education groups – Initial state (baseline scenario)*

Education groups, as defined in the study, are also very evenly distributed throughout the HMA. The global $\tilde{H}$-value of 0.04 arises from many thousands of areal units that are slightly less diverse than the overall region (Figure 18). Some higher values (darker blue) are being produced from less-diverse areas in the dense inner city of Helsinki; however, these are offset by many smaller patches which are more integrated than the overall region.

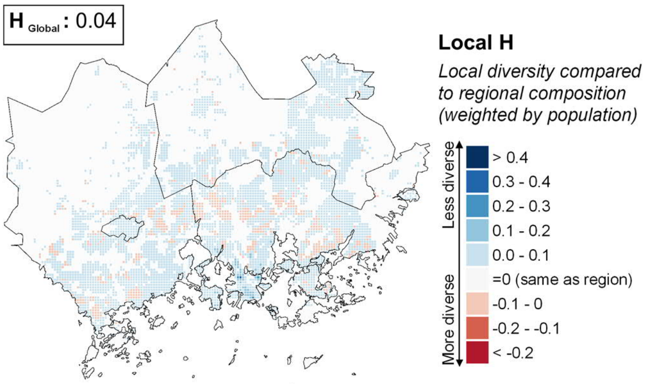


*Figure 18* - *Information Theory Index ($\tilde{H}$) indicators for the baseline state of education groups. The sum of the local $\tilde{H}$ measures as mapped are equal to the global $\tilde{H}$ figure. The local units' values take both the regional population composition and density into account. Local values in the legend are multiplied by 1000 for legibility.*

The relative sizes of the education groups are the most even of the three grouping variables tested in this paper; however, it is important to remember that the basic education group (21% of the population) is still significantly smaller than secondary education (43%) and tertiary education groups (36%). This needs to be taken into consideration when considering the $\tilde{Q}_m$ scores. Higher education residents appear to be more likely to encounter members of the same group in the dense city centre of Helsinki, whilst those with basic education have a much lower chance of residing in the same area as another resident with only basic education (Figure 19).

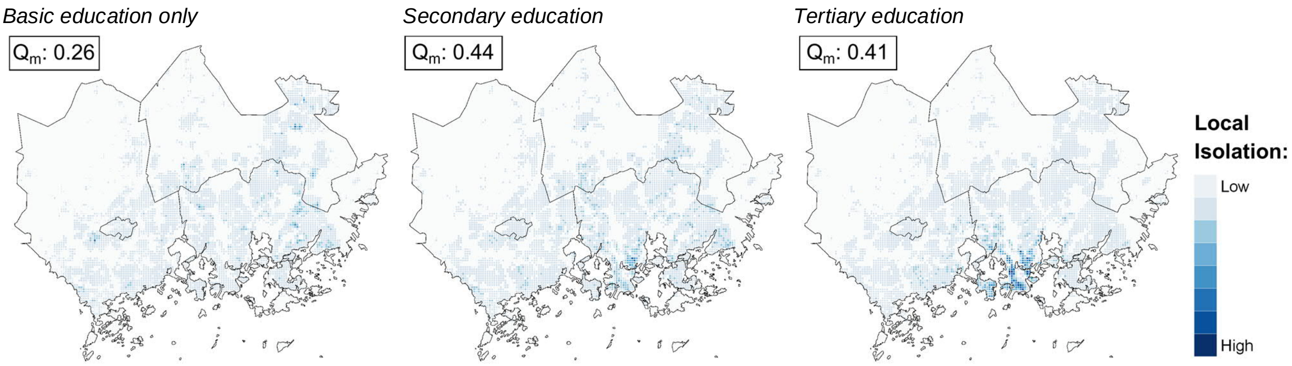


*Figure 19* - *Isolation index ($\tilde{Q}_m$) values for the baseline state of education groups. The sum of the local $\tilde{Q}_m$ measures as mapped are equal to the global $\tilde{Q}_m$ figure indicated above each map. As $\tilde{Q}_m$ is affected by the regional proportions of each group, the global values cannot be directly compared across groups.*

***Language groups – Initial state (baseline scenario)***

With an $\tilde{H}$-value of 0.08, the different language groups are evenly distributed throughout the HMA. A less diverse (more segregated) patch is observed near the Kallio and Vallila areas of Helsinki, however there are many areas which are more diverse which act to negate this in the global figure (Figure 20). In effect, these areas that are more diverse represent an overrepresentation of the minority groups compared to their regional proportions, and their existence necessitates that there are less diverse regions elsewhere in the region, and vice versa.

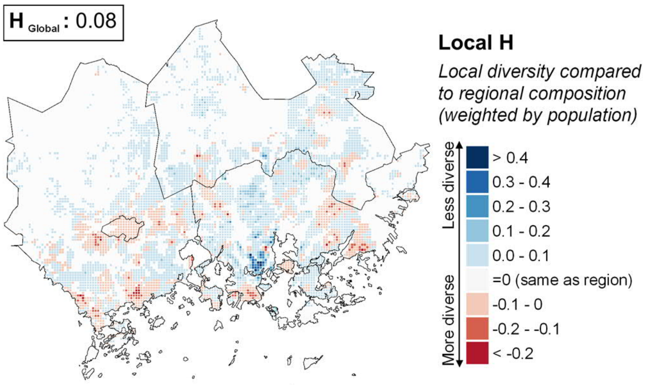


*Figure 20* - *Information Theory Index ($\tilde{H}$) indicators for the baseline state of language groups. The sum of the local $\tilde{H}$ measures as mapped are equal to the global $\tilde{H}$ figure. The local units' values take both the regional population composition and density into account. Local values in the legend are multiplied by 1000 for legibility.*

Language groups are the most imbalanced of the three sorting variables in terms of regional proportions. Whilst Finnish mother tongue agents represent over three-quarters of the population of HMA, those with Swedish mother tongue represent only 5%. This accounts for the high isolation index of the Finnish-speaking agents of 0.79, indicating that they are more likely to live in an area with other Finnish-speaking agents (Figure 21). The Swedish-speakers isolation index is increased by concentrations in Kauniainen and downtown Helsinki, whilst the 'other' language groups appear to be more widely dispersed, perhaps reflecting the pluralistic nature of this classification.

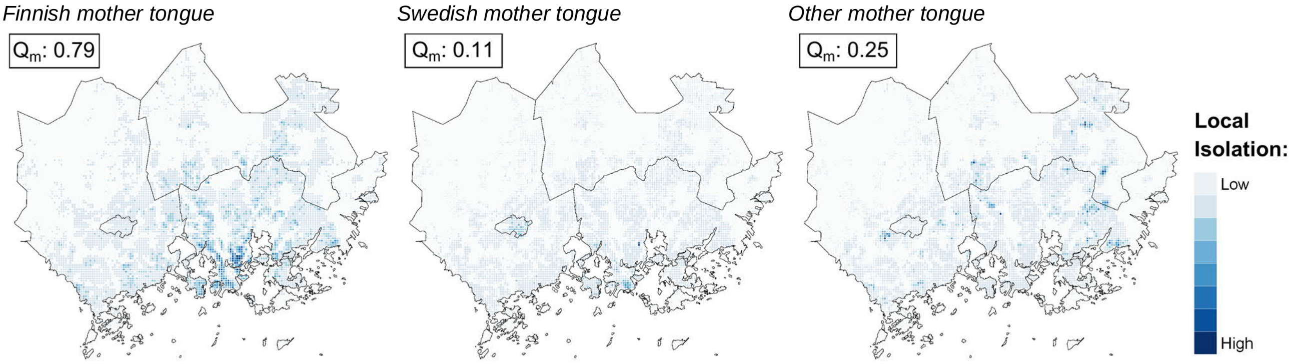


*Figure 21* - *Isolation index ($\tilde{Q}m$) values for the baseline state of language groups. The sum of the local $\tilde{Q}m$ measures as mapped are equal to the global $\tilde{Q}m$ figure indicated above each map. As $\tilde{Q}m$ is affected by the regional proportions of each group, the global values cannot be directly compared across groups.*

## 7.2 Results of simulation experiments

A summary of the simulation results for each of the different scenarios is presented in Table 7.

*Table 7 - Summary of simulation results after 20 cycles. Each scenario was run 10 times, and the mean aggregate values reported. Measures of evenness ($\tilde{H}$) and isolation ($\tilde{Q}_m$) are reported for each group*

| | **Baseline measures** | **Scenario A 30% pref** | **Scenario B 50% pref** | **Scenario C 30/70% pref** |
|---|---|---|---|---|
| ***Education groups*** | | | | |
| $\tilde{H}$ - *Global* | 0.04 | 0.36 | 0.56 | 0.25 |
| $\tilde{Q}$ - *Basic education* | 0.26 | 0.50 | 0.57 | 0.41 |
| $\tilde{Q}$ - *Secondary education* | 0.44 | 0.50 | 0.80 | 0.49 |
| $\tilde{Q}$ - *Tertiary education** | 0.41 | 0.64 | 0.78 | 0.57 |
| ***Income groups*** | | | | |
| $\tilde{H}$ - *Global* | 0.06 | 0.26 | 0.74 | 0.14 |
| $\tilde{Q}$ - *Low-income* | 0.24 | 0.35 | 0.70 | 0.31 |
| $\tilde{Q}$ - *Medium-income* | 0.55 | 0.59 | 0.91 | 0.58 |
| $\tilde{Q}$ - *High-income** | 0.34 | 0.53 | 0.93 | 0.39 |
| ***Language groups*** | | | | |
| $\tilde{H}$ - *Global* | 0.08 | 0.08 | 0.15 | 0.81 |
| $\tilde{Q}$ - *Finnish ** | 0.79 | 0.79 | 0.82 | 0.97 |
| $\tilde{Q}$ - *Swedish* | 0.11 | 0.08 | 0.08 | 0.74 |
| $\tilde{Q}$ - *Other* | 0.25 | 0.25 | 0.37 | 0.85 |

** Assigned the higher 70% prerference in Scenario C*

### *7.2.1 Scenario A: 30% preference for co-group neighbours*

In this first simulation, agents are willing to live in mixed neighbourhoods as long as 30% of their neighbours are of the same group as themselves. With three groups in each of the simulations, if an agent lives in a neighbourhood with less than 30% of its own group, it will necessarily be a minority within the neighbourhood. Under these parameters, segregation can be seen to increase rather significantly for both education and income variables, whilst spatial differentiation on the basis of residents' mother tongue remains largely unchanged from the baseline.

At the beginning of the simulation, almost 100% of the medium-income agents and secondary education agents are satisfied with their neighbourhood and therefore do not need to migrate (Figures 24b and 27b). Accordingly, the increased segregation of the income and education groups, indicated by increasing levels of $\tilde{H}$ (Figures 22b and 25b), are brought about largely through the movements of the other groups. Over 75% of the agents with a tertiary education are satisfied at the beginning of the simulation, and within about 5 cycles all agents of this group find a suitable neighbourhood. After 20 cycles, high-income and low-income agents increase their isolation (increasing $\tilde{Q}_m$) through migration (Figure 24a), however a large number of agents never find a suitable location. The same is true for basic education agents.

A key factor in the apparent stability of the language group's situation is the overall group proportions in the region. The relative sizes of the education or income groups are more evenly distributed than that of language groups, whereas those who have Finnish as their mother tongue represent 78% of the population. This clear majority means that all agents within this group already live in neighbourhoods which have over 30% Finnish speakers (Figure 30b), and therefore do not need to migrate. However, compared to the education and income simulations, the majority status of Finnish-speakers, and the rather integrated starting position of the group, means that it is difficult for Swedish-speaking and 'other-language' agents from finding suitable neighbourhoods. As the migrations are random, the low chances of arriving in a neighbourhood with 30% of the same group for these minority groups means that they repeatedly continue to migrate each cycle.[1] Indeed, rather

---

[1] Non-reported early simulation tests with a lower preference threshold of 20% actually resulted in higher segregation for language groups ($\tilde{H}$ value of 0.19) due to the fact that minority groups have more chance to find a suitable neighbourhood. The Finnish-speaking agents still had no need to move as they already lived in neighbourhoods with over 20% of other Finnish-speakers, however the 'other language' agents moved towards increasing isolation ($\tilde{Q}_m$ =0.36), with around 80% of agents finding suitable neighbourhoods. This is more than twice as many as under the 30% simulation. Swedish-speaking agents also increased their isolation ($\tilde{Q}_m$ = 0.13) and as a result Finnish-speakers also ended up marginally more isolated compared to the baseline ($\tilde{Q}_m$ = 0.80). Income groups at 20% preference threshold

than remaining stagnant, as the lack of evolution in the $\tilde{H}$ and $\tilde{Q}_m$ values would suggest, agents of these groups are hyper-mobile, attempting anew each cycle to try and find a satisfactory location. At the end of 20 cycles, the ‘other-language’ speakers (17% of regional population) marginally increase the total number of agents who are living in neighbourhoods which meet their preference thresholds (Figure28b). The Swedish-speaking agents (5% of regional population), however, end the simulations in a more integrated postion than at cycle 0 (Figure 29).

---

were also slightly more segregated ($\tilde{H}$ value of 0.29) than at 30% for a similar reason, whilst education groups (the most evenly proportionate variable) were less segregated at 20% preference ($\tilde{H}$ value of 0.29) than at 30% preference.

## *Income groups – Scenario A (30%)*

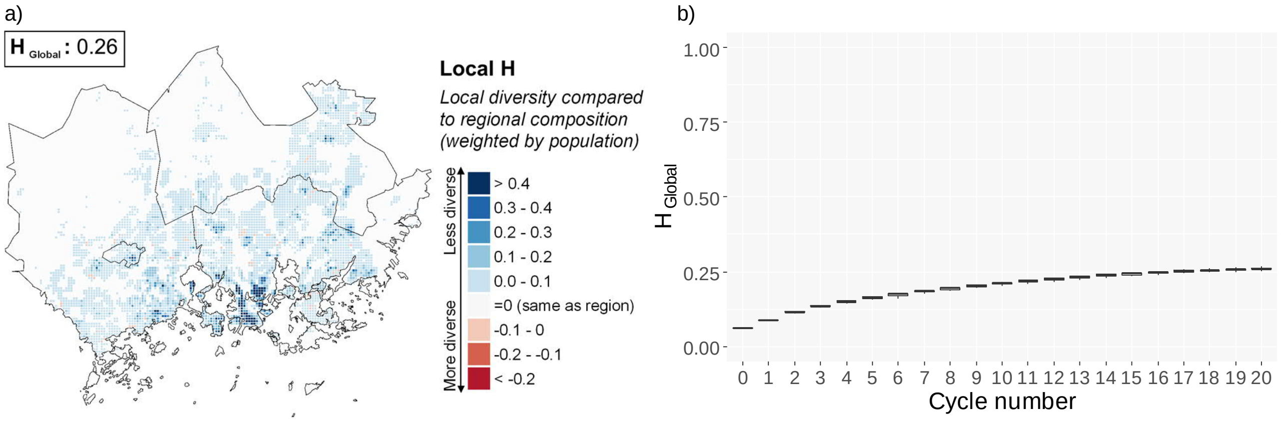


*Figure 22 – Information theory index ($\tilde{H}$) for income groups with 30% preference thresholds after 20 cycles. The sum of the local $\tilde{H}$ values as mapped (a) is equal to the global $\tilde{H}$ value. Local values are multiplied by 1000 in the legend. Simulations were repeated 10 times and the evolution of the global $\tilde{H}$ values over the 20 cycles reported (b).*

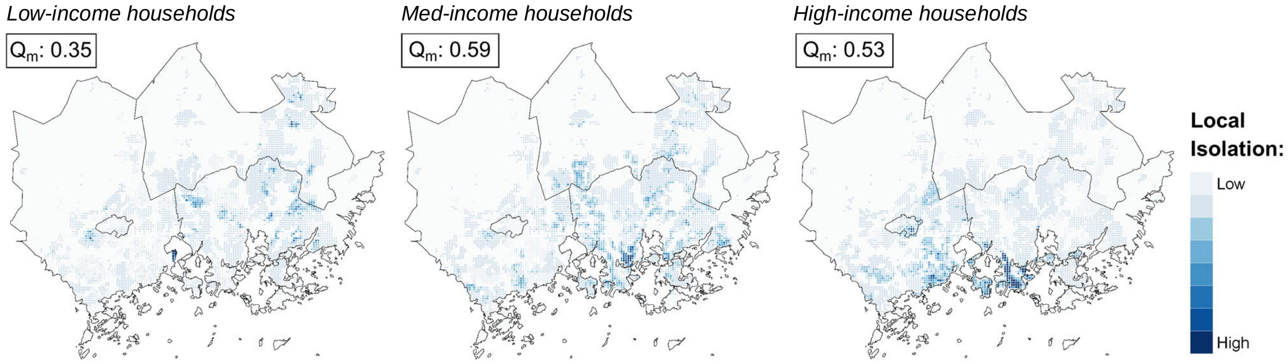


*Figure 23 – Global and local isolation index ($\tilde{Q}_m$) for income groups with 30% preference thresholds after 20 cycles. The sum of the local $\tilde{Q}_m$ values as mapped is equal to the global $\tilde{Q}_m$ figure indicated above each map. As the calculation of $\tilde{Q}_m$ considers the overall proportions of each group in the region, it cannot be directly compared across groups, but can be compared to their baseline.*

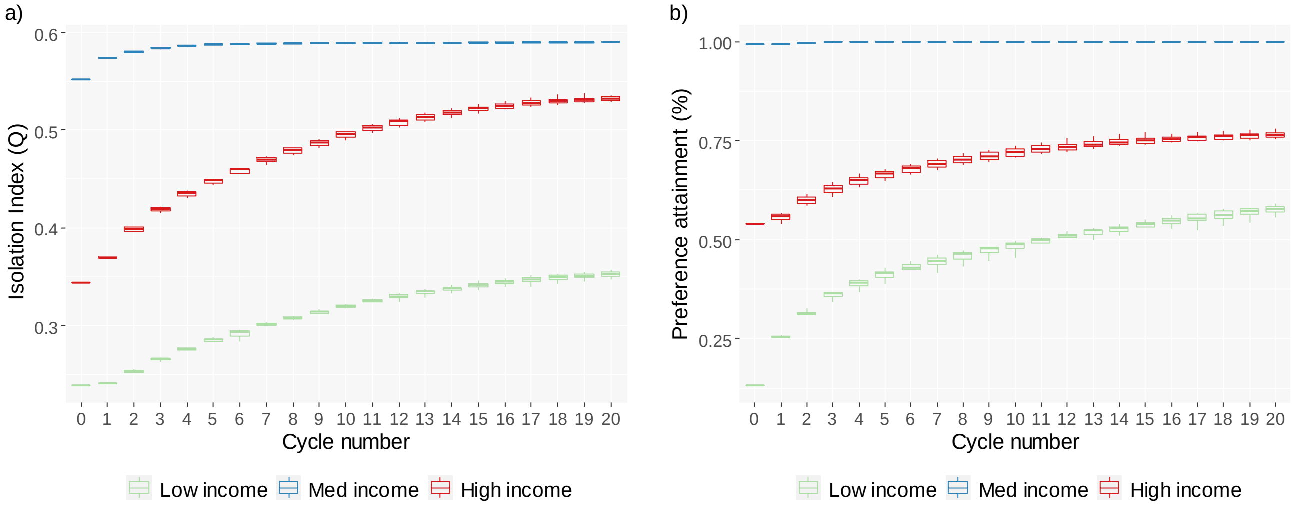


*Figure 24 – Evolution of the isolation index ($\tilde{Q}_m$) for income groups with 30% preference thresholds (a) and the proportion of agents which are attaining their preference threshold for same-group neighbours (b). 10 simulations were run to account for the stochastic nature of the model, the aggregate results are presented here as a box plot.*

### Education groups – Scenario A (30%)

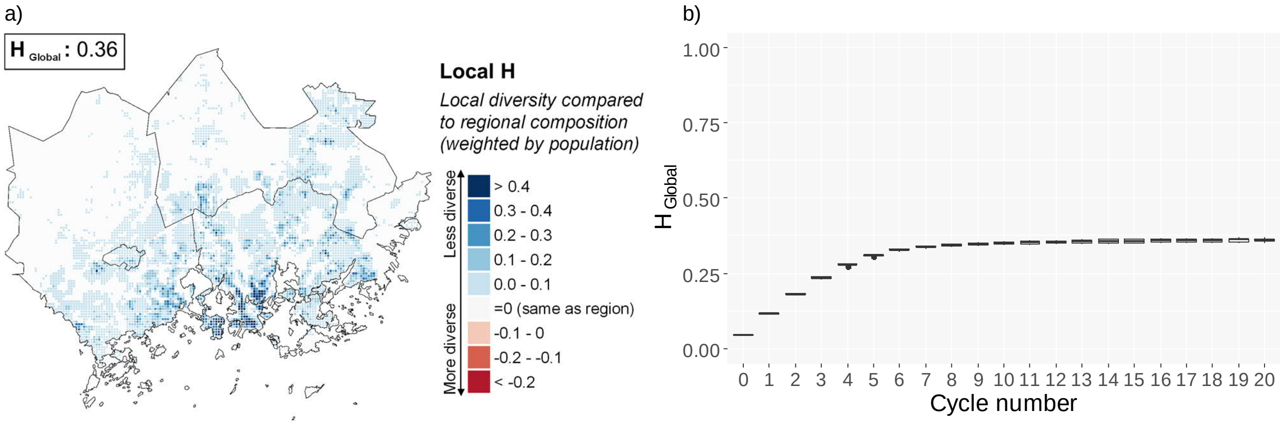


*Figure 25 - Information theory index ($\tilde{H}$) for education groups with 30% preference thresholds after 20 cycles. The sum of the local $\tilde{H}$ values as mapped (a) is equal to the global $\tilde{H}$ value. Local values are multiplied by 1000 in the legend. Simulations were repeated 10 times and the evolution of the global $\tilde{H}$ values over the 20 cycles reported (b).*

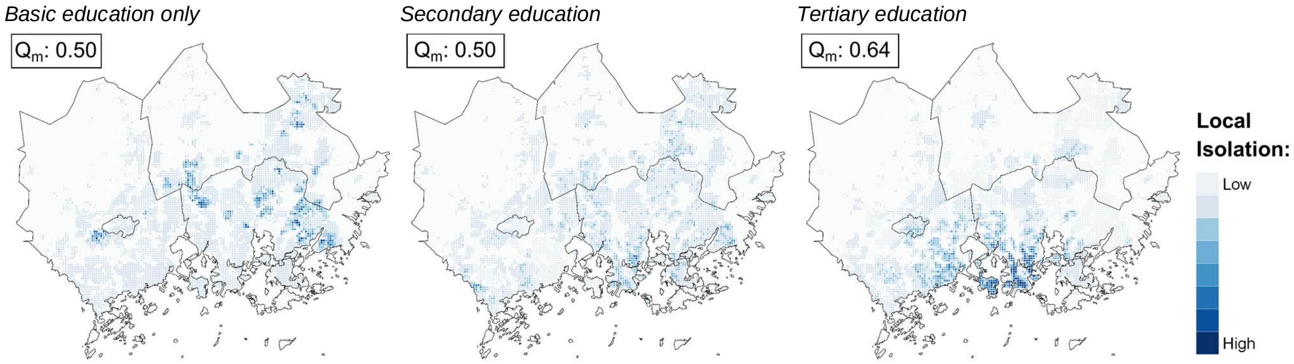


*Figure 26 - Global and local isolation index ($\tilde{Q}_m$) for education groups with 30% preference thresholds after 20 cycles. The sum of the local $\tilde{Q}_m$ values as mapped is equal to the global $\tilde{Q}_m$ figure indicated above each map. As the calculation of $\tilde{Q}_m$ considers the overall proportions of each group in the region, it cannot be directly compared across groups, but can be compared to their baseline.*

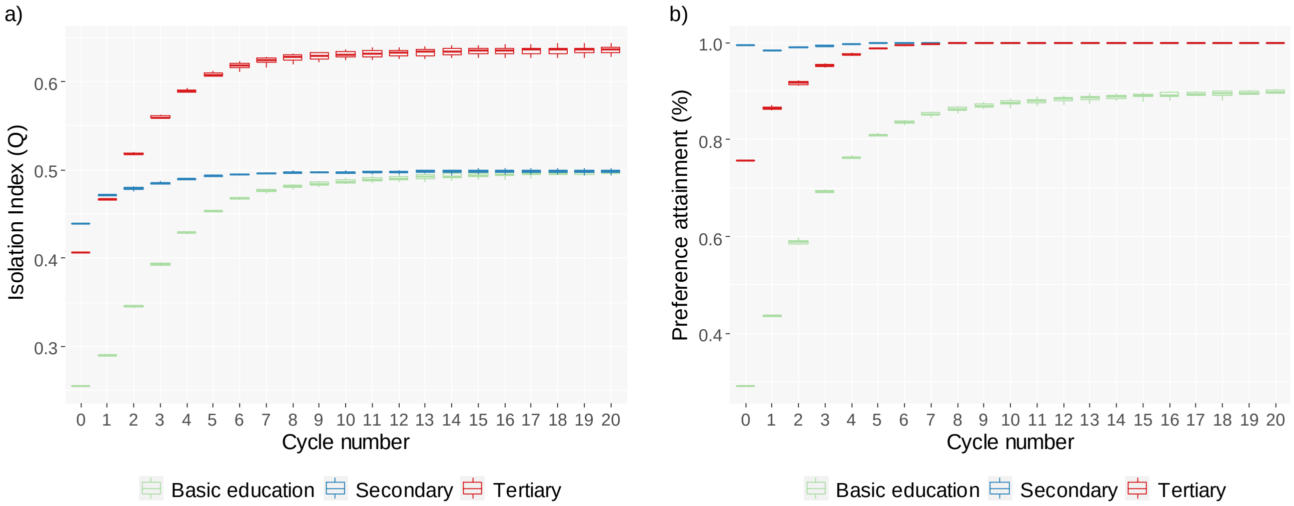


*Figure 27 - Evolution of the isolation index ($\tilde{Q}_m$) for education groups with 30% preference thresholds (a) and the proportion of agents which are attaining their preference threshold for same-group neighbours (b). 10 simulations were run to account for the stochastic nature of the model, the aggregate results are presented here as a box plot.*

## *Language groups – Scenario A (30%)*

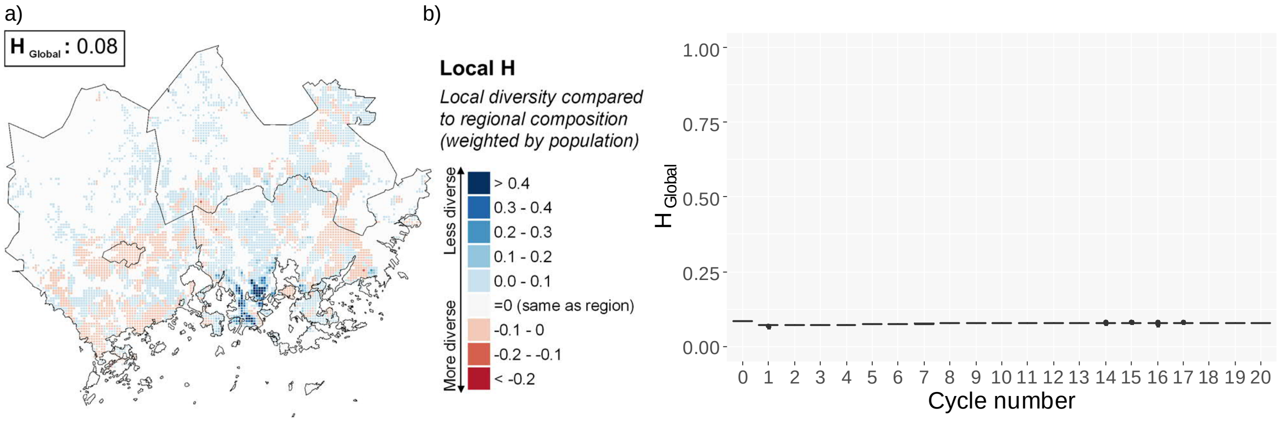


*Figure 28 - Information theory index ($\tilde{H}$) for language groups with 30% preference thresholds after 20 cycles. The sum of the local $\tilde{H}$ values as mapped (a) is equal to the global $\tilde{H}$ value. Local values are multiplied by 1000 in the legend. Simulations were repeated 10 times and the evolution of the global $\tilde{H}$ values over the 20 cycles reported (b).*

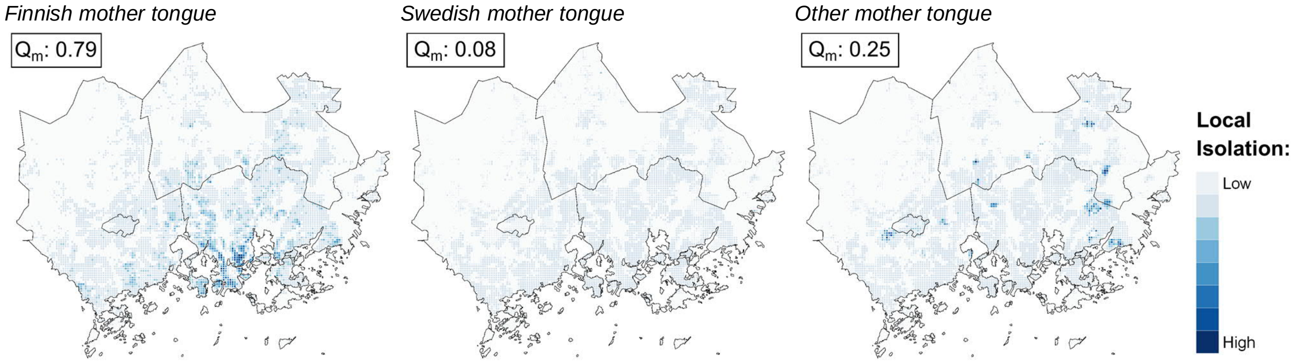


*Figure 29 - Global and local isolation index ($\tilde{Q}_m$) for language groups with 30% preference thresholds after 20 cycles. The sum of the local $\tilde{Q}_m$ values as mapped is equal to the global $\tilde{Q}_m$ figure indicated above each map. As the calculation of $\tilde{Q}_m$ considers the overall proportions of each group in the region, it cannot be directly compared across groups, but can be compared to their baseline.*

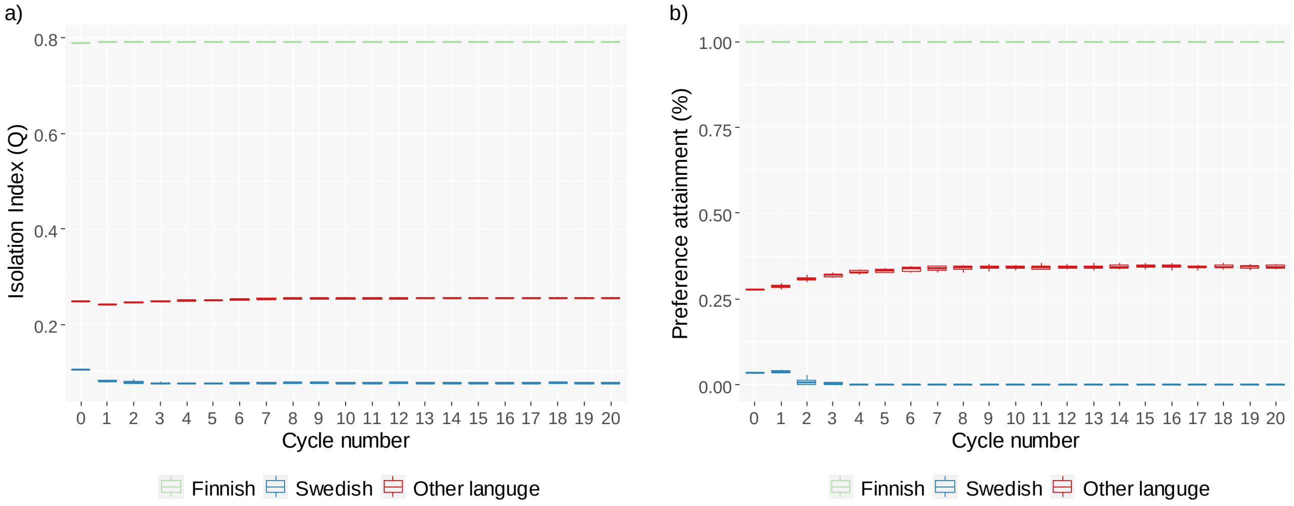


*Figure 30 - Evolution of the isolation index ($\tilde{Q}_m$) for language groups with 30% preference thresholds (a) and the proportion of agents which are attaining their preference threshold for same-group neighbours (b). 10 simulations were run to account for the stochastic nature of the model, the aggregate results are presented here as a box plot.*

### *7.2.2 Scenario B: 50% preference for co-group neighbours*

A preference for residing in a neighbourhood with at least 50% of co-group agents suggests that agents seek to be the majority group in the neighbourhood. With a 50% preference, segregation, measured with $\tilde{H}$, increases across all of the variables, and increases beyond the levels of segregation produced under the 30% preference scenario. The rate of increase, however, is not consistent across all groups.

The $\tilde{H}$-value for income groups increases from a baseline of 0.06, to 0.74 over 20 cycles (Figure 31), significantly more than under the 30% threshold. Over 75% of medium-income agents are already living in a neighbourhood which meets the 50% theshold. There is an indication of neighbourhood tipping in Figure 33b when the proportion of medium-income agents residing in a suitable neighbourhood drops slightly after the first cycle. As the neighbourhoods do not share common boundaries, it is unclear whether this is due to the departure of medium-income agents, or the arrival of new high-income and low-income households into these agents neighbourhoods. Whichever the case, migration of other agents has evidently tipped the composition unfavourably for some medium-income agents. A consistent pattern across the 10 duplicate simulations, is that it is not until the medium and high-income groups become more settled that the low-income agents begin to find suitable neighbourhoods. This appears to be due to the fact that the low-income group is very well integrated, and therefore none of their agents are living in a 50% majority neighbourhood upon initialisation. The formation of a sufficent cluster seemingly requires agents of the other groups to first move away in order to facilite this. At the end of 20 cycles the isolation index of the medium-income and high-income groups approaches complete isolation (Figure 32).

Education groups increased from a baseline $\tilde{H}$-value of 0.04 to 0.56 with a 50% preference (Figure 34). The secondary and tertiary-educated agents increase their isolation at an almost identical rate, both arriving at a $\tilde{Q}_m$ value of around 0.80 after 20 cycles (Figure 35). None of the basic education agents, representing the smallest share of the population at around 20% and starting from a more intergrated position, find a suitable neighbourhood until around halfway through the simulation, when the other two groups begin to settle. The stochastic nature of the neighbourhood selection is most visible here, with large variances in $\tilde{Q}_m$ for basic education agents after cycle 10 recorded for the 10 identical simulation run (Figure 36).

The distribution of language groups in the HMA becomes less even (more segregated) than the baseline level, with $\tilde{H}$ increasing from 0.08 to 0.15 after 20 cycles (Figure 37). The reasons for this modest increase are consistent with results of the 30% preferences, where there was no increase in the value of $\tilde{H}$. All agents of the Finnish-speaking majority are already living in neighbourhoods where they have a 50% share or or higher of the local population (Figure 39b). None of the Swedish-speaking agents are unable to find a suitable neighbourhood where they can achieve a 50% majority due to their significant minority status, and as a consequence of their random relocations to try and find a new neighbourhood their $\tilde{Q}_m$ value actually decreases from 0.11 to 0.08. As ‘other-language’ agents increase their isolation within the HMA, this also further increases the isolation of the Finnish-speaking agents, producing the slight increase in $\tilde{H}$.

## *Income groups – Scenario B (50%)*

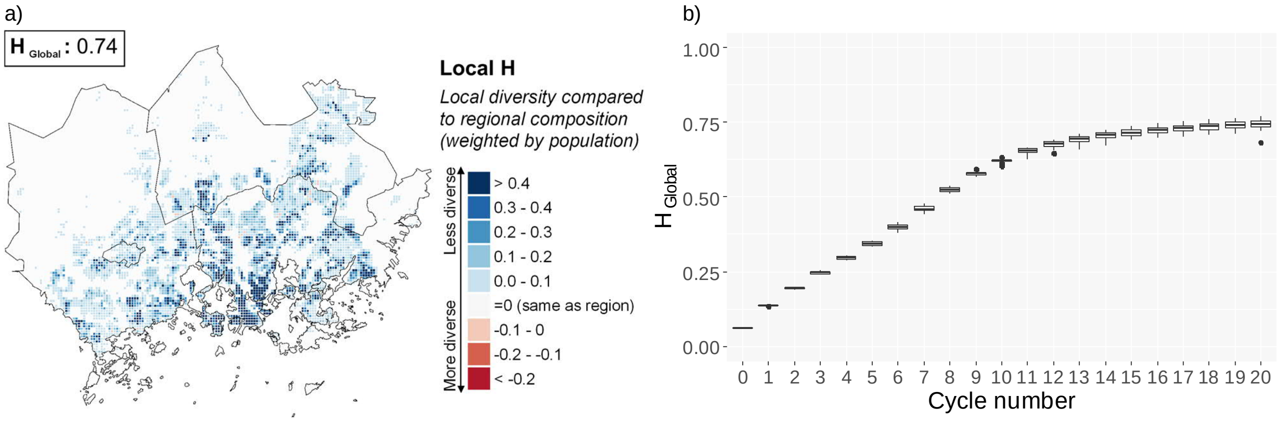


*Figure 31 - Information theory index ($\tilde{H}$) for income groups with 50% preference thresholds after 20 cycles. The sum of the local $\tilde{H}$ values as mapped (a) is equal to the global $\tilde{H}$ value. Local values are multiplied by 1000 in the legend. Simulations were repeated 10 times and the evolution of the global $\tilde{H}$ values over the 20 cycles reported (b).*

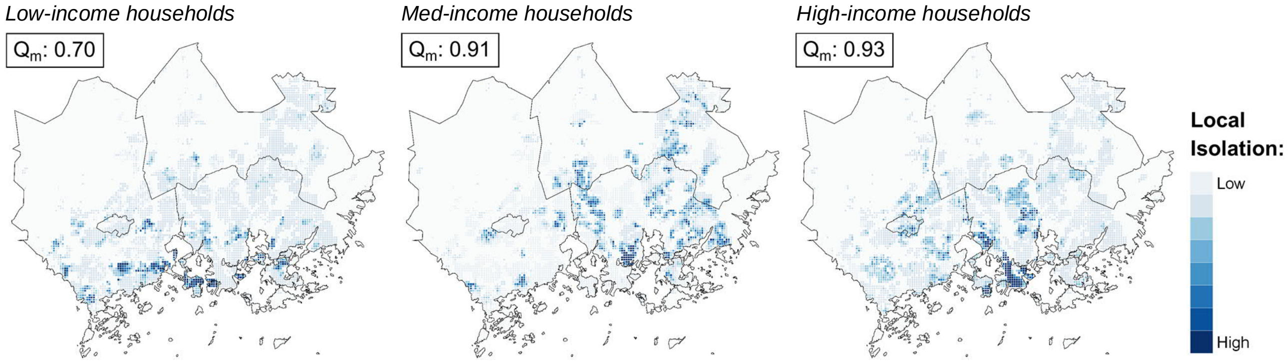


*Figure 32 - Global and local isolation index ($\tilde{Q}_m$) for income groups with 50% preference thresholds after 20 cycles. The sum of the local $\tilde{Q}_m$ values as mapped is equal to the global $\tilde{Q}_m$ figure indicated above each map. As the calculation of $\tilde{Q}_m$ considers the overall proportions of each group in the region, it cannot be directly compared across groups, but can be compared to their baseline.*

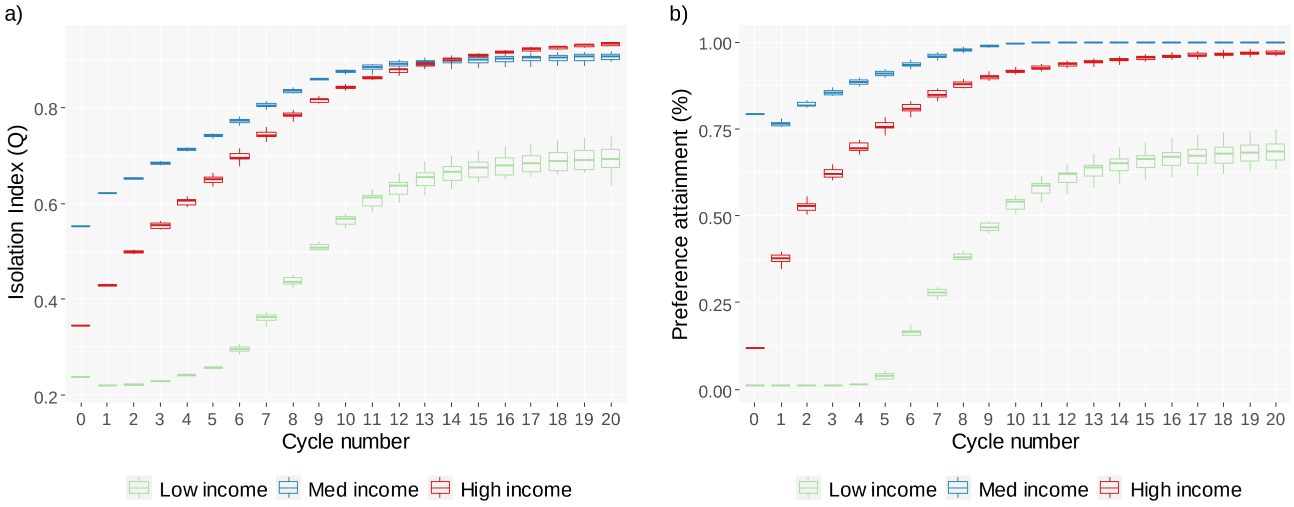


*Figure 33 - Evolution of the isolation index ($\tilde{Q}_m$) for income groups with 50% preference thresholds (a) and the proportion of agents which are attaining their preference threshold for same-group neighbours (b). 10 simulations were run to account for the stochastic nature of the model, the aggregate results are presented here as a box plot.*

### *Education groups – Scenario B (50%)*

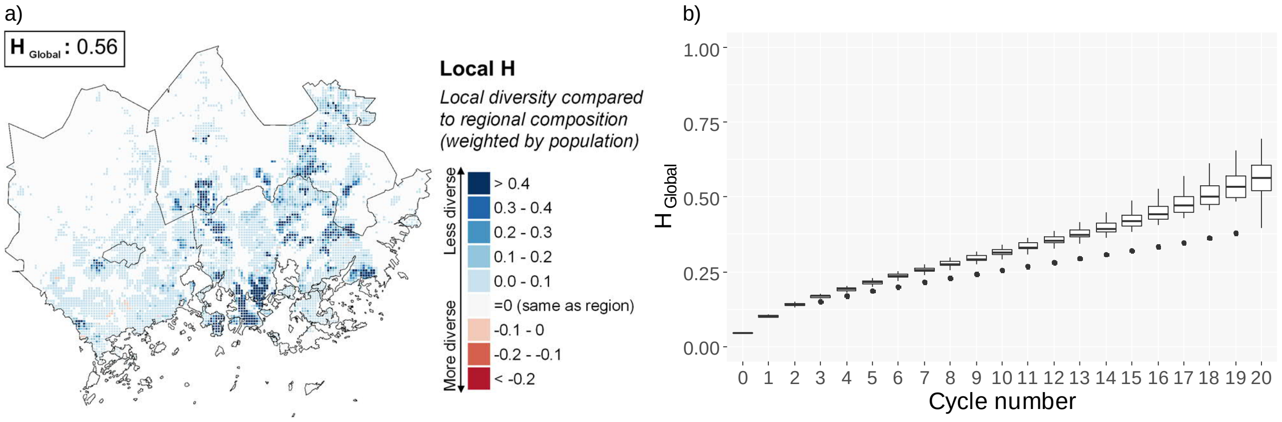


*Figure 34 - Information theory index ($\tilde{H}$) for education groups with 50% preference thresholds after 20 cycles. The sum of the local $\tilde{H}$ values as mapped (a) is equal to the global $\tilde{H}$ value. Local values are multiplied by 1000 in the legend. Simulations were repeated 10 times and the evolution of the global $\tilde{H}$ values over the 20 cycles reported (b).*

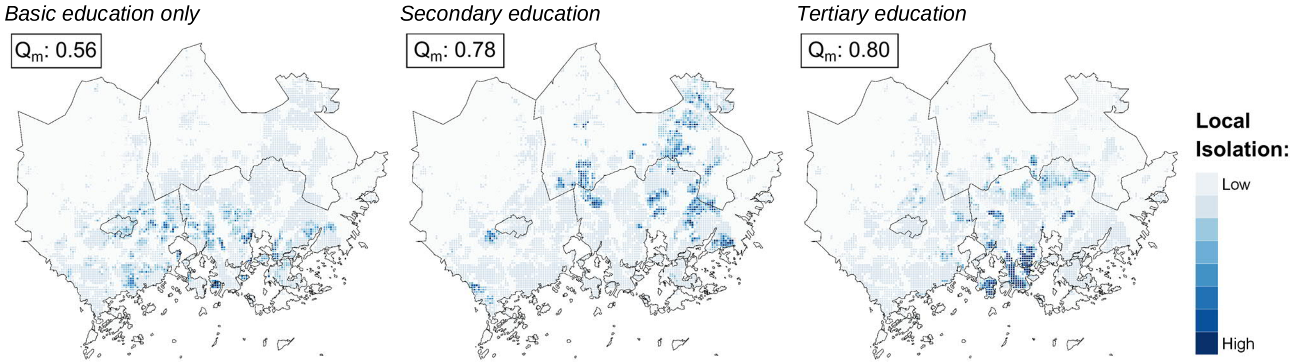


*Figure 35 - Global and local isolation index ($\tilde{Q}_m$) for education groups with 50% preference thresholds after 20 cycles. The sum of the local $\tilde{Q}_m$ values as mapped is equal to the global $\tilde{Q}_m$ figure indicated above each map. As the calculation of $\tilde{Q}_m$ considers the overall proportions of each group in the region, it cannot be directly compared across groups, but can be compared to their baseline.*

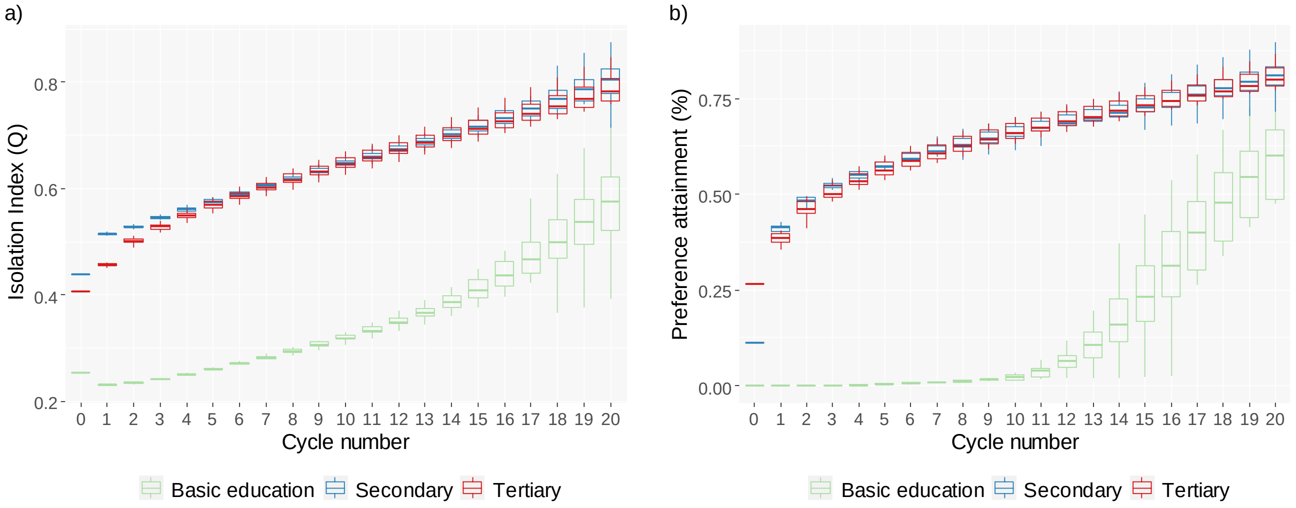


*Figure 36 - Evolution of the isolation index ($\tilde{Q}_m$) for education groups with 50% preference thresholds (a) and the proportion of agents which are attaining their preference threshold for same-group neighbours (b). 10 simulations were run to account for the stochastic nature of the model, the aggregate results are presented here as a box plot.*

## *Language groups – Scenario B (50%)*

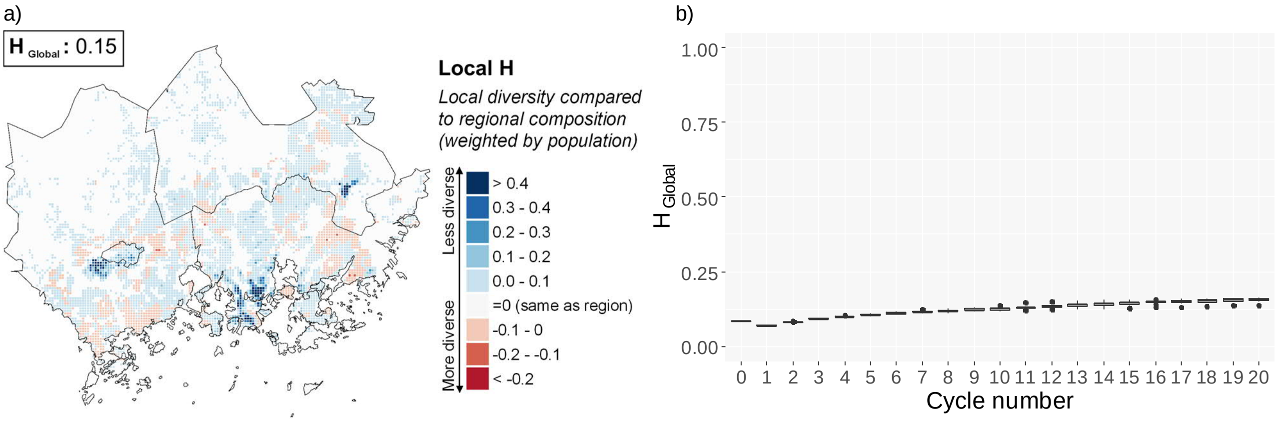


*Figure 37 - Information theory index ($\tilde{H}$) for language groups with 50% preference thresholds after 20 cycles. The sum of the local $\tilde{H}$ values as mapped (a) is equal to the global $\tilde{H}$ value. Local values are multiplied by 1000 in the legend. Simulations were repeated 10 times and the evolution of the global $\tilde{H}$ values over the 20 cycles reported (b).*

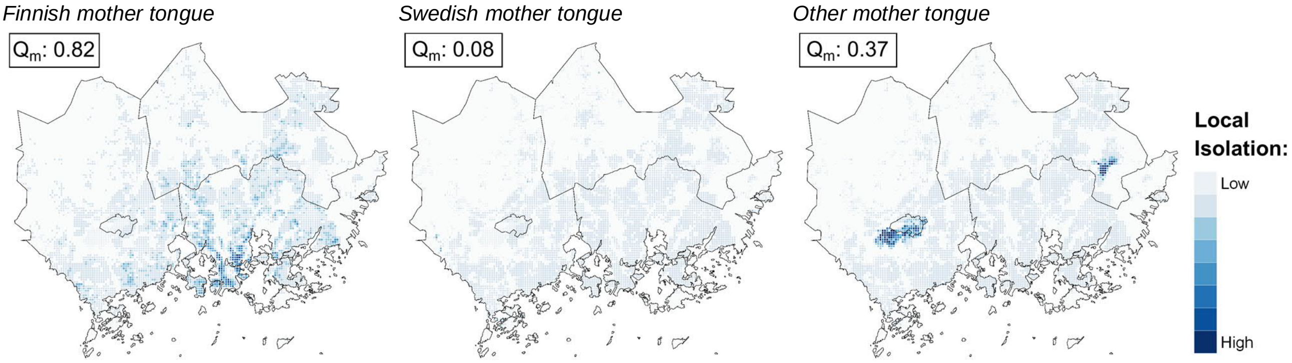


*Figure 38 - Global and local isolation index ($\tilde{Q}_m$) for language groups with 50% preference thresholds after 20 cycles. The sum of the local $\tilde{Q}_m$ values as mapped is equal to the global $\tilde{Q}_m$ figure indicated above each map. As the calculation of $\tilde{Q}_m$ considers the overall proportions of each group in the region, it cannot be directly compared across groups, but can be compared to their baseline.*

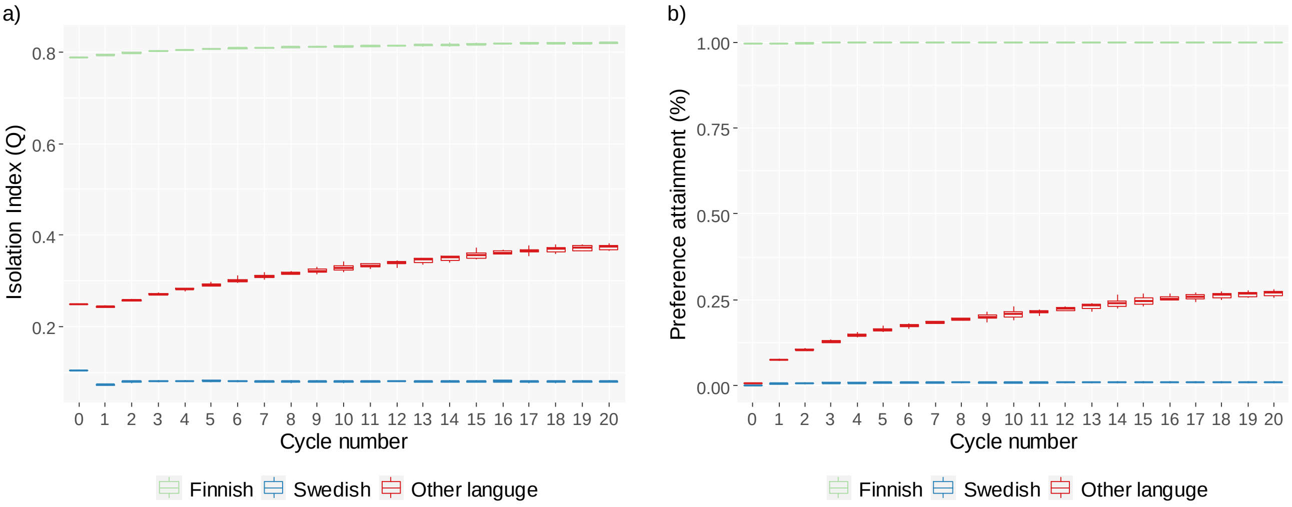


*Figure 39 - Evolution of the isolation index ($\tilde{Q}_m$) for language groups with 50% preference thresholds (a) and the proportion of agents which are attaining their preference threshold for same-group neighbours (b). 10 simulations were run to account for the stochastic nature of the model, the aggregate results are presented here as a box plot.*

### *7.2.3 Scenario C: 70%/30% asymmetric co-group preferences*

The purpose of this final scenario is to test the potential for one group with a higher preference for homophilious clustering to impact the overall segregation levels in a city. The literature review suggested that the mobility of highly-educated and high-income may hold a heightened role in residential segregation dynamics in European cities. On this basis, high-income agents and tertiary-education agents have been assigned the 70% preference threshold. For language groups, agents with Finnish as their mother tongue were asigned the higher preference, due to their considerable majority status in the region. In contrast, the high-income and tertiary-educated agents are both minority groups. This distinction becomes important in the results.

For the education and income groups, whilst segregation increases compared to the baseline, the final $\tilde{H}$-values under this scenario are actually lower (groups are more evenly distributed) than even under the 30% symmetric preference scenario. This is due to the groups with the highest preference beginning the simulations with none of their agents living in a 70% majority neighbourhood, and agents then finding it difficult to find a suitable neighbourhood (Figure 42 and 45). As in Scenario A (30% symmetric preferences), medium-income agents and agents with secondary education, both the majority groups in their respective simulations, begin with almost 100% of their agents already in a suitable neighbourhood. Accordingly they do not need to migrate, making it extremely difficult for the minority groups with the higher preferences to achieve their desired neighbourhood composition. The continued migration of these groups appears to also inhibit the other minority groups (i.e. basic education, low-income) from finding suitable neighbourhood. In this scenario, after 20 cycles the majority groups maintain similar $\tilde{Q}_m$ values to the results of the symmetric 30% tests, however both the minority groups end more integrated, with lower $\tilde{Q}_m$ values.

The simulations for language groups tell a different story. For the first time a significantly large number of Finnish-speaking agents (17%) begin the simulation living in a neighbourhood with less than the prescribed preference threshold, and need to relocate (Figure 48b). In the first cycle, these agents move towards increasing the isolation of the Finnish-speaking agents. As a consequence, this increases the possibility of Swedish-speaking and ‘other’ language agents sorting into neighbourhoods which achieve their own goals which was previously unnatainable under the 30% scenario. The combined impact of this is a signficiant move towards complete segregation. The $\tilde{H}$-value increases from its baseline measure of 0.08 to 0.81 after 20 cycles (Figure 46), and there is a significant increase in the isolation of all three groups (Figure 47).

### *Income groups – Scenario C (30%/70%)*

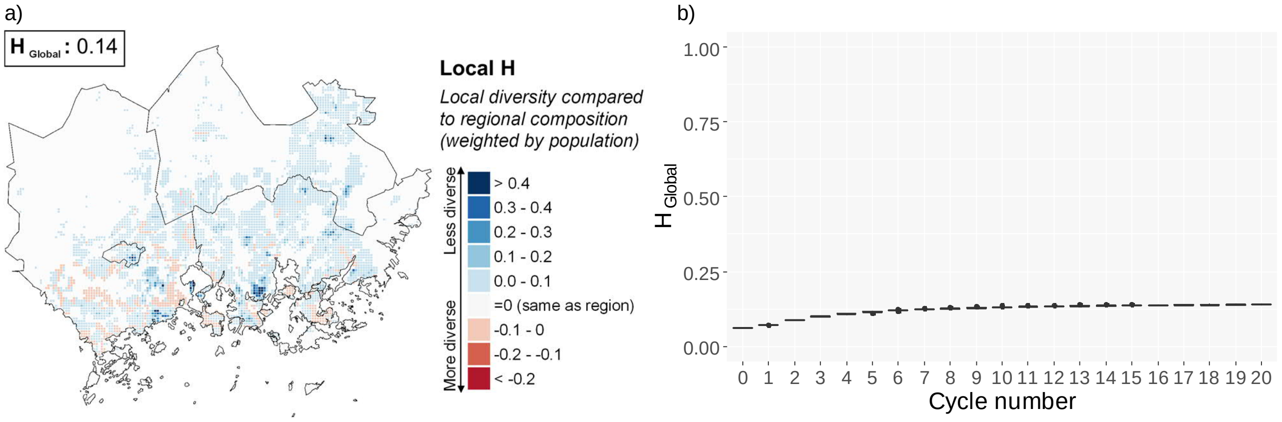
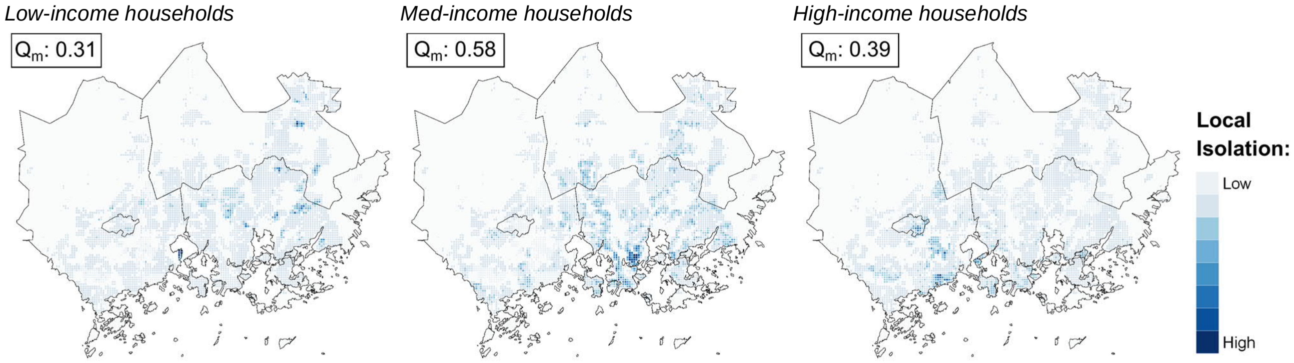


*Figure 40 - Information theory index ($\tilde{H}$) for income groups with 30%/70% preference thresholds after 20 cycles. The sum of the local $\tilde{H}$ values as mapped (a) is equal to the global $\tilde{H}$ value. Local values are multiplied by 1000 in the legend. Simulations were repeated 10 times and the evolution of the global $\tilde{H}$ values over the 20 cycles reported (b). High-income agents were assigned the higher 70% preference threshold; all other agents have 30%.*


*Low-income households*

*Med-income households*

*High-income households*

*Figure 41 - Global and local isolation index ($\tilde{Q}_m$) for income groups with 30%/70% preference thresholds after 20 cycles. The sum of the local $\tilde{Q}_m$ values as mapped is equal to the global $\tilde{Q}_m$ figure indicated above each map. As the calculation of $\tilde{Q}_m$ considers the overall proportions of each group in the region, it cannot be directly compared across groups, but can be compared to their baseline. High-income agents were assigned the higher 70% preference threshold; all other agents have 30%.*

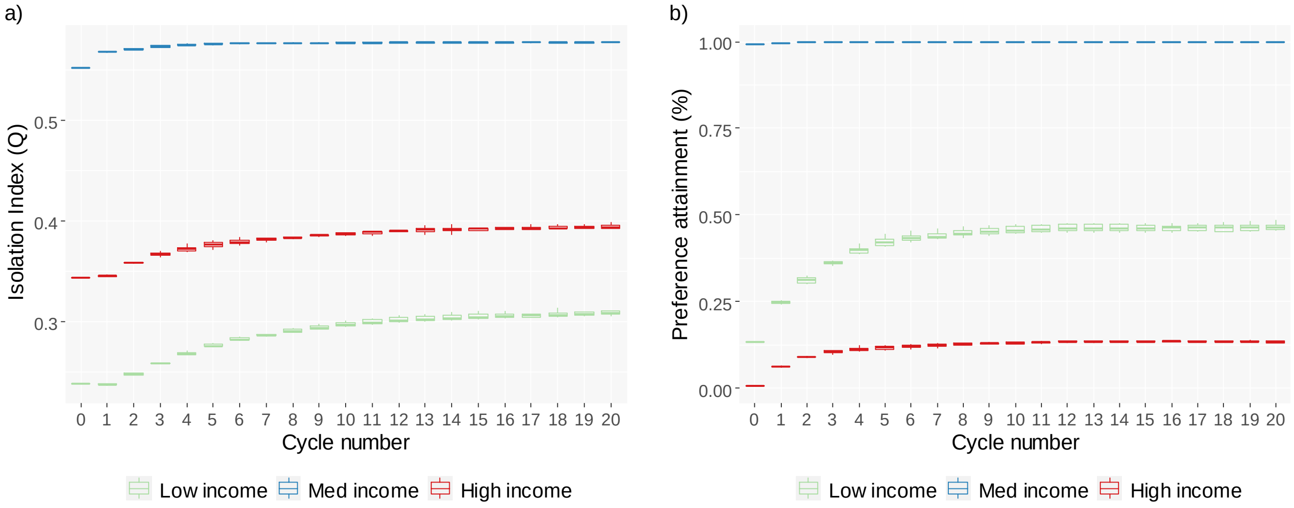


*Figure 42 - Evolution of the isolation index ($\tilde{Q}_m$) for income groups with 30%/70% preference thresholds (a) and the proportion of agents which are attaining their preference threshold for same-group neighbours (b). 10 simulations were run to account for the stochastic nature of the model, the aggregate results are presented here as a box plot. High-income agents were assigned the higher 70% preference threshold; all other agents have 30%.*

## *Education groups – Scenario C (30%/70%)*

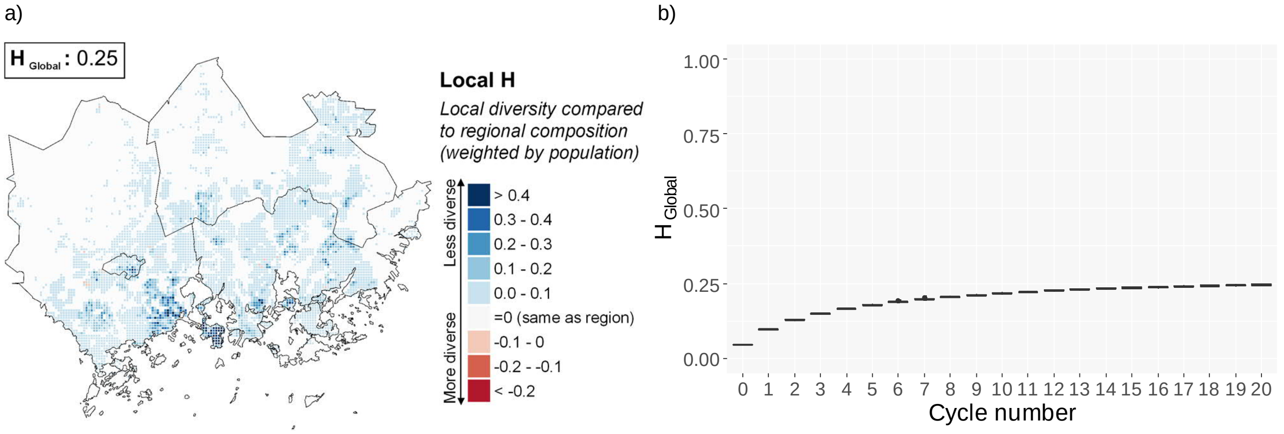
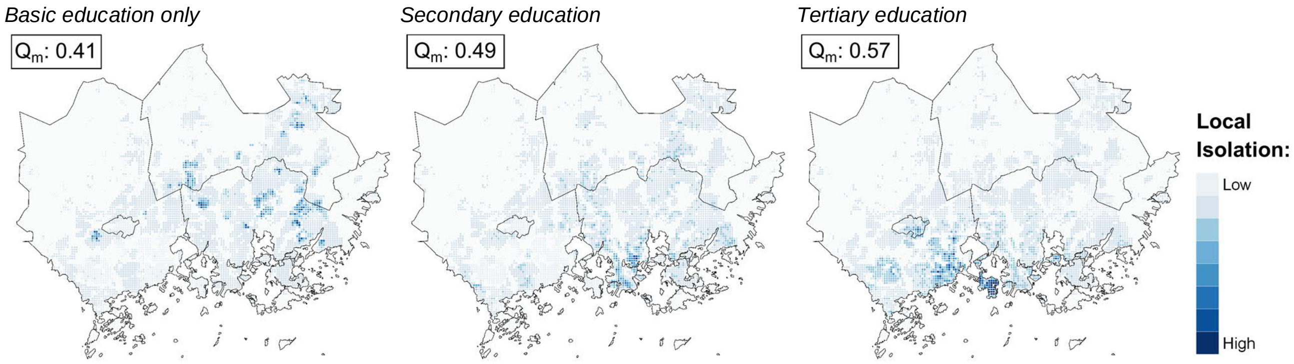


*Figure 43 - Information theory index ($\tilde{H}$) for education groups with 30%/70% preference thresholds after 20 cycles. The sum of the local $\tilde{H}$ values as mapped (a) is equal to the global $\tilde{H}$ value. Local values are multiplied by 1000 in the legend. Simulations were repeated 10 times and the evolution of the global $\tilde{H}$ values over the 20 cycles reported (b). Agents with a tertiary education were assigned the higher 70% preference threshold; all other agents have 30%.*


*Basic education only*

*Secondary education*

*Tertiary education*

*Figure 44 - Global and local isolation index ($\tilde{Q}_m$) for education groups with 30%/70% preference thresholds after 20 cycles. The sum of the local $\tilde{Q}_m$ values as mapped is equal to the global $\tilde{Q}_m$ figure indicated above each map. As the calculation of $\tilde{Q}_m$ considers the overall proportions of each group in the region, it cannot be directly compared across groups, but can be compared to their baseline. Agents with a tertiary education (highest level) were assigned the higher 70% preference threshold; all other agents have 30%.*

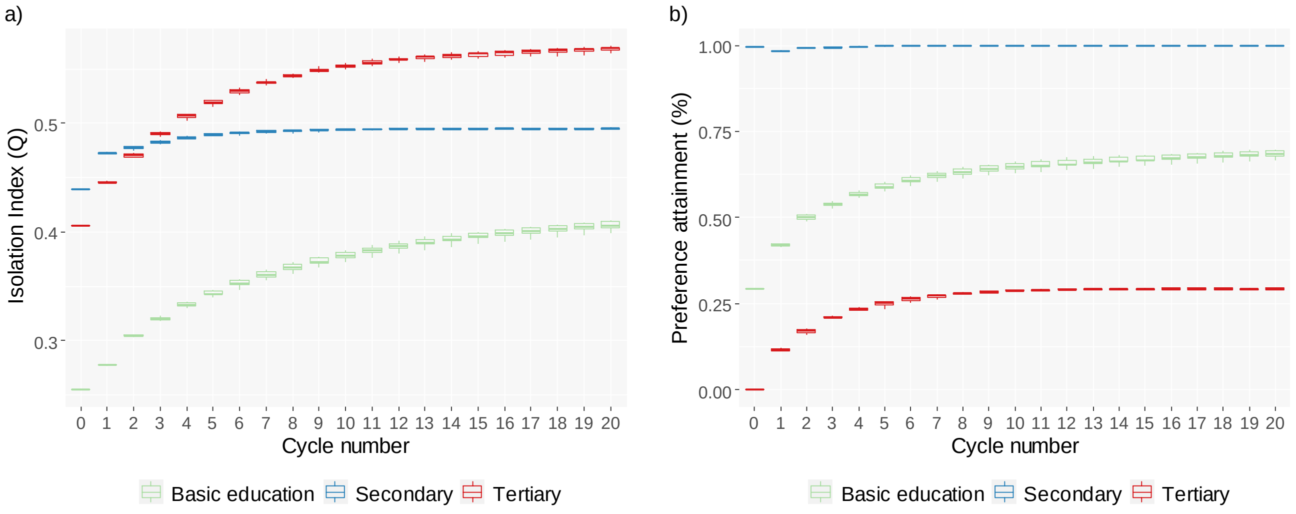


*Figure 45 - Evolution of the isolation index ($\tilde{Q}_m$) for education groups with 30%/70% preference thresholds (a) and the proportion of agents which are attaining their preference threshold for same-group neighbours (b). 10 simulations were run to account for the stochastic nature of the model, the aggregate results are presented here as a box plot. Agents with a tertiary education (highest level) were assigned the higher 70% preference threshold; all other agents have 30%.*

## *Language groups – Scenario C (30%/70%)*

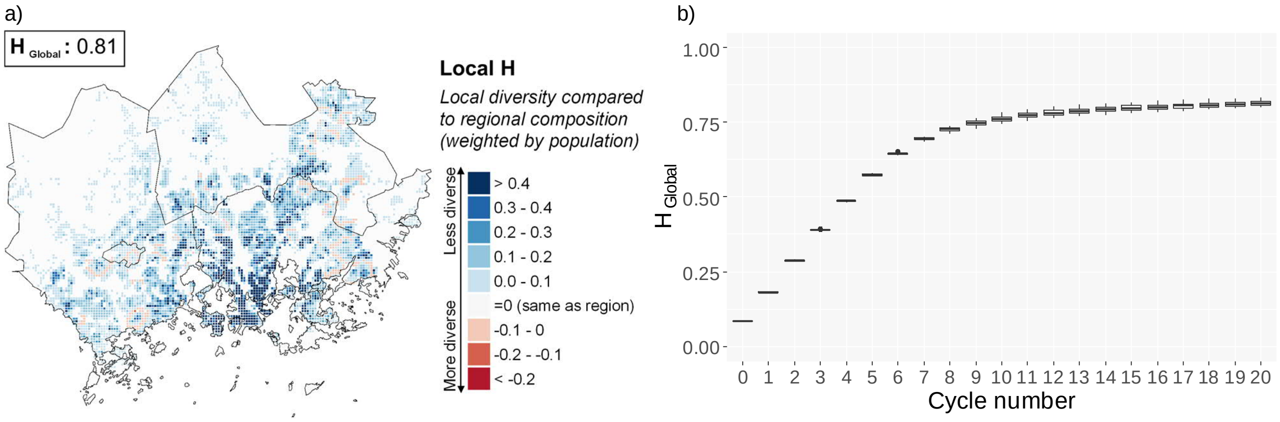


*Figure 46 - Information theory index ($\tilde{H}$) for language groups with 30%/70% preference thresholds after 20 cycles. The sum of the local $\tilde{H}$ values as mapped (a) is equal to the global $\tilde{H}$ value. Local values are multiplied by 1000 in the legend. Simulations were repeated 10 times and the evolution of the global $\tilde{H}$ values over the 20 cycles reported (b). Agents with Finnish mother tongue were assigned the higher 70% preference threshold; all other agents have 30%.*

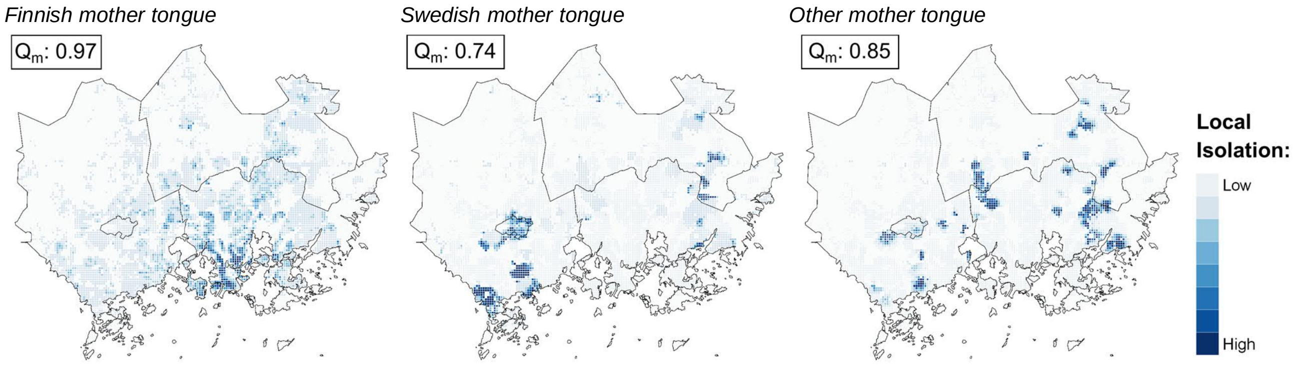


*Figure 47 - Global and local isolation index ($\tilde{Q}_m$) for language groups with 30%/70% preference thresholds after 20 cycles. The sum of the local $\tilde{Q}_m$ values as mapped is equal to the global $\tilde{Q}_m$ figure indicated above each map. As the calculation of $\tilde{Q}_m$ considers the overall proportions of each group in the region, it cannot be directly compared across groups, but can be compared to their baseline. Agents with Finnish as their mother tongue were assigned the higher 70% preference threshold; all other agents have 30%.*

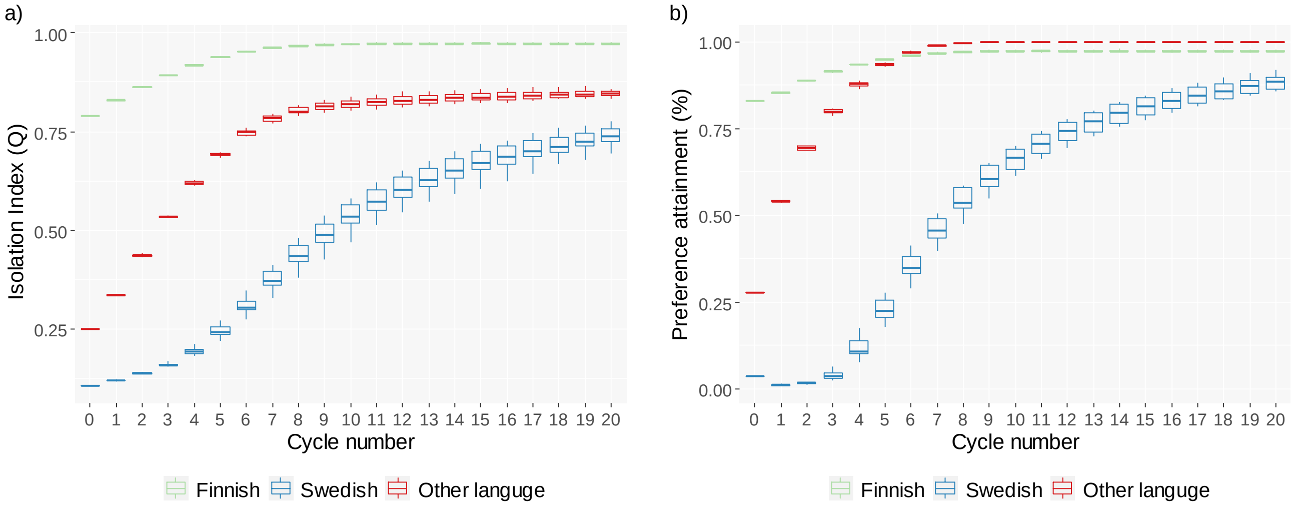


*Figure 48 - Evolution of the isolation index ($\tilde{Q}m$) for language groups with 30%/70% preference thresholds (a) and the proportion of agents which are attaining their preference threshold for same-group neighbours (b). 10 simulations were run to account for the stochastic nature of the model, the aggregate results are presented here as a box plot. Agents with Finnish as their mother tongue were assigned the higher 70% preference threshold; all other agents have 30%.*

# 8 Discussion

Schelling's early individual-based models revealed how highly segregated neighbourhoods can emerge from the uncoordinated mobility decisions of individual households. These results were achieved despite all residents being satisfied with living in mixed neighbourhoods to varying degrees. This study has sought to extend this basic model, to investigate whether similar results are achieved when the system being modelled is representative of a real urban area, both in terms of the physical urban structure, and the structure of its population. An empirically informed geospatial ABM for the HMA was developed for this purpose, which has the potential to test multiple scenarios concerning population dynamics in the region. Only one key parameter was tested for this study, that of preferences for living amongst like-neighbours. The results of the simulations run in this study have supported the potential for individual preferences and social distance dynamics to produce residential segregation, in the theoretical absence of any housing discrimination, institutional structures or concerted centralised organisation. Agent's mobility was only restricted to the extent that there must be vacant housing available in any chosen location. It is important to reiterate that these results only describe the functioning of the system within the modelling environment, and cannot be extrapolated to be explanative of the real processes at play in the HMA. There is no assertion that households acting on individual preferences is the major, nor even a significant force in the HMA. As a theoretical exercise, however, the results suggest that preferences and local self-organisation cannot be disregarded as factors in residential segregation. This study has additionally demonstrated the advantages of simulation, and the use of ABMs to investigate the possible dynamics of complex segregation processes. Such examination of the interplay between local and higher-level processes cannot be properly examined through traditional methods. The results demonstrate that inferring "micromotives" for the individaul households from the observed "macrobehavior" in such complex systems can be difficult (Schelling, 1978). The same is true for predicting macro outcomes by having an understanding of the rules of local interaction. Even with the extremely simplified system parameters of these simulations, non-linear behaviour can be observed.

The use of educational attainment, income and language groups as possible sources of homophily was derived from the literature review and suggestions of their role in residential sorting in European cities. However, these classifications also permitted testing of the impact of relative groups size on the sorting procedures. Group size became important in several ways. If groups have a preference for like-neighbours which is in excess of their relative size, the results of successfully acting on these preferences necessarily entails some degree of isolation of this group which can also reduce overall eveness of the region. With three groups of equal size, preferences of 30% are theoretically compatible with complete integration, however above 33.33% some degree of segregation must occur if preferences are fulfilled. The different group sizes within this study add additional layers to this equation. Increasing the intensity of preference for a certain percentage of co-group neighbours was demonstrated to be far from linear in its outcomes.

Whilst this study has broadly supported Schelling's original findings, the operational model departs from Schelling in several ways, with notable impacts. Firstly, Schelling continues the migration of agents until they are all satisfied with their neighbourhood or all possible neighbourhoods are exhausted. As the simulations conducted for this paper were terminated after 20 cycles, the system remains in a far from stable state at the end of many simulations. Extending the simulations over a much longer period would likely produce even more acute levels of segregation than were recorded.

One recurrent issue in the simulations was the inability of minority groups to find a suitable location when a majority group is evenly distributed throughout the region, a feature not found in Schelling's simulation, where asymmetrical group sizes resulted in higher levels of segregation. Several factors may be at play here. The realistic modelling environment may have been partly responsible. The initial conditions can have a large impact on the outcomes due to path dependency. In all of the simulations, the initial state is one of a very integrated population, reflecting the low levels of segregation in the HMA, and the level of vacancy is quite low (8%). This may make it more difficult for minority groups to find suitable neighbourhoods. Secondly, Schelling moved agents directly to suitable neighbourhoods, assuming perfect information about the neighbourhood composition of any chosen location. In this study, agents who decide to migrate move to a random location and do not consider their neighbourhood composition until they are called upon in the following cycle. The stochastic nature of the relocations particularly affects the ability of minority groups to find a suitable neighbourhood, leading to repeated migrations. Moving directly to a suitable neighbourhood, if it exists, would certainly have resulted in increased segregation. Some degree of randomness in residential choice is not inconsistent with real human behaviour and indeed this can be a strength of ABMs. Humans do not always behave in economically rational ways. Households may hastily purchase a property or make residential decisions that are formed on incomplete information. However, whilst randomness is inherent, parametrising randomness is difficult, and in the case of this model, the lack of any vetting processes by agents before accepting their new location created unrealistic behaviour and inefficiencies in neighbourhood selection. A final factor, which is present also in Schelling and many other ABM studies, is the simplistic coding of a neighbourhoods utility value. From an agent's perspective, a neighbourhood has a utility value of either 0 or 1. With a 30% preference threshold defined, an agent residing in a neighbourhood with 29% of the same group will leave their current location and migrate, even if the chances of finding a better location are statistically low. As the migrations in this model are random, there is a real chance the relocation may be to a neighbourhood with a lower proportion of co-group neighbours, particularly for minority groups.

### *Limitations and reflections on the model specification and operation*

Agents never consider more than one possible location at a time. They have already left their current location before they assess their new neighbourhood. This could be improved by allowing agents to sample multiple properties before choosing a new location, or at least comparing the new location to the existing location before deciding to move. Several studies have adopted such rule-sets, which prescribe, for example, that agents sample a fixed number of locations, and then use an ordering and ranking system to select the best fit (e.g. Ardestani et al., 2018; Benenson et al., 2002; Feitosa et al., 2011; Perez et al., 2019). All of these additional vetting processes necessarily add computing time to the migration process of each agent. Calculating the composition of 10 neighbourhoods may take 10 times as long as just one neighbourhood, for example. Due to the size of the model, and the fact that all simulations were conducted on a standard laptop computer, additional vetting processes were deemed too computationally expensive for the purpose of this study. The model had 100,000 agents and over 5,000 possible spaces for them to move to. Each simulation was run for 20 cycles, and repeated 10 times for three different grouping variables and three different scenarios. This represents 1,800 simulation cycles. Small additions in computing time for individual agents decision-making can thus quickly scale up.

One notable departure from reality in the parametrising of the model, which became apparent in the simulations for income groups most notably, was the assumption that all agents are of equal power in choosing locations, as preferences are the only decision-making parameter. This permitted exploration of dynamics in a theoretical system where such inequalities do not exist, however in reality some population groups regularly have more power, or choices available, than others. High-income agents with a desire to live in high-income neighbourhoods, for example, can be facilitated in realising this through competition-driven real-estate markets that allow them to outbid those with less economic resources. The types and structure of housing available can also be important. In the theoretical world of this model, spaces have no qualities other than their occupancy status. The price, quality or size of housing is not included for example, neither is proximity to amenities, nature, views etc. A few models have included a simple real-estate market, either by allocating spaces of different real estate value, or by using a supply and demand model. In this scenario, agents may have the additional attribute of an income level, which would determine their ability to move into a space. Costs may also be factored into agents decision-making considerations through utility values, such that higher income agents may see a higher utility value in living near other high-income agents despite areas being more ‘expensive’ which would not be shared by agents of other groups (Crooks, 2010, p. e.g.; Zhang, 2011). Similarly, if theories of households with high cultural capital preferring to cluster in more cosmopolitan areas were to be tested, proximity to cultural venues, for example, could be included as a parameter for the education groups.

Another likely factor in the apparent stability of certain population groups in some simulations is due to the system being closed in nature. There are no interactions with environments outside the system (for example, inter-regional or international migrations) and the population is static, in that agents are ‘immortal’. If a household is satisfied with their current location, they can occupy the same residence forever if the composition of their neighbourhood never decreases below their preference threshold. This was seen particularly acutely in the language group’s simulations for scenarios A and B (30% and 50% preferences), where the Finnish-speaking agents command such a large majority that they were all exceeding these preferences upon initialisation and had no reason to migrate at all during the simulations. This differs significantly from real population dynamics, which have regular residential turnover prompted by life-cycle events. This has been addressed in other studies (e.g. Ardestani et al., 2018; Crooks, 2010) by randomly selecting agents for forced migration, removing agents from the model once they reach a certain ‘age’, periodically creating new agents, and using empirical data on in and out migration to trigger changes each cycle.

The neighbourhood definition is a key parameter in ABMs, as with most types of spatial analysis. Definitions of what constitutes a neighbourhood, and how neighbourhoods are perceived by residents, are ongoing concerns for researchers. This model adopted a spatial proximity definition that is personalised from each agent’s respective location in space. The agents’ neighbourhood ‘vision’ was prescribed to be sufficiently large to include at least the immediately adjacent grid cells. This means that the neighbourhoods of agents are not commonly defined and overlap with one another. However, it may be that in some situations arbitrary administrative boundaries do become more important than proximity, if a socialised meaning has been imbued onto a postcode for example. Different neighbourhood definitions may result in vastly different outcomes at the macro scale. Geographical features may also play a role in shaping neighbourhood perception. Whilst the urban form and geographical footprint of the HMA was represented in the modelling environment, and only the inhabited areas of the city were included, the potential impact of barriers like highways and rivers on neighbourhood perception (see Crooks, 2010) are excluded. More broadly, the relative importance of the neighbourhood as a factor in residential decision-making may additionally be challenged. Mobility decisions in Schelling-type models assume that residents are concerned about their neighbours, and that residents make decisions having complete information about the composition of the neighbourhood in which they live. In different cultural and geographic contexts, the importance of having relations with neighbours or the role of the neighbourhood in structuring the lives of residents may be of higher or lower value.

### *Data limitations*

With any model, there are several limitations which must be acknowledged. Cities are complex entities, and models are always simplistic abstractions of reality. Any model is only as good as the data it are based upon. Despite the model being constructed using empirical data from a real urban area, the agents, environment and rule-set are all necessarily simplifications of the real systems. The availability of geospatial population data in the Helsinki Metropolitan Area, aggregated to a 250m x 250m grid in this case, provides quite accurate geographic precision over a large region. However, as the exact positions of agents within these grid cells is unknown, it was approximated using grid centroids in this model, which affects neighbourhood perceptions. There are also many assumptions required when converting any population to representative agents. Income groups are recorded at the household level, the most likely mobility unit in terms of migration; however, education level and mother tongue are both individual-level statistics without any possibility to reconstitute households. Even if other data would permit this, family members may have different educational or ethnic backgrounds and income can be a fluid measure, evolving over time. Educational attainment is only recorded for those over 18 years of age, which is a arguably a reasonable age for treating an individual as a potential migrating unit, however the language statistics are recorded for residents of all ages, including minors. Finally, whilst the education, language and income group of individuals in a grid cell are known, it is not possible to reconstruct a multi-dimensional individual by connecting these statistics due to data privacy. This prevents the possibility of analysing residential decisions that involve multiple homophilious sorting criteria.

### *Policy implications*

It is beyond the scope of this thesis to consider the implications of segregation produced through preferences and social distance dynamics. This study has not suggested that social distance dynamics is a major, or even important, force driving segregation in the Helsinki region, however there is evidence that it may play a role in some cities (e.g. Boterman et al., 2020). Such tendencies would create an apparent tension with the residential social mixing policies of egalitarian Nordic welfare states such as Finland. Whilst governments regularly intervene directly concerning socio-economic spatial inequalities and segregation, and often indirectly address ethnic segregation where associations between income and tenure type exist, imposing restrictions on social-cultural practices is unlikely (Boterman et al., 2020, p. 19).

### *Potential for further research*

Residential segregation is a complex socio-spatial process, involving many multi-dimensional individuals whose interactions contribute to shaping the global properties of the system. Simulation provides an avenue to explore this system through the construction of hypothetical experiments such as those conducted in this paper. Rather than merely describing or measuring patterns, simulation allows the study of the underlying dynamics and relationships between micromotives and macro outcomes. Barriers to the use of modelling as a research technique continue to fall, and simulation as a tool in the social sciences is gaining in popularity. In the 50 years since Schelling's models, social ABMs have evolved from manually moving pennies and dimes around a chessboard, to permitting autonomous vector-based simulation with millions of agents and few limits to possible parameters. The study only considered one potential factor in segregation dynamics; however, the developed model can be parametrised to include any number of additional decision-making parameters, macro conditions and housing practices. Indeed, there are few limits to how far agent based models can be parametrised (see, for example, Ardestani et al., 2018; Feitosa et al., 2011). As more variables and rules are introduced into the model, however, the ability to interpret why a particular outcome arises becomes diminished. Additionally, while models may be used for forecasting and predictive purposes, the complex nature of the system means that the mechanisms by which to steer the real-world system towards this outcome may remain unclear (Clarke, 2014). Ultimately, it is the simplicity of Schelling's model which provides its explanatory power. ABMs are perhaps most useful when conducting exploratory simulations such as in this study, making discrete changes to the local rules and observing the impact on the system.

The model could have benefits from more detailed individual-level data. Particularly when seeking to calibrate models with historical population movements, there is a need for longitudinal individual-level data that enables anonymised residents to be tracked throughout time and space.

Whilst data of this type does exist in some cities, access to the most detailed individual-level data often requires usage in controlled environments due to privacy regulations and have access methods that can provide barriers to their use for novel applications such as ABM research. As a key parameter in the study, the specification of preference thresholds could also have benefited from empirical studies and survey data on this topic from the HMA. Whilst several studies in the US have collected data on the stated preferences of different neighbourhood compositions, the author is not aware of any similar surveys that have been carried out in the Helsinki region.

# 9 Conclusions

Cities are complex entities, with socio-spatial structures that are constantly evolving as households migrate within and outside of their boundaries. The degree of residential segregation visible in today's cities is the product of the historical residential mobility of many individual households over many years. Households' residential choices are constrained at least by the state of supply and structure of the housing market, and the economic resources which they possess. Macro factors such as real estate markets, labour markets, housing policies and institutional discrimination in the housing market can thus play a large role in residential mobility and segregation dynamics. This study has sought to explore whether segregation may also arise as an unintended consequence of individual households making residential decisions entirely upon the basis of preferences.

Thomas Schelling was pioneering in using simple agent-based models to demonstrate how individual household's residential choices can lead to unanticipated and unexpected segregation in a city, even when this was not explicitly desired by any households. Despite increasing interest in ABM as a tool for the study of complex social phenomena such as segregation in recent years, there have been limited attempts to model a realistic urban area. This study presents a geospatial agent-based model, developed using empirical data from the Helsinki Metropolitan Area in Finland. The functioning of the model remains very simple to keep the processes transparent, however several features, not present in Schelling's early work were introduced into the model. The physical urban structure of the HMA, and the population composition, density and distribution of agents is represented in themodel through the inclusion of geospatial register data. Three variables representing potential sources of homophily in residential sorting were tested; namely, education level, income and language. By holding all parameters constant, except for the preference threshold for like-neighbours, the simulations were able to explore the effect that different preference structures, group sizes and initial conditions can have on segregation processes and their outcomes.

Whilst the results of this study make no claims that homophilious individual preferences are a dominant, or even important, force in driving segregation in the HMA, they do suggest that such uncoordinated local interactions are capable of producing segregation under certain theoretical conditions. Complex behavioural dynamics including self-organisation, emergence, nonlinearity and path dependence are all visible within the simulation results. These properties present significant challenges for traditional research methods. ABMs are one method which may help us interrogate these processes underlying complex social dynamics such as residential segregation.

# 10 Bibliography

Ærø, T. (2006). Residential Choice from a Lifestyle Perspective. *Housing, Theory and Society, 23*(2), 109–130.

Anderson, T., Leung, A., Perez, L., & Dragićević, S. (2020). Investigating the Effects of Panethnicity in Geospatial Models of Segregation. *Applied Spatial Analysis and Policy*.

Andersson, R., Brattbakk, I., & Vaattovaara, M. (2017). Natives' opinions on ethnic residential segregation and neighbourhood diversity in Helsinki, Oslo and Stockholm. *Housing Studies, 32*(4), 491–516.

Andersson, R., Musterd, S., & Galster, G. (2019). Port-of-Entry Neighborhood and Its Effects on the Economic Success of Refugees in Sweden. *International Migration Review*, *53*(3), 671–705.

Andrus, D. (2005). The Wiki and the Blog: Toward a Complex Adaptive Intelligence Community. *Studies in Intelligence*, *49*(3), 63–70.

Ardestani, B., O'Sullivan, D., & Davis, P. (2018). A multi-scaled agent-based model of residential segregation applied to a real metropolitan area. *Computers, Environment and Urban Systems, 69*, 1–16.

Bakens, J., & Pryce, G. (2019). Homophily horizons and ethnic mover flows among homeowners in Scotland. *Housing Studies*, *34*(6), 925–945.

Bar-Yam, Y. (2002). General features of complex systems. In *Encyclopedia of Life Support Systems (EOLSS)*. UNESCO EOLSS Publishers.

Barros, J., & Feitosa, F. F. (2018). Uneven geographies: Exploring the sensitivity of spatial indices of residential segregation. *Environment and Planning B: Urban Analytics and City Science, 45*(6), 1073–1089.

Benard, S., & Willer, R. (2007). A wealth and status-based model of residential segregation. *Journal of Mathematical Sociology*, *31*(2), 149–174.

Benenson, I., Hatna, E., & Or, E. (2009). From schelling to spatially explicit modeling of urban ethnic and economic residential dynamics. *Sociological Methods and Research*, *37*(4), 463–497.

Benenson, I., Omer, I., & Hatna, E. (2002). Entity-based modeling of urban residential dynamics: The case of Yaffo, Tel Aviv. *Environment and Planning B: Planning and Design*, *29*(4), 491–512.

Bernelius, V., & Vaattovaara, M. (2016). Choice and segregation in the 'most egalitarian' schools: Cumulative decline in urban schools and neighbourhoods of Helsinki, Finland. *Urban Studies, 53*(15), 3155–3171.

Bernelius, V., & Vilkama, K. (2019). Pupils on the move: School catchment area segregation and residential mobility of urban families. *Urban Studies*, *56*(15), 3095–3116.

Boterman, W. R., Musterd, S., & Manting, D. (2020). Multiple dimensions of residential segregation. The case of the metropolitan area of Amsterdam. *Urban Geography*, 1–26.

Bourdieu, P. (1984). *Distinction : a social critique of the judgement of taste* (P. Bourdieu (ed.)). Harvard University Press.

Bourdieu, P. (2005). *The social structures of the economy* (C. Turner (trans.); Issue 1). Polity Press.

Bruch, E. E. (2014). How Population Structure Shapes Neighborhood Segregation 1. *American Journal of Sociology*, *119*(5), 1221–1278.

Bruch, E. E., & Mare, R. D. (2006). Neighborhood choice and neighborhood change. *American Journal of Sociology*, *112*(3), 667–709.

Brun, J., & Fagnani, J. (1994). Lifestyles and Locational Choices—Trade-offs and Compromises: A Case-study of Middle-class Couples Living in the Ile-de-France Region. *Urban Studies*, *31*(6), 921–934.

Cheshire, P. (2012). Are mixed community policies evidence based? A review of the research on neighbourhood effects. In M van Ham, D. Manley, N. Bailey, L. Simpson, & D. Maclennan (Eds.), *Neighbourhood effects research: New perspectives* (1st ed., pp. 267–294). Springer Netherlands.

City of Helsinki. (2017). *The most functional city in the world–Helsinki City Strategy 2017–2021*. https://www.hel.fi/helsinki/en/administration/strategy/strategy/city-strategy/

City of Helsinki. (2019a). *Greater Helsinki's foreign-language population to double by year 2035*. https://www.hel.fi/uutiset/en/kaupunginkanslia/Greater-Helsinkis-foreign-language-population-to-double-by-year-2035

City of Helsinki. (2019b). *Helsingin seudun vieraskielisen väestön ennuste 2018–2035*. Helsingin kaupunginkanslian tilastojulkaisuja 2019:3. https://www.hel.fi/hel2/tietokeskus/julkaisut/pdf/19_03_14_Tilastoja_3_Vuori.pdf

City of Helsinki. (2019c). *Population with foreign background in Helsinki: Population with foreign background*. https://ulkomaalaistaustaisethelsingissa.fi/en/content/population-foreign-background

City of Helsinki. (2019d). *Population with foreign background in Helsinki: The current situation*. https://ulkomaalaistaustaisethelsingissa.fi/en/content/current-situation

City of Helsinki. (2020). *Helsinki population to exceed 700,000 in 2027*. https://www.hel.fi/uutiset/en/kaupunginkanslia/helsinki-population-in-2027

Clark, W. A. V. (1991). Residential Preferences and Neighborhood Racial Segregation: A Test of the Schelling Segregation Model. *Demography*, *28*(1), 1–19.

Clark, W. A. V. (2009). Changing Residential Preferences across Income, Education, and Age: Findings from the Multi-City Study of Urban Inequality. *Urban Affairs Review*, *44*(3), 334–355.

Clark, W. A. V, & Fossett, M. (2008). Understanding the social context of the Schelling segregation model. *Proceedings of the National Academy of Sciences of the United States of America*, *105*(11), 4109–4114.

Clarke, K. (2014). Cellular Automata and Agent-Based Models. In M. M. Fischer & P. Nijkamp (Eds.), *Handbook of Regional Science* (pp. 1217–1233). Springer Berlin Heidelberg.

Crooks, A. T. (2010). Constructing and implementing an agent-based model of residential segregation through vector GIS. *International Journal of Geographical Information Science*, *24*(5), 661–675.

Crooks, A. T., & Heppenstall, A. J. (2012). Introduction to Agent-Based Modelling. In A. J. Heppenstall, A. T. Crooks, L. M. See, & M. Batty (Eds.), *Agent-Based Models of Geographical Systems* (pp. 85–105). Springer Netherlands.

Dhalmann, H. (2013). Explaining Ethnic Residential Preferences-The Case of Somalis and Russians in the Helsinki Metropolitan Area. *Housing Studies*, *28*(3), 389–408.

Edin, P., Fredriksson, P., & Åslund, O. (2003). Ethnic Enclaves and the Economic Success of Immigrants: Evidence from a Natural Experiment. *The Quarterly Journal of Economics*, *118*(1), 329–357.

ESRI. (2020). *How Kernel Density works*. https://pro.arcgis.com/en/pro-app/latest/tool-reference/spatial-analyst/how-kernel-density-works.htm

Feitosa, F. F. (2010). *Urban Segregation as a Complex System: An Agent-Based Simulation Approach*. Dissertation. University of Bonn. Ecology and Development Series 70.

Feitosa, F. F., Câmara, G., Monteiro, A. M. V, Koschitzki, T., & Silva, M. P. S. (2007). Global and local spatial indices of urban segregation. *International Journal of Geographical Information Science*, *21*(3), 299–323.

Feitosa, F. F., Le, Q. B., & Vlek, P. L. G. (2011). Multi-agent simulator for urban segregation (MASUS): A tool to explore alternatives for promoting inclusive cities. *Computers, Environment and Urban Systems*, *35*(2), 104–115.

Finnish National Agency for Education. (2020). *Finnish Vocational Education and Training*. https://www.oph.fi/en/education-system/finnish-vocational-education-and-training

Fossett, M. (2006). Ethnic preferences, social distance dynamics, and residential segregation: Theoretical explorations using simulation analysis. *Journal of Mathematical Sociology*, *30*(3–4), 185–273.

Fossett, M., & Waren, W. (2005). Overlooked implications of ethnic preferences for residential segregation in agent-based models. *Urban Studies*, *42*(11), 1893–1917.

Hall, A., & Virrantaus, K. (2016). Visualizing the workings of agent-based models: Diagrams as a tool for communication and knowledge acquisition. *Computers, Environment and Urban Systems*, *58*, 1–11.

Hanquinet, L., Savage, M., & Callier, L. (2012). Elaborating Bourdieu's Field Analysis in Urban Studies: Cultural Dynamics in Brussels. *Urban Geography*, *33*(4), 508–529.

Hasanzadeh, K., Kyttä, M., & Brown, G. (2019). Beyond Housing Preferences: Urban Structure and Actualisation of Residential Area Preferences. *Urban Science*, *3*(1), 21.

Hedman, L., van Ham, M., & Manley, D. (2011). Neighbourhood Choice and Neighbourhood Reproduction. *Environment and Planning A: Economy and Space*, *43*(6), 1381–1399.

HSY. (2020). *SeutuData'19*. https://www.hsy.fi/en/air-quality-and-climate/geographic-information/seutudata/

Huynh, Y. (2019). *The "second generation" as actors of desegregation in Finnish cities? A study of the residential mobility patterns of the descendants of immigrants in Helsinki*. Master's Thesis. University of Helsinki.

Iceland, J. (2004). The multigroup entropy index (also known as Theil's H or the information theory index). In *US Census Bureau*. http://www.census.gov/housing/patterns/about/multigroup_entropy.pdf

Jayaprakash, C., Warren, K., Irwin, E., & Chen, K. (2009). The interaction of segregation and suburbanization in an agent-based model of residential location. *Environment and Planning B: Planning and Design*, *36*(6), 989–1007.

Karsten, L. (2007). Housing as a Way of Life: Towards an Understanding of Middle-Class Families' Preference for an Urban Residential Location. *Housing Studies*, *22*(1), 83–98.

Kauppinen, T. M. (2007). Neighborhood effects in a European city: Secondary education of young people in Helsinki. *Social Science Research*, *36*(1), 421–444.

Kauppinen, T. M., & van Ham, M. (2019). Unravelling the demographic dynamics of ethnic residential segregation. *Population, Space and Place*, *25*(2).

Kukk, K., van Ham, M., & Tammaru, T. (2019). EthniCity of Leisure: A Domains Approach to Ethnic Integration During Free Time Activities. *Tijdschrift Voor Economische En Sociale Geografie*, *110*(3), 289–302.

Laurie, A. J., & Jaggi, N. K. (2003). Role of “vision” in neighbourhood racial segregation: A variant of the schelling segregation model. *Urban Studies*, *40*(13), 2687–2704.

Lee, J.-S., Son, W.-J., Lee, H.-T., & Cho, I.-S. (2020). Verification of Novel Maritime Route Extraction Using Kernel Density Estimation Analysis with Automatic Identification System Data. In *Journal of Marine Science and Engineering* (Vol. 8, Issue 5).

Marcińczak, S., Tammaru, T., Strömgren, M., & Lindgren, U. (2015). Changing patterns of residential and workplace segregation in the Stockholm metropolitan area. *Urban Geography*, *36*(7), 969–992.

Massey, D. S., & Denton, N. A. (1988). The dimensions of residential segregation. *Social Forces*, *67*(2), 281–315.

Massey, D. S., & Denton, N. A. (1993). *American apartheid : segregation and the making of the underclass*. Harvard University Press.

McPherson, M., Smith-lovin, L., & Cook, J. M. (2001). Birds of a feather: Homophily in social networks. *Annual Review of Sociology*, *27*, 415–444.

Mulder, C. H. (2007). The family context and residential choice: A challenge for new research. *Population, Space and Place*, *13*(4), 265–278.

Musterd, S., van Gent, W. P. C., Das, M., & Latten, J. (2016). Adaptive behaviour in urban space: Residential mobility in response to social distance. *Urban Studies*, *53*(2), 227–246.

Nordregio. (2017). *Foreign-born inhabitants as a share of total population in 2016.* https://nordregio.org/maps/foreign-born-inhabitants-as-a-share-of-total-population-in-2016/

O’Sullivan, D. (2009). Changing neighborhoodsĝ€"neighborhoods changing: A framework for spatially explicit agent-based models of social systems. *Sociological Methods and Research*, *37*(4), 498–530.

O’Sullivan, D., MacGill, J., & Yu, C. (2003). Agent-based residential segregation: a hierarchically structured spatial model. *Proceedings of Agent 2003 Conference on Challenges in Social Simulation*, 493–507.

Odland, T. (2020). *KDEpy*. https://github.com/tommyod/KDEpy

OECD. (2018). *Divided Cities Understanding Intra-urban Inequalities*. OECD.

Official Statistics of Finland (OSF). (2019). *Population structure 2019 [e-publication]. Appendix table 2. Population according to language 1980–2019* . Helsinki: Statistics Finland. https://tilastokeskus.fi/til/vaerak/2019/vaerak_2019_2020-03-24_tau_002_en.html

Official Statistics of Finland (OSF). (2020). *Population structure 2019 [e-publication]*. Helsinki: Statistics Finland. http://www.stat.fi/til/vaerak/tau_en.html

Park, R. E. (1926). The urban community as a spatial pattern and a moral order. In E. W. Burgess (Ed.), *The Urban Community* (pp. 3–20). University of Chicago Press.

Park, Y. M., & Kwan, M. P. (2018). Beyond residential segregation: A spatiotemporal approach to examining multi-contextual segregation. *Computers, Environment and Urban Systems*, *71*, 98–108.

Perez, L., Dragicevic, S., & Gaudreau, J. (2019). A geospatial agent-based model of the spatial urban dynamics of immigrant population: A study of the island of Montreal, Canada. *PLoS ONE*, *14*(7).

Reardon, S. F., & Bischoff, K. (2011). Income Inequality and Income Segregation. *American Journal of Sociology*, *116*(4), 1092–1153.

Reardon, S. F., & O'Sullivan, D. (2004). Measures of Spatial Segregation. *Sociological Methodology*, *34*, 121–162.

Sawyer, R. K. (2005). *Social Emergence : Societies As Complex Systems*. Cambridge University Press.

Schelling, T. C. (1969). Models of Segregation. *The American Economic Review*, *59*(2), 488–493.

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

Schelling, T. C. (1972). A Process of Residential Segregation: Neighborhood Tipping. In A. Pascal (Ed.), *Racial Discrimination in Economic Life* (pp. 157–184). Lexington Books.

Schelling, T. C. (1978). *Micromotives and macrobehavior*. W. W. Norton Company.

Skifter Andersen, H. (2015). Explanations for special neighbourhood preferences among ethnic minorities. *Housing, Theory and Society*, *32*(2), 196–217.

Statistics Finland. (2018). *Numbers and income of dwelling population and household-dwelling units by municipality in 1995-2017*. http://pxnet2.stat.fi/PXWeb/pxweb/en/StatFin_Passiivi/StatFin_Passiivi__zzz__tjkt/statfin_tjkt_pxt_001.px/table/tableViewLayout1/

Statistics Finland. (2019a). *Grid Database 2019: User Manual*. https://www.tilastokeskus.fi/static/media/uploads/tup/ruututietokanta/rttk2019_kuvaus_en.pdf

Statistics Finland. (2019b). *Grid Database 2019*. https://www.stat.fi/tup/ruututietokanta/index_en.html

Statistics Finland. (2019c). *Väestötilastopalvelu*. https://www.stat.fi/tup/tilastotietokannat/index.html

Statistics Finland. (2020). *Asuntokunnat henkilöluvun ja talotyypin mukaan 31.12.* https://api.aluesarjat.fi/pxweb/fi/Helsingin seudun tilastot/Helsingin seudun tilastot__Helsingin seutu__Asuminen__Asuntokunnat/Hginseutu_AS_AK01_Asuntokunnat_talot_hloluku.px/

Taillandier, P., Gaudou, B., Grignard, A., Huynh, Q., Marilleau, N., Caillou, P., Philippon, D., & Drogoul, A. (2019). Building, composing and experimenting complex spatial models with the GAMA platform. *GeoInformatica*, *23*(2), 299–322.

Theil, H. (1972). *Statistical decomposition analysis. With applications in the social and administrative sciences*. North-Holland Publishing Company.

Vaalavuo, M., van Ham, M., & Kauppinen, T. M. (2019). Income Mobility and Moving to a Better Neighbourhood: An Enquiry into Ethnic Differences in Finland. *European Sociological Review*, *35*(4), 538–551.

Vaattovaara, M., Joutsiniemi, A., Kortteinen, M., Stjernberg, M., & Kemppainen, T. (2018).

Experience of a Preventive Experiment: Spatial Social Mixing in Post-World War II Housing Estates in Helsinki, Finland. In D. B. Hess, T. Tammaru, & M. van Ham (Eds.), *Housing Estates in Europe: Poverty, Ethnic Segregation and Policy Challenges* (pp. 215–240). Springer, Cham.

Vaattovaara, M., & Kortteinen, M. (2003). Beyond Polarisation versus Professionalisation? A Case Study of the Development of the Helsinki Region, Finland. *Urban Studies*, *40*(11), 2127–2145.

van Diepen, A. M. L., & Musterd, S. (2009). Lifestyles and the city: Connecting daily life to urbanity. *Journal of Housing and the Built Environment*, *24*(3), 331–345.

van Gent, W., Das, M., & Musterd, S. (2019). Sociocultural, economic and ethnic homogeneity in residential mobility and spatial sorting among couples. *Environment and Planning A*, *51*(4), 891–912.

van Ham, Maarten, & Tammaru, T. (2016a). *Increasing Socio-Economic Segregation in European Cities: East meets West*. Urbanisation, mobilities and immigration (URMI) Policy Brief, November 2016.

van Ham, Maarten, & Tammaru, T. (2016b). New perspectives on ethnic segregation over time and space. A domains approach. *Urban Geography*, *37*(7), 953–962.

Vasanen, A. (2012). Beyond stated and revealed preferences: the relationship between residential preferences and housing choices in the urban region of Turku, Finland. *Journal of Housing and the Built Environment*, *27*(3), 301–315.

White, M. J. (1983). The Measurement of Spatial Segregation. *American Journal of Sociology*, *88*(5), 1008–1018.

White, M. J. (1986). Segregation and Diversity Measures in Population Distribution. *Population Index*, *52*(2), 198–221.

Xie, Y., & Zhou, X. (2012). Modeling individual-level heterogeneity in racial residential segregation. *Proceedings of the National Academy of Sciences of the United States of America*, *109*(29), 11646–11651.

Yin, L. (2009). The dynamics of residential segregation in Buffalo: An agent-based simulation. *Urban Studies*, *46*(13), 2749–2770.

Yinger, J. (1995). *Closed Doors, Opportunities Lost*. Russell Sage Foundation.

Zhang, J. (2011). Tipping and residential segregation: A unified schelling model. *Journal of Regional Science*, *51*(1), 167–193.

# 11 Acknowledgements

This thesis emerged from a joint project for the complex adaptive systems course run by Aalto University. I owe a special thanks to Babak Firoozi Fooladi who was instrumental in the early development of the GAMA model which later became the backbone of this project. Thank you to my supervisors, Mari Vaattovaara, from University of Helsinki, and Anssi Joutsiniemi, from Aalto University, for encouraging me to pursue this work as a thesis topic, and for the constructive discussions and guidance. I also need to thank Katja Vilkama, Research Director at the City of Helsinki, for her support for this work, and particularly in assisting with sourcing the data for the model. This thesis benefited at an early stage from many insightful comments from my friends and peers within the Urban Studies and Planning Master's programme for which I am very grateful. Finally, a big thank you to my friends and family for their emotional support, and for pushing me to get this thesis finished. This thesis was largely written from home under Covid-19 restrictions, which has not been the easiest task.

# 12 Appendices

## Appendix 1 – Major language groups in the Helsinki Metropolitan Area (OSF, 2020)

| | **Espoo** | **Helsinki** | **Kauniainen** | **Vantaa** | **HMA** | ***HMA %*** |
|---|---|---|---|---|---|---|
| **Official languages** | | | | | | |
| Finnish | 217,486 | 511,043 | 5,770 | 180,904 | 915,203 | *77.1%* |
| Swedish | 20,033 | 36,665 | 3,197 | 5,575 | 65,470 | *5.5%* |
| Total official languages | 237,519 | 547,708 | 8,967 | 186,479 | 980,673 | *82.6%* |
| **Foreign languages** | | | | | | |
| Russian | 7,409 | 18,869 | 73 | 8,673 | 35,024 | *3.0%* |
| Estonian | 5,861 | 10,620 | 117 | 8,354 | 24,952 | *2.1%* |
| Somali | 2,874 | 11,466 | - | 2,874 | 17,214 | *1.5%* |
| Arabic | 4,335 | 7,963 | 60 | 3,572 | 15,930 | *1.3%* |
| English | 3,347 | 7,121 | 98 | 1,765 | 12,331 | *1.0%* |
| Chinese | 2,939 | 3,820 | 73 | 1,142 | 7,974 | *0.7%* |
| Kurdish | 1,877 | 3,580 | 34 | 1,578 | 7,069 | *0.6%* |
| Persian, Farsi | 2,098 | 3,139 | 73 | 1,272 | 6,582 | *0.6%* |
| Albanian | 2,210 | 1,588 | 11 | 2,677 | 6,486 | *0.5%* |
| Vietnamese | 1,307 | 2,559 | - | 1,713 | 5,579 | *0.5%* |
| Other | 17,955 | 35,402 | 291 | 13,676 | 67,324 | *5.7%* |
| Total foreign languages | 52,212 | 106,127 | 830 | 47,296 | 206,465 | *17.4%* |
| **Total residents** | **289,731** | **653,835** | **9,797** | **233,775** | **1,187,138** | ***100.0%*** |