Measuring activity-based Social Segregation using Public Transport Smart Card Data

M.Sc. TIL Thesis

Lukas Kolkowski

TU Delft

MEASURING ACTIVITY-BASED SOCIAL SEGREGATION USING PUBLIC TRANSPORT SMART CARD DATA

A thesis submitted to the Delft University of Technology in partial fulfillment of the requirements for the degree of

Master of Science in Transport, Infrastructure and Logistics

by

Lukas Kolkowski

November 2021

Lukas Kolkowski: *Measuring activity-based Social Segregation using Public Transport Smart Card Data* (2021)

The work in this thesis was performed at
The Faculty of Civil Engineering and Geosciences, Delft University of Technology

In collaboration with
KTH Royal Institute of Technology & AB Storstockholms Lokaltrafik (SL)

An electronic version of this thesis is available at repository.tudelft.nl
Cover photo: *Citybanan tunnel construction*. SL

Supervisors:	Dr.ir. O. (Oded) Cats	Delft University of Technology
	Dr.ir. T. (Trivik) Verma	Delft University of Technology
	Dr. E. (Erik) Jenelius	KTH Royal Institute of Technology
	M. (Malvika) Dixit	Delft University of Technology

PREFACE

A bit more than two years of inspiring lectures and projects at various faculties and my personal interest in public transport led me to this research topic. To round off my M.Sc. studies with this thesis fits the interdisciplinary concept of my "Transport, Infrastructure & Logistics" master. I am grateful for the excellent education I received at TU Delft and will strive to use it for the benefit of society.

I would like to thank my thesis committee members for their support on this thesis project. First of all, I thank Malvika Dixit for her extensive support, feedback, and listening ear. I am also very grateful for the support in data analysis and her advice during the writing of my thesis. Further, I appreciate the wide-ranging advice and guidance of Oded Cats. I also thank you for the impulse towards connecting public transport-related and socially-relevant research fields for my thesis project. I appreciate Trivik Verma for his advice on research outlining, data analysis, and visualization. I would also like to thank Erik Jenelius for his conceptual feedback and advice along the way.

I would like to thank Matej Cebecauer from KTH for his great support with the data setup. You were a great help and I really appreciate your flexibility and willingness to answer my questions. Further, Isak Jarlebring Rubensson from the transport administration of Stockholm was a great help in putting my study results into perspective, thank you very much.

Finally, I would like to express my gratitude to all those who have supported me during this time. Friends and family have supported me in a warm-hearted way and made it easy for me to turn my thoughts away once in a while.

Lukas Kolkowski
Rotterdam, November 2021

CONTENTS

1 INTRODUCTION 1
1.1 Problem statement & research gap . . . 1
1.2 Research questions . . . 4
1.3 Document structure . . . 4
2 LITERATURE REVIEW – ACTIVITY-BASED SOCIAL SEGREGATION 7
2.1 Residential segregation . . . 7
2.2 Activity-based segregation . . . 9
2.3 Segregation in relation to mobility and public transport . . . 11
2.3.1 Social segregation, transport disadvantage and public transport 12
2.3.2 Public transport smart card data . . . 13
2.4 Segregation measures . . . 16
2.5 Measuring activity-based social segregation using mobility data . . . 18
3 METHODOLOGY – MEASURING SOCIAL SEGREGATION USING ACTIVITY-BASED MOBILITY DATA 21
3.1 Framework . . . 21
3.2 Modeling steps . . . 22
3.2.1 Requirements for residential social groups . . . 22
3.2.2 Mobility data requirements . . . 23
3.2.3 Connecting residential social groups to mobility data . . . 23
3.2.4 Measuring segregation from mobility data . . . 24
3.3 Possible applications - measuring the evolution of segregation . . . 24
3.4 Ordinal information theory index . . . 25
4 MEASURING SOCIAL SEGREGATION WITH STOCKHOLM COUNTY'S PUBLIC TRANSPORT DATA 29
4.1 Segregation and mobility in Stockholm County . . . 29
4.1.1 Stockholm County's public transport . . . 29
4.1.2 The segregation impact of the new commuter train tunnel . . . 31
4.2 Choice of social groups and geographical scale . . . 32
4.3 Connecting social groups to public transport smart card data . . . 34
4.3.1 Data collection, inference and cleaning . . . 34
4.3.2 Connecting income groups of Stockholm to PT data . . . 36
5 RESULTS AND DISCUSSION 41
5.1 Temporal segregation analysis . . . 41
5.2 Zonal segregation analysis . . . 42
5.3 Analyzing the evolution of segregation . . . 43
5.4 Segregation changes in focus areas . . . 46
5.5 Discussion . . . 49
5.6 Limitations . . . 52
6 CONCLUSIONS 55
6.1 Answering the research questions . . . 55
6.2 Recommendations . . . 56
6.2.1 Application recommendations . . . 56
6.2.2 Future studies . . . 57
Bibliography 59
A APPENDIX A: SCIENTIFIC PAPER 65
B APPENDIX B: SCD JOURNEY SUCCESS RATES 73
C APPENDIX C: WEIGHTED AND ABSOLUTE SEGREGATION CONTRIBUTIONS 2018-2020 75
D APPENDIX D: SIGNIFICANCE TESTS 77

LIST OF FIGURES

Figure 2.1 Dimensions of segregation studies 10
Figure 2.2 Inferences towards observed SCD journeys with home stations 14
Figure 3.1 Framework for measuring social segregation by connecting mobility data to socioeconomic data 22
Figure 3.2 Distribution function f 27
Figure 4.1 Rail-bound PT in Stockholm County 30
Figure 4.2 Citybanan tunnel Stockholm 31
Figure 4.3 Residential income quantiles in Stockholm County DeSo zones 34
Figure 4.4 Stop area assignment - bus station Hallandsgatan 36
Figure 4.5 Public transport destinations 2020 37
Figure 4.6 Income group's PT destinations 2020 38
Figure 4.7 Differences in numbers of 2020 PT journey destinations between income group 1 and 4 39
Figure 5.1 Segregation index levels over all days of week 5 41
Figure 5.2 Weighted and absolute segregation contribution 2017 42
Figure 5.3 Segregation changes 2017-2018 44
Figure 5.4 Distribution of Segregation contribution changes 2017-2018 . 44
Figure 5.5 Segregation changes 2017-2020 45
Figure 5.6 Distribution of Segregation contribution changes 2017-2020 . 45
Figure 5.7 Stockholm city center segregation changes 46
Figure 5.8 Segregation changes along the northwest corridor 47
Figure 5.9 Segregation change in northwest focus areas 48
Figure 5.10 Segregation change in east/southwest focus areas 49
Figure C.1 Weighted and absolute segregation contribution 2018 75
Figure C.2 Weighted and absolute segregation contribution 2019 75
Figure C.3 Weighted and absolute segregation contribution 2020 75

LIST OF TABLES

Table 2.1	Activity-based segregation and activity-space studies	11
Table 2.2	Multi-group segregation indices	17
Table 4.1	Income quantiles in Stockholm County	33
Table 4.2	2020 PT Journeys per income group	38
Table B.1	Home zone and destination inference success rates	73
Table D.1	t-test on zone's weighted segregation contribution 2018 compared to 2017	77
Table D.2	t-test on zone's weighted segregation contribution 2020 compared to 2017	77

ACRONYMS

DeSo Demografiska statistikområden (Demographic statistics areas)

GPS Global Positioning System

ID identification number

K 1,000

PT public transport

SAN Stop Area Number

SCD smart card data

SL Storstockholms Lokaltrafik

1 INTRODUCTION

Social segregation is the spatial, temporal and access-related distance between individuals and groups with different social backgrounds. A well-known example of social segregation is the uneven concentration of rich and poor in urban areas (Paddison and Hamnett, 2000). This study focuses on measuring segregation of social groups using activity-based data. The context in which this study is embedded is outlined below.

1.1 PROBLEM STATEMENT & RESEARCH GAP

Social segregation can cause and enlarge inequality and thereby jeopardize reaching the United Nations' 17 goals of sustainable development (United Nations, 2020). Segregation often leads to disparities in essential living conditions (Leonard, 1987; Acevedo-Garcia and Lochner, 2003; Marques, 2012). The resulting disparities in access to necessities lead to further extremes such as poverty. Furthermore, segregation can cause isolation limiting contact, communication, and social relations (Hunt and Walker, 1974). Current trends indicate widening gaps within societies whereof chances for equality are less and society is driven apart from each other. Less mixing implies more separation and thus often exclusion.

Segregation is hard to capture since it is individually experienced, partly unnoticed, and quantifying it requires considerable efforts (see e.g. Zhang and Zheng (2015)). A better understanding of segregation is needed to mitigate its effects. In addition, capturing segregation and disentangling its causes facilitates the analysis of policy and infrastructure changes. Mitigating the effects of social segregation is a major challenge of today's society and its policymakers, while capturing it remains a hard-to-grasp phenomenon and a much-discussed research topic.

Since the early 20th century, researchers have tried to capture segregation, usually by measuring it with indices (Bell, 1954; Theil and Finizza, 1971; James and Taeuber, 1985; Massey and Denton, 1988; Reardon and Firebaugh, 2002). Many efforts have been devoted to exploring different types of segregation such as by nationality, race, ethnicity, religion, gender as well as economic and social status. In this study social segregation is focused exclusively.

Social segregation can be defined as the uneven spatial distribution of social groups according to Le Roux et al. (2017). Further, social segregation can be determined from differences in the temporal usage of space and access to amenities and services for different social groups. These social groups are usually inhabitants of a study area classified by socially relevant characteristics such as income, age, educational level, and migration background.

Until recently, spatial segregation of social groups was measured using segregation indices applied on mostly residential socioeconomic data (Bischoff and Reardon, 2014). These data sets are generated at the residence of people for instance with every census. Some studies centered on other socio-geographical spaces such as places of work, shopping, or leisure. Main findings indicate that income,

educational level, housing types as well as spatial distance between groups are key drivers for segregation (Tan et al., 2019; United Nations, 2020).

Nevertheless, focusing on just one socio-geographical space such as residence or place of work leads to a static view on the uneven spatial distribution of social groups (Wong and Shaw, 2011; Xu et al., 2019). Naturally, human life can not only be described by residential characteristics. There is a common understanding that people tend to be activity-driven and therefore move. A key part of activities is mobility since it represents individual choices of moving and thereby records the dynamics of daily life (Galiana and Sakarovitch, 2020). Using only static data can lead to a partial view and capturing only fractions of social segregation.

As a consequence, recent studies utilize activity-based data to measure segregation (e.g. Farber et al. (2015)). These studies integrate an activity perspective by using data from social networks, mobile phones, Global Positioning System (GPS), travel surveys, or diaries to measure segregation (Ureta, 2008; Silm and Ahas, 2014; Tan et al., 2019; Tao et al., 2020). By doing so researchers were able to cover more than just one socio-geographical space using activity-based data (Wong and Shaw, 2011).

Often, mobility data is used to measure such activity-based segregation since there is a clear link between segregation and the so-called transport disadvantage. Transport disadvantage incorporates the lack of access to transportation and is a cause for segregation (Church et al., 2000; Currie et al., 2010). A crucial role in these missing opportunities can be linked to public transport. As a public service, it is supposed to provide access to infrastructure, service and amenities for everybody. Among others, Kaufmann (2004) found that segregation can be caused by the public transport system design.

Activity-based segregation studies often rely on the aforementioned data sources. Commonly used mobility data sources such as GPS, mobile phone, or travel survey data can include accuracy, privacy, and availability issues, as well as incomplete data sets (Bagchi and White, 2005; Pelletier et al., 2011). In addition, it can require immense efforts and high costs to obtain sufficient data sets from data sources such as travel surveys and diaries.

Therefore, more attention is given to public transport data, which have two originally independent functions, both of which can be related to the measurement of segregation. Observed public transport data provides valuable mobility traces which can be used to measure activity-based segregation. At the same time, public transport should facilitate access and diminish barriers between social groups. More equal access should then lead to less segregation.

Research in the public transport sector currently focuses on broadening the use of public transport smart card data, a well-established chip card technology for fare collection (Pelletier et al., 2011). Since public transport has a key role in people's daily dynamics, especially in urban environments, its data might be key to overcome other data sources' shortcomings. Smart cards offer unprecedented large data sets of real transactions, thus observed travel data (Utsunomiya et al., 2006). It unravels mobility and travel patterns and hence could offer an activity-based perspective on social segregation. Although smart card data is restricted to public transport users' activities within the existing system's options, it offers a unique data source for measuring mobility-based segregation.

Combining socially relevant data, as used for residential segregation studies, with activity-based mobility data could combine the strengths of the two currently prevalent approaches to segregation measurement. While socio-demographic or

socioeconomic data can be used to distinguish social groups, mobility data reveals activity patterns and the resulting mix of groups.

In addition, there is a need of concurrently measuring multiple social groups' segregation between each other. The segregation studies originally conducted, comparing only two groups, are insufficient because of the different layers and complexity of today's society. Therefore, so-called multi-groups segregation measurement is focused.

This begs the question of how large-scale disaggregate mobility data, such as smart card data, could be used to measure activity-based multi-group segregation. Based on the achievements of earlier studies this suggests a multi-group segregation study using both mobility data and residential socioeconomic data.

So far, only Abbasi et al. (2021) measured two-group and multi-group social segregation using public transport smart card data. This study successfully proved how mobility traces obtained from smart card data facilitate assessing multi-group, mobility-based social segregation.

The study of Abbasi et al. (2021) was able to extract social characteristics and thereby created social groups from the smart card data itself. For many transport authorities and countries, this kind of personal information would not be available or extracting it would raise data privacy concerns. As a result, social information often cannot be retrieved directly from smart cards. In addition, even richly equipped smart cards often do not contain the desired social information.

As a result, there is a lack of a method to measure activity-based multi-group social segregation on disaggregate large-scale mobility data sets such as smart card data, when the required social information cannot be extracted from the mobility data. Mobility data such as smart card data has been linked to residential socioeconomic data before, just not specifically in segregation applications. Since large-scale disaggregate mobility data and data that indicates social groups are initially unrelated, there are no default connections established. Therefore, segregation-related research currently faces:

The lack of a method to measure activity-based multi-group social segregation using socioeconomic and large-scale disaggregated mobility data such as public transport smart cards

To measure activity-based multi-group social segregation using large-scale disaggregate mobility data it lacks a general method describing the process of linking data sources. In particular, mobility data does not contain social information which needs to be linked to socioeconomic data before calculating segregation indices.

Consequently, this study develops a method to link social groups, large-scale mobility data, and multi-group segregation measures to quantify segregation. The aim of this study is to measure activity-based multi-group segregation, which is neither limited to just one type of mobility data nor to a specific socioeconomic data set or socioeconomic information coupled to mobility data.

Therefore, this study formulates general requirements regarding mobility data as well as socioeconomic data sets for segregation studies with disaggregate mobility data. In a second step, socioeconomic residential data is linked to each disaggregate element of the large-scale mobility data. Lastly, multi-group segregation measures can be applied to the enriched disaggregated mobility data. This yields **a method for measuring social segregation using large-scale disaggregate mobility data**

and socioeconomic data.

Finally, this method is applied for smart card data sets of the public transport system of Stockholm County, Sweden. To measure segregation of, in this case, ordinal social groups, Reardon's "ordinal information theory index" is applied (Reardon, 2009). More specifically, the income segregation of public transport passengers at their journey destination is measured. Thereby, the evenness of the social mix of travelers at the destination level is quantified. The evolution of segregation is then analyzed with public transport smart card data sets of comparable time intervals. One-week data sets of each year's week 5 from 2017 till 2020 are used. Assessing the evolution of social segregation over time also leads to conclusions about the effect of the "Citybanan" railway tunnel project in Stockholm which was finished in July 2017 and aimed at reducing segregation.

With the Stockholm case study a public transport destination-based approach is chosen for social segregation study. Thereby, this study contributes to the field of mobility- and destination-based segregation studies. It also adds up to the discussion on activity-based segregation, especially due to the unique connection of social data to activity traces. In addition, the information theory index has not been applied yet in this context of ordered social groups.

In the following, this chapter sets out the research questions according to the research gap addressed earlier. After, the document structure of this study is outlined.

1.2 RESEARCH QUESTIONS

Based on the research gap indicated, this study aims to answer the following main research question:

How can multi-group activity-based social segregation be measured using large-scale disaggregated mobility data such as public transport smart card data?

To answer the above there are four sub-research questions defined below:

1. How is social segregation measured and how can it be represented using mobility-based data?
2. What are the requirements from large-scale disaggregated mobility data for enabling the measurement of multi-group social segregation?
3. How can socio-demographic characteristics of smart card users be inferred?
4. How does activity-based social segregation evolve and how does disaggregated mobility data facilitate its analysis?

The first three sub-research questions incorporate finding a method to link social groups to large-scale disaggregated mobility data and quantify social segregation. Lastly, the fourth sub-research question covers the time-dependent part to analyze the evolution of segregation.

1.3 DOCUMENT STRUCTURE

Following up on this introduction, a review of literature on segregation and public transport data is given in chapter 2. As a result, a connection is made between the measurement of activity-based segregation and public transport smart card data. In chapter 3, a method is developed for connecting residential social groups to

disaggregate mobility data and thereby calculating activity-based segregation measures. Introducing the case study of Stockholm County, the method is applied to the public transport authority SL's smart card data sets in chapter 4. Results of the method's application are presented in chapter 5 and discussed thereafter including limitations of the study. Lastly, conclusions are made in the final chapter 6.

2 LITERATURE REVIEW – ACTIVITY-BASED SOCIAL SEGREGATION

Segregation can be approached from many perspectives and is an early observed phenomenon of concentration and inequality. The most well-known example is the division of poor and rich within cities (Paddison and Hamnett, 2000; Xu et al., 2019). Segregation is caused by the unintentional as well as intentional "isolation of people from those unlike themselves" (Li and Wang, 2017). It leads to experiencing inequality regarding essential life factors such as food, income, educational level, housing quality, and many more. Moreover, segregation can be experienced by the distance between social groups. Among other approaches, social segregation is measured as the spatial distance between groups since it is acknowledged that spatial distance drives segregation (Ellis et al., 2012).

Segregation is also caused by factors other than spatial distance. It can be related to the lack of access to necessities or infrastructure as well as to temporal factors that prevent social groups from mixing with each other. For example, segregation can vary by time of day, such as discovered by Le Roux et al. (2017) who assessed segregation in the Paris region around the clock. Further, access to amenities and the resulting opportunities shape segregation. For instance Logan and Burdick-Will (2016) examined significant differences in segregation as a function of access to education. Differences in ethnicity, gender, and nationality also lead to segregation and fields of research that have been widely discussed for some time; see, for example, Leonard (1987).

For this study, only social segregation is considered. Social segregation is hereby defined as **the uneven spatial distribution of social groups** according to Le Roux et al. (2017). This study does not cover other interpretations of segregation, such as racial or gender segregation. The terms social segregation and segregation are in respect to the study's context used interchangeably.

In this chapter both the measurement of segregation as well as activity data collection, specifically mobility data, are discussed. First, an overview is made in section 2.1 and 2.2 on the current state of research on social segregation, focusing on activity-based segregation. Thereafter, the increasingly important role of mobility and public transport (PT) data and its relation to segregation is discussed in section 2.3. After, measures of segregation are discussed in 2.4. This precedes the detailed description of the study's focus on measuring activity-based social segregation using mobility data, such as from public transport usage in section 2.5.

2.1 RESIDENTIAL SEGREGATION

Understanding social segregation has been a frequently investigated research topic in many respects. Social segregation is found to correlate with socio-demographic characteristics and is therefore often analyzed using these kinds of static information. Most commonly, segregation is explored using spatially aggregated data at the place of residence, for instance, see Bischoff and Reardon (2014) as well as Musterd et al. (2017). Residential segregation studies are the most common form

of focusing on one socio-geographical space. Other approaches use data from the location of work, shopping, or leisure activity.

Residential data is usually obtained from census or other register data including socio-demographic and socioeconomic statistics. Typically, relevant information such as income, educational status, and age groups is allocated to smaller spatial units. Either these are estimates for the population living in the zones or individual data is present. Before data like this is made publicly available, often a single person's data from e.g. registers is aggregated to protect privacy. Thereby, categories are built or averages and medians are calculated to indicate the characteristics of the zone inhabitants. Thereby, inhabitants are often grouped on the basis of geographical but fictitious governmental zones.

Social segregation is a complex research topic including many relevant aspects. Therefore, a high number of socioeconomic variables are used for segregation studies. Commonly used socioeconomic variables provide information about education, age, and often income such as in the study of Bischoff and Reardon (2014). Others consider housing type, educational level, or ethnicity as the indicating factors for segregation (Ivaniushina et al., 2019; Logan and Burdick-Will, 2016). Overall, income is found to be a major factor influencing segregation. Disparities in income levels lead to inequality in access to infrastructure and amenities and could therefore lead to less mixing and what is called income segregation.

Xu et al. (2019) state that segregation is weakly correlated at the individual but highly correlated at the level of bigger groups. Therefore, often social groups or clusters are built summarizing individuals with the same socioeconomic values or geographic locations, like the place of residence. Various studies use a combination of relevant socioeconomic variables to build clusters, such as Almlöf et al. (2021). The aggregation towards a manageable amount of groups eases the analysis and enables depicting significant differences between groups. On the other hand, grouping summarizes socio-demographic values and might lead to less detailed analyses.

Due to the relevance of combining socioeconomic variables to understand segregation, often, social groups or clusters are built using socioeconomic variables. A population is assigned to categories with thereof categorical characteristics. In some cases, clusters could incorporate an order which would make them ordinal. This is due to the different distances between social groups as described by Reardon (2009). For example, a low-income group tends to have fewer differences from a middle-income class than a social group consisting of the richest percent of the population. Therefore, such groups would imply an order, and one would expect segregation to be higher between low- and high-income classes.

Nevertheless, social groups made from socio-demographic and socioeconomic variables carry valuable information about the inhabitant group's life circumstances. Applying segregation measures such as defined by Massey and Denton (1988) quantifies segregation from residential data. More on the function of segregation measures, which are mostly indices, is covered in section 2.4.

Initially, segregation studies covered the comparison between two groups. These studies often assessed big groups in the traditional understanding of segregation, that primarily focused on capturing segregation between different races or ethnicities. Mostly, these segregation indices are restricted to only measure two groups segregation and can not be used for multi-group comparisons. As segregation analyses developed towards also assessing other characteristics, the two-group measurements were not made to compare for instance multiple age

groups concurrently to each other.

This explains the differentiation made between two-group and multi-group segregation studies. Reardon and Firebaugh (2002) describe the need to measure multiple groups' segregation between each of them in one measure. This brings up so-called multi-group segregation indices which cover the concurrent segregation measuring of multiple groups. A discussion on the difference between two- and multi-group approaches is given the following as well when discussing segregation indices in section 2.4.

Despite the accomplishment of segregation studies using residential data, more and more research focuses on including data from activities. This approach is discussed in the following section.

2.2 ACTIVITY-BASED SEGREGATION

Next to residential factors such as income, segregation often is caused and influenced by the spatial and temporal distance between people as well as access to resources. This leads to the concept that activities and thereby mobility should also be considered as valuable indicators of segregation (Reardon and O'Sullivan, 2004; Kwan, 2009; Wong and Shaw, 2011). Xu et al. (2019) point out that the static character of residential data hinders obtaining a dynamic view of social mixing in cities. Residential segregation studies do not capture the manifold social interactions a human is experiencing throughout the day (Moro et al., 2021). In addition, residential data only leads to conclusions on one socio-geographical space, namely housing, ignoring others such as work or social activities.

There are many causes of segregation, such as spatial distance, temporal differences, or differences in access to resources, all of which are related to the activities of residents. This is due to the nature of people moving from their homes or other locations to places where they can access services and facilities. Activity-based data offer the opportunity to better assess temporal and access-related factors than studies using only a socio-geographic space.

Ellis et al. (2004) found significant differences in work and home-related segregation patterns. In accordance with this, Wong and Shaw (2011) declare that segregation is a function of exposure to other groups within the individual activity space. Also, the above mentioned factors such as different access to amenities and temporal differences stress the point of going beyond one socio-geographic space/residential segregation. This leads to a new perspective of segregation studies, so-called activity-based segregation.

A general belief regarding activity-based segregation is that looking at activity locations or destinations reveals more moderating effects about segregation than only considering residential data (Ellis et al., 2004; Wong and Shaw, 2011). This implies that residential segregation studies would overestimate the level of segregation. As a result of these findings, many recent studies focus on including the activity aspect when trying to capture segregation.

Nevertheless, studies that focus only on one socio-geographic area are of great interest regarding the segregation impact of socio-demographic or socioeconomic factors. When meaningful variables such as gender, age and income are used, they are often associated and linked to the living or working space. Therefore, these studies are still very relevant when examining those life factors and socio-

geographic spaces.

However, the use of activity-based data expands the spaces studied and improves accuracy in measuring more dynamic segregation factors such as spatial, temporal, and access-related causes. A key component to enhance segregation analysis it to include the dynamics of human life. Therefore, this study explores more than just the static approach of residential studies.

Activity-based segregation measurement can be found in recent studies such as Tan et al. (2019); Li and Wang (2017). These projects aim to incorporate the dynamics of mainly cities as a key element of segregation (Le Roux et al., 2017). Among others, Silm and Ahas (2014), as well as Athey et al. (2020), point out that that this *experienced segregation* from activities is significantly less than residential segregation. This explains the recent research focus on activity-based segregation since it appears to state a more realistic image of segregation.

Figure 2.1 shows the two different dimensions of segregation studies discussed so far. On the one side studies measure segregation between two or multiple groups, as discussed in 2.1. The other distinction made, sets out residential versus activity-based approaches. While activity-based segregation is focused by current research efforts only a few consider the assessment of multiple groups' segregation. This research field is colored in the figure.

Figure 2.1: Dimensions of segregation studies

Looking into activity-based segregation, there are different types of activity data used. Among other data sources, recent activity-related segregation studies make extensive use of mobility data as it describes the way of moving around and potentially mingle. Utilizing mobility data for activity-based segregation studies offers valuable insights in the mixing of inhabitants. Table 2.1 provides an overview of the key papers of activity-based segregation. As it can be seen, activity-based segregation studies use mobility-capturing data such as GPS, mobile phone, social network, travel survey, and travel diary data (Järv et al., 2015; Farber et al., 2015; Xu et al., 2019; Le Roux et al., 2017).

Travel surveys and diaries are popular data sources for mobility studies. These are rich databases for segregation studies, such as the ones mentioned in Table 2.1, since they include socio-demographic information next to mobility-related data. Further, mobile phone data sets are getting more popular to observe people's activity. Using the GPS signal enables continuous and accurate tracking of individuals and thereby obtaining disaggregated activity data sets.

Many studies chose mobility data to analyze activity-based segregation. In the following, the connection from segregation to mobility data is made, starting with a review of activity-based and mobility-related studies and their data sources. The section below discusses the availability and characteristics of the data sources mentioned in Table 2.1. Several shortcomings are revealed for some of the mobility-

Table 2.1: Activity-based segregation and activity-space studies

Authors & Year	**Measure type**	**Grouping criteria**	**Measure used**	**Data source**
Wong & Shaw, 2011	two-group	racial-ethnic data	Activity space-bounded exposure measures	travel diaries
Silm et al., 2014	two-group	ethnic	dissimilarity index, modified index of isolation	mobile phones
Järv et al., 2015	two-group	ethnic	activity-space study	mobile phones
Le Roux et al., 2017	multi-group	socio-demographic	Gini index, information theory index	travel surveys
Li & Wang, 2017	individual	socio-demographic	regression model	activity diaries
Xu et al., 2019	individual	socio-demographic	social similarity measure	mobile phones, socioeconomic
Tan et al., 2019	two-group	ethnic	exposure indices, multilevel regression model	activity diaries
Tao et al., 2020	two-group	income	activity-space measures	travel surveys
Abbasi et al., 2021	two-group/ multigroup	socio-demographic	two-group dissimilarity and exposure indices /multigroup entropy index	PT smart cards

related data sources. As a consequence, upcoming possibilities using emerging data-collecting technologies such as public transport smart card data are discussed.

2.3 SEGREGATION IN RELATION TO MOBILITY AND PUBLIC TRANSPORT

As concluded from section 2.2, there is a clear connection of activity-based segregation to mobility since mobility is representing big parts of inhabitants' activities. Of particular interest when considering mobility patterns are the destinations of journeys. These are means to an end, as people move for a reason and can be expected to engage in activities near the destination of their journey. Therefore, people can be expected to mix near journey destinations, which is interesting for segregation studies.

A major component of mobility and activity spaces is public transport (PT), especially in urban environments. Not only does it shape transportation but also urban structures, for example, due to transit-oriented development. Dawkins and Moeckel (2016) found that public transport-oriented development could even lead to transit-induced gentrification which in turn leads to segregation.

As touched upon in chapter 1, there are two different roles public transport incorporates. PT plays a major role in facilitating access and diminish barriers between social groups while on the other side it shows traces of mobility. As a result, public transport data is of high interest when capturing segregation and

analyzing access to resources.

Up to this date, obtaining accurately observed mobility traces remains a challenge. Therefore, public transport data gets more and more considered when capturing social segregation. Still, measuring segregation with PT data is not yet explored extensively since only a few recent segregation studies make use of this type of data, for instance, see Abbasi et al. (2021).

Digitalization of public transport systems leads to extended opportunities in obtaining observed, disaggregated PT data. Assessing observed public transport mobility patterns can lead to new insights into passenger behavior and thereby in understanding PT users' activities. Recent studies such as Verma et al. (2021) point out the need to look beyond residential data for PT travel pattern studies. Since both public transit demand and segregation studies are heading towards more activity-based approaches as discussed in section 2.2 the combination of measuring segregation with public transport data is explored in this section.

The following subsection continues looking into the relationship between segregation and public transport. It is explained why observed PT data is a valuable data source to measure segregation.

2.3.1 Social segregation, transport disadvantage and public transport

Activity-based segregation studies predominantly use mobility data. This is not only due to its availability but also related to the role of mobility regarding segregation. Research found a direct link between segregation and the so-called transport disadvantage (Li and Wang, 2017).

Transport disadvantage studies describe the limited access to transportation (Delbosc and Currie, 2011). It results in less access to essential infrastructure to participate both socially and economically (Lucas, 2011). Transport disadvantage is strongly correlated to social exclusion as found by studies such as Church et al. (2000). This then entails less mixing of social groups, which implies higher levels of social segregation and can lower the quality of life (Delbosc and Currie, 2011).

Especially in urban environments, public transport is a major component of mobility and activity spaces. As a public service, public transport is supposed to provide accessibility to everybody and therewith every social group. Not only is public transportation a key component of transportation disadvantage, but studies have found a direct link of PT's causing segregation (Rokem and Vaughan, 2018). Less access to public transport causes higher levels of segregation. Looking at the problem from the opposite perspective, Kaufmann (2004) found that social segregation is "a product of the mobility" of certain social groups, in this case, wealthy urban inhabitants.

Further, public transport is not only crucial in urban but also rural areas. The study of Kamruzzaman and Hine (2012) showed that in rural areas low-income groups' activities are located along public transport corridors being highly correlated to the use of PT systems.

Subsequently, access and use of public transport are directly affecting segregation. Vice versa, segregation levels can be set into relation with the use of PT systems. From the few current studies such as Abbasi et al. (2021), it can be concluded that segregation is analyzable from public transport data since PT systems are key in providing mobility and access and therefore influencing segregation. As mentioned earlier, mobility-based segregation studies can have different sources such as GPS

and mobile phone data. For public transport data collection, these two sources are less significant since its provided data about the system's usage is hardly distinguishable from other modes or forms of moving. The following subsection provides an overview of PT data collection for ex-post evaluation studies such as segregation studies. It focuses on the upcoming use of observed disaggregate travel data from smart cards.

2.3.2 Public transport smart card data

Public transport plays a key role in people's daily dynamics, especially in urban environments. To measure segregation from public transport data large-scale observed disaggregate data is required. This section starts by presenting the need for observed disaggregated travel data in public transport. It continues elaborating on the key findings towards such large-scale disaggregated journey data sets and points out the use of public transport smart card data.

Public transport passenger data collection

Traditionally, mass transit users' travel demand is estimated by stated preference surveys, travel diaries, or aggregated cross-sectional and time-series analyses (Viallard et al., 2019). Surveys have their strengths in assessing preferences. Travel diaries show complete mobility traces for a small population. However, the traditional methods either lack actual observation, accuracy or completeness. Cheaper and more complete data sources, such as GPS data, provide accurate mobility traces that are not linked to public transportation and often come with privacy concerns.

The above mentioned data collection methods can imply disadvantages because of insufficient data collection, immense required efforts, and high costs (Utsunomiya et al., 2006) as well as privacy issues. Due to these shortcomings, the public transport sector has been limited in accurately analyzing observed mobility patterns.

To obtain more valuable insights into public transport travel patterns and demand there is a need for observed passenger behavior data. The public transport sector, therefore, has an interest in collecting data of observed travel dynamics and is catching up using emerging technologies such as GPS and automatic fare collection systems (Pelletier et al., 2011). Over the past decade, a significant advance in PT data collection of observed user behavior has been achieved (Luo et al., 2018). One key contribution is made by so-called public transport smart cards, an established automated fare collection technology that incorporates user-specific chip cards to tap in and sometimes also tap out at stations and/or in vehicles.

Smart cards are expanding since technology arose to supply public transport users with chip cards and transform infrastructure and vehicles accordingly. For more background information see e.g. Cheung (2006) describing the implementation of smart cards in The Netherlands. Worldwide introductions of public transport smart cards affected the sector positively since it is perceived as a reliable fare collection method and convenient for passengers (Tuncer, 2018). Among other advantages, smart cards facilitate the non-physical integration of different transport modes which adds up to customer convenience. As a side effect, smart card implementation enables one of the most complete methods of data collection by dynamically collecting timestamps and geotags (Viallard et al., 2019). There is a big potential in smart cards collecting real-user behavior to overcome the shortcomings of traditional public transport data collection and analysis (Bagchi and White, 2005; Pelletier et al., 2011).

While there have been early studies validating smart card data (SCD) as reliable, accurate sources for public transportation studies (Bagchi and White, 2005; Barry et al., 2002; Park et al., 2008), there are some challenges for processing and interpreting the data. Data-related issues include incompleteness, alighting inferences for tap-in-only-systems as well as the amount of data and computational challenges to overcome. Significant progress has been made and researchers developed several mining techniques according to the literature review on smart card data use in public transport by Pelletier et al. (2011). For incomplete journey data, caused by e.g. only tap-ins, transfer and destination inferences are developed. For instance Ma et al. (2013) tested different algorithms for what is called trip chaining.

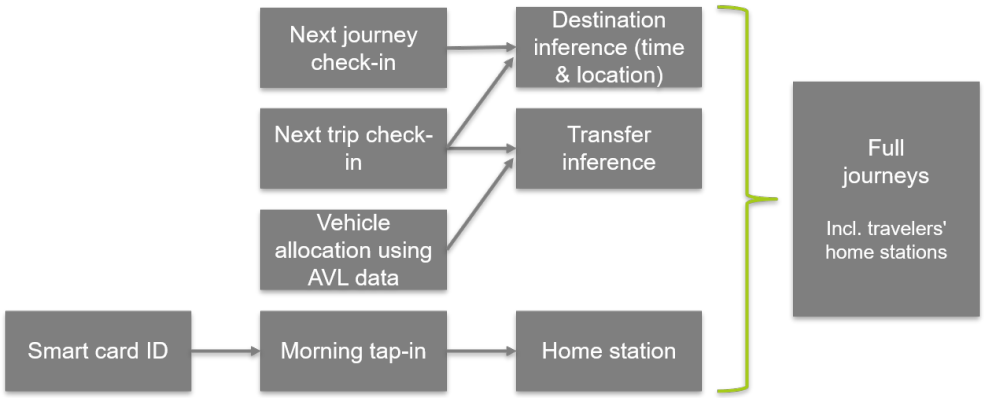

Figure 2.2: Inferences towards observed SCD journeys with home stations

Assuming successful inferences, according to the scheme in Figure 2.2, full observed journey data sets can be obtained. As a result of the aforementioned work, observed travel diaries can be revealed per smart card identification number (ID). Recent research like Dixit et al. (2019) work with inferred travel diaries as well as observed origin-destination matrices with the help of zonal aggregation. Deriving travel diaries and OD-matrices from smart card data imply obtaining data on the majoritarian parts of networks' real usages. This progress on observing disaggregated travel demand enables further analysis which is discussed in the following chapter.

Smart card data applications for ex-post evaluation

There is a wide range of applications for public transport smart card data. Especially relevant for the context of this study are ex-post evaluation studies. Some relevant before-and-after evaluations of public transport network-related changes use smart card data. The study of Arbex et al. (2019) analyzes changes in travel times, transfers, and accessibility of a bus network improvement while pointing out the advantage of continuous smart card data over household travel surveys. Further, Fu and Gu (2018) assess a metro network change with the help of smart card data. Brands et al. (2020) evaluated multiple changes to the urban multi-modal network of Amsterdam underlining the opportunities smart card data yields for public transport assessment and planning.

Inter alia Viallard et al. (2019) and Devillaine et al. (2012) mention that for most countries smart card data is not linking travel data to users' socio-demographic data due to data privacy. In many cases, smart card data just reveals the pure actions of anonymous system users since public transport authorities can only provide encrypted card IDs. This mostly happens due to privacy policies in the respective regions or countries.

Only a few studies accessed additional socially relevant information such as age groups of cardholders via their fare subscription (Abbasi et al., 2021). To overcome the drawback of no direct access to smart card holders' social attributes, researchers inferred home zones per smart cardholder, see for instance Tamblay et al. (2016); Kholodov (2019). Home zone inference algorithms such as Sari Aslam et al. (2019) are used for this purpose. It allows to estimate home stations and thereby home zones by looking at the most frequently used morning trip origin. This enables linking each card to a home zone, as shown in Figure 2.2. These can potentially be used to derive zonal characteristics, for instance spatially aggregated census data. The home zones' socio-demographic statistics are thereby linked to a card ID and therefore a user.

This technique is applied in some studies such as Goulet-Langlois et al. (2016) where clustering methods are used to extract passenger's temporal and geographical activity to link the identified patterns to socioeconomic attributes. As Goulet-Langlois et al. (2016) conclude, current research concentrates on segmenting the travel patterns of public transport users. Optimizing and testing cluster algorithms and cluster composition has been done by many studies such as Weng et al. (2018); Viallard et al. (2019); Kieu et al. (2015); Briand et al. (2017). These studies use smart card data to segment users on temporal or spatial travel patterns and therewith assessing the public transport passenger population. Other authors build upon the efforts made on clustering, for instance, Briand et al. (2017) developed a model to regroup passengers in clusters based on their temporal habits and transit usage.

Originally, SCD of public transport users has disadvantages over other mobility data sets, such as incompleteness and lack of socioeconomic data in comparison to travel surveys or diaries. Comparing smart card data with phone data leads to the conclusion that phone data sets include bigger parts of the population simply due to that in many societies more people are regular mobile phone users than public transport users. That leads to phone data revealing a more realistic profile of society while on the same hand including more accurate mobility profiles since people are mostly moving with their phones. A potential pitfall of using mobile phone data is that the observation of mode and destination might get lost by the continuous motion sensor. Additionally, the usage of individuals' phone data raises privacy concerns and could be impossible in some countries. The collection of PT smart card data also raises privacy concerns but on a smaller scale as long as data is protected within the public transport authority and/or the data collector (which can be an external service provider, see Cheung (2006)).

Public transport smart card data offers various benefits, also regarding the measurement of segregation. Smart cards reveal more accurate and complete data sets than travel surveys or diaries as discussed in section 2.3.2. Compared to the usage of GPS and mobile phone data, smart cards are easier accessible and less critical for data protection. SCD is restricted to the public transport system but is, therefore, able to identify changes in PT service (Wei et al., 2015). Especially, over other residential-based data, smart card data offers unprecedented advantages. The use of SCD facilitates the progression from "potential mixing" using residence data to "actual exposure/possible contact" because it indicates, to some extent, how people are brought together. Nevertheless, it does not ensure measurement of actual contact, since traveling together does not mean coming into contact.

In general, there is potential to link smart card data with other research areas to broaden the application range and use observed travel data to gain deeper insights into passenger behavior. One of these is the application of public transit data, partic-

ularly smart card data, to the investigation of segregation, which is discussed below.

To understand the measurement of segregation and its possible application on public transport smart card data the commonly used segregation indices are introduced. In the following section 2.4 both two-group and multi-group segregation measures are introduced. Further, ordinal segregation is focused since social groups are often indicated using ordinal variables as concluded before in section 2.3.

2.4 SEGREGATION MEASURES

Since the first segregation studies in the mid-20th century, a wide range of primarily indices was developed to capture segregation. There are several dimensions of segregation characterized with different approaches to measure segregation. Massey and Denton (1988) classified segregation measures into five dimensions: evenness, exposure, concentration, centralization, and clustering.

Most commonly used are measures of evenness and exposure, in particular the dissimilarity and exposure indices (Massey, 2012; Reardon and Firebaugh, 2002). These measures describe the distribution of groups among organizational units without regard for their spatial proximity (Reardon and Firebaugh, 2002). Traditionally, such measures are used for residential segregation studies.

As touched upon before, segregation studies were initially made to compare two groups of different races or ethnic groups to each other. These so-called two-group measures calculate the segregation of one group to the other.

Despite the groundbreaking achievements of early developed segregation measures calculating two groups' segregation from each other, there is some criticism regarding the two-group measures. Measuring two-group segregation is applicable for comparing larger groups of ethnics to each other such as in the studies of Silm and Ahas (2014); Järv et al. (2015) regarding the Russian-speaking minority in Estonia. For many contexts of social segregation studies, multiple groups' segregation from each other needs to be captured.

Among other criticism, Reardon and Firebaugh (2002) described two-group measures as "increasingly inadequate", especially in diversified societies. Since two-group measures are restricted to only measuring the segregation of two groups, there are several approaches of developing multi-group measures. Meng et al. (2006) developed a social difference coefficient that modifies the multi-group spatial segregation indices from previous research. These are the Local Getis index, the Proximity index, and the Exposure index.

Most renowned is the early research of Reardon and Firebaugh (2002) who developed six multi-group segregation indices. Later, some of the indices were extended towards the use of ordinal variables (Reardon, 2009). Table 2.2 includes the multi-group indices developed by Reardon and Firebaugh (2002).

Further, the characteristics of social variables play an important role. Often, variables such as educational level or age are grouped and put in order and are therefore ordinal. In many cases, this type of data is a result of e.g. census data collection and implies dealing with ordinal variables. As discussed before in section 2.2, often continuous and interval variables are used to form clusters which are then indicated by ordinal variables as well. Therefore, this thesis project focuses on ordinal variables. To measure segregation from ordinal variables, including an inherent ordering the multi-group measures presented in Table 2.2 cannot be

Table 2.2: Multi-group segregation indices developed by Reardon and Firebaugh (2002)

Multigroup index name	**Function**	**based on**
Dissimilarity index (Duncan's index)	can be interpreted as the percentage of all individuals who would have to transfer among units to equalize the group proportions across units, divided by the percentage who would have to transfer if the system started in a state of complete segregation.	Duncan and Duncan (1955); Morgan (1975); Reardon (1998); Sakoda (1981)
Exposure index	Assessing the social homogeneity: Contact probability of one social group to another one	Bell (1954)
Variance ratio index (Relative Variation Index)	exposure-based and based on each observation's deviation from a central value such as the standard deviation	Bell (1954); James and Taeuber (1985)
Entropy score	Measure of diversity: the extent to which multiple groups are evenly distributed	Massey and Denton (1988)
Gini index	Scale of inequality (e.g. on income) - measure of disproportionality that emphasizes how groups are disproportionately represented in each spatial unit. Based on each observation's deviation from all other observations such as the mean difference	Reardon and Firebaugh (2002)
Information theory index	Measure of diversity that assesses the degree of social mixing within the spatial units,	Theil (1972); Theil and Finizza (1971)

applied right away.

Hence, adaptions were made to overcome these drawbacks. Reardon (2009), as well as Meng et al. (2006), developed ordinal segregation measures based on the above-mentioned multi-group measures. Meng et al. (2006) extends exposure indices by a factor that accounts for the order of categories. But as Reardon (2009) states, this approach relies on arbitrary intervals assigned to ordinal values and only exposure- but no evenness-related measures were derived. Therefore, Reardon (2009) developed an evenness-related approach based on the multi-group measures of Reardon and Firebaugh (2002). The resulting segregation measures rely on the thought that segregation can be measured as "the extent to which variation within organizational units (unordered categories) is less than total variation in the population" (Reardon, 2009).

Looking at these measures, the so-called "information theory index", initially developed by Theil and Finizza (1971); Theil (1972), has been found superior for multi-group application compared to other multi-group measures. This was determined by Reardon and Firebaugh (2002) also using criteria developed by James and Taeuber (1985). Only the information theory index satisfies both the organizational and group decomposition properties. For more information on evaluating measures of multi-group segregation see Reardon and Firebaugh (2002). In addition, its measurement of evenness fits the definition of segregation given by Le Roux et al. (2017). Looking at both factors, the ordinal version of the information theory index is discussed more in detail below.

The *ordinal information theory index* is a measure of the "evenness" dimension that assesses the degree of social mixing within spatial units. It is based on the generalized form of the information theory index by Theil and Finizza (1971); Theil (1972) which Reardon and Firebaugh (2002) refined for measuring multi-group segregation of ordinal groups. It aims to measure "the extent to which ordered groups are evenly distributed across unordered categories" (Reardon, 2009). Minimum segregation is reached when the distribution of ordered groups is mirroring that of the population. Maximum segregation is reached once the relative distribution among categories has no variance. The ordinal information theory index's exact mathematical formulation and its function are introduced in section 3.4.

Looking at applications, the ordinal version of the information theory index, also called Theil's index, is utilized in some studies such as Ivaniushina et al. (2019) and Monkkonen et al. (2018). To assess school segregation in St. Petersburg, Russia in relation to socioeconomic statuses, Ivaniushina et al. (2019) uses individual-level school data. Monkkonen et al. (2018) analyses the urban sprawl and the growing geographic scale of segregation in Mexico. Therefore, census data sets aggregated per city are used. From 1990–2010 the income and education levels of cities are taken as an input for the segregation metrics.

The above studies prove the metrics to meaningfully measure segregation. So far, there is no application of the ordinal information theory index using disaggregated public transport smart card data. Since the metrics is assessed to be capable of measuring the segregation experienced by public transport travelers, a connection between the ordinal information theory index, social clusters, and PT smart card data is made in the following chapter.

2.5 MEASURING ACTIVITY-BASED SOCIAL SEGREGATION USING MOBILITY DATA

As it can be seen from the mobility-related segregation studies discussed in section 2.3, the segregation measures described in section 2.4 are well applicable on activity-based data sets. However, there is a shortcoming of activity-based data regarding the social aspect. In most cases, activity-based data does not reveal any social background information which makes it hard to depict social groups and therefore measure their segregation.

To overcome this, it needs a connection between mobility data and socially relevant information. Sometimes, the activity-based data is connected to socio-demographic residential data to indicate social backgrounds. This is mostly done via a geotag of the mobility data. This can then be matched with the location information available for census data or other comparable sources of socio-demographic data. Afterwards, segregation measures are applied to the disaggregated mobility data sets.

Le Roux et al. (2017) concluded that crossing residential and activity-based data leads to an improved view on social characteristics of populations. Summing up, there are three components of a successful measurement of social segregation with mobility data. It includes socioeconomic data or predefined groups to obtain social groups and large-scale disaggregate mobility data representing activities. Combining the two data sets, a segregation measure that is capable to calculate multi-group mobility-based segregation index is needed.

Of the studies mentioned in Table 2.1, only the recent research of Abbasi et al. (2021) uses the three above-mentioned for the context of measuring mobility-based multi-group social segregation. This study utilizes disaggregated smart card data journeys from public transport and the two-group dissimilarity and exposure indices to measure the segregation between two groups. Applying this two-group measure, the segregation of seniors, children/youth, and passengers with disabilities in relation to all other passengers is separately calculated. Additionally, a multi-group entropy index is calculated. The information on this kind of social status is taken from the public transport smart card itself.

However, not many PT smart cards include data on the socioeconomic situation of the passenger. Particularly in countries with high data privacy restrictions, social information cannot be extracted from or linked immediately to these types of mobility data sets. So far, for activity-based studies, there is no connection made to residential socioeconomic data as used in many other segregation studies.

In addition, due to the categorical, sometimes ordinal, and very often multi-dimensional nature of socially segregated groups, there is a need to compare multiple groups at the same time. Hence, multi-group activity-based social segregation measurement is needed which is less dependent on prevalent socioeconomic data.

Combining these findings leads to the question of whether connecting residential socioeconomic data to large-scale data of multiple groups' activities would yield new opportunities to measure social segregation. To facilitate this, a methodology to measure activity-based social segregation from large-scale disaggregated mobility data is required. This is also formulated in the research gap in section 1.1. and consequently leads to the research questions from section 1.2. To answer the research question on measuring social segregation from large-scale disaggregated mobility data, a methodology is developed in the following chapter 3.

3 METHODOLOGY – MEASURING SOCIAL SEGREGATION USING ACTIVITY-BASED MOBILITY DATA

The literature review established a distinct connection between initially residential-measured segregation and activity-based approaches as well as the motivation to use mobility data. In the following, the combination of the above is made, developing a method that enables measuring multi-group segregation of social groups using their socioeconomic information and observed mobility data.

This should answer the main research question on how social segregation can be measured at places where people are meeting due to their mobility behavior. Social groups' disaggregated large-scale mobility data could yield new insights measuring activity-based multi-group social segregation. Combining residential socioeconomic data with activity-based mobility data could merge the strengths of both currently prevalent segregation measurement approaches.

Hereafter, the methodology is presented in 3.1. It describes how large-scale mobility and residential data can be used to measure social segregation. Each process step is further explained in 3.2.

In the first subsection 3.2.1 requirements for residential social data are given. After, in 3.2.2, requirements for mobility data are outlined. Both scopes down the data sets that can be considered. In the third step, the process is described to enrich mobility data in such a way that it connects to travelers' social characteristics. This is done by connecting social data and its implied groups to mobility data. After, the initial setup the last step in 3.2.4 implies calculating social segregation from the compiled data set. As indicated in the literature review in section 2.2 and 2.4, this is usually done using segregation indices. Therefore, this methodology is designed to measure (multi-group) segregation indices as a final step and introduces the segregation measure "ordinal information theory index" developed by Reardon (2009).

3.1 FRAMEWORK

To operationalize the concept according to the indication in 2.5, a process is defined and presented in 3.1. It shows the modeling steps towards calculating segregation indices from mobility data and its requirements for possibly connected data sets. Residential socioeconomic data and its abstracted groups are connected to observed disaggregated mobility data by using the same spatial organizational units, usually a type of administrative zones. Once the socioeconomic data is assigned to the mobility data via travelers' home zones, different segregation measures can be applied.

This approach is fairly independent of the type of organizational units and the number of social groups used. However, it was developed taking into account administrative zones and more than two groups, thus a multi-group approach. Further, the use of large-scale disaggregated observed mobility data is suggested. Theoretically, it could also use disaggregated non-observed data such as from surveys. Though it would have to have a certain size to calculate the distribution of travelers.

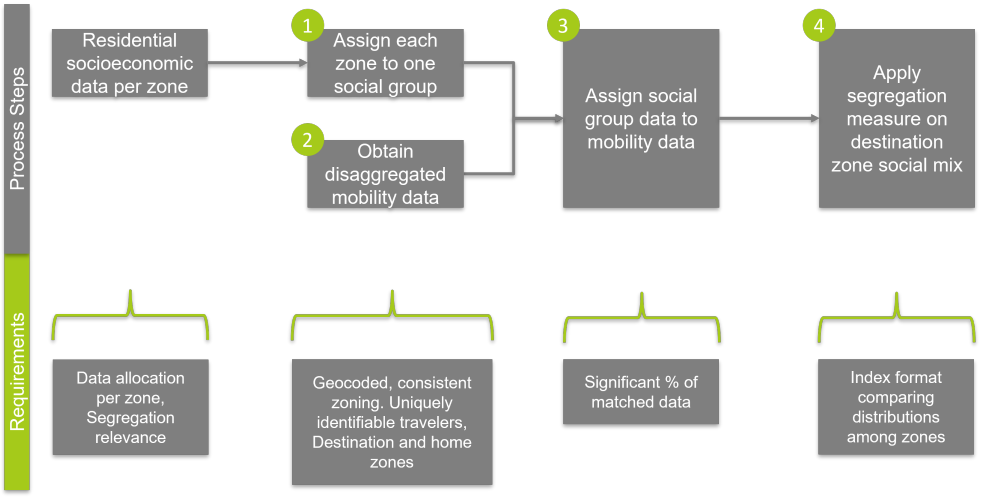

Figure 3.1: Framework for measuring social segregation by connecting mobility data to socioeconomic data

3.2 MODELING STEPS

In the following, the process steps illustrated in Figure 3.1 and its data requirements are outlined.

3.2.1 Requirements for residential social groups

As described in section 2.1, segregation is commonly measured using socio-demographic and socioeconomic data obtained from the place of residency. All criticism of looking at only residential data set aside, the residence is still a highly valuable factor when trying to understand the social composition of society. People spend major parts of their day at their home location. Also, residency indicates where to geographically locate socioeconomic data since the data is assigned to administrative spatial units such as zones.

Assuming a considerable amount of zones and social differences within a society, the following describes requirements for residential social data. Most commonly socially relevant information can be obtained from socio-demographic data as retrieved from census or other governmental databases. The first requirement for this data is that it is or can be allocated to the same zoning as the mobility data. Next to the data allocation per zone, the socioeconomic or socio-demographic data should be relevant for segregation. As analyzed earlier, this entails the use of variables such as income.

As known from studies discussed earlier in 2.1, commonly used social variables are income, ethnicity, housing type, educational level, and more. These variables can be of many fashions. To mention some: average income levels are continuous variables, ethnicities are categorical and educational levels are rather ordinal. Consequently, the methodology is designed to be independent of the type of social data.

Since continuous, categorical, and ordinal socioeconomic data can be numerous and widely scattered throughout zones, often groups or clusters are built. This leads to the advantage of aggregating social information into more meaningful social groups. This common approach is used in many studies to be able to conclude on and relate conclusions to bigger parts of society.

The grouping can either imply using only one variable defining social groups or build social groups from multiple variables which is mostly done by clustering. To measure segregation, this methodology is independent of the way of forming social groups. In the following, all of these approaches are referred to when mentioning social groups.

Once these requirements are fulfilled, the first process step can be executed. This means assigning each zone to a social group. How these zones are assigned is up to the user of this methodology and can be deterministic or probabilistic also depending on the approach used to build the social groups. Next, the mobility data requirements are introduced to enable their connection.

3.2.2 Mobility data requirements

As mentioned, coherent zoning is crucial to connect social data to mobility data. Available and valid home zones, as well as activity locations, are crucial regarding the connection to the social group covered in the following section 3.2.3. This means that especially for destination-based approaches a significant amount of mobility data should include the destination points of the travel so that these can be traversed into destination zones.

Further, in disaggregate mobility data sets, travelers should be distinguishable and traceable. Thereby, individuals can be recognized and home bases (buildings, neighborhoods or stations close to expected home) can be inferred from their travel patterns. Each traveler's zone in which this home base is located is needed to match the residential social data to the traveler. This allows to link each disaggregate transaction to a social group and is a crucial part of the method.

As mentioned earlier in section 2.2, destinations are the key to a more activity-based approach of measuring segregation. It indicates where people are traveling to. Therefore, this methodology suggests using observed destination-based mobility data, though is not restricted to it. In principle, any type of disaggregated large-scale mobility data can be used, as long as it fulfills the above requirements. If this is the case, the second process step shown in Figure 3.1 can be addressed to obtain mobility data.

3.2.3 Connecting residential social groups to mobility data

In a third step, social groups are assigned to mobility data as described in Figure 3.1. Important here is to match a significant % of mobility data with social data. Due to the earlier requirements such as assigning a home zone to each mobility transaction, this could be a challenge depending on the data set. But to obtain significant results for the segregation measure a majority part of the mobility data should be connectable with social data. Depending on the case more than 70 % of mobility traces should be connectable to social data in order to obtain a significant number of matches.

This approach links aggregated socioeconomic data to disaggregated mobility data. This incorporates the assumption that travelers out of a zone to some extent represent a homogeneous group.

Assuming both the requirements are met and the assumptions can be made, the data can be matched. This is done by including the home zones' information of each traveler into the mobility data set. Every journey is made by one distinctive user.

This user has a home base and a respective home zone. Via the home zone, the social information gets connected to each transaction the user makes. Knowing every transactions' social classification enables calculating segregation measures which is done in the next step.

3.2.4 Measuring segregation from mobility data

The measure is required to have an index format that calculates some relative distribution of groups within spatial zones. The index format facilitates transferability to different contexts and facilitates comparability.

Further, this methodology is designed for the application on multi-group cases. This relies on the analysis of activity-based segregation studies in 2.2. In fact, it could be used for two-group measurements as well.

Once the above requirements are fulfilled, the calculations can be made. Generally, this method is designed to assess the social mix at the journey destination. Therefore, in the following destination-based approaches are focused. In general, the activity-based approach also works with different activity locations such as transfer stations or similar points of mobility activities.

It should be noted, that even with destination-based segregation calculation aimed at, the social data is assigned to the home zone of each journey since it relates to the residential background of the traveler. With this approach, the social mix at the destination is determined by each traveler's home zone's social group. In other words: travelers' experienced segregation at the journey destination zone, assessed by the segregation measure, depends on their home zone's social status.The compatibility of individuals with the social characteristics of their home region is the main assumption of this method.

The time frame on which the social mix and thereby the segregation can be measured depends on the mobility data available. The smaller the time units, the more realistic segregation measures get. A shorter time span implies a higher probability of meeting or interacting in a spatial unit.

Segregation indices are calculated by time unit. Depending on the function of the segregation measure, it also offers to look into each geographical unit's contribution to the segregation index. Thereby, a zonal analysis is enabled.

When calculating results, the characteristics of time units should be considered. For instance, to obtain one segregation value per zone, an average over all time units could be taken. Such aggregated zonal calculations could lead to a more compromised view on and understanding of segregation.

3.3 POSSIBLE APPLICATIONS – MEASURING THE EVOLUTION OF SEGREGATION

To understand how segregation develops over time, mobility data sets containing disaggregate big data can be of great value. This implicates a straightforward approach of comparing segregation index results over time. For public transport data, some smart card data setups and the connected privacy policies even facilitate tracking a user ID over time. This means that the transaction's card ID is recorded and kept in the database so that via the card ID all transactions of a certain period can be matched to a user. Other mobility data such as GPS could do the same via

a device ID. Using these mobility traces, even an individuals' segregation could be assessed over time.

But there are also some constraints to studying segregation evolution. To set up an experiment over time, it is reasonable to take same types of time spans into account. These could be same weeks or months over several years, ensuring that the results are somewhat comparable. Thereby some control is given over seasonal effects. Although, the control over other effects is very little. A wide range of factors influences the change in use of public transport such as infrastructure and service changes or car ownership. Therefore, interpretations regarding the segregation index over time have to be made carefully.

As the segregation measure introduced hereafter in 3.4 shows, segregation can be measured on a small-scale zonal level. Therefore, the evolution of segregation can be measured even for a single zone. This allows observing a zone's social mix of travelers over time.

This could be precious in the public transport context of this study. As known from accessibility studies such as Arbex et al. (2019), the PT system's supply significantly influences travel activities. Hence, any change to the public transport system might show effects on the segregation measure. Still, this approach does not offer to control for a single variable or change so that no direct interrelations with segregation can be diagnosed.

To measure social segregation and possibly its evolution, the process steps and requirements presented in 3.2.4 must be followed. It includes choosing a appropriate segregation measure.

As touched upon in section 2.1, social segregation-relevant data often incorporates an order such as age, income, educational levels. These data sets are therefore likely to be ordinal which needs to be accounted for when choosing a segregation metrics. The following introduces the *ordinal information theory index* by Reardon (2009). The index incorporates a multi-group ordinal measure which is based on the information theory index of Theil and Finizza (1971); Theil (1972).

3.4 ORDINAL INFORMATION THEORY INDEX

With the income groups being connected to the disaggregated data, the journeys, a segregation index is needed to calculate the segregation experienced at the destination level. Since the income groups inherent a specific order from low to high-income zones, they are considered as ordinal. As discussed in section 2.1, there are different distances between ordinal groups, thus different experienced segregation levels depending on these distances. Therefore, a segregation measure is needed for this case study to incorporate the ordinal nature of the groups.

According to Reardon (2009), an ordinal segregation measure is introduced in the following. The **ordinal information theory index** is measuring segregation as the ratio of between-category variation to total variation. First, the following notations are introduced.

- k = ordered categories (social groups)
- m = unordered categories (neighborhoods, zones)
- t_m = total population in m
- T = Total Population
- c_m = $[K-1]$-tuple of cumulative population distribution in m
- v = ordinal variation

v is measuring the ordinal variation and indicates how close the distribution of a population t is to the minimum and maximum variation (Reardon, 2009). For every unordered category m, the population distribution within the ordered groups is given in the $[K-1]$-tuple c_m. This yields a tuple of cumulative proportions such as (0.2, 0.5, 0.8) over all but the last ordinal class k, which value always adds up to 1.

In the following the metrics to calculate the ordinal information theory index is given:

$$\Lambda = \sum_{m=1}^{M} \frac{t_m}{Tv}(v - v_m) \tag{3.1}$$

Reardon (2009) developed four ordinal segregation measures, of which originally the ordinal information theory index is indicated by using the subscript 1. For this study, only the ordinal information theory index is considered. Consequently the subscript 1 is left out for the measure function Λ or inputs such as v.

As it can be seen from equation 3.1, the variance ratio is set into relation to the ratio of the population in m. While the left part of the function gives the proportion of population t in m to the total population T multiplied by the overall ordinal variation. In the right part of the function $v - v_m$, the difference between the overall ordinal variation v and the ordinal variation v in m is calculated. To obtain the segregation index the sum is taken over all considered unordered categories M. The index results of Λ are ranging from 0 (no segregation) to 1 (max. segregation).

To determine the ordinal variation v_m and subsequently v over all M, the following equation is introduced.

$$v = \frac{1}{K-1} \sum_{j=1}^{K-1} f(c_j) \tag{3.2}$$

The ratio of $[K-1]$ ordinal classes is multiplied with the sum of $K-1$ values of f, the distribution function defined below in equation 3.3. The closer v_m is to 1 the less homogeneity there is in the unordered group m. Contrarily, v_m=0 indicates the maximum amount of homogeneity in m.

As the differences between v and v_m are crucial for the segregation measure Λ, the just mentioned extreme cases of v_m would be expected to have a big influence on the segregation metrics. Though, this would happen only when the overall v is not tending to one or the other side. In addition, every difference between v and v_m is set into relation with the population within the segregation measure Λ which overall could diminish those effects according to the population it is affecting.

The distribution function f uses the cumulative population c calculated for each zone. The below defined f represents the distribution function for the ordinal information theory index.

$$f(c) = -[clog_2(c) + (1-c)log_2(1-c)] \tag{3.3}$$

Figure 3.2 visualizes function f for c ranging from the extremes 0 to 1. f is a continuous function in the interval [0,1] with its peak at $f(0,5) = 1$ and its boundaries at $f(0) = 0$ and $f(1) = 0$. Further, f is increasing within [0,0.5] and decreasing between [0.5,1].

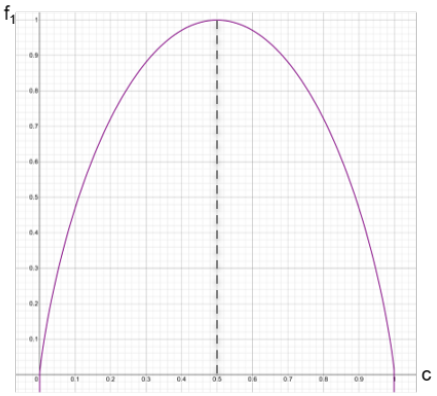

Figure 3.2: Distribution function f

To calculate the segregation index Λ, the cumulative population c is crucial. A data set applicable for this metric would therefore need to be able to indicate cumulative populations in the unordered groups (zones) m. This implies that the data set would need disaggregated population data of which every element can be assigned to one unordered group m and one ordinal group k.

To apply this index on PT smart card data, let Λ_m indicate the contribution of the in-scope zone (=unordered group) m to the overall Λ of all zones M. Further, the income groups are included as ordered groups K. The ordinal information theory index Λ can be calculated even on a small scale with at least two zones taken into account.

After, the connection is made to link the PT smart card data sets of 2017-2020 to the income groups, the above-introduced index is applied. This reveals the results of the Stockholm County segregation study which are shown in the next chapter.

4 MEASURING SOCIAL SEGREGATION WITH STOCKHOLM COUNTY'S PUBLIC TRANSPORT DATA

The framework developed in chapter 3 is applied on the public transport smart card data of Stockholm County, Sweden. At first the case study environment is introduced including its public transport system and income groups. Then, public transport smart card data is connected to the home-zone related income group to measure segregation in Stockholm. Thereafter, challenges of contextual data processing are documented. Finally, the application is set to obtain segregation index results.

4.1 SEGREGATION AND MOBILITY IN STOCKHOLM COUNTY

This subchapter introduces the case study context of Stockholm County, the most densely populated area in Sweden with 2.4 million inhabitants (SCB, 2021). The wealthy northern European country was found to face increased levels of segregation, 20 to 25 years after the shift from "the Swedish model", a comprehensive welfare system, to a neoliberal housing policy (Hedin et al., 2012; Andersson and Turner, 2014; Grundström and Molina, 2016).

Segregation in the Swedish capital Stockholm is mostly connected to findings on residents' ethnics and income (Andersson and Kährik, 2015). The study of Hedin et al. (2012) found a "growth of supergentrification and low-income filtering" between 1986–2001. While the gentrification took place mostly in the northern and eastern islands of the metropolitan area, the low-income filtering tended towards the northwest and southwest corridors. In recent years, especially low-income groups seem to be segregated towards the outskirts (Grundström and Molina, 2016). As a result, Stockholm has more residential poverty segregation than other European metropolises (Haandrikman et al., 2021).

Recent years showed several endeavors of the authorities to tackle segregation such as the Citybanan project discussed in 4.1.2. Thereby, segregation could potentially have been diminished which requires reassessment. Since this study intends to incorporate an activity-based perspective, it goes beyond residential factors. As concluded earlier in section 2.3.1, segregation can be set into relation with more activity-based views by including mobility data.

In the following, the County's public transport system is introduced. Thereafter, the segregation case study is set into context with recent changes to the public transport system which are expected to have impact on the segregation results.

4.1.1 Stockholm County's public transport

The transport authority Storstockholms Lokaltrafik (SL) is in charge of the public transport system in Stockholm County, Sweden. Its public transport system is using the smart card technology for automatic fare collection of the metro, light

rail, bus, ferry, and commuter train transport.

Stockholm's public transport network has a radial structure according to the splattered islands the city is built on. An overview of the inner Stockholm public transport system is given in Figure 4.1. Its central station and its counterpart Slussen on the south side represent the bottleneck of many (rail-related) transport activities. Outside of the city center of Stockholm, the public transport system depends on the regional train traffic and commuter trains as well as busses and ferries.

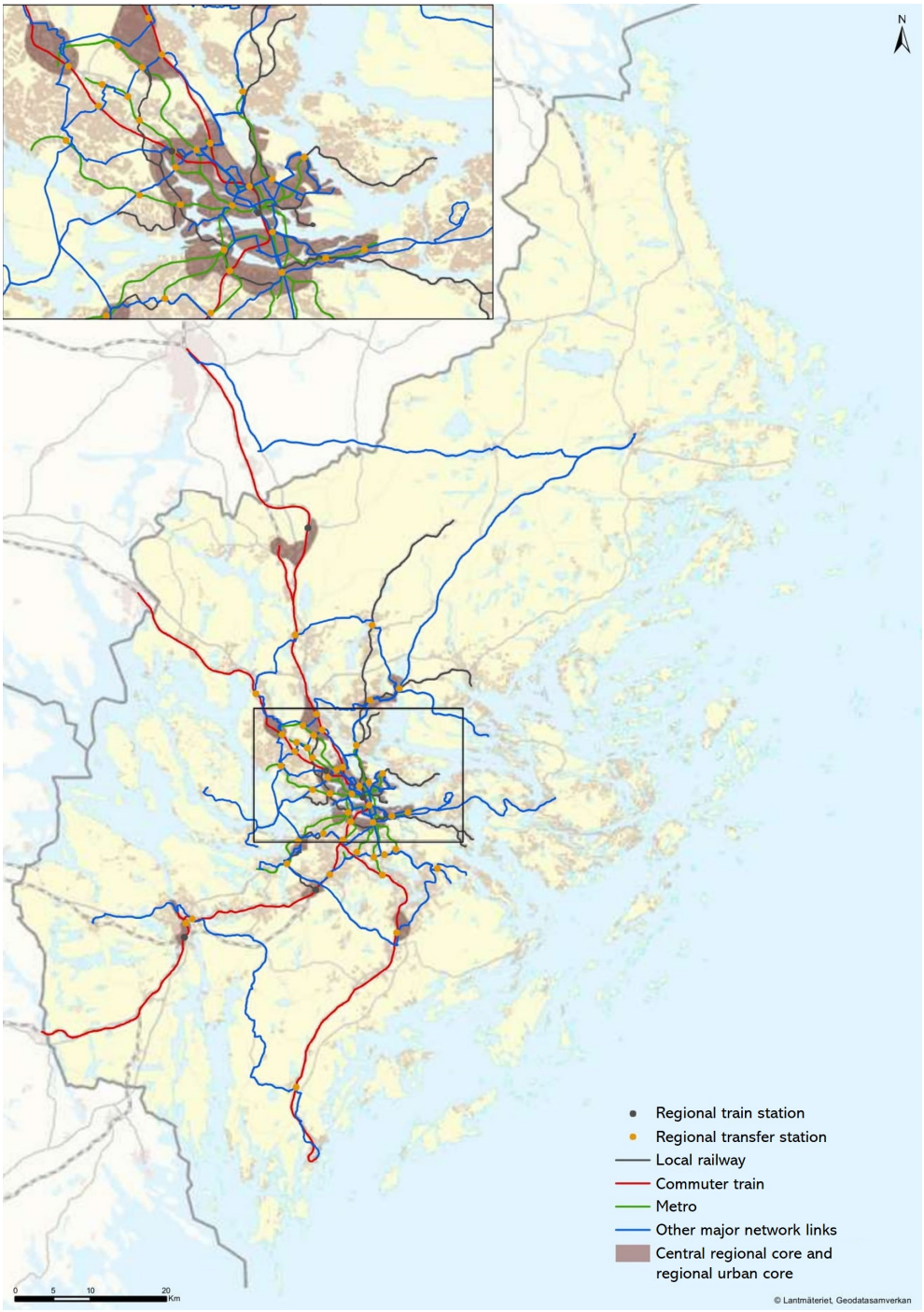

Figure 4.1: Rail-bound PT in Stockholm County (Stockholms läns landsting (SLL), 2017)

In general, the Stockholm region has a high public transport usage in a dense network. Since Stockholm is built on multiple islands, the transport system

depends on bridges and few built connections over the water. As a result, the car traffic in Stockholm can be heavily congested. The municipality tries to distress that with a congestion toll each inner-city car user has to pay on weekdays. These costs motivate additional public transport usage.

In general, Stockholm County constitutes a densely populated and crowded city where the public transport system entails a crucial role. Like many European urban centers, The Stockholm region continues to grow. While the PT system continues to grow in usage, transport-related and in particular PT-related investments, are even more expensive due to the geographical situation.

Since segregation is found to be a fundamental issue in the Stockholm region, the following broaches the connection of a major public transport improvement and possible segregation effects.

4.1.2 The segregation impact of the new commuter train tunnel

The Swedish transportation authority Trafikverket and the transportation authority of Stockholm County Storstockholms Lokaltrafik (SL) implemented a major change to Stockholm County's public transport network with the 'Citybanan' project. The core of the project, the new commuter train tunnel in Stockholm's city center, opened in July 2017. It was build to distress both a national and regional bottleneck in the corridor of the inner Stockholm City (Trafikverket, 2020). This major network change by the so-called "Stockholm City Line" tunnel, led to the separation of commuter trains and regional/national train tracks in the inner-city, see Figure 4.2.

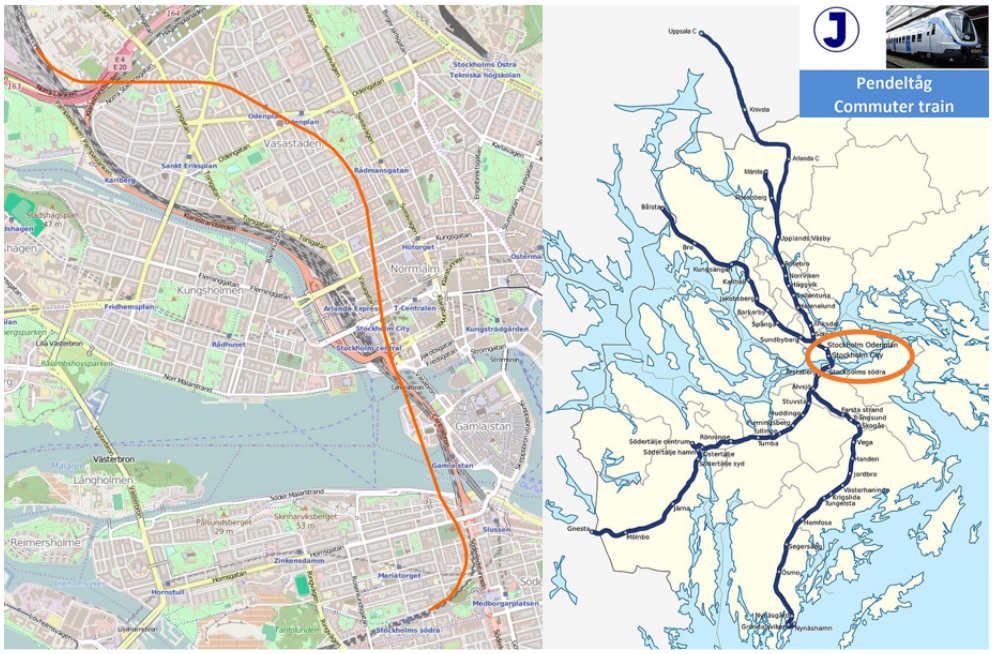

Figure 4.2: Left: Citybanan tunnel Stockholm (regional rail tracks: black-white lines, commuter train tunnel: orange), Ellgaard (2009). right: Commuter train bottleneck in Stockholm County, Frohne (2014)

Therewith it enabled higher frequencies for commuter trains and introduced a new station at "Odenplan" which substitutes the old station "Karlberg Station" (SL, 2020). Another station is built close to the central station "Centralstation", called "City Station". It allows the commuter trains to not interfere with the central station tracks while still stopping in walking distance to each other. In addition,

both newly built commuter train stations are located right under the respective metro stations of the same name.

Since its opening on 10th July 2017, the public transport of Stockholm County changed accordingly in the adaptation of the commuter train lines, so-called Pendeltåg. The project was considered to both significantly enhance punctuality and higher traffic throughput for urban, regional, and even national transport and thereby increase the level of service for the traveler. Part of the investment motivation was to reduce segregation and improve accessibility for outer suburbs. Since the project not only redesigns the existing public transport network but also aims at reducing segregation, Stockholm County is a suitable case study to apply the designed methodology.

In addition, there has been a fare change to the Stockholm County transport system in January 2017. This flat fair policy was introduced with a similar background and its effects are studied by Kholodov (2019). Other than the Citybanan and the fare change, there were no remarkable changes to the public transport system in the time span 2016-2019. With the start of the Covid-related pandemic measures in March 2020 Bus related tap-ins were restricted. Therefore, no data beyond the beginning of 2020 is considered for this segregation study.

4.2 CHOICE OF SOCIAL GROUPS AND GEOGRAPHICAL SCALE

To determine the groups to use for a social segregation study the context of the study has to be taken into account. Common practice is to derive the social status of an individual from socioeconomic variables. These are determined at the point of residency.

In the case of Stockholm County, collecting socioeconomic data is based on the Swedish demographic statistics areas, so-called "Demografiska statistikområden" (DeSo) zones (Sandberg and Palmelius, 2018). These 1287 DeSo zones are made on a small scale of usually a couple of urban blocks up to representing frictions of a neighborhood and even several widespread rural municipalities. They are used to record and allocate Swedish socioeconomic data from censuses or registers. DeSo zones usually include 1500 inhabitants with boundaries set at 600 and 3500 inhabitants per DeSo zone (Sandberg and Palmelius, 2018).

From the official Swedish statistics SCB (2017) there is socioeconomic data available regarding gender, age, educational level, and more. Per DeSo zone, databases often indicate one value per variable. It includes rich data from different sources. For economic variables, the SCB's income and tax register data is collected on the Stockholm County population in 2017.

As mentioned in 2.3.2, the field of building clusters and finding the ideal composition of social groups has been well researched. There can be single socioeconomic variables determining social groups such as income levels or ethnic differences as done by Ivaniushina et al. (2019). Further, socioeconomic variables could be mixed with spatial characteristics so that not only the socioeconomic factors determine the cluster building but also the location of residency.

For the case of Stockholm County's public transport users groups are made from the median total earned income per zone of the 20+ years old population (Befolkning 20+ år efter sammanräknad förvärvsinkomst, Median_ink, Medianinkomst).

Income plays a major role not only for segregation as found out earlier in section 2.1. As emphasized in 4.1.1, there is evidence that segregation in Stockholm is partly driven by different levels of income. Therefore, this study looks into income groups of Stockholm County.

For this study, income quantiles are build which results in four different categories of income. As segregation is proven to be related to low-income groups, the amount of income groups is kept small. This is done with the intention to potentially extract findings regarding the lowest income group. Each DeSo zone is assigned to one of four income quantiles using the median income of each zone obtained from the 2017 SCB income and tax register (SCB, 2017). In total there are four groups splitting up the population as it can be seen in Table 4.1.

Table 4.1: Income quantiles in Stockholm County

Group	Name	median annual income of 20+ years population	Population	Share
1	Lower income zones	below 274K SEK	611,963	26.2 %
2	Lower-middle income zones	274-327K SEK	568,799	24.3 %
3	Upper-middle income zones	327-372K SEK	572,519	24.5 %
4	Higher income zones	above 372K SEK	586,006	25.1 %

Due to using income variables, automatically there is an order imposed on the groups. There are different distances between the income groups. Group 1 and 2 have a smaller gap between each other than groups 1 and 4. From a segregation perspective, these different distances between the groups are crucial to incorporate into the analysis. Groups that are closer income-wise are usually socially less driven apart from each other.

In Figure 4.3, the distribution of income quantiles in Stockholm County is shown. It states a splattered distribution of every income group. Nevertheless, patterns can be seen such as the highest income group centering in and around especially the northern part of the city center as well as the close suburbs of Stockholm City. Particularly in the North and East, rich suburbs are identifiable in Figure 4.3. Income groups 3 and 4 determine the cityscape of the northern inner-city. Contrarily, the South-West, North-West, and some parts of the South of the city belong to the lowest income quantile.

Figure 4.3 illustrates the mosaic-style structure of income distribution. While the highest income group 4 resides central, lower-income groups 1 and 2 are mostly spread around these centers. As a bigger part of the southwest corridor of group 1 zones, the cities of Botkyrka and Södertalje have a number of low-income zones.

Most rural areas of Stockholm County either belong to income quantile 2 and in the North and East also to group 1. Some rural areas are inhabited by high-mid income groups. Overall the low-mid income areas (group 2) are spread outside of the city center especially in the South.

To connect the social groups to Stockholm's public transport data, the groups described in Table 4.1 need to be connected to the home zones of the travelers. In the following section the data collection, inference, and connection to the social clusters are described.

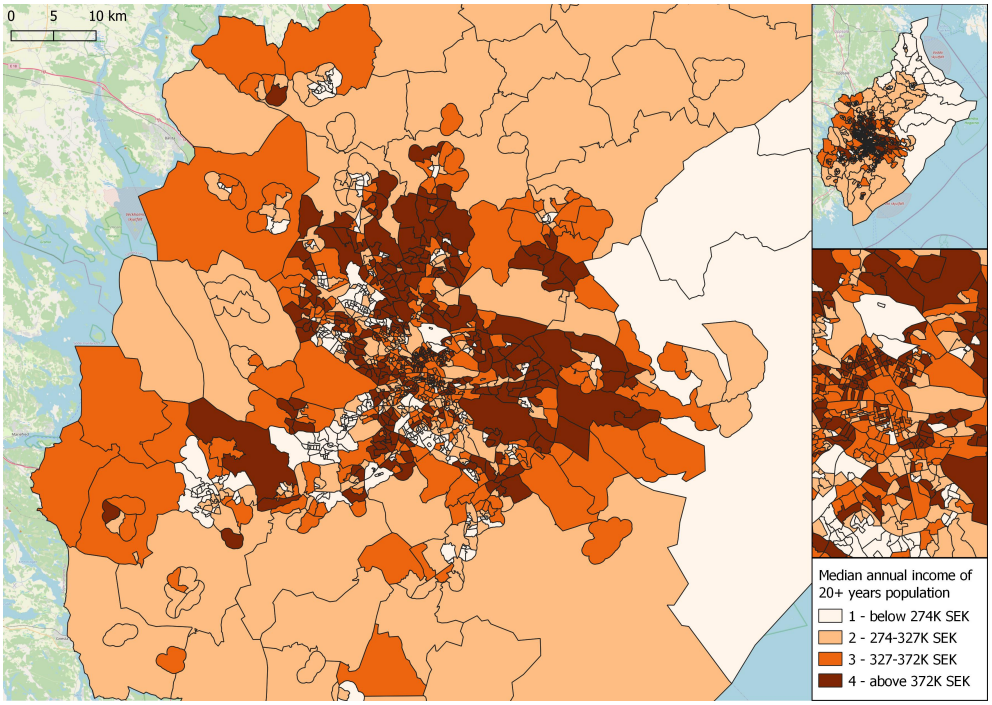

Figure 4.3: Residential income quantiles in Stockholm County DeSo zones

4.3 CONNECTING SOCIAL GROUPS TO PUBLIC TRANSPORT SMART CARD DATA

Connecting the social groups formed before is independent of whether multi-dimension variable clusters or, like in this case, one-dimensional groups are present. As the requirements for mobility data in subsection 3.2.2 already introduced, using mobility data to measure destination-based segregation needs home zones and destinations. In the case of Stockholm County, big smart card data sets are available which are rich but need few inferences and cleaning. These preparation steps are described in the subsection hereafter and followed by the actual connection of the social data to the smart card transactions in subsection 4.3.2.

4.3.1 Data collection, inference and cleaning

The calculation of segregation indexes from SCD is building up on earlier achievements with these data sets as mentioned in section 2.3.2. Before being able to apply an index segregation measure for in this case ordinal groups on this type of large-scale disaggregated mobility data set, the Stockholm data set is prepared.

For this study, disaggregated public transport smart card data is made available by the transport authority SL for all days of the year 2017 till 2020. The time period from the end of January to the beginning of February is indicated as stable by the transport authority. Stability is referring to the regularity of operations itself as well as passenger numbers. This period incorporates very few events that could cause deviations in passenger numbers such as holidays. To obtain a manageable data set, one week is chosen for each year from 2017 to 2020. Therefore, journey smart card data sets of each year's week 5 are considered for measuring Stockholm's segregation of social groups.

These smart card data sets include already matched transactions. This means that the raw transaction data sets of check-in and -out timestamps have been matched

via card keys (user card IDs) to build up journeys. Thereby e.g. transfers are inferred.

As a result, 10.5 to 10.8 million journeys for week 5 of 2017-2020 are present. In the following, these are prepared, towards the specific application for segregation index calculation. As described in section 3.2.2, the first steps most likely imply carrying out multiple inferences to obtain the required full journey data sets. For the case of Stockholm this implies the following three steps:

1. Alighting inference: Since in the case study context of Stockholm County there are no tap-outs, the data set does not include alighting stops. Therefore, an alighting stop inference (location and time) is carried out in accordance with the algorithms presented in 2.3.2. For the regarded years 2017-2020 about 77-83 % of journeys can be linked to a destination.

2. Home zone inference: The method of Sari Aslam et al. (2019) is adopted for home zone inferences in Stockholm. Applying this home zone inference methodology to the study context of Stockholm County was found applicable by the study of Kholodov (2019). Consequently, for week 5 in February 2020 home zones are inferred by the ratio of 9 zone boardings within 4 months. Since the encasing smart card data sets in Stockholm include about 10 weeks of data for instance from 01 January 2020 till 16 March 2020, a threshold of 5 morning tap-ins is determined. Only tap-ins after 04:00 AM are considered for first day-tapins. In this case, about 44 % of all PT users have an inferred home zone. The low inference rate is not too alarming since the rest of the card keys only account for a smaller amount of journeys, thus occasional users. In total, about 73-80 % of week 5 journeys within 2017-2020 can be matched to a destination and a home zone, see appendix B for more detailed information.

3. Stop area allocation: Alighting stop areas of each journey are recorded with the corresponding card ID. Due to the data setup, the destination inference algorithm does not assign to a single stop platform but to so-called stop area numbers. This means that the passengers are accurately estimated to alight in that area, but not on which exact platform. A stop area number (SAN) is an allocation of same-mode platforms to a stop name. It is preset by the transport authority and can represent a single platform and most commonly the classic station with opposite platforms up to terminals with several platforms. The transport authority uses SANs to allocate transactions of mostly similar-named platforms and thereby gain an overview in terms of stop-related passenger numbers. In this case all inferred alightings are made on the level of SANs. Thereby, many stops of the same name/in the same small-size area are geo-coded to one X,Y coordinate.
   This eventually leads to passengers of multiple platforms being assigned to one side of the street or square while the other side also has at least one platform. In the case of Stockholm, it often happens that a street divides two different DeSo zones. Therefore, all passengers get assigned to only one of the zones. Hence, all passenger load gets allocated to one coordinate. This is part of a limitation of this study which is further discussed in 5.6. In Figure 4.4, this phenomenon is shown given the example of the bus station "Hallandsgatan" where two-direction bus platforms of the same stop name are assigned to one stop area. While the platforms are split over two DeSo zones, the SAN of Hallandsgatan is located at the northern platform. The passenger load of both platforms is therefore only assigned to one street side and thereby only to one DeSo zone.

For week 5 in 2020, 8.45 out of 10.85 million journeys include a destination stop area and a home zone. In total, more than 21 % of smart card transactions are taken

Figure 4.4: Stop area assignment - bus station Hallandsgatan

out of the data set due to inferences. The remaining ca. 78 % includes reliable inferences and a significant amount of journeys to proceed with the calculation of the segregation index. Similar conversion rates are given for the years 2016-2019 which can be seen in appendix B.

After the inferences cleared the data set, there are several factors that potentially reduce a data set depending on the case study. For the Stockholm data, this is recorded below.

The data set can include multiple journeys of the same card key on one day to the same zone. This is not an uncommon case of a traveler visiting a zone at least twice a day. Since it represents the number of opportunities to experience segregation, these cases are kept in the data set.

About 2-3 % of all zones are visited by one person per day down to only one person per week. This could be due to the fact that some stops in often rural destinations are so little frequented. Since these individuals are experiencing segregation as well, their journeys are kept in the data set. Their extreme segregation experience is relativized by the small impact on the overall segregation measure. Still, a single zone perspective considering only one of these zones would probably display less realistic segregation experiences.

Circa 0.02 % of all journeys in the 2020 week 5 data set begin on Sunday, day 7 but end on day 8, a Monday morning. But due to no other data for this day, the segregation measure of the 8th day would not be valid and therefore these journeys are taken out of the segregation study. For the other years, these conditions and interventions look the same.

Similarly, less than 50,000 journeys have destinations with a destination DeSo zone outside of the Stockholm County borders. These journeys are left in as long as the card user has a home zone inside of Stockholm County. On the other hand, it should be considered, that only the DeSo zones within Stockholm County are grouped y income for this segregation study. This means that passengers with home bases outside of the County are dropped after the home zone inferences, as explained above.

4.3.2 Connecting income groups of Stockholm to PT data

To connect the residential-based social data according to subsection 3.2.3, the income groups are assigned to the journeys' home zones. Due to the inferences and filtering before, all destination and home zone inferred journeys can be assigned to an income group. This builds the basis for the segregation index calculation in the section hereafter.

Figure 4.5 provides a first overview of 2020 journeys to destination zones. This includes all income groups. It can be seen that the load of journeys varies

significantly from 1 passenger arriving at a zone to more than 800,000 arriving per week in the busiest zone, the central station area. It is clearly displayed that the city center of Stockholm is the busiest. Still, even in this area the passenger loads arriving vary highly. In general, a splattered structure of public transport usage is revealed.

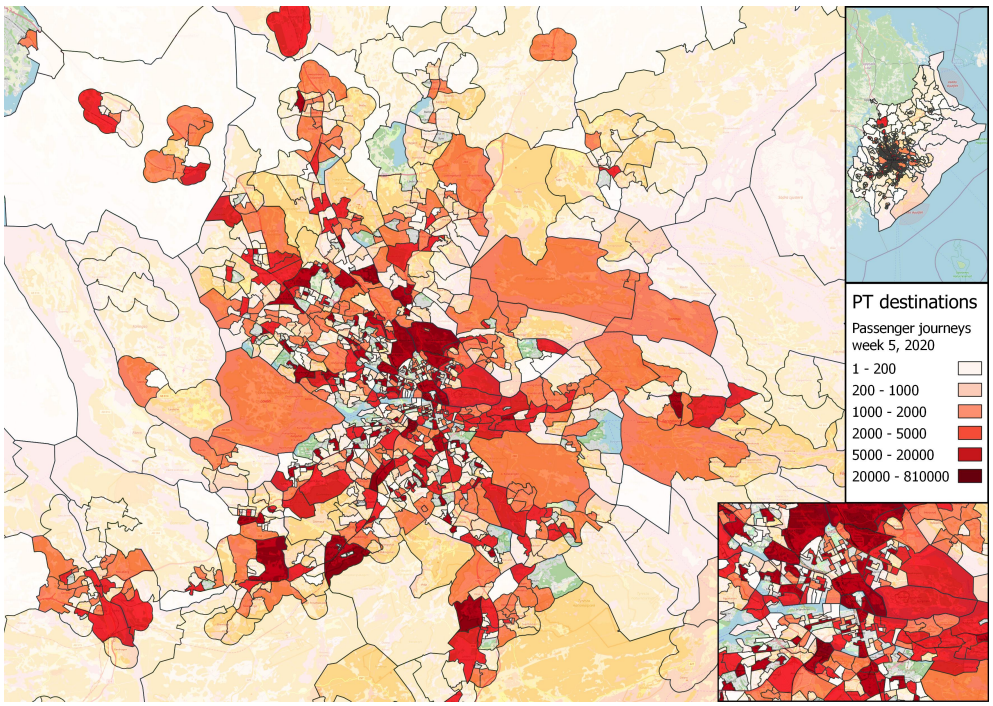

Figure 4.5: Public transport destinations 2020 - Number of journey destinations within DeSo zones

In addition, Figure 4.5 shows the radial structure of the public transport system of Stockholm. Easily, the suburban cities can be detected looking at the colorful spots outside of the center. In addition, a tendency of public transport passengers towards the North-West and South-West can be seen. This reveals the structure of inhabited areas around the Swedish capital.

Figure 4.6 illustrates income group's destinations in week 5 of 2020. It aims to answer where income groups are traveling to using the public transport system. In total, 8.45 million journeys are identified including a home and destination zone. These split up over all groups as presented in Table 4.2.

Group 1 appears to travel through the northwest and southwest corridors. Also, this group of travelers coming from the lowest income zone appears to have the most centralized travel profile. Group 2 tends to be more widely spread in space. The same counts for group 3 but with more tendencies to the East compared to group 2. Income group 4 has a more urban travel profile as well. Looking at the travel directions, group 4 tends more towards the South-East and North of the city. Income group 4 also creates a large share of journeys to inner-city zones.

As Table 4.2 shows, there are differences between all income groups in the number of journeys made. Contrarily to the somewhat similar population in every group, the total journeys are split unevenly. It can be seen that the journeys made per passenger decrease with higher income groups. While the lowest income group combines 31.8 % of all journeys, the highest income group only accounts for about 20 %. Breaking it down to an individual level, the lowest-income group inhabitants

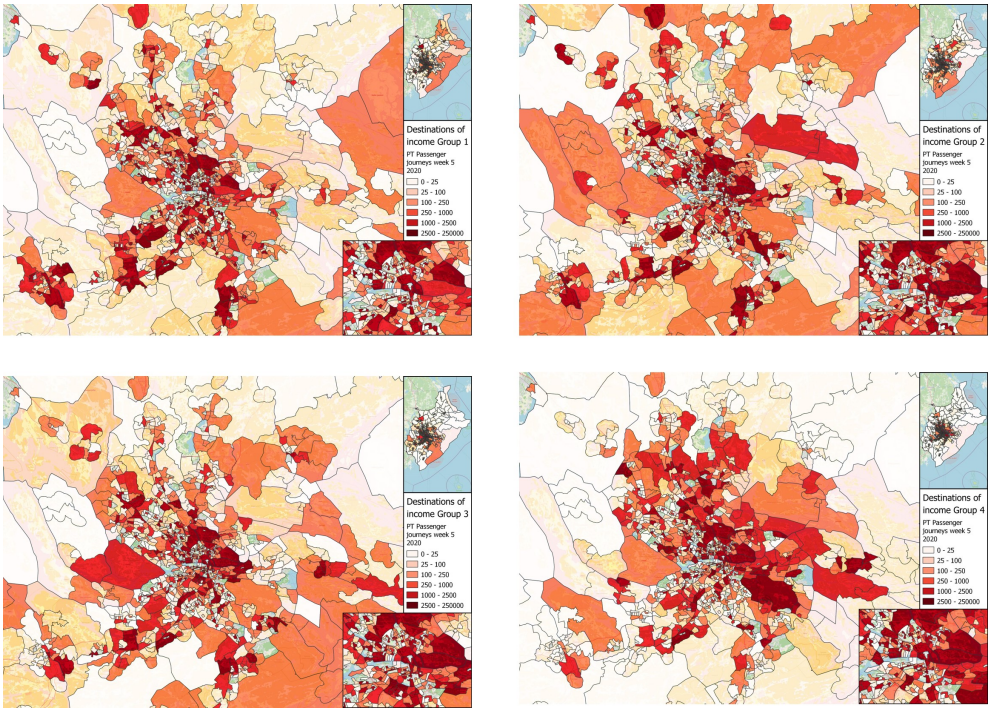

Figure 4.6: Income group's PT destinations 2020 - Number of journey destinations within DeSo zones

Table 4.2: 2020 PT Journeys per income group

Income group	Share of total journeys	Journeys per inhabitant
1	31.8 %	4.4
2	25.9 %	3.8
3	22.2 %	3.3
4	20.1 %	2.9

have an average of 4.4 journeys per week while the highest-income group is close to three weekly PT journeys. This can be explained by people from higher income areas travelling by private modes instead of using public transit.

To illustrate this example, differences in traveling of the most contrary groups' journeys are compared. Therefore, Figure 4.7 shows the subtraction of PT journeys of group 4 from group 1 journeys. This means that a positive balance indicates more group 1 travelers and a negative evinces more group 4 travelers.

A deficit can be spotted in many urban areas, where significantly more inhabitants of one income group travel. Rural areas show minor differences between the groups' amount of travel while some indicate slightly more group 1 travelers. It should be considered that when looking at the differences, group 1 has almost 12 % more share of total journeys.

As differences between groups 1 and 4 can be seen in the Figure 4.7, the latter travels more to the East and North of the city while the lowest income group produces more journeys in the northwest and especially southwest corridor. Also, parts of the South and central areas are dominated by income group 1 over group 4. Especially, in central zones it could relate more to the advantage in passenger numbers of group 1 rather than really indicating a low-income occupied area. On the other side, some parts of the city center appear to be higher frequented by

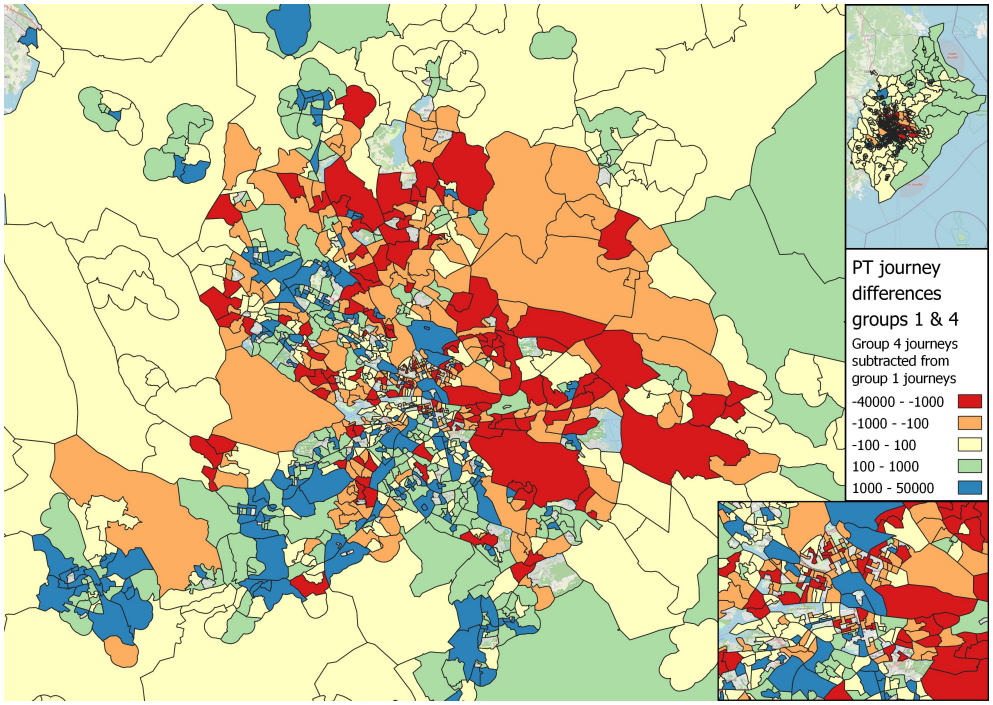

Figure 4.7: Differences in numbers of 2020 PT journey destinations between income group 1 and 4

income group 4 travelers.

Summarising the above analyses, it can be stated that there are significant differences in the public transport usage of the four different income groups. Not only are there substantial differences in the numbers of journeys but also in the spatial distribution of these across the groups. This allows making initial assumptions that there is a certain degree of income separation within the public transport system. To confirm correlations, however, it requires a differentiated measurement of segregation. In the next step, the analysis of PT destination-based segregation should be enabled. For the income groups, the segregation index per day is calculated using a suitable index measure which is introduced in the next section.

After applying the framework from 3.1 and exploring the data set up, now the segregation index introduced in 3.4 can be applied. Doing so reveals segregation results for the Stockholm case study as presented and discussed in the next chapter.

5 RESULTS AND DISCUSSION

In this chapter, the Stockholm case study results are presented and discussed. First an analysis is made over the days of the week. Zonal segregation and its evolution are then presented. In looking at segregation changes over time, the focus is on some zones of particular interest. Following up, the results of applying the ordinal information theory index on smart card data are discussed including a summary of the approach's limitations.

5.1 TEMPORAL SEGREGATION ANALYSIS

First, a high-level view using all DeSo zones is made by calculating the daily segregation level with the ordinal information theory index for each year's week 5. As discussed at the beginning of chapter 2, temporal factors are associated with social segregation and can be of great importance in the analysis of segregation composition. Segregation levels often vary over time. The latter is assessed below by looking at segregation differences over weekdays.

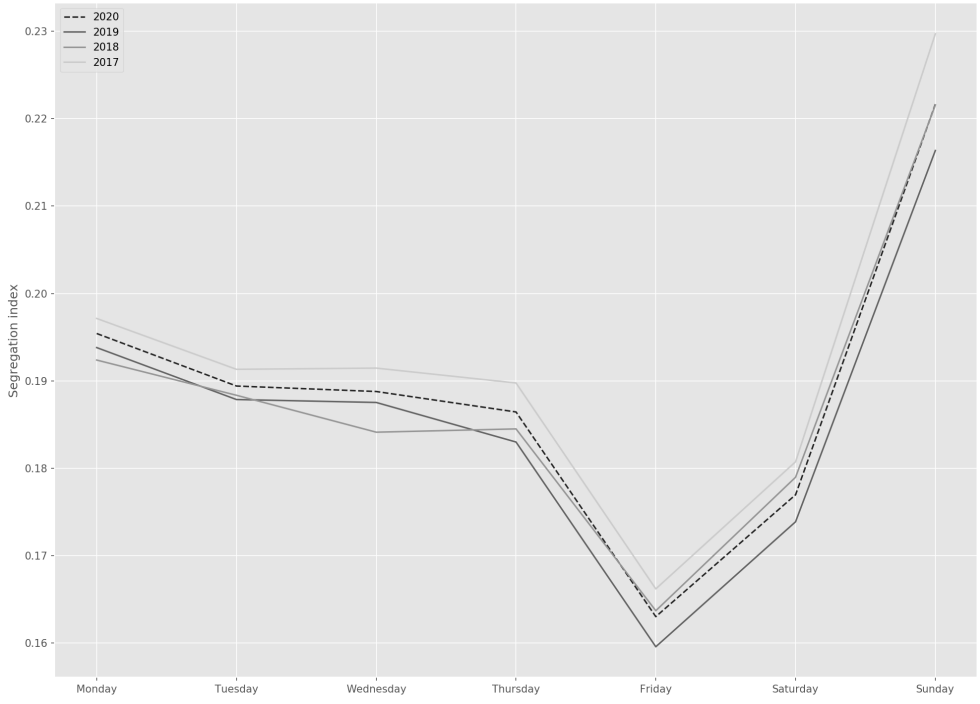

Figure 5.1: Segregation index levels over all days of week 5

As shown in Figure 5.1, the segregation indices calculated vary between 0.16 and 0.23 over the days of the week. After 2017, declining segregation levels are found until 2019 with a fallback in 2020. The social segregation index score for each day of the week in 2017 averages 0.1923. Compared to 2017, the average segregation drops by 2.4% to 0.1877 in 2018 and by 3.3% to 0.1856 in 2019. In 2020, the segregation level averages 0.1888, up slightly from 2019 and 2018 but still 1.8% lower than in

2017.

Figure 5.1 shows small index declination between 2017 and 2018 and for most days even more to 2019. This implies that segregation levels were the lowest in 2019. Looking at 2020 the index displays lower levels than 2017 but higher segregation than in 2019 and for Monday to Thursday in 2018.

Looking at the results in Figure 5.1, it shows that people mix on a similar level on Monday to Thursday. Even less segregation is indicated for Fridays and Saturdays. On Sundays, travelers mix less and experience more segregation as the index is leaning more towards 1 than on other days.

The temporal assessment shows the general development over years and days of the week. It can be seen that the index values of each year have very similar patterns over the weekdays. In the following the interest is to find out where these drops in overall segregation originate from. First, the segregation index is split up into a zonal level. Afterwards, the differences of zonal segregation are assessed in section 5.3.

5.2 ZONAL SEGREGATION ANALYSIS

Figure 5.2 shows the "absolute" and weighted segregation contribution of each zone in 2017. The weighted value implies each zones contribution to the segregation index calculated by $\frac{t_m}{Tv}(v - v_m)$. The absolute contribution is the result of $v - v_m$. In other words, the absolute value expresses the contribution before being set into relation with the number of passengers affected while the weighted value accounts for the number of passengers affected compared to the overall amount of passengers. For both visualizations shown, the zones are split into quantiles, equally counted based on their segregation contribution.

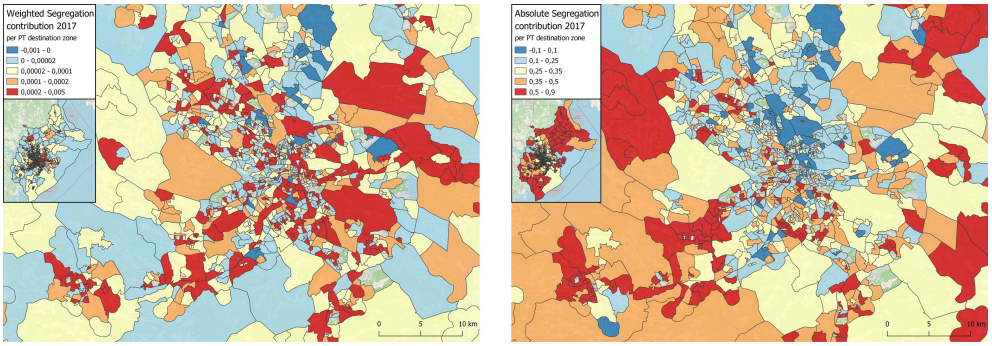

Figure 5.2: Weighted and absolute segregation contribution 2017 - each DeSo zone's arriving PT passenger mix contribution to the segregation index level

Non-colored fields in Figure 5.2 imply that there is no segregation contribution calculated for this zone. This means that no passenger arrived in the zone in the respective time span. Especially for central zones, this also relates to the allocation of stops as explained in 4.3.1.

It can be seen, that the weighted segregation contribution is the highest in central zones and suburban centers. Also, Figure 5.2 illustrates low weighted segregation contribution in rural areas but high absolute segregation. Conversely, many central zones are low in absolute but high in weighted contribution. This means that the

actually experienced segregation in these zones is rather low but because so many passengers experience it, the actual (weighted) contribution to the segregation index is higher than average.

What stands out is the general pattern of zones with both high absolute and weighted contributions to the segregation index. This means that many passengers experience segregation at these destinations. In particular, passengers traveling to northwest and south as well as southwest locations might experience segregation. For 2017, outskirt neighborhoods such as Tensta, Rinkeby, and Fagersjö and suburbs like Vårby, Botkyrka, and Södertalje are indicated to have high absolute and weighted contributions to the segregation index. This implies that a substantial amount of public transport passengers arriving in these zones have less mixed income groups. Also, some inner-city districts like Hammarbyhöjden show high levels for both weighted and absolute segregation.

For the years 2018-2020, appendix C presents both absolute and weighted segregation contributions. Equal to the visualization in Figure 5.2, the rest of the years' segregation is visualized. Similar phenomenons are observed when comparing 2018-2020 with the 2017 visualizations. The absolute contribution over the years shows mostly the same structure, with the rural areas showing high levels of segregation as well as the southeastern suburbs and the northwestern city and suburbs. Patterns of weighted contribution stay similar over the years. Contribution to the segregation index mostly comes from the load of passengers in the city and suburban centers. Compared to 2017, 2020 shows little more segregation indication in the outskirts and overall a more scattered distribution.

5.3 ANALYZING THE EVOLUTION OF SEGREGATION

In the next step, the development of the zone's segregation is focused to find the exact spots of declining or increasing segregation. After the Citybanan change in July 2017, the following years' segregation levels might show effects. As the temporal analysis in 5.1 shows, no significant differences were found with respect to segregation patterns during the week. Therefore, looking at the days of the week plays a less important role, so that weekly averages are taken and compared across years. This leads to a temporal and spatial analysis of weekly zonal segregation changes.

This can be derived from the segregation index results presented in 5.2 without any recomputation by looking at the weighted segregation contribution $\frac{t_m}{Tv}(v - v_m)$ per zone m and its change over time. As explained in the section before, the weighted segregation contribution provides the contribution of every zone to the segregation index.

By taking differences of weighted segregation contribution between years, it can be seen whether the specific zone contributed to a decline or rise of the segregation index. Technically, for every zone and day of the week, the difference is calculated by taking the more recent year's contribution and subtracting the 2017 contribution. Then, the average of differences is determined per zone over all days of the week. Thereby, the evolution of segregation can be assessed on a zonal level. A negative difference will indicate a decline in segregation. Contrarily, a positive difference shows an increase in contribution to segregation.

First, the years 2017 and 2018 are compared. Figure 5.3 visualizes some dark green spots point out a strong, decreasing change in segregation. The more red a zone is colored, the more the change leans towards an increase in segregation. Most of the

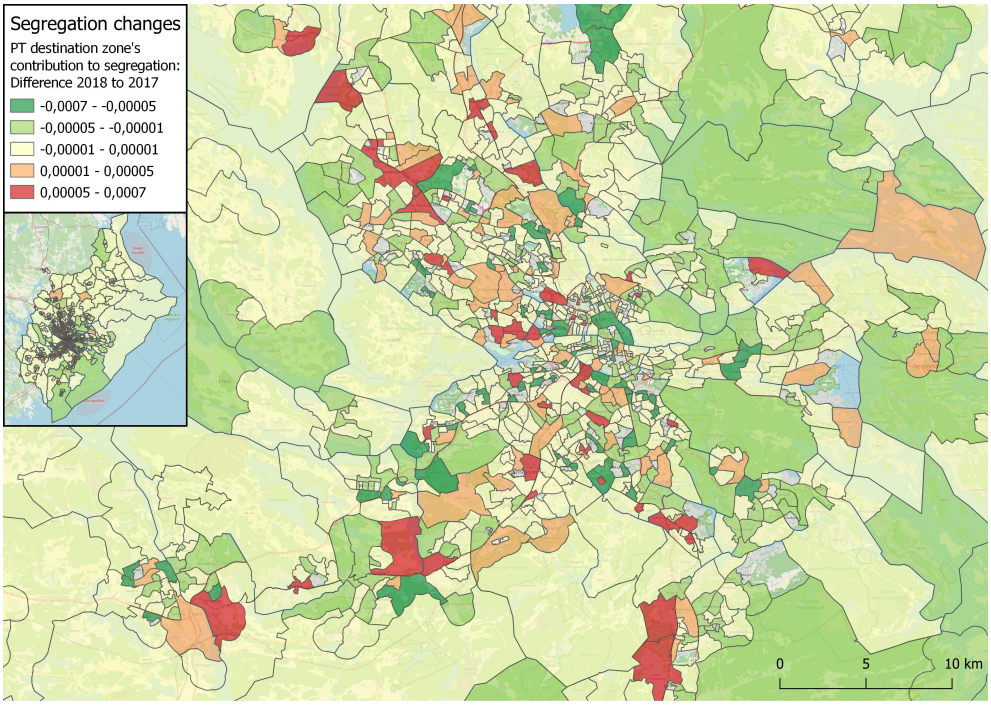

Figure 5.3: Segregation changes 2017-2018 - Change of each DeSo zone's arriving PT passenger mix contribution to the segregation index level. Decreasing levels indicate less segregation contribution

zones indicate a minor change within -0.05 to 0.05.

Comparing 2018 segregation contribution with 2017, Figure 5.3 displays mixed, scattered effects with decreasing trends in the North-West and South-/South-West of the city. Also, there are small effects of declining segregation in the Eastern suburbs. Rural areas' segregation decreased especially towards the South of the county.

The dispersed structure and inner-city zone size make it difficult to immediately spot patterns and estimate effects. In addition, the zone sizes have an impact on the impression of segregation levels, though it does not indicate the number of passengers affected. In the city center, few sharply increased segregation zones can be detected accompanied by strong decreasing and slightly to not decreasing segregation levels. Urban zones with segregation reductions outnumber the ones with rises for 2018.

Mentioning the levels of segregation change, the histogram in Figure 5.4 helps to pick out the highlights.

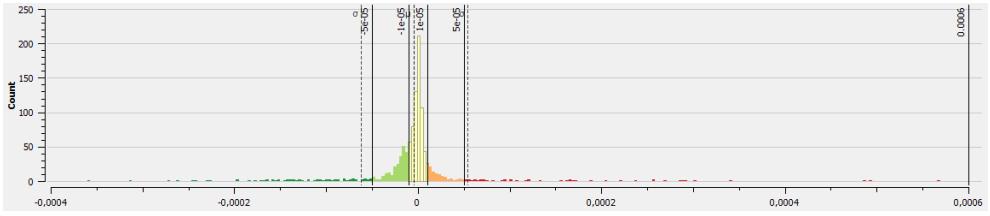

Figure 5.4: Distribution of Segregation contribution changes 2017-2018

The histogram in Figure 5.4 shows normally distributed segregation changes per zones. Counting up the values of segregation contribution into bins, the graph clearly displays that the majority of zones' segregation contribution changes barely.

Further, it shows a higher number of changes towards less segregation than increased segregation.

Figure 5.5 shows the difference in contribution to the segregation index of the years 2017 and 2020. Assessing the 2020 changes might offer a long-term perspective on the effects of the Citybanan project.

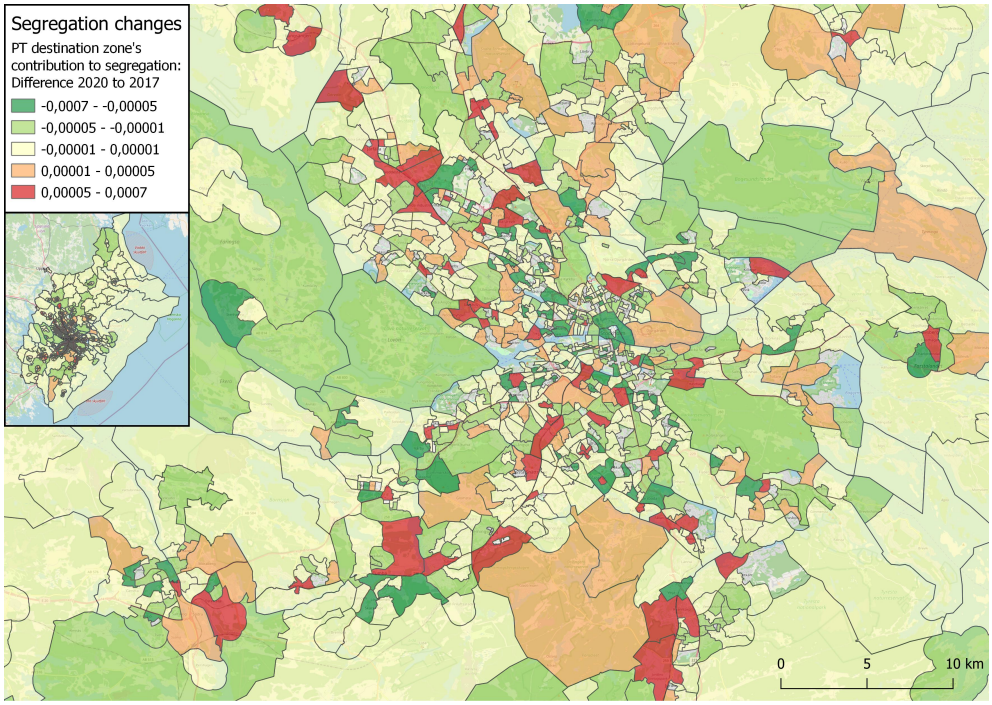

Figure 5.5: Segregation changes 2017-2020 - Change of each DeSo zone's arriving PT passenger mix contribution to the segregation index level. Decreasing levels indicate less segregation contribution

It can be seen that the illustration shows similar patterns as the 2018 comparison, but with slightly fewer decreasing effects in total. Again, a histogram in Figure 5.6 helps to look into the distribution of changes.

The histogram in Figure 5.6 shows the count of zones' segregation changes. These are similar to the 2018 changes as well. While many zones remain fairly unchanged, diminishing change towards less segregation outweighs increases in segregation.

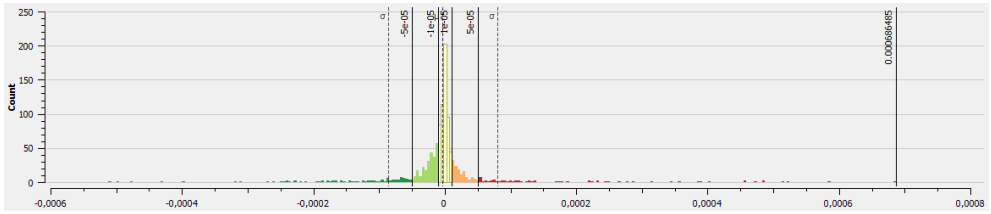

Figure 5.6: Distribution of Segregation contribution changes 2017-2020

For both 2018 as well as the 2020 comparisons to 2017 there are more zones where less segregation is experienced than zones that experienced more segregation. Even though for both years there are some zones' which experienced stronger inclination in segregation (given in dark red), the amount of substantial declines (given in dark green) outweighs it as the mean change of zonal contribution for 2018 is -0.0000043

and for 2020 is -0.0000032. The overall change in 2020 is about 25% less compared to 2018 changes which matches the overall fallback trend stated earlier.

To find out what these changes mean for the social segregation situation in Stockholm, a test comparing the means of segregation indices can be helpful. This significance test is performed on the data sets of zone's weighted segregation contribution of both 2017-2018 and 2017-2020 comparisons. Two t-tests are performed to determine whether the means of the respective data sets are significantly different from each other.

In appendix D these significance tests performed on the weighted zonal segregation contribution are shown. The segregation per zone of both years 2018 and 2020 are compared to the base year 2017. For the same years, the t-tests did not reject the null hypothesis so that the difference between the sample means is not convincing enough to say that the data sets differ significantly. This implies that the general segregation development is not found significantly different to the base year 2017 which is further discussed in section 5.5.

5.4 SEGREGATION CHANGES IN FOCUS AREAS

Some remarkable zones' change of segregation index is presented in this section. Potentially, urban neighborhoods and neighborhoods around the new commuter train line could show decreasing segregation after July 2017. The Citybanan project facilitated higher frequencies and an overall improved public transport system. This should lead to more accessibility to reach and mix in downtown and other central-urban areas. Also, accessibility towards the suburbs is improved. Therefore, passengers could potentially experience more mixing at these destinations as well.

First, the city center is presented in Figure 5.7. Similar to earlier stated results, the effects are mixed as well as scattered throughout the zones. In the city center, some stronger decreases can be spotted in the central, western and southern zones.

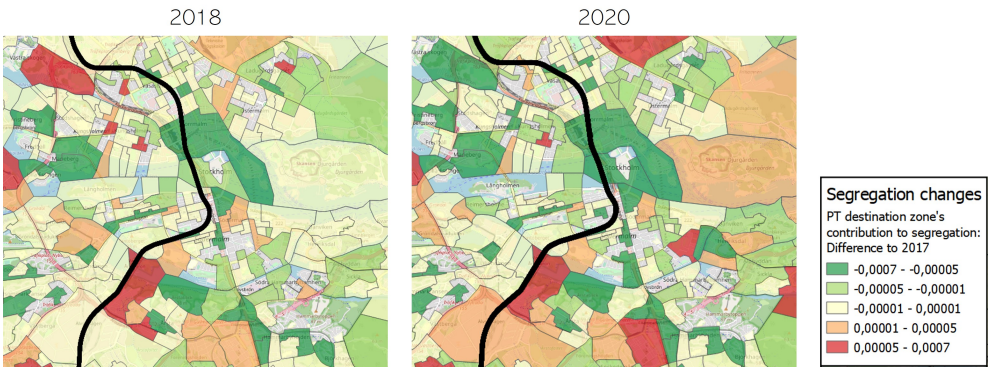

Figure 5.7: Stockholm city center segregation changes - Differences in zonal segregation contribution from both 2018 and 2020 compared to 2017. Decreasing levels indicate less segregation contribution. Commuter train lines are drawn as black lines

The segregation of urban zones, which incorporate the central station, Stockholms södra, as well as the new Odenplan commuter train station, strongly decreases for both years. In addition, the central northwestern Sundbyberg and Solna station zones show declined segregation levels for 2018 and 2020. Figure 5.7 also shows more substantial inner-city segregation reduction in 2020 than in 2018.

But especially when looking at the sprinkled red zones in Figure 5.3 and 5.5, the segregation increases seem to correlate with the radial structure of the public transport system. Therefore, the commuter train lines are drawn as black lines to provide an overview of where PT is enhanced after 2017. This is included for both Figures 5.7 and 5.8 so that the following analysis zooms out to zones connected to commuter train services.

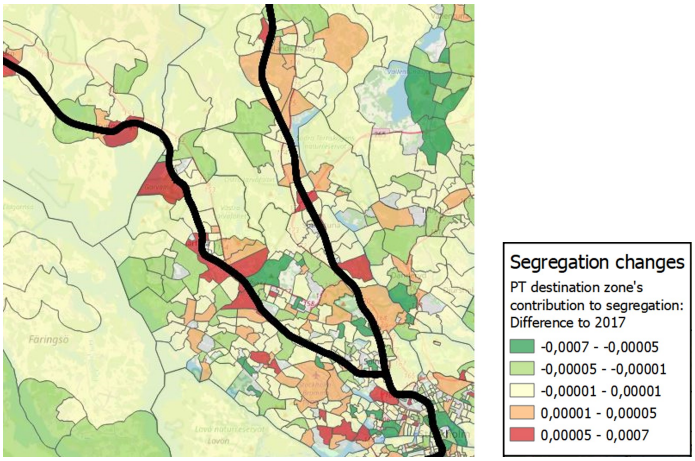

Figure 5.8: Segregation changes along the northwest corridor - Differences in zonal segregation contribution from 2018 compared to 2017. Decreasing levels indicate less segregation contribution. Commuter train lines are drawn as black lines

Outside of these very central located zones, zones with commuter train stations display a small to mostly large increase in segregation. Figure 5.8 displays the occurrence of increased segregation between 2017 and 2018 along the commuter train line for the northwest corridor. The commuter train station zones are actually found responsible for the radially spread dark red zones. This leads to the depiction of partly decreasing segregation levels in suburbs with one strongly segregation-increasing zone - the zone of the commuter train station and local PT hub.

There are few exceptions to the above-discussed phenomenon such as the southern rural station zones Ösmo, Gröndalsviken, and Nynäshamnn which display rather decreasing segregation changes. This matches the earlier observed segregation decline in the South of the County. Surprisingly, also the 2019 opened southern station Vega indicates increasing segregation development in 2020.

So far, the segregation changes next to commuter train stations are exemplified. In the following, more context is added to this perspective as well as other zones somewhat further away from the commuter train stations are considered. Since Stockholm public transport improved as a whole, segregation developments are interesting to observe in various areas where social mixing is not given by default.

For Stockholm County, the neighborhoods Hammerby sjöstad, Djurgårdsstaden, Kista, Vällingby, Skärholmen are indicated as interesting areas to observe segregation developments. These PT-connected neighborhoods include various DeSo zones which partly have specific backgrounds and points of interest that could attract or repel different income groups.

The northwestern neighborhood Kista incorporates a high-tech business district in a historically low-income area. Next to its northeast border it connects to the commuter train station of Helenelund and has its own metro station. Only two

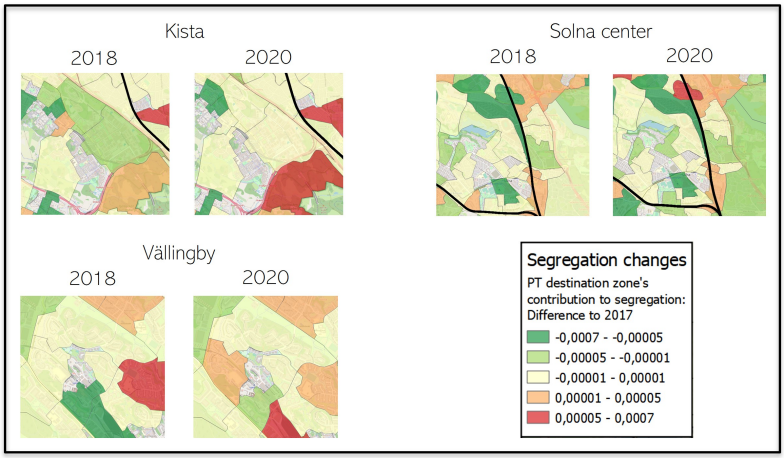

Figure 5.9: Segregation change in northwest focus areas - Differences in zonal segregation contribution from 2018 and 2020 compared to 2017. Decreasing levels indicate less segregation contribution. Commuter train lines are drawn as black lines

commuter train stations apart, the well-connected Solna can be found with its diverse income group structure. It represents a bigger suburb right next to the city with various points of interest including a stadium and large shopping center. Vällingby is the third focused area in the northwest corridor. It does not directly connect to the commuter train but is served by several metro lines and buses. Vällingby is a rather remote suburb with extensive shopping opportunities and zones with the highest as well as lowest income groups living there. All three zones show high public transport usage, especially in the zone including the mentioned PT hubs, see Figure 4.6. Zooming in on what is illustrated in section 5.3, Figure 5.9 presents these three areas' segregation changes for both 2017-2018 and 2017-2020 comparisons.

Next, Figure 5.10 visualizes two more zones southwest of the city, and a developing neighborhood close to the center. Skärholmen is a low-income area located in the southwest fringe of the city. It holds a shopping center and a metro station next to it. Djurgårdsstaden is a relatively new-developed high-income in the northeast of the city center. These two areas are not directly related to the Citybanan tracks. Contrarily, Södertalje is connected multiple times to several commuter train lines. The working-class mid-sized city outside of Stockholm incorporates a remarkable amount of industrial areas. Most of its zones are classified as low-income areas.

For both figures 5.9 and 5.10 mixed effects with partly less segregation for 2018 compared to 2020 can be observed. Interesting is that some neighborhoods' segregation contribution strongly decreases in the zone where PT stations are located that are not related to the commuter train. This can be observed in Kista, Vällingby and Djurgårdsstaden and for 2018 in Skärholmen.

As mentioned earlier, commuter train station zones further away from the center have increasing segregation levels, like Södertälje's stations and the Helenelund station which can be found in the Kista visualization.

Skärholmen seem to have varying effects with the West of the area having increasing segregation levels for both 2018 and 2020. Overall, Södertälje shows more zones with decreasing than increasing segregation. Particularly the low-income central Western and northern parts of the city are decreasing in segregation. The Eastern part Östertälje and its commuter train station increased in segregation levels for

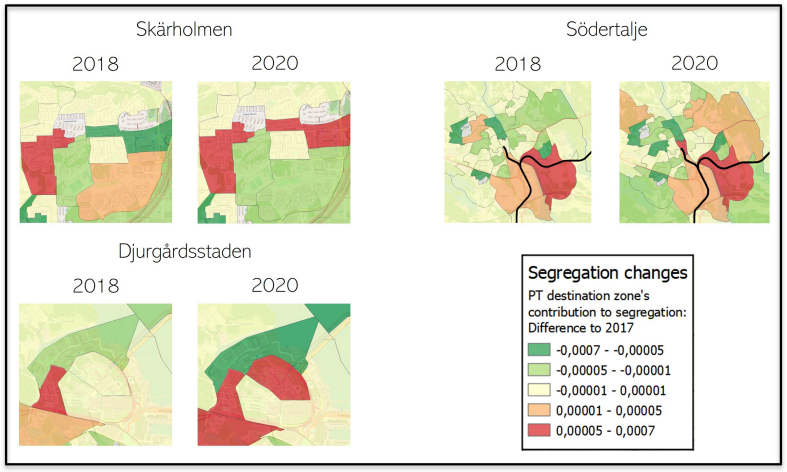

Figure 5.10: Segregation change in east/southwest focus areas - Differences in zonal segregation contribution from 2018 and 2020 compared to 2017. Decreasing levels indicate less segregation contribution. Commuter train lines are drawn as black lines

both years compared to 2017.

To some extent, the zones in Kista show decreased segregation. This can be observed in the office areas and the nearby neighborhood of Husby. Djurgårdsstaden shows decreased segregation for the waterfront development zones in 2020 while at the same time the zones of Hjorthagen reveal higher segregation levels.

To conclude, double-edged results are obtained for most focused zones. Since these findings are based on public transport data, the results are examined regarding PT locations to identify the origins of segregation trends. Looking at the development of segregation, central zones with public transport hubs mostly have declining segregation levels over the years. Zones outside of the city center show increasing segregation for commuter train stations and, to some extent, decreasing segregation levels for other PT hubs.

The following chapter discusses the results of sections 5.1-5.4 and sets them into context with the earlier findings of this study.

5.5 DISCUSSION

In summary, the results presented in the sections above, provide substantial insights into the features of the segregation measure. Overall, the results proved the functionality of the method developed before in chapter 3. In this section, the results are discussed and put into context both methodologically and empirically.

Segregation levels shown in Figure 5.1 indicate a rather low level on the scale of 0 to 1 the index is operating on, with 1 indicating maximum segregation and no mixing. Still, to interpret the index values it lacks context by e.g. comparing it to other cities' mobility-based segregation results. However, the weighted differences between the years' weekly average index allow accurate conclusions to be drawn on the segregation development. For the Stockholm case, the segregation index reported less segregation in the years after 2017.

The results presented in 5.1 indicate stable segregation levels on the usual workdays Monday to Thursday. Figure 5.1 shows the lowest segregation on Fridays which can be related to the combination of work, leisure, and shopping activities. A similar, but less pronounced phenomenon could explain the second-least segregation on Saturdays. What stands out is Sunday since it displays the highest segregation levels. This could be related to lower usage of public transport and the circumstances that a Sunday is considered as a rest day with the least working activity which leads inhabitants to stay more within their home zone. Compared to home-based calculation, work-related activities diminish segregation (Ellis et al., 2004). The segregation results of this study, displayed in Figure 5.1, potentially match these findings but also point to the importance of leisure, shopping, and other, often weekend-related activities. Contrarily to the findings of Abbasi et al. (2021), this study cannot confirm low interaction levels on weekends.

Further, the weighted segregation results from Figure 5.2 are consistent with the public transport usage described in Figure 4.5. This can partly be explained by the population-related index calculation used in this study, see equation 3.1. In short, it is assumed that where no passengers arrive, no segregation can be experienced. Also-relating to the population-dependent index calculation, it shows how the initially segregated outskirts do not play a key part in the calculation of the segregation index since fewer passengers experience it. In turn, the lower experienced segregation levels of passengers traveling to the city center outweigh the higher suburban and rural segregation experiences, due to the higher number of passengers affected.

The zonal segregation changes calculated in section 5.3, depict mixed effects between both 2017-2018 and 2017-2020. While the majority of zones remain almost unchanged or indicate small changes in segregation levels, on both extremes there are zones with a substantial change in segregation. Decreasing segregation is found in the city center in particular for 2020. In addition, urban zones with public transport hubs and primarily commuter train stations display less segregation over the years. The zone of the newly created Odenplan station is part of this. Both stronger increasing and decreasing effects are indicated for the northwest and southwest corridor.

To assess the actual strength of the changes in each zone in the overall context, the significance test from the 5.3 section is used. The overall non-significant changes in mean segregation indicate that the declines in segregation do not fundamentally change the scene. There are both decreasing and increasing levels in different zones, possibly offsetting the effects so that the overall changes hardly differ.

The analysis of focused areas in 5.4 confirms the patterns found in the evolution study of 5.3. Moreover, increasing segregation is found in zones with commuter trains outside of the city. This leads to the conclusion, that after 2017 the suburban station zones carried less income-diverse passengers. On the other hand, suburban zones outside of the commuter train station areas slightly tend towards decreasing segregation levels, especially when including other PT hubs. Setting this into context with the way the segregation contribution is calculated, the question arises on why the segregation levels differ so much in adjacent areas.

As the initial case study introduction given in 4.1 suggests, there might be some effects related to the enhanced public transport system after July 2017. When looking at the commuter train corridor where essential improvements were made due to the Citybanan project, several increased- but also decreased-segregation zones can be found. By the setup of this case study, decreasing segregation levels are related to a more diverse use of the PT system. It is also possible to spot the

zones where this increased social mixing originates from. Still, the cause for greater social mixing remains unrevealed and attempts at explanation can only be made cautiously.

In the case of Stockholm, one approach could be to set the results into context with local circumstances and general developments. While Stockholm is one of the fastest-growing metropolitan areas of Europe, a lot of housing- and traffic-related pressure lays on the city center. Urban gentrification leads to a segmentation of higher income groups in the city center and pushes low-income groups to the fringes.

But in accordance with the above, the general trends of disparities in housing and mobility opportunities might lead to more segregation in suburbs and peri-urban areas. Suburbans not traveling by car can often be related to low-income groups. As referred to in section 2.3.1, low-income groups' activities in more rural areas are located along public transport corridors (Kamruzzaman and Hine, 2012). This could explain Stockholm's higher segregation levels at commuter train stations in the suburbs since potentially higher income groups would either travel by car to the same places or initially choose different neighborhoods to live in according to their mobility behavior. Policies such as the congestion fee in Stockholm might intensify the dependency on public transport for low-income groups. Increasing segregation levels in these suburban and peri-urban zones could be linked to general trends of urbanization and gentrification, as well as PT dependency and the transport disadvantage of low-income groups.

Still, when assessing mobility-based segregation in Stockholm, the absolute segregation values as shown in Figure 5.2 are lowest in the city center. This is probably caused by various activity-related social mixings such as work, leisure and shopping, and public services. Also, the segregation levels decrease in the city center. This could potentially indicate that due to the enhanced PT connection of particularly suburbs, the city center is easier reachable by all income classes. To summarize it from a public transport perspective, city center inbound PT passengers are found to be more income-diverse in both 2018 and 2020 than in 2017, while outbound passengers towards the suburbs have more and more uniform income backgrounds, especially when traveling to commuter train stations.

However, partly decreasing segregation is found in some fringes of the city, suburban zones outside of commuter train stations, and especially in the southwest corridor. This might relate to other modes than the commuter train, which carry more diverse passengers since for instance some metro station zones are found to have decreased segregation levels.

The increase in segregation just mentioned is certainly not intended by the authorities. But the decreasing average segregation level between 2017 and 2019 and the decreasing segregation level in the city center mentioned above might be desirable for the decision makers in Stockholm County. A greater social mix of public transport users could lead to conclusions about facilitating access. It could imply improved access to services and facilities in the city center. This increased mix and potentially improved access is what the decision maker wanted to enhance, especially for low-income groups.

It is crucial to notice that the segregation findings of this study cannot be confirmed as caused by one of the above reasons nor can segregation changes be directly explained by the public transport upgrade after the Citybanan implementation. There can be no direct effects concluded without further disentangling. The segregation calculated for this case study depends on the usage of the public

transport system. Every day's public transport usage depends on many factors, which are not only PT operations-related, such as policies, economic and labor market situation, or just simply the weather. This should be cautiously considered when trying to interpret the results of segregation changes on what the Citybanan project might have affected. Especially when discussing the changes in 2018, the flat fare policy change in January 2017 should also be considered. It could have had effects on the general but especially the usage from outskirt inhabitants and thereby affect segregation results for 2018.

The same caution applies when interpreting the results of the focus areas presented in section 5.4. Depending on the zoning used, the methodology proves to be designed for assessing developments on smaller scales such as these neighborhoods. Still, the Stockholm results show, that segregation analyzed from mobility patterns can vary substantially and can indicate but does not disclose local causes.

All in all, the present study raises and emphasizes the possibility to combine socioeconomic data and mobility data to analyze segregation. This contributes to the ongoing discussion on segregation, transport disadvantages, and in a broader sense, the widening gaps of many societies.

The method is flexibly designed for the use of different types of groups. In this case, the measure incorporates both the ordinal aspect as well as the fact that the group sizes differ slightly and the PT journey's made are distributed unequally across the groups. For these cases a segregation measure that accounts for relative shares is essential.

As a side effect, this study contributes to ex-post evaluation of transport (policy) changes. Among others, Graham (2014) concluded that there is less-observed ex-post transport appraisal which is "presumably because we are generally more interested in predicting how our future investments will fare than in assessing how well we have allocated resources in the past.". Despite its restrictions, this study provides support to a more specific analysis of tax-heavy investments such as the Citybanan commuter train tunnel.

Nevertheless, the remaining questions on potential causes and implications of the observed segregation lead to the main limitation of this study of not indicating direct causal effects. The following section discusses these and further limitations.

5.6 LIMITATIONS

In this section, the methodology's limitations are discussed. While the discussion section outlines the functionality and applicability of the methodology developed, there are some constraints to consider when interpreting the results of this study as well as when applying the method to other cases.

As mentioned before, no direct causal effects can be determined from the segregation changes for instance regarding changes to the public transport system. Though, this method could give valuable indications depending on the case and its data sets.

For the interpretation of results the spatial dependency of segregation indices plays a role, see Wong (1997). Zonal aggregations are often made when there is no disaggregated data of home locations and socioeconomic status. When using such, presumptions are made about the homogeneity of a zone's group. For the Stockholm case, the dispersed structure of public transport usage is partly related to the general setup of the radial public transport system with its transport hubs

serving certain areas. But, especially for small-sized zones, thus in the city center and other more central zones, the algorithm of stop area allocation causes some zones to look empty while others appear as highly frequented. Allocating all these activities to specific zones distorts the picture of inner-city use and the mixing of different social groups. As a result, the segregation results obtained with the presented method depend on the spatial data setup.

Furthermore, this study only measures potential interaction and thereby only potential segregation. The real experienced segregation stays hidden. Although this approach happens to be flexible, the main limitations remain: Any segregation index using smart card data is restricted to only measuring potential interaction/mixing of public transport users. People might travel together to the same zones but do not necessarily get in contact with each other.

Another limitation of this study's application using public transport smart card data measuring potential interaction on a daily basis. Actually, passengers are less probable to meet if they visit a zone at different periods of the day. Also, mixing throughout zones is limited to the PT travel options given at that time of day. It implies that zones (temporarily) not reachable by PT are excluded from the segregation study as it can be detected in this study, see e.g. the zones not colored according to the legend scheme in Figure 5.3.

Further, the data used in this case study can only assess the segregation between public transport users. No active modes or car users are included in this study. Also, national and regional train traffic is not assessed since the data set is restricted to the modes within the transport authority of Stockholm County. Looking at segregation this plays a significant role in the assessment since public transport usage already appeared to split the social groups, as seen in Table 4.2. Especially when applying the method to other contexts with even less PT usage of higher-income groups there could be a risk of only analyzing the segregation between certain social groups. This would make findings on segregation less transferable to society as a whole.

Adding up to this, inferences and estimations are never totally reliable, while still being significant. In fact, the true locations of passenger homes and destinations remain unknown. Further, the home zone inference algorithm explained in section 2.3.2 takes out the inhabitants that use public transport sporadically. People who are not dependent on PT might have higher incomes since they can afford other modes. This could mean that the segregation measurement within the PT system excludes occasional, more wealthy travelers. Consequently, this adds up to the point discussed above, where a segregation analysis of PT users might evaluate an already segregated community.

Also, the socioeconomic data used in this study has shortcomings regarding its level of aggregation. Within a DeSo zone all inhabitants of that zone are treated as one big mass of the same median income value. Within one zone all inhabitants are assigned to one social group which for some could be a misjudgment. As a consequence, they are part of a group's segregation evaluation, which they actually do not belong to. In general, these drawbacks could occur by any form of aggregation used regarding social group building, zoning, and other forms of clustering.

Finally, it remains to mention that the chosen winter period travel data might indicate a more busy period to travel by public transport. Many (European) public transport systems have their busiest months according to weather and vacation periods. While there are fewer busy periods in the summer months as well as

during other vacation periods like the winter/Christmas vacations, active mode users could be inclined to use PT in winter. Less usage could lead to less mixing and therefore deteriorate segregation levels.

Overall, this study successfully analyzed segregation levels. Still, there are limitations of which most relate to the case study and its data. Conclusions are given in the next chapter, including suggestions for future studies.

6 CONCLUSIONS

The aim of the present research was to examine the possibilities of measuring social segregation of multiple groups with large-scale disaggregated mobility data. In particular, public transport smart card data has been chosen to explore.

The following looks back on the initially formulated research questions and evaluates the results of this study in this regard. After, recommendations are given in two fashions. First, the method developed and case study results obtained are transferred towards their practical applicability. Lastly, further research is recommended, retrospecting the conclusions and limitations of this study.

6.1 ANSWERING THE RESEARCH QUESTIONS

This study aimed to answer the main research question on how multi-group activity-based social segregation can be measured using large-scale disaggregated mobility data. Therefore, this study utilized existing methods of activity-based social segregation measures and combines them with socioeconomic data. Doing so this study successfully conceptualizes and applies process steps towards the measurement of segregation from mobility data. This enables measuring social segregation at an activity-end level and adds up to the current research perspective on activity-based segregation. The Stockholm case study proved the applicability of the designed method and provided valuable insights into income segregation analyzed from public transport smart card data.

To capture segregation index measures for multiple social groups were investigated. Multiple indices are found to be applicable and requirements were defined to measure multi-group activity-based social segregation. Thereby, an answer was found to sub-research question 1. The ordinal information theory index presented in 3.4 was found applicable for the Stockholm case study and allows to assess the diversity/social mix of public transport passengers at the journey destination. Further, the methodology is flexible enough to be applied using other multi-group segregation indices.

To answer research question 2, a methodology was developed to define process steps to link social groups to mobility data. If the mobility data is connected to home zones as described in chapter 3.1, social groups built from residential data can be matched. Even different data types could be used as long as there can be some aggregation into groups or clusters. Those could even incorporate non-social factors, as long as the connection to the residence is given.

Defining the requirements of the two sub-research questions mentioned above also leads to answer the third question on requirements. As it was asked what requirements there are for multi-group segregation assessment from mobility data, it turns out that home zones, as well as destinations or other activity locations, need to be present.

Lastly, segregation can be tracked over time using continuous and traceable smart card data sets. This facilitates analyzing the evolution of segregation per social group, answering sub-research question 4. Further, it depicts the spatial locations where changes originate. As it can be seen from the Stockholm case study, this approach is straightforward but requires the use of comparable data sets and time spans.

6.2 RECOMMENDATIONS

As this study is found to answer the initially posed research questions successfully, this section looks into what can be recommended for potential users of the methodology. In addition, there is much to be discovered in future studies, some of which are discussed in this chapter.

6.2.1 Application recommendations

The insights gained from this study may be of assistance to both policymakers and transport authorities as well as other organizations interested in segregation and its relation to mobility. Using the method to connect mobility data to social data could potentially lead to a more realistic depiction of social segregation due to assessing it from a more activity-based perspective. An implication of this is the possibility to examine segregation development in relation to a policy or other changes, possibly transport-related.

When analyzing results, attention should be paid to the potential shortcomings of the data set. The interpretation of segregation index outcomes depends strongly on the data type used. In the case of public transport data, only segregation among public transport users is assessed, as discussed in section 5.6.

When applicants of the methodology are interested in the precise levels or changes of segregation and the social mix in zones, it is possible to look into each zone's composition using the absolute and weighted zonal segregation contribution. Particularly for urban planners and policymakers it could be of interest to measure social segregation effects. Still, when interpreting results, attention should be paid to causal effects which leads to possibilities of future research.

In the case of Stockholm, analysts and decision makers can assess the segregation situation using the results of this study. Daily or weekly segregation levels help evaluate overall levels and trends. Weighted segregation levels are suitable for analyses in which the relation of zone segregation plays a role. The absolute segregation contribution should be used for detailed, intra-zonal assessment, as it is more informative when other zones and the total population are not in focus. However, consideration should be given to how many people are affected by the results.

The results help evaluating the segregation situation in Stockholm and at the same time raise the question of why segregation is appearing more or less in certain areas. One goal could be to understand why commuter train related suburban and rural station zones increased in segregation. To answer the question on why these zones see less mixed passengers arriving, housing data or more traditional transport collection methods such as surveys could be used to explore the context and set the findings into relation with simultaneous processes such as gentrification and housing developments.

The results obtained show how social segregation depends on the urban context. Similarly densely populated urban areas could show similar developments. The assessment of segregation based on mobility data also showed the clear dependence on, in this case, public transport, which should be taken into account at all times in the analysis.

It is also in the interest of decision makers to explore how this method can be used to evaluate the impact of infrastructure or service improvements. These and other issues are addressed in the next section, which focuses on future studies.

6.2.2 Future studies

The first approach for further research could be to disentangle causal effects of the segregation index and zonal segregation contributions. By disentangling what led to for instance decreasing levels of segregation, it could be assessed whether direct effects can be linked to specific changes. For the Stockholm case, this would be interesting regarding the causal effects of changes to the public transport system, specifically the Citybanan project.

As referred to in the limitations, using public transport data comes with some boundaries regarding the analysis of a whole society's actual social mixing. To overcome this, the developed method could be applied using other large-scale mobility data set such as GPS data. Probably hard-to-obtain data sets for future research could include actual interactions instead of only potential interactions. Since there is a trend towards combining multiple data sources (Wei et al., 2015), this might reveal new opportunities for segregation analysis as well.

As mentioned in the limitations in section 5.6, the time-scale could be changed to for instance an hourly focus to obtain more realistic results regarding potential interaction. Though, this approach could split activities into very small amounts. This could mean that there would be many zones and time periods with no arriving travelers. Probably this study would then need to focus on busy areas only.

Regarding the shortcomings of the aggregation of zonal inhabitants to one deterministic social variable, future studies could look into more comprehensive socioeconomic data sets. This approach might work best with using even smaller zones than the ones with on average 1500 people who are all allocated to the same social group.

When using traceable data such as public transport smart card data, one could even break down the analysis to look into tracking an individual's segregation experience and changes.

In addition one could conduct this methodology on regions with more expected segregation since Stockholm County, Sweden is a comparably wealthy region with for instance high levels of employment and university education (Almlöf et al., 2021). Another possibility is to look into the segregation influence of certain modes. For instance, ferries could reduce segregation. Often, cities are split by rivers which creates a natural distance since only a few bridges exist. Ferries could potentially overcome it.

Finally, most important for authorities and policymakers could be a study looking into monetizing the segregation effects. This could be implemented with, for instance, using cost-benefit analyses regarding the population experiencing significant segregation changes.

Summing up, this study successfully proves to assess social segregation by using large-scale disaggregated mobility data. These findings contribute in several ways to the understanding of segregation as both an activity-based and residential phenomenon. A basis is provided for diverse applications as well as further research.

BIBLIOGRAPHY

Abbasi, S., Ko, J., and Min, J. (2021). Measuring destination-based segregation through mobility patterns: Application of transport card data. *Journal of Transport Geography*, 92.

Acevedo-Garcia, D. and Lochner, K. A. (2003). Residential segregation and health. *Neighborhoods and health*, pages 265–87.

Almlöf, E., Rubensson, I., Cebecauer, M., and Jenelius, E. (2021). Who continued travelling by public transport during COVID-19? Socioeconomic factors explaining travel behaviour in Stockholm 2020 based on smart card data. *European Transport Research Review*, 13(1):1–13.

Andersson, R. and Kährik, A. (2015). Widening gaps : Segregation dynamics during two decades of economic and institutional change in stockholm. In *Socio-Economic Segregation in European Capital Cities*, pages 134–155. New York: Routledge.

Andersson, R. and Turner, L. M. (2014). Segregation, gentrification, and residualisation: From public housing to market-driven housing allocation in inner city stockholm. *International Journal of Housing Policy*, 14:3–29.

Arbex, R., da Cunha, C. B., and Speicys, R. (2019). Before-and-after evaluation of a bus network improvement using performance indicators from historical smart card data. *Public Transport*, pages 1–19.

Athey, S., Ferguson, B. A., Gentzkow, M., and Schmidt, T. (2020). Experienced segregation. Technical report, National Bureau of Economic Research.

Bagchi, M. and White, P. R. (2005). The potential of public transport smart card data. *Transport Policy*, 12:464–474.

Barry, J. J., Newhouser, R., Sayeda, S., Barry, J. J., Newhouser, R., Rahbee, A., and Sayeda, S. (2002). Origin and destination estimation in New York City with automated fare system data. *Transportation Research Record*, 1817(1):183–187.

Bell, W. (1954). A Probability Model for the Measurement of Ecological Segregation. *Social Forces*, 32(4):357–364.

Bischoff, K. and Reardon, S. F. (2014). Residential segregation by income, 1970-2009. *Diversity and disparities: America enters a new century*, 43.

Brands, T., Dixit, M., and Oort, N. V. (2020). Impact of a new metro line in amsterdam on ridership, travel times, reliability and societal costs and benefits. *Issue*, 20:335–353.

Briand, A. S., Côme, E., Trépanier, M., and Oukhellou, L. (2017). Analyzing year-to-year changes in public transport passenger behaviour using smart card data. *Transportation Research Part C: Emerging Technologies*, 79:274–289.

Cheung, F. (2006). Implementation of Nationwide Public Transport Smart Card in the Netherlands: Cost–Benefit Analysis. *Transportation research record*, 1971(1):127–132.

Church, A., Frost, M., and Sullivan, K. (2000). Transport and social exclusion in London. *Transport policy*, 7(3):195–205.

Currie, G., Richardson, T., Smyth, P., Vella-Brodrick, D., Hine, J., Lucas, K., Stanley, J., Morris, J., Kinnear, R., and Stanley, J. (2010). Investigating links between transport disadvantage, social exclusion and well-being in melbourne - updated results. *Research in Transportation Economics*, 29:287–295.

Dawkins, C. and Moeckel, R. (2016). Transit-induced gentrification: Who will stay, and who will go? *Housing Policy Debate*, 26:801–818.

Delbosc, A. and Currie, G. (2011). Transport problems that matter - social and psychological links to transport disadvantage. *Journal of Transport Geography*, 19:170–178.

Devillaine, F., Munizaga, M., and Trépanier, M. (2012). Detection of activities of public transport users by analyzing smart card data. *Transportation Research Record*, pages 48–55.

Dixit, M., Brands, T., van Oort, N., Cats, O., and Hoogendoorn, S. (2019). Passenger travel time reliability for multimodal public transport journeys. *Transportation Research Record*, 2673:149–160.

Duncan, O. D. and Duncan, B. (1955). A methodological analysis of segregation indexes. *American sociological review*, 20(2):210–217.

Ellgaard, H. (2009). *Citybanan Stockholm op de kaart*. Wikipedia. Retrieved from https://nl.wikipedia.org/wiki/Citybanan_Stockholm#/media/Bestand:Citybanan_karta.jpg on October 11, 2021.

Ellis, M., Holloway, S. R., Wright, R., and Fowler, C. S. (2012). Agents of change: Mixed-race households and the dynamics of neighborhood segregation in the united states. *Annals of the Association of American Geographers*, 102:549–570.

Ellis, M., Wright, R., and Parks, V. (2004). Work together, live apart? geographies of racial and ethnic segregation at home and at work. *Annals of the Association of American Geographers*, 94:620–637.

Farber, S., O'Kelly, M., Miller, H. J., and Neutens, T. (2015). Measuring segregation using patterns of daily travel behavior: A social interaction based model of exposure. *Journal of Transport Geography*, 49:26–38.

Frohne, E. (2014). *Strecke der Pendeltåg (Stockholm)*. Wikipedia. Retrieved from https://de.wikipedia.org/wiki/Pendelt%C3%A5g_(Stockholm)#/media/Datei:Linjekarta_f%C3%B6r_Stockholms_pendelt%C3%A5g.jpg on October 11, 2021.

Fu, X. and Gu, Y. (2018). Impact of a new metro line: Analysis of metro passenger flow and travel time based on smart card data. *Journal of Advanced Transportation*, 2018.

Galiana, L. and Sakarovitch, B. (2020). Residential segregation, daytime segregation and spatial frictions : an analysis from mobile phone data. Documents de Travail de l'Insee - INSEE Working Papers g2020-12, Institut National de la Statistique et des Etudes Economiques.

Goulet-Langlois, G., Koutsopoulos, H. N., and Zhao, J. (2016). Inferring patterns in the multi-week activity sequences of public transport users. *Transportation Research Part C: Emerging Technologies*, 64:1–16.

Graham, D. (2014). Causal influence for ex-post evaluation of transport interventions. *International Transport Forum Discussion Papers*, No. 2014/13, OECD Publishing, Paris.

Grundström, K. and Molina, I. (2016). From folkhem to lifestyle housing in sweden: segregation and urban form, 1930s–2010s. *International Journal of Housing Policy*, 16:316–336.

Haandrikman, K., Costa, R., Malmberg, B., Rogne, A. F., and Sleutjes, B. (2021). Socio-economic segregation in european cities. a comparative study of brussels, copenhagen, amsterdam, oslo and stockholm. *Urban Geography*.

Hedin, K., Clark, E., Lundholm, E., and Malmberg, G. (2012). Neoliberalization of housing in sweden: Gentrification, filtering, and social polarization. *Annals of the Association of American Geographers*, 102:443–463.

Hunt, C. and Walker, L. (1974). *Ethnic Dynamics: Patterns of Inter-Group Relations in Various Societies*. Homewood, IL: The Dorsey Press.

Ivaniushina, V., Makles, A. M., Schneider, K., and Alexandrov, D. (2019). School segregation in st. petersburg–the role of socioeconomic status. *Education Economics*, 27:166–185.

James, D. R. and Taeuber, K. E. (1985). Measures of segregation. *Sociological Methodology*, 15:1–32.

Järv, O., Müürisepp, K., Ahas, R., Derudder, B., and Witlox, F. (2015). Ethnic differences in activity spaces as a characteristic of segregation: A study based on mobile phone usage in tallinn, estonia. *Urban Studies*, 52:2680–2698.

Kamruzzaman, M. and Hine, J. (2012). Analysis of rural activity spaces and transport disadvantage using a multi-method approach. *Transport Policy*, 19:105–120.

Kaufmann, V. (2004). Social and political segregation of urban transportation: the merits and limitations of the Swiss cities model. *Built Environment*, 30(2):146–152.

Kholodov, Y. (2019). Evaluation of public transport fare policy using smartcard data travel patterns change and distributional effects in stockholm county. [Master's Thesis. Delft University of Technology].

Kieu, L. M., Bhaskar, A., and Chung, E. (2015). Passenger segmentation using smart card data. *IEEE Transactions on Intelligent Transportation Systems*, 16:1537–1548.

Kwan, M. P. (2009). From place-based to people-based exposure measures. *Social Science and Medicine*, 69:1311–1313.

Le Roux, G., Vallée, J., and Commenges, H. (2017). Social segregation around the clock in the paris region (france). *Journal of Transport Geography*, 59:134–145.

Leonard, J. S. (1987). The interaction of residential segregation and employment discrimination. *Journal of Urban Economics*, 21(3):323–346.

Li, F. and Wang, D. (2017). Measuring urban segregation based on individuals' daily activity patterns: A multidimensional approach. *Environment and Planning A*, 49:467–486.

Logan, J. R. and Burdick-Will, J. (2016). School segregation, charter schools, and access to quality education. *Journal of Urban Affairs*, 38:323–343.

Lucas, K. (2011). Making the connections between transport disadvantage and the social exclusion of low income populations in the tshwane region of south africa. *Journal of Transport Geography*, 19:1320–1334.

Luo, D., Bonnetain, L., Cats, O., and van Lint, H. (2018). Constructing spatiotemporal load profiles of transit vehicles with multiple data sources. *Transportation Research Record*, 2672:175–186.

Ma, X., Wu, Y. J., Wang, Y., Chen, F., and Liu, J. (2013). Mining smart card data for transit riders' travel patterns. *Transportation Research Part C: Emerging Technologies*, 36:1–12.

Marques, E. (2012). Social networks, segregation and poverty in são paulo. *International Journal of Urban and Regional Research*, 36:958–979.

Massey, D. S. (2012). Reflections on the dimensions of segregation. *Social Forces*, 91:39–43.

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

Meng, G., Hall, G. B., and Roberts, S. (2006). Multi-group segregation indices for measuring ordinal classes. *Computers, Environment and Urban Systems*, 30:275–299.

Monkkonen, P., Comandon, A., Escamilla, J. A. M., and Guerra, E. (2018). Urban sprawl and the growing geographic scale of segregation in mexico, 1990–2010. *Habitat International*, 73:89–95.

Morgan, B. S. (1975). The segregation of socio-economic groups in urban areas: a comparative analysis. *Urban Studies*, 12(1):47–60.

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

Musterd, S., Marcińczak, S., van Ham, M., and Tammaru, T. (2017). Socioeconomic segregation in european capital cities. increasing separation between poor and rich. *Urban Geography*, 38:1062–1083.

Paddison, R. and Hamnett, C. (2000). *Handbook of Urban Studies*. SAGE Publications.

Park, J. Y., Kim, D. J., and Lim, Y. (2008). Use of smart card data to define public transit use in seoul, south korea. *Transportation Research Record*, pages 3–9.

Pelletier, M. P., Trépanier, M., and Morency, C. (2011). Smart card data use in public transit: A literature review. *Transportation Research Part C: Emerging Technologies*, 19:557–568.

Reardon, S. F. (1998). Measures of racial diversity and segregation in multigroup and hierarchically structured populations. In *annual meeting of the Eastern Sociological Society, Philadelphia, PA*.

Reardon, S. F. (2009). Measures of ordinal segregation. In *Occupational and residential segregation*. Emerald Group Publishing Limited.

Reardon, S. F. and Firebaugh, G. (2002). Measures of multigroup segregation. *Sociological methodology*, 32(1):33–67.

Reardon, S. F. and O'Sullivan, D. (2004). Measures of spatial segregation. *Sociological methodology*, 34(1):121–162.

Rokem, J. and Vaughan, L. (2018). Segregation, mobility and encounters in jerusalem: The role of public transport infrastructure in connecting the 'divided city'. *Urban Studies*, 55:3454–3473.

Sakoda, J. M. (1981). A generalized index of dissimilarity. *Demography*, 18(2):245–250.

Sandberg, G. and Palmelius, S. (2018). *Demografiska statistikområden, en ny regional indelning under kommuner*. Statistikmyndigheten SCB. Retrieved from https://www.scb.se/hitta-statistik/artiklar/2018/demografiska-statistikomraden-en-ny-regional-indelning-under-kommuner/.

Sari Aslam, N., Cheng, T., and Cheshire, J. (2019). A high-precision heuristic model to detect home and work locations from smart card data. *Geo-Spatial Information Science*, 22:1–11.

SCB (2017). SCB:s Open data for DeSO – Demographic Statistical Areas. Retrieved from https://www.geodata.se/geodataportalen/srv/swe/catalog.search;jsessionid=42B5AAC3339638A205A27724ECF960BF#/search?resultType=swe-details&_schema=iso19139*&type=dataset%20or%20series&from=1&to=20 using the explanation from https://www.scb.se/en/services/open-data-api/open-geodata/deso--demographic-statistical-areas/.

SCB (2021). Population by region and year. Retrieved from http://www.statistikdatabasen.scb.se/pxweb/en/ssd/START__BE__BE0101__BE0101A/BefolkningNy/.

Silm, S. and Ahas, R. (2014). The temporal variation of ethnic segregation in a city: Evidence from a mobile phone use dataset. *Social Science Research*, 47:30–43.

Stockholms läns landsting (SLL) (2017). Regional utvecklingsplan för stockholmsregionen: Rufs 2050.

Tamblay, S., Galilea, P., Iglesias, P., Raveau, S., and Muñoz, J. C. (2016). A zonal inference model based on observed smart-card transactions for santiago de chile. *Transportation Research Part A: Policy and Practice*, 84:44–54.

Tan, Y., Chai, Y., and Chen, Z. (2019). Social-contextual exposure of ethnic groups in urban china: From residential place to activity space. *Population, Space and Place*, 25.

Tao, S., He, S. Y., Kwan, M. P., and Luo, S. (2020). Does low income translate into lower mobility? an investigation of activity space in hong kong between 2002 and 2011. *Journal of Transport Geography*, 82.

Theil, H. (1972). *Statistical decomposition analysis; with applications in the social and administrative sciences*, volume 14. Amsterdam: North-Holland Publishing Company.

Theil, H. and Finizza, A. J. (1971). A note on the measurement of racial integration of schools by means of informational concepts. *The Journal of Mathematical Sociology*, 1(2):187–193.

Tuncer, U. A. (2018). The role of technology in public transport integration and governance–smart card use in istanbul and mexico city brt systems. *IGLUS Quarterly*, 4:7–13.

United Nations (2020). *World Social Report 2020: inequality in a rapidly changing world.* United Nations Department of Economic and Social Affairs. New York: United Nations publication. Sales No. E.20.IV.1.

Ureta, S. (2008). Mobilising poverty?: Mobile phone use and everyday spatial mobility among low-income families in santiago, chile. *Information Society*, 24:83–92.

Utsunomiya, M., Attanucci, J., and Wilson, N. (2006). Potential uses of transit smart card registration and transaction data to improve transit planning. *Transportation research record*, 1971(1):118–126.

Verma, T., Sirenko, M., Kornecki, I., Cunningham, S., and Araújo, N. A. (2021). Extracting spatiotemporal commuting patterns from public transit data. *Journal of Urban Mobility*, 1:100004.

Viallard, A., Trépanier, M., and Morency, C. (2019). Assessing the evolution of transit user behavior from smart card data. *Transportation Research Record*, 2673:184–194.

Wei, M., Liu, Y., and Sigler, T. J. (2015). An exploratory analysis of Brisbane's commuter travel patterns using smart card data. In *State of Australian Cities National Conference, 2015, Gold Coast, Queensland, Australia*.

Weng, X., Liu, Y., Song, H., Yao, S., and Zhang, P. (2018). Mining urban passengers' travel patterns from incomplete data with use cases. *Computer Networks*, 134:116–126.

Wong, D. W. and Shaw, S. L. (2011). Measuring segregation: An activity space approach. *Journal of Geographical Systems*, 13:127–145.

Wong, W. S. (1997). Spatial dependency of segregation indices. *The Canadian Geographer / Le Geographe canadien*, 41:128–136.

Xu, Y., Belyi, A., Santi, P., and Ratti, C. (2019). Quantifying segregation in an integrated urban physical-social space. *Journal of the Royal Society Interface*, 16.

Zhang, J. and Zheng, L. (2015). Are people willing to pay for less segregation? evidence from u.s. internal migration. *Regional Science and Urban Economics*, 53:97–112.

A APPENDIX A: SCIENTIFIC PAPER

Measuring activity-based Social Segregation using Public Transport Smart Card Data

L. Kolkowski[1], M. Dixit[2], O. Cats[2], T. Verma[3] and E. Jenelius[4]

[1] *Graduate student MSc. Transport, Infrastructure and Logistics, Delft University of Technology, The Netherlands*
[2] *Delft University of Technology, Faculty of Civil Engineering and Geosciences, The Netherlands*
[3] *Delft University of Technology, Faculty of Technology, Policy and Management, The Netherlands*
[4] *KTH Royal Institute of Technology, Division of Transport Planning, Sweden*

Abstract— While social segregation is often assessed in terms of one socio-geographic space, usually place of residence, more recent approaches also incorporate activity-based and, in particular, mobility-based data. This study extends the use of mobility data to measure social segregation between multiple groups by developing a method to connect socio-economic data from the place of residence to mobility data. The method gets applied on the public transport smart card data of Stockholm County, Sweden, using the ordinal information theory index. Applying the index on the destination mix of 2017-2020 smart card data sets for week 5, shows significant differences between income groups' segregation along the radial public transport corridor. The findings also enable to assess the evolution of segregation. In Stockholm, the overall slight decrease in income segregation can be linked to declining segregation in the city center and its public transport hubs. Increasing zonal segregation is related to suburban and rural zones with commuter train stations. This method helps to quantify and thus better understand segregation based on the dynamics of social life. It also allows an evaluation of public transport, which should facilitate potential interaction between social groups.

Keywords— Social segregation, Income segregation, Activity-based, Mobility-based, Public transport smart card data, Ex-post transport appraisal, Smart card data analysis

I. Introduction & Background

Social segregation is the spatial, temporal and access-related distance between individuals and groups with different social backgrounds. Less mixing of people from different social backgrounds means more segregation and thus often exclusion. Segregation often leads to disparities in essential living conditions [1, 2, 3]. Mitigating the effects of social segregation is a major challenge of today's society and its policymakers, while capturing it remains a hard-to-grasp phenomenon and a much-discussed research topic.

Until recently, spatial segregation of social groups was measured using segregation indices applied on mostly residential socioeconomic data [4], thus static data of one socio-geographic space. Main findings indicate that income, educational level, housing types as well as spatial distance between groups are key drivers for segregation [5, 6].

Using only static data of one socio-geographical space can lead to a partial view and capturing only fractions of social segregation. Recent studies utilize activity-based data to measure segregation [7]. These studies integrate an activity perspective by using data from social networks, mobile phones, GPS, travel surveys, or diaries to measure segregation [8, 9, 5, 10]. Often, mobility data is used to measure such activity-based segregation which can include accuracy, privacy, and availability issues, as well as incomplete data sets [11, 12]. In addition, it can require immense efforts and high costs to obtain sufficient data sets.

This brought up using public transport data which has not only the role to facilitate access and diminish barriers between social groups but could also provide valuable mobility traces to measure activity-based segregation. Limited access to transportation results in less access to essential infrastructure to participate both socially and economically [13]. Transport disadvantage is strongly correlated to social exclusion as found by studies such as [14].

Especially in urban environments, public transport (PT) is a major component of mobility and activity spaces. As a public service, public transport is supposed to provide accessibility to everybody and therewith every social group. Not only is public transportation a key component of transportation disadvantage, but studies have found a direct link of PT causing segregation [15]. Less access to public transport causes higher levels of segregation. Looking at the problem from the opposite perspective, [16] found that social segregation is "a product of the mobility" of certain social groups, in this case, wealthy urban inhabitants. Therefore, this study focuses on making use of PT data to measure social segregation.

Research in the public transport sector currently focuses on broadening the use of public transport smart card data, a well-established chip card technology for fare collection [12]. Since public transport has a key role in people's daily

Corresponding author: Lukas Kolkowski, lukas@kolkowski.de

dynamics, especially in urban environments, its data might be key to overcome other data sources' shortcomings. Smart cards offer unprecedented large data sets of real transactions, thus observed mobility traces [17].

When assessing social segregation often segregation indices used were restricted to measure segregation levels between two groups. Regarding the many different layers and complexity of today's society, these indices were extended to incorporate multiple groups. Therefore, this study focuses on measuring what is known as multi-group segregation.

So far, only Abbasi et al. (2021) measured two-group and multi-group social segregation using public transport smart card data [18]. They were able to extract social characteristics and thereby created social groups from the smart card data itself. For many transport authorities and countries, this kind of personal information would not be available or extracting it would raise data privacy concerns. As a result, social information often cannot be retrieved directly from smart cards. In addition, even richly equipped smart cards often do not contain the desired social information.

Mobility data such as smart card data has been linked to residential socioeconomic data before, just not specifically in segregation applications. However, if the required social information cannot be extracted from the mobility data, there is a **lack of a method to measure activity-based multi-group social segregation on disaggregate large-scale mobility data sets such as public transport smart card data**. Based on the research gap indicated, this study aims to answer the following main research question:

How can multi-group activity-based social segregation be measured using large-scale disaggregated mobility data such as public transport smart card data?

To answer the above there are four sub-research questions defined below.

1. *How is social segregation measured and how can it be represented using mobility-based data?*
2. *What are the requirements from large-scale disaggregated mobility data for enabling the measurement of multi-group social segregation?*
3. *How can socio-demographic characteristics of smart card users be inferred?*
4. *How does activity-based social segregation evolve and how does disaggregated mobility data facilitate its analysis?*

Combining socially relevant data, as used for residential segregation studies, with activity-based mobility data could combine the strengths of the two currently prevalent approaches to segregation measurement. While socio-demographic or socioeconomic data can be used to distinguish social groups, mobility data reveals activity patterns and the resulting mix of groups. Therefore, this study develops a method to link social groups, large-scale mobility data, and multi-group segregation measures to quantify segregation.

In the following section II, a method is developed for connecting residential social groups to disaggregate mobility data and thereby calculating activity-based segregation measures. Introducing the case study of Stockholm County, the method is applied to the transport authority Storstockholms Lokaltrafik (SL)'s public transport smart card data sets in section III. Results of the method's application are presented in section IV and discussed thereafter including limitations of the study. Lastly, conclusions are made in the final section V.

II. METHODOLOGY

This study develops a method to enrich mobility data in such a way that it connects to travelers' social characteristics. General requirements regarding mobility data as well as socioeconomic data sets for segregation studies with disaggregate mobility data are formulated. In a second step, socioeconomic residential data is linked to each disaggregate element of the large-scale mobility data. Lastly, multi-group segregation measures can be applied to the enriched disaggregated mobility data. This yields a method for measuring social segregation using large-scale disaggregate mobility data and socioeconomic data.

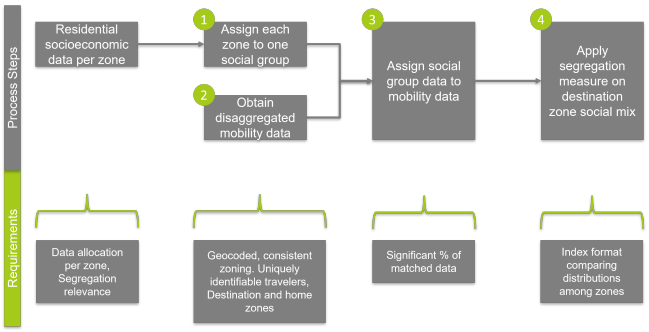

Fig. 1: Framework for measuring social segregation by connecting mobility data to socioeconomic data

The modelling steps are described in Figure 1. Residential socioeconomic data and its abstracted groups are connected to observed disaggregated mobility data by using the same spatial organizational units, usually a type of administrative zones. Once the socioeconomic data is assigned to the mobility data via the travelers' home zones, different segregation measures can be applied. While the actual connection is somewhat straight-forward, attention should be paid to the requirements of each process step outlined below.

a. Requirements for residential social groups

The residence is still a highly valuable factor when trying to understand the social composition of society. Also, residency indicates where to geographically locate socioeconomic data since the data is assigned to administrative spatial units such as zones. Assuming a considerable amount of zones and social differences within a society, the following describes requirements for residential social data. Most commonly, socially relevant information can be obtained from socio-demographic data as retrieved from census or other governmental databases. The main requirement for this data is that it is segregation-relevant and can be allocated to the same zoning as the mobility data. As analyzed earlier, this entails the use of variables such as income, housing type or educational level. These can be continuous, categorical or ordinal. Consequently, the methodology is designed to be independent of the type of social data as well as the way of forming

social groups, for instance by clustering.

b. Mobility data requirements

Coherent zoning is crucial to connect social data to mobility data. Available and valid home zones, as well as activity locations, are crucial regarding the connection to the social group covered in the following. This means that especially for destination-based approaches a substantial amount of mobility data should include the destination points of the travel so that these can be traversed into destination zones.

Travelers should be distinguishable and traceable. Thereby, individuals can be recognized and home base (buildings, neighborhoods or stations close to expected home) can be inferred from their travel patterns. This home base can then be matched to a home zone coherent to the zoning used for social groups.

c. Connecting residential social groups to mobility data

Using the home zones, this approach links aggregated socioeconomic data to disaggregated mobility data. It is assumed that travelers from one zone represent a homogeneous group to a certain extent. Depending on the case, more than 70% of the mobility traces should be able to be linked to social data to obtain a significant number of matches.

The matching is done by including the home zones' information of each traveler into the mobility data set. Every journey is made by one distinctive user. This user has a home base and a respective home zone. Via the home zone, the social information gets connected to each transaction the user makes. Knowing every transactions' social classification enables calculating segregation measures which is done in the next step.

d. Measuring segregation index from mobility data

The measure is required to have an index format that calculates some relative distribution of groups within spatial zones. The index format facilitates transferability to different contexts and facilitates comparability.

As many segregation relevant variables lead to ordinal social groups, the "ordinal information theory index" developed by Reardon (2009) is used [19]. The ordinal information theory index measures segregation as the ratio of between-category variation to total variation. As a result, travelers' experienced segregation at the journey destination zone, assessed by the segregation measure, depends on their home zone's social status.

The following notations are introduced to calculate the ordinal information theory index given in Equation 1. The index is based on the ordinal variation function v shown in Equation 2 which relies on the distribution function f presented in Equation 3.

- k = ordered categories (social groups)
- m = unordered categories (neighborhoods, zones)
- t_m = total population in m
- T = Total Population
- c_m = $[K-1]$-tuple of cumulative population distribution in m
- v = ordinal variation

$$\Lambda = \sum_{m=1}^{M} \frac{t_m}{Tv}(v - v_m) \tag{1}$$

$$v = \frac{1}{K-1} \sum_{j=1}^{K-1} f(c_j) \tag{2}$$

$$f(c) = -[clog_2(c) + (1-c)log_2(1-c)] \tag{3}$$

By tracking mobility users over time and comparing similar time spans, there is the possibility to measure evolution of segregation. Since the index allows to calculate contributions to the segregation index on a zonal level, the evolution of segregation can be measured even for a single zone. This allows observing a zone's social mix of travelers over time. Still, no causal effects can be derived directly.

III. APPLICATION

The process steps developed in section II are applied on the public transport smart card data of Stockholm County, which is the most densely populated area in Sweden with 2.4 million inhabitants. Like many European metropolitan areas, it is a densely populated and crowded city where the public transport system plays a crucial role. The wealthy northern European country was found to face increased levels of segregation, 20 to 25 years after the shift from "the Swedish model", a comprehensive welfare system, to a neoliberal housing policy [20, 21, 22].

Segregation in Stockholm is mostly connected to findings on residents' ethnics and income [23]. Particularly low-income groups seem to be segregated towards the outskirts [22] along the northwest and southwest corridors. Recent years showed several endeavors of the authorities to tackle segregation such as the Citybanan project including a new commuter train tunnel in Stockholm's city center, opened in July 2017. It was build to distress both a national and regional bottleneck and led to the separation of commuter trains and regional/national train tracks in the inner-city. Next to operational gains, part of the investment motivation was to reduce segregation and improve accessibility for outer suburbs.

To determine social groups, each so-called DeSo (Demographic statistics areas) zone is assigned to one of four income quantiles using the median total earned income per zone of the 20+ years old population of each zone obtained from the 2017 Swedish income and tax register [24]. As a result, a mosaic-style structure of income distribution is revealed. While the highest income group 4 resides central, lower-income groups 1 and 2 are mostly spread around these centers. In particular, the group build from the areas with lowest median income, group 1, resides within the northwest and southwest suburban corridors.

The segregation index is measured at the destination of journeys, where destinations are indicated as stop area, an allocation of same-mode platforms based on a stop name. For

week 5 in 2020, 8.45 million journeys including a destination stop area and an inferred home zone are obtained. Similar journey amounts are derived for the same week in 2017, 2018 and 2019. To connect the residential-based social data to public transport smart card data, the income groups are assigned to the journeys' home zones. Therefore, smart card users' home zones are inferred using the method of Aslam et al. (2019) [25]. Once the social information is connected to every smart card transaction via card IDs, the smart card data set is enriched to apply a destination-based measurement of the social mixture.

Connecting the income groups shows differences in amount and direction of journeys. Contrarily to the somewhat similar population in every group, the total journeys are split unevenly. It can be seen that the journeys made per passenger decrease with higher income groups. While the lowest income group 1 combines 31.8% of all journeys, the highest income group 4 only accounts for about 20%.

Group 1 appears to travel in the northwest and southwest corridors, while Groups 2 and 3 are more spatially dispersed. Income group 4 has a more urban travel profile, traveling to the Southeast and North of the city.

The ordinal nature of the income groups distinguishes the measure to use for segregation. Therefore, the following results show the outcome of applying the ordinal information theory index by Reardon (2009) [19].

IV. Results and Discussion

The social segregation index score for each day of the week in 2017 averages 0.1923. Compared to 2017, the average segregation drops by 2.4% to 0.1877 in 2018 and by 3.3% to 0.1856 in 2019. In 2020, the segregation level averages 0.1888, up slightly from 2019 and 2018 but still 1.8% lower than in 2017.

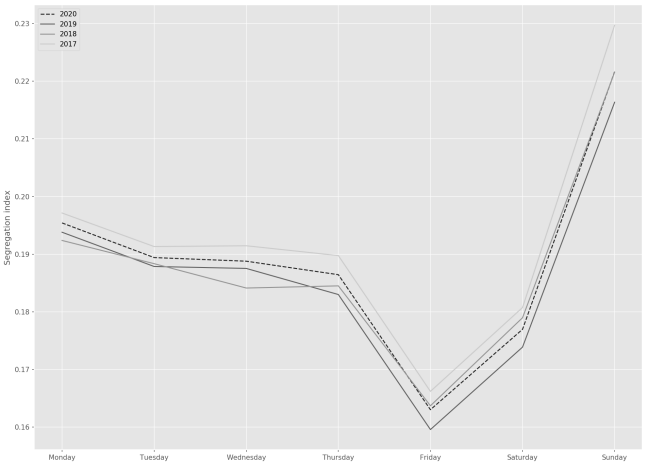

Fig. 2: Segregation throughout week 5

Figure 2 shows that people mix on a similar level on Monday to Thursday. Even less segregation is indicated for Fridays and Saturdays. On Sundays, travelers mix less and experience more segregation as the index is leaning more towards 1 than on other days.

Figure 3 and Figure 4 show respectively the weighted and "absolute" segregation contribution of each zone in 2017. The absolute contribution is the result of the differences between the total ordinal variation and the zone-specific variation $v - v_m$. The weighted value implies each zone's contribution to the segregation index calculated by the absolute contribution in relation to the population affected $\frac{t_m}{T_v}(v - v_m)$. In other words, the absolute value expresses the contribution before being set into relation with the number of passengers affected, while the weighted value accounts for the number of passengers affected compared to the overall amount of passengers.

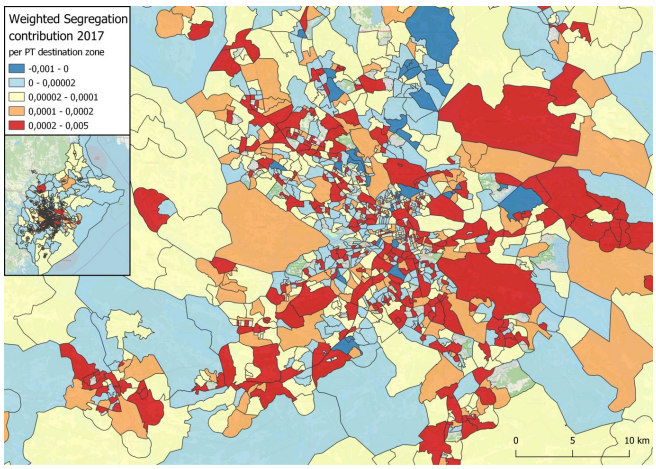

Fig. 3: Weighted segregation contribution 2017

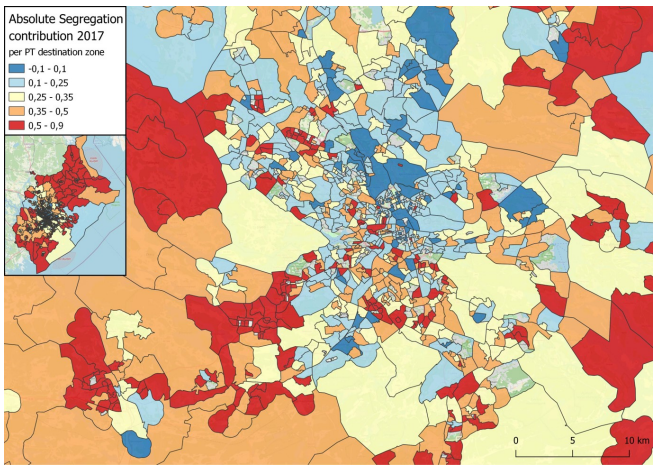

Fig. 4: Absolute segregation contribution 2017

As can be seen, the weighted segregation contribution is highest in central zones and suburban centers. What stands out, is the general pattern of zones with both high absolute and weighted contributions to the segregation index, which indicates that many passengers experience segregation at these destinations. For 2017, outskirt neighborhoods have high absolute and weighted contributions to the segregation index. Contribution to the segregation index mostly comes from the load of passengers in the city and suburban centers.

a. Analyzing segregation evolution

By taking differences of weighted segregation contributions between years, it can be seen whether a specific zone contributed to a decline or rise of the segregation index. Technically, for every zone and day of the week, the difference is calculated by taking the more recent year's contribution and

subtracting the 2017 contribution. Then, the average of differences is determined per zone over all days of the week. Thereby, the evolution of segregation can be assessed on a zonal level. A negative difference will indicate a decline in segregation. Contrarily, a positive difference shows an increase in contribution to segregation.

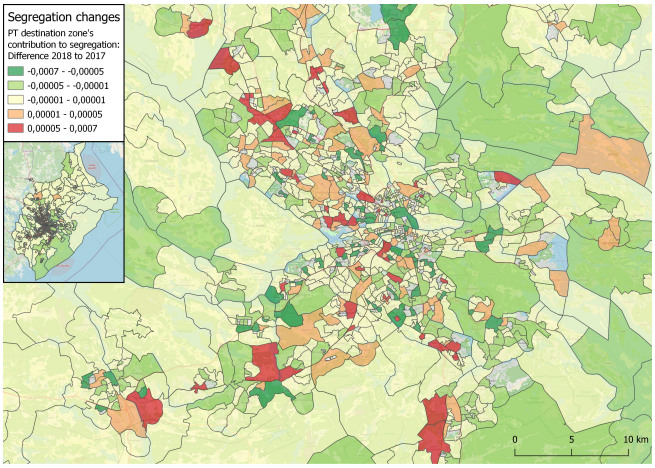

Fig. 5: Changes in contribution to segregation index 2017-2018

The dispersed structure and inner-city zone size make it difficult to immediately spot patterns and estimate effects in Figure 5. In addition, the zone sizes have an impact on the impression of segregation levels, though it does not indicate the number of passengers affected. In the city center, few sharply increased segregation zones can be detected, accompanied by strong decreasing and slightly to not decreasing segregation levels. Urban zones with segregation reductions outnumber the ones with rises for 2018.

For both 2018, as well as the 2020 comparisons to 2017, there are more zones where less segregation is experienced than zones that experienced more segregation. Even though for both years there are some zones which experienced stronger inclination in segregation (given in dark red), the amount of substantial declines (given in dark green) outweighs it as the mean change of zonal contribution for 2018 is -0.0000043 and for 2020 is -0.0000032. The overall change in 2020 is about 25% less compared to 2018 changes which matches the overall fallback trend stated earlier.

Potentially, urban neighborhoods and neighborhoods around the new commuter train line could show decreasing segregation after July 2017. Similar to earlier stated results, the effects are mixed as well as scattered throughout the zones. In the city center, some stronger decreases can be spotted in the central, western and southern zones, as shown in Figure 6 and Figure 7.

The segregation of urban zones, which incorporate the central station, Stockholms södra, as well as the new Odenplan commuter train station, strongly decreases for both years. In addition, the central northwestern Sundbyberg and Solna station zones show declined segregation levels for 2018 and 2020.

Increasing segregation levels are found in suburban and rural zones with commuter train stations. This leads to the depiction of partly decreasing segregation levels in suburbs with one strongly segregation-increasing zone: the zone of the commuter train station and local PT hub. Contrarily,

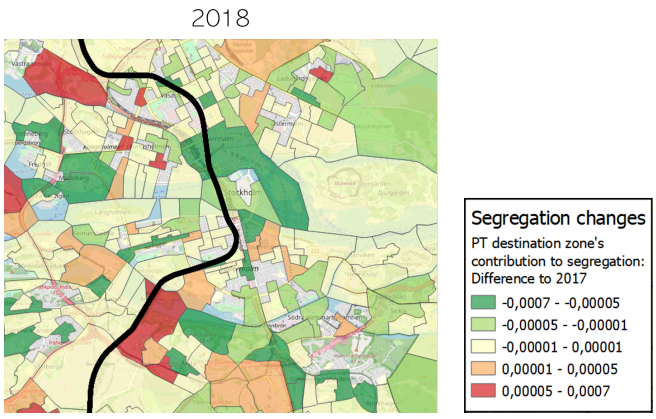

Fig. 6: Stockholm city center zonal segregation changes 2017-2018

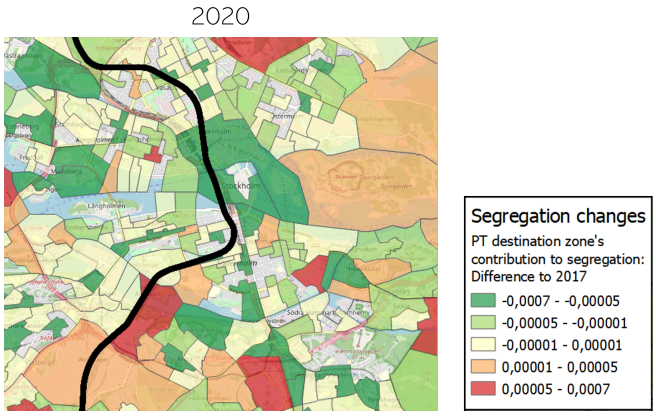

Fig. 7: Stockholm city center zonal segregation changes 2017-2020

suburban zones outside of the commuter train station areas slightly tend towards decreasing segregation levels, especially when including other PT hubs.

To conclude, double-edged results are obtained. Since these findings are based on public transport data, the results are examined regarding PT locations to identify the origins of segregation trends. Looking at the development of segregation, central zones with public transport hubs mostly have declining segregation levels over the years. Zones outside of the city center show increasing segregation for commuter train stations and, to some extent, decreasing segregation levels for other PT hubs.

b. Discussion

Figure 2 shows stable segregation levels on workdays with the lowest segregation on Saturdays and especially Fridays. These can be related to the combination of work, leisure, and shopping activities. Sunday is considered as a rest day with the least working activity which leads inhabitants to stay more within their home zone. These findings match other activity-based conclusions, such as that work-related activities diminish segregation [26].

By the setup of this case study, changes in segregation levels are related to a more diverse use of the PT system and to the population affected. Lower experienced segregation levels of passengers traveling to the city center outweigh the higher suburban and rural segregation experiences, due to the

higher number of passengers affected in central areas.

Weighted differences between the years' weekly average index allow accurate conclusions to be drawn on the segregation development. For the Stockholm case, the segregation index reported less segregation in the years after 2017, especially in the city center. These effects potentially relate to the enhanced public transport system after July 2017.

City center inbound PT passengers are found to be more income-diverse in 2018 and 2020 than in 2017, while outbound passengers towards the suburbs have more and more uniform income backgrounds, especially when traveling to commuter train stations. Both stronger increasing and decreasing effects are indicated for the northwest and southwest corridors. Increasing segregation levels in these suburban and peri-urban zones could be linked to general trends of urbanization and gentrification, as well as PT dependency and the transport disadvantage of low-income groups.

Main limitation of this study are the unrevealed direct causal effects as well as only assessing potential interaction within only PT users, which represent an already segmented group. In addition, the assumptions made about the homogeneity of groups in an area could lead to inadequacies in capturing the actual social composition.

V. CONCLUSIONS

This study utilized existing methods of activity-based social segregation measures and combines them with socioeconomic data to answer the main research question posted. Conceptualizing and applying process steps towards the measurement of segregation from mobility data enables measuring social segregation of public transport users at an activity-end level. Using the method to connect mobility data to social data could potentially lead to a more realistic depiction of social segregation and examine segregation developments in relation to transport or policy changes.

Daily or weekly segregation levels help evaluate overall levels and trends. Weighted segregation levels are suitable for analyses in which the relation of zone segregation plays a role. The absolute segregation contribution should be used for detailed, intra-zonal assessment, as it is more informative when other zones and the total population are not in focus. Particularly for urban planners and policymakers, it could be of interest to measure social segregation effects.

The results help evaluating the segregation situation in Stockholm and at the same time raise the question of why segregation is appearing more or less in certain areas. One goal could be to understand why commuter train related suburban and rural station zones increased in segregation. To answer the question on why these zones see less mixed passengers arriving, housing data or more traditional transport collection methods such as surveys could be used to explore the context and set the findings into relation with simultaneous processes such as gentrification and housing developments.

Finally, most important for authorities and policymakers could be a study looking into monetizing the segregation effects. This could be implemented with, for instance, using cost-benefit analyses regarding the population experiencing significant segregation changes.

Summing up, this study successfully proves to assess social segregation by using large-scale disaggregated mobility data. The method is flexibly designed for the use of different types of groups, indices, and mobility data. The present study raises and emphasizes the possibility to combine socioeconomic data and mobility data to analyze segregation. This contributes to the ongoing discussion on segregation, transport disadvantages, and ex-post transport appraisal.

REFERENCES

[1] J. S. Leonard, "The interaction of residential segregation and employment discrimination," *Journal of Urban Economics*, vol. 21, no. 3, pp. 323–346, 1987.

[2] D. Acevedo-Garcia and K. A. Lochner, "Residential segregation and health," *Neighborhoods and health*, pp. 265–87, 2003.

[3] E. Marques, "Social networks, segregation and poverty in são paulo," *International Journal of Urban and Regional Research*, vol. 36, pp. 958–979, 9 2012.

[4] K. Bischoff and S. F. Reardon, "Residential segregation by income, 1970-2009," *Diversity and disparities: America enters a new century*, vol. 43, 2014.

[5] Y. Tan, Y. Chai, and Z. Chen, "Social-contextual exposure of ethnic groups in urban china: From residential place to activity space," *Population, Space and Place*, vol. 25, 10 2019.

[6] United Nations, *World Social Report 2020: inequality in a rapidly changing world.* United Nations Department of Economic and Social Affairs. New York: United Nations publication. Sales No. E.20.IV.1, 2020.

[7] S. Farber, M. O'Kelly, H. J. Miller, and T. Neutens, "Measuring segregation using patterns of daily travel behavior: A social interaction based model of exposure," *Journal of Transport Geography*, vol. 49, pp. 26–38, 12 2015.

[8] S. Ureta, "Mobilising poverty?: Mobile phone use and everyday spatial mobility among low-income families in santiago, chile," *Information Society*, vol. 24, pp. 83–92, 3 2008.

[9] S. Silm and R. Ahas, "The temporal variation of ethnic segregation in a city: Evidence from a mobile phone use dataset," *Social Science Research*, vol. 47, pp. 30–43, 2014.

[10] S. Tao, S. Y. He, M. P. Kwan, and S. Luo, "Does low income translate into lower mobility? an investigation of activity space in hong kong between 2002 and 2011," *Journal of Transport Geography*, vol. 82, 1 2020.

[11] M. Bagchi and P. R. White, "The potential of public transport smart card data," *Transport Policy*, vol. 12, pp. 464–474, 9 2005.

[12] M. P. Pelletier, M. Trépanier, and C. Morency, "Smart card data use in public transit: A literature review," *Transportation Research Part C: Emerging Technologies*, vol. 19, pp. 557–568, 2011.

[13] K. Lucas, "Making the connections between transport disadvantage and the social exclusion of low income populations in the tshwane region of south africa," *Journal of Transport Geography*, vol. 19, pp. 1320–1334, 11 2011.

[14] A. Church, M. Frost, and K. Sullivan, "Transport and social exclusion in London," *Transport policy*, vol. 7, no. 3, pp. 195–205, 2000. [Online]. Available: www.elsevier.com/locate/tranpol

[15] J. Rokem and L. Vaughan, "Segregation, mobility and encounters in jerusalem: The role of public transport infrastructure in connecting the 'divided city'," *Urban Studies*, vol. 55, pp. 3454–3473, 11 2018.

[16] V. Kaufmann, "Social and political segregation of urban transportation: the merits and limitations of the Swiss cities model," *Built Environment*, vol. 30, no. 2, pp. 146–152, 2004.

[17] M. Utsunomiya, J. Attanucci, and N. Wilson, "Potential uses of transit smart card registration and transaction data to improve transit planning," *Transportation research record*, vol. 1971, no. 1, pp. 118–126, 2006.

[18] S. Abbasi, J. Ko, and J. Min, "Measuring destination-based segregation through mobility patterns: Application of transport card data," *Journal of Transport Geography*, vol. 92, 4 2021.

[19] S. F. Reardon, "Measures of ordinal segregation," in *Occupational and residential segregation*. Emerald Group Publishing Limited, 2009.

[20] K. Hedin, E. Clark, E. Lundholm, and G. Malmberg, "Neoliberalization of housing in sweden: Gentrification, filtering, and social polarization," *Annals of the Association of American Geographers*, vol. 102, pp. 443–463, 3 2012.

[21] R. Andersson and L. M. Turner, "Segregation, gentrification, and residualisation: From public housing to market-driven housing allocation in inner city stockholm," *International Journal of Housing Policy*, vol. 14, pp. 3–29, 1 2014.

[22] K. Grundström and I. Molina, "From folkhem to lifestyle housing in sweden: segregation and urban form, 1930s–2010s," *International Journal of Housing Policy*, vol. 16, pp. 316–336, 7 2016.

[23] R. Andersson and A. Kährik, "Widening gaps : Segregation dynamics during two decades of economic and institutional change in stockholm," in *Socio-Economic Segregation in European Capital Cities*. New York: Routledge, 2015, pp. 134–155.

[24] SCB, "SCB:s Open data for DeSO – Demographic Statistical Areas," 2017, retrieved from https : / / www . geodata . se / geodataportalen / srv / swe / catalog . search ; jsessionid = 42B5AAC3339638A205A27724ECF960BF # /search ? resultType = swe-details & _schema = iso19139* & type = dataset % 20or % 20series & from = 1 & to = 20 using the explanation from https : / / www . scb . se / en / services / open-data-api / open-geodata / deso--demographic-statistical-areas/.

[25] N. Sari Aslam, T. Cheng, and J. Cheshire, "A high-precision heuristic model to detect home and work locations from smart card data," *GeoSpatial Information Science*, vol. 22, pp. 1–11, 1 2019.

[26] M. Ellis, R. Wright, and V. Parks, "Work together, live apart? geographies of racial and ethnic segregation at home and at work," *Annals of the Association of American Geographers*, vol. 94, pp. 620–637, 9 2004.

APPENDIX

Table 1 gives an overview on the Income groups used in this study made from 2017 tax and income register data of SCB [24].

TABLE 1: INCOME QUANTILES IN STOCKHOLM COUNTY

Group	Name	median annual income of 20+ years population	Population	Share
1	Lower income zones	below 274K SEK	611,963	26.2 %
2	Lower-middle income zones	274-327K SEK	568,799	24.3 %
3	Upper-middle income zones	327-372K SEK	572,519	24.5 %
4	Higher income zones	above 372K SEK	586,006	25.1 %

B APPENDIX B: SCD JOURNEY SUCCESS RATES

Table B.1: Home zone and destination inference success rates

Year	Number of journeys	Journeys with destinations	Journeys with destinations share	Journeys with destinations and home zone	Journeys with destinations and home zone share
2017	10516394	8119890	77.2%	7704362	73.3%
2018	10594007	8721720	82.3%	8388876	79.2%
2019	10683207	8858317	82.9%	8509129	79.6%
2020	10850917	8870936	81.8%	8450172	77.9%

C APPENDIX C: WEIGHTED AND ABSOLUTE SEGREGATION CONTRIBUTIONS 2018-2020

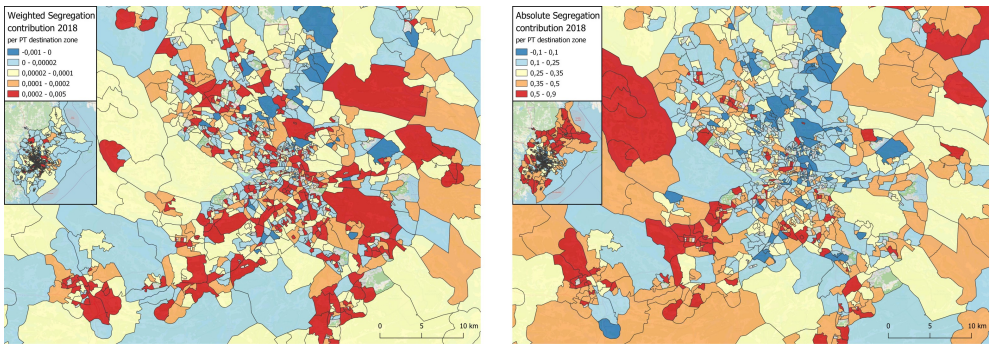

Figure C.1: Weighted and absolute segregation contribution 2018 - each DeSo zone's arriving PT passenger mix contribution to the segregation index level

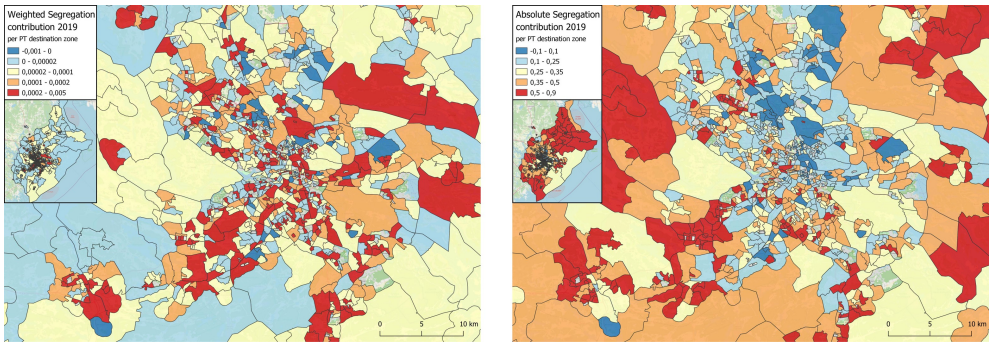

Figure C.2: Weighted and absolute segregation contribution 2019 - each DeSo zone's arriving PT passenger mix contribution to the segregation index level

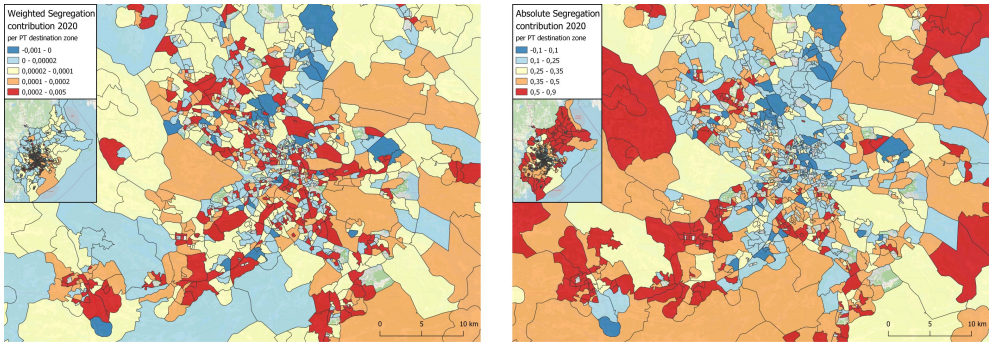

Figure C.3: Weighted and absolute segregation contribution 2020 - each DeSo zone's arriving PT passenger mix contribution to the segregation index level

D APPENDIX D: SIGNIFICANCE TESTS

Two t-tests are performed to determine whether the means of the respective data sets are significantly different from each other.

Table D.1: t-test on zone's weighted segregation contribution 2018 compared to 2017

2017-2018	Variable 1	Variable 2
Mean	0.000227853	0.000232711
Variance	0.0000002429	0.0000002405
Observations	1087	1087
Hypothesized Mean Difference	0	
df	2172	
t Stat	-0.2304	
P(T≦t) one-tail	0.4089	
t Critical one-tail	1.6456	
P(T≦t) two-tail	0.8178	
t Critical two-tail	1.9611	

Table D.2: t-test on zone's weighted segregation contribution 2020 compared to 2017

2017-2020	Variable 1	Variable 2
Mean	0.000227209	0.000232711
Variance	0.0000002427	0.0000002405
Observations	1087	1087
Hypothesized Mean Difference	0	
df	2172	
t Stat	-0.2610	
P(T≦t) one-tail	0.3971	
t Critical one-tail	1.6456	
P(T≦t) two-tail	0.7941	
t Critical two-tail	1.9611	

Note that 6 zones visited in 2017 were not visited in 2018. The same zones where not visited in 2020 plus additional 3 zones, adding up to a total of 9 zones excluded from this t-test due to no data available.