UNIVERSITY OF COPENHAGEN
FACULTY OF SOCIAL SCIENCES

Master Thesis in Social Data Science

Mobility and Income Segregation in Madrid, Spain

Faculty of Social Sciences, University of Copenhagen

Carolina Brañas Soria

Supervised by Sandro Sousa and Yevgeniy Golovchenko

Number of Characters (with spaces): 95,898

Abstract

This study explores the relationship between income segregation and urban mobility in the city of Madrid. Literature shows that bringing mobility into the study of inequality and segregation can bring an extra dimension to its study. By using a granular mobility dataset of trips throughout a month of 2022, a series of methods are developed to measure which districts play a central role in pulling and pushing travellers, and to measure which residents travel more, and longer distances within the city, based on the median income per consumption unit and Gini index of their district of residence. This research provides insights on the ongoing income segregation patterns in Madrid, and identifies slightly homophilic travelling tendencies based on the socioeconomic background of the district of residence of individuals. It also reveals that lower income groups tend to perform more trips, and longer on average, than higher income groups. This research contributes to the field of urban studies and emerging fields like Science of Cities or Urban Data Science, and aims to aid policymakers to identify mobility isolated areas in the urban space to develop effective policies to reduce the negative effects of gentrification and segregation.

Contents

1 Introduction

With an increasing urbanisation rate across the globe, understanding how cities evolve and how we interact with them is a key aspect to improve the overall citizens' satisfaction and quality of life. Segregation and inequality are inherent to the existence of complex societies, and it has become 'a crucial subject across dozens of scientific fields, from Sociology and Urban Studies to Environmental Science and Gender Studies' (Netto et al., 2024).

Unfortunately, 'due to its pervasive nature, segregation is naturally hard to pinpoint' (Netto et al., 2024). Because of its multifactorial nature and its causes, these patterns, including those happening in urban areas, are still far from being understood. A new line of research expands on the concept of mobility to further understand the complexity of different kinds of segregation. After this emergence of a 'mobilities' paradigm (Park and Kwan, 2017), segregation is assessed as a dynamic concept, considering activities happening beyond just residential areas. As Candipan et al. state, the foundations of the study on (racial) segregation still remain at the analysis of neighbourhoods of residence, rather than the individuals' everyday travels. These authors focus on trips across neighbourhoods instead of just the residence of individuals to study racial segregation patterns in the United States. Combined with this new line of research, there is now a higher availability of mobility data, increasing the possibilities to study mobility patterns and demographic characteristics of individuals with high granularity. This extra dimension of study can bring many possibilities and angles to apprehend segregation patterns.

The ultimate goal of this study is to contribute to the growing body of research on income segregation and inequality patterns in Madrid, Spain, with the aid of computational tools to focus on mobility patterns. I argue that there is a need to add an extra dimension to the study of segregation by focusing on these daily trips, as this approach helps grasping the aforementioned complex inequality and segregation causes better than just focusing on residential areas. This research contributes to the field of urban studies, and emerging fields like Science of Cities or Urban Data Science, which 'exploit new large-scale urban datasets with quantitative tools' (Alessandretti et al., 2023). This research aims to provide insights and methods to identify mobility patterns based on income in Madrid. These insights can be used by policymakers to identify mobility isolated areas in the urban space, all with the focus on developing more effective policies for contributing to societal changes (Park and Kwan, 2017). I aim to answer the following research question:

> *RQ: To what extent do residents from economically segregated districts in Madrid move differently than those in other districts?*

To address the main question, I focus on testing whether the following hypotheses hold. Hypothesis 1 is rooted in the pull-push concept developed by Ravenstein (Ravenstein, 1885) and whether these pull-push patterns are mimicked in the urban system of Madrid:

Hypothesis 1 (H1): *Low-income and highly segregated districts push residents out for daily mobility purposes, like work or recreational, whereas high-income and low segregated districts pull residents, as these districts tend to offer higher amenities and work possibilities.*

To test hypothesis 1, I measure which districts play a central role in terms of receiving (in-weight) or pushing (out-weight) residents from them and the relationship to income. In addition, I measure mixing patterns across districts based on income.

Hypothesis 2 and hypothesis 3 are based on the continuous expansion of urban cities into large metropolitan areas and the consequent gentrification patterns exacerbated by contexts of economic crisis and global tourism (Ardura Urquiaga et al., 2020). As cities tend to grow, and low-income residents tend to get displaced to the peripheral areas (Ardura Urquiaga et al., 2020), these might need to travel more often, and longer distances, to central districts offering work opportunities and a high number of amenities:

Hypothesis 2 (H2): *Residents from low-income and highly segregated districts tend to make a higher number of trips to high-income and low segregated districts than vice versa.*

Hypothesis 3 (H3): *Residents from low-income and highly segregated districts tend to perform, on average, longer distance trips than those living in high-income and low segregated districts.*

To test hypothesis 2, I quantify to what extent residents from lower income districts tend to travel to higher income districts compared to the rest of residents in terms of number of trips. To test hypothesis 3, I measure to what extent the average travel distance differs between residents from low and high income districts.

The *literature review* touches upon the previous attempts of defining and measuring segregation, and the contextual situation of Madrid as an object of study. It also delves into the reasoning in depth behind the formulation of the hypotheses that guide the research. The *data* section evolves around the dataset used for this project, and the choice of income variables to guide the process. The *Methods* section describes the initial exploratory analysis used to address the state of income distribution in Madrid during 2021. This allows me to select the appropriate income variables for the study. To do so, (1) I plot and study the distribution of six income estimators to understand which are more accurate for representing the income and segregation levels of the districts. Based on this analysis, I select two key variables to understand the interactions between mobility and income: *the median income per consumption unit* and the *Gini index* by district. (2) I then calculate Global and Local Moran indices for each district based on the median income per consumption unit to illustrate the clustered patterns of income distribution in the city. After the initial exploratory analysis that indeed reveals clustering patterns, I develop methods to study the relationship between income and mobility to test the previously formulated hypotheses. The *Results* and *Discussion* sections present the findings and their implications.

2 Literature Review

2.1 Segregation: A Complex and Multi-Dimensional Process

Despite its complexity, segregation can be defined as the 'processes, practices or situations in which individuals or social groups are separated, or their interactions restricted, based on distinguishing characteristics such as race, ethnicity, income, gender, occupation, education, geography, age, or other social attributes' (Netto et al., 2024). Most researchers agree on the idea that it is a complex process with diverse causal factors, which vary based on the different social processes taking place at a certain time and location (Marcuse, 2022). In addition, Netto et al. argue that 'its complexity is also evident in the frequent conflation of terminology'. As a result, such multi-dimensional process requires a multidisciplinary approach (Vaughan and Arbaci, 2011). Due to this intricacy, the existing literature on segregation is very extensive, and provides many different tools and perspectives to measure it.

Back in 1988, Massey and Denton attempted to capture such complexity by defining the five dimensions of segregation: evenness, exposure, clustering, centralization, and concentration (Massey and Denton, 1988). As Feitosa, F. F. et al. summarise, Massey and Denton defined evenness and exposure as aspatial dimensions of segregation: 'evenness concerns the differential distribution of population groups. Exposure involves the potential contact between different groups'. On the other hand, clustering, centralization and concentration were classified as spatial dimensions, since they require information about location, shape and/or size of areal units: 'clustering refers to the degree to which members of a certain group live disproportionately in contiguous areas. Centralization measures the degree to which a group is located near the centre of an urban area. Concentration indicates the relative amount of physical space occupied' (Feitosa, F. F. et al., 2007).

It was in 2004 when Reardon and O'Sullivan reviewed Massey and Denton's work and argued that segregation has no aspatial dimension. Hence, the difference between evenness (aspatial) and clustering (spatial) is most likely an effect of data aggregation at different scales. As a result, they proposed the spatial exposure/isolation dimension, which measures the chance of having individuals from different (or the same) groups living side by side (Reardon and O'Sullivan, 2004).

2.2 Spatial Segregation Indices

Along with several definitions of segregation, many indices were developed in an attempt to measure it. As Feitosa, F. F. et al. mention, segregation measures range from the earliest approaches that focus on calculating the differentiation between two population groups, like the dissimilarity index D (Duncan and Duncan, 1955) or the exposure/isolation index (Bell, 1954) to more recent and refined methods. A second generation of indices aimed to capture the segregation between more than two groups (Morgan, 1975; Sakoda, 1981). However, this family of indices is susceptible to a key problem, highlighted by White: the checkerboard landscape problem. Let us imagine that the neighbourhoods of a metropolis are represented as squares, which are coloured as black if the neighbourhood has a high income value, and as white otherwise. Using these indices, segregation measures for a metropolis in which the squares are placed like a checkerboard turn out to be the same as the ones for a metropolis in which all the white squares are moved to one side and all the black squares are moved to the other. This is clearly a limitation, as we would expect a higher segregation index when the values are highly clustered. This problem is represented in Figure 1.

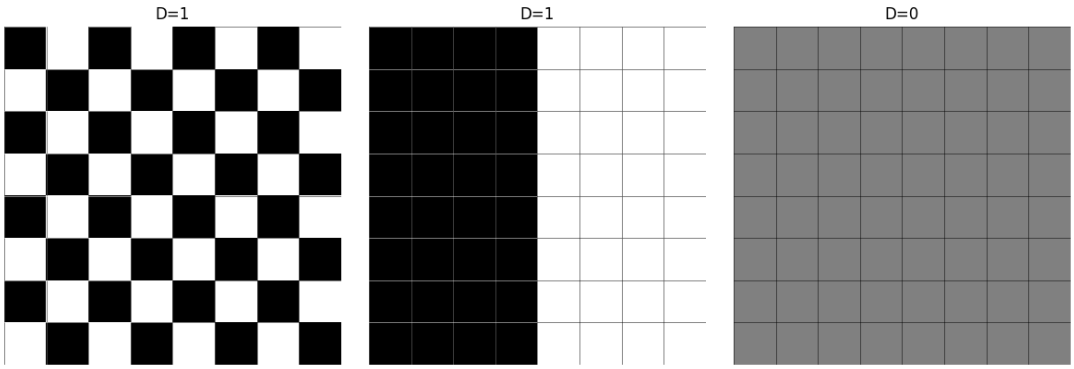

Figure 1: The checkerboard problem occurring in traditional measures. The central and left squares in the figure obtain the same segregation index value ($D = 1$), regarding the spatial arrangement of the spatial units. The right one obtains a different value ($D = 0$).

To solve this problem, new indices that took into account the spatial arrangement of the

population were developed, like the index of spatial proximity (White, 1983) or the distance-based index of dissimilarity (Jakubs, 1981). Within this group of spatial indices it is important to mention others like the Boundary Perimeter/Area ratio, the spatial Gini index, or the Moran's I Global Auto-correlation (Moran, 1950).

The literature on spatial indices quickly evolved so researchers can specify their own definition of interaction across groups based on certain spatial features. For example, Wong proposed a spatial version of the previously mentioned dissimilarity index, named the spatial dissimilarity index. Reardon and O'Sullivan replaced population counts of tracts by weighted population density values.

However, it is important to remember that reducing such a multifactorial problem to a set of numerical measures has its limitations in itself. As Yao et al. explain, spatial segregation indices intrinsically have a geographic nature, hence they are subject to two fundamental problems: how spatial interaction is represented and the choice of a spatial scale. Segregation indices are a key tool in the exploratory data analysis process, but they are always approximations of the truth, and further measurements must be completed to try to understand segregation patterns.

2.3 Urban Segregation and Mobility

Urban segregation is the kind of segregation occurring within urban areas. It 'transfers social inequalities to the physical space and, in turn, the characteristics of urban areas can accentuate processes of inequality. Ultimately, these dynamics can be understood as mutually reinforcing' (González-García et al., 2024). As it occurs with all types of segregation, many authors have tried to identify its causes. Some argue that urban segregation 'has different meanings and effects depending on the specific form and structure of the metropolis, as well as the cultural and historical context. Its categories include income, class, race, and ethnical-spatial segregation' (White, 1983; Jargowsky, 1996; Reardon and O'Sullivan, 2004). In particular, Marcuse indicates that urban segregation has been historically related to three aspects: 'cultural elements (racial, ethnic, religious), status (income level, power relations) and that related to the division of urban functions (zoning according to urban use)'.

In recent years, there has been a growing body of research emerging thanks to the ability to obtain and study big databases, which supports that, to measure segregation, 'urban researchers should take into account mobility behaviour and not only residential patterns' (Moro et al., 2021), questioning classic segregation concepts such as *social isolation* (Wilson, 1987). Despite its limitations, mobile phone data has recently emerged as a great source to study these granular mobility patterns (González et al., 2008) that Moro et al. are referring to. Neira et al. argue that this access to phone data allows to quantify segregation in urban areas by 'analysing individual mobility patterns and the places people visit'. For example, Bassolas et al. studied why some socio-economic minorities get an abnormal COVID-19 impact in the US by focusing on mobility and commuting patterns. Kazmina et al. argue that, 'while the spatial focus on neighbourhoods in both literature and policymaking is understandable, it may underestimate actual patterns of social segregation'. To avoid these pitfalls, they follow a population-scale network analysis approach to disintegrate socio-economic segregation in the Netherlands. Müürisepp et al. focused on segregation in activity spaces, and mobility flows to analyse ethnic, racial, religious, linguistic, socioeconomic and demographic dimensions of segregation.

I follow this new line of research and argue that there is a strong need to focus on mobility patterns to understand segregation in depth, adding an additional dimension to the research process.

2.4 Madrid: An Example of a Gentrified, Segregated Southern European City

There is extensive literature focusing on understanding the state of segregation in cities, especially in the United States and Northern Europe. For example, Haandrikman et al. compare the state of socio-economic segregation in Brussels, Copenhagen, Amsterdam, Oslo and Stockholm. Previously, Musterd et al. concluded that there was an increase in social inequality and spatial segregation in European capitals between 2001 and 2011, suggesting that the typical European city is leaning toward the polarised urban model characteristic of North American and Latin American cities (Borja and Castells, 1997). Unfortunately, 'given the complex social mix of European cities, it is not possible to identify a single model of segregation' (Zambon et al., 2017) that could be extrapolated to all cities. For example, Mediterranean urban geographies often differ widely from American or Northern European settlements. These Southern European contexts are typically 'associated with economic informality, planning deregulation, family-oriented welfare regimes and weak (and partly ineffective) housing policies' (Zambon et al., 2017). In addition, the effects of the Second World War, the dictatorship status throughout the XX century, and the 2008 crisis have had a widely different impact in Southern Europe than in the rest of the continent, shaping urban morphology and segregation patterns differently. Due to these contextual disparities, I argue that it is not feasible to extrapolate the study of segregation to a wide area, but rather there is a need to study segregation patterns within areas with similar cultural and contextual characteristics. As a result, I focus in particular on Madrid, capital of Spain, to study its income segregation patterns.

Madrid is divided into twenty-one districts with a total population of over three million residents, and a foreign population (non Spanish nationality) of 17% (Municipio en Cifras, 2023). It has been found that 'residents of large cosmopolitan areas have less exposure to a socio-economically diverse range of individuals' (Nilforoshan et al., 2023), meaning that there might be an increasing pattern of residential segregation within big, cosmopolitan cities. Madrid is big enough to be considered a large, cosmopolitan area in which citizens tend to cluster socio-economically. As it often happens with such big metropolises, it is surrounded by many commuter towns that do not officially depend on the city's administrative scope, but that influence it in many ways. For the scope of this project, I have decided to focus only on the official twenty-one districts that build up Madrid, and these commuter towns have been left out of the analysis, as most of them have high populations and would need a specific analysis on their own. Figure 2 shows a map of the official districts of study, whereas Table 1 displays their population.

To help answer the main research question, I develop three hypotheses rooted in Ravenstein's pull-push concept (Ravenstein, 1885), the continuous expansion of cities and the consequent gentrification patterns.

Hypothesis 1: Pull-Push Patterns in Madrid

The pull-push concept was first introduced by Ravenstein in his Laws of Migration (Ravenstein, 1885), where he recognised the existence of an absorption process through which people surrounding a rapidly growing city move into it. This process continues to happen until the pull factor is spent. The Laws of Migration (Ravenstein, 1885) comprehend a global understanding of migration flows than just those happening inside urban areas. However, based on the segregation patterns expressed in Madrid and the large size of the city, I suggest that there might exist pull-push factors that are being reflected in daily mobility patterns, with high income, low segregated areas pulling citizens in for work or recreational purposes, and low income, high segregated areas pushing citizens out.

Hypotheses 2 and 3: Gentrification in Madrid

These pull-push forces might be exacerbated by the amenity distribution and gentrification patterns in the city. Like it happens in many Southern European cities, 'the central district of Madrid and its surroundings are showing patterns of a new wave of gentrification' (Ardura Urquiaga et al., 2020). Due to its cultural appeal and lower housing costs, the city has seen an influx of workers and tourists, which has 'displaced lower-income populations from the city's centre' (Ardura Urquiaga et al., 2020) towards peripheral areas. As a result, 'poor residents who cannot afford to relocate' must 'remain in under-served neighbourhoods' (Kaufmann et al., 2022). This creates a vicious cycle in which inhabitants from the poorly-connected areas face 'ever-more tenuous (and often increasingly expensive) links to even basic access to essential services like energy, transportation, communications, even certain urban streets' (Graham and Marvin, 2002). These difficulties have been observed by Park and Kwan, who estate that socially marginalized groups often show more restricted mobility patterns than other groups due to a lower share of private vehicle ownership in addition to the deprivation of adequate public transportation, entrapping them in a resource-poor area. Despite these reduced mobility patterns, it is often the lowest income demographics that need to travel longer distances to reach thriving areas in the city: 'Social exclusion is manifest in the fact that these family members usually have to travel long distances in order to reach important destinations, especially their workplaces. This fact alone would not necessarily mean social exclusion if it wasn't connected with the second aspect of their daily travels: their limited access to modes of mobility' (Ureta, 2008).

All these gentrification relocation processes are fed by many sociological changes in the recent years, exacerbating segregation, like the 2008 recession, unemployment and rising house prices (González-García et al., 2024). The social and income segregation in the city also has a clear spatial manifestation, where most of the 'rich' districts are located in the north-west part of the city, and the 'poor' districts are mostly clustered in the south-east (González-García et al., 2024), divided by an imaginary 'poverty line'. According to these authors, socio-spatial mobility is affected by these inequalities, and the urban spaces are a reflection of them.

It is the advances in transportation modes, urban sprawl, the aforementioned gentrification processes and pushing of low-income residents to peripheral areas, and the geographic differences between home and work locations that support the formulation of the hypotheses driving this research.

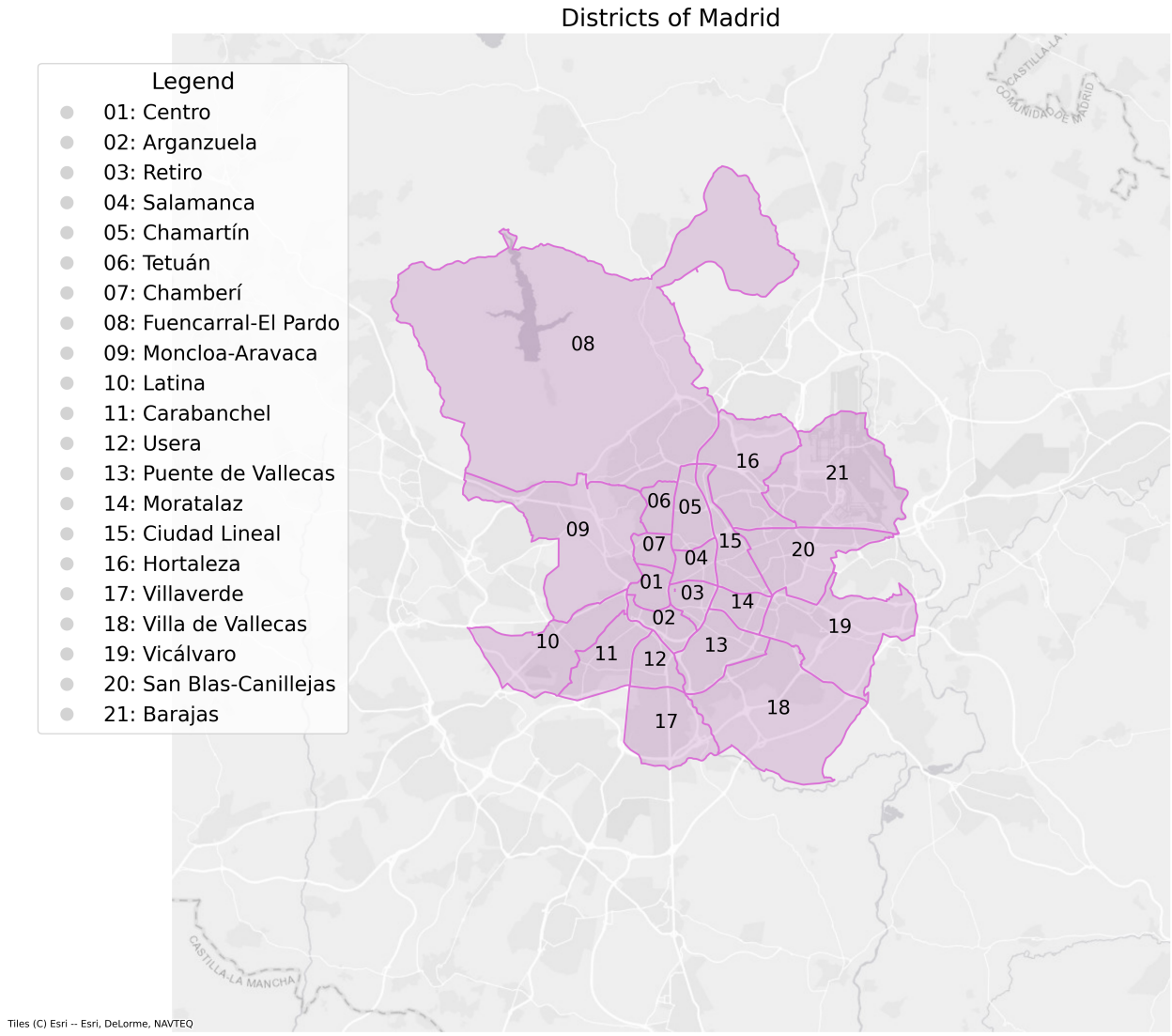

Figure 2: The official twenty-one districts comprising Madrid.

Name	Population
Centro	140,644
Arganzuela	153,982
Retiro	118,335
Salamanca	145,579
Chamartín	145,444
Tetuán	159,564
Chamberí	138,335
Fuencarral-El Pardo	247,327
Moncloa-Aravaca	120,589
Latina	239,693
Carabanchel	258,064
Usera	142,324
Puente de Vallecas	238,577
Moratalaz	93,671
Ciudad Lineal	216,400
Hortaleza	192,809
Villaverde	154,464
Villa de Vallecas	114,469
Vicálvaro	75,283
San Blas-Canillejas	159,900
Barajas	49,955

Table 1: Population by district in Madrid in January of 2021.

3 Data

This section delves on the processes to extract the adequate data for the study. To perform this research, I retrieve a dataset of over four million rows containing home-origin trips and demographic information during March of 2022 within the official twenty-one districts of study, provided by The Ministry of Transport and Sustainable Mobility (Ministerio de Transportes y Movilidad Sostenible (MITMA), 2022). The data presented is aggregated and anonymised. This mobility dataset is complemented with a selected income indicator, the median income by unit of consumption by district in 2021, and an intra-district inequality indicator, the Gini index by district in 2021.

All the code and thorough steps to extract the data and conduct the study are detailed in the following GitHub repository: `https://github.com/carobs9/segregation-madrid`.

3.1 Mobility Data

The Ministry of Transport and Sustainable Mobility (Ministerio de Transportes y Movilidad Sostenible (MITMA), 2022) published open data on national mobility patterns retrieved from mobile phone positioning collected by a national mobile network operator in Spain (Ponce-de-Leon et al., 2021). The trips were aggregated using users' movements between consecutive stays of at least 20 minutes in the same area, disregarding trips of less than 500 metres (Ponce-de-Leon et al., 2021). Finally, the data has been aggregated based on origin-destination (OD) terms at hourly time scale, encoding trips occurred during a given hour between two districts. Lastly, 'for each origin and destination, the activities at origin and destination are classified as home, work/study place, frequently and infrequently visited place. This data collection is based on individuals' active events, e.g., users' calls together with passive events, in which the user's device position is registered due to changes in the cell tower of connection' (Duran-Sala et al., 2024). In essence, a trip is defined as the movement of an individual between two consecutive activities, which are the reasons that motivate a trip (work, home, and other).

The original data encompasses trips from January 1st of 2022 until today, and is updated regularly. The smallest units of study are districts, hence this is the unit I use to analyse mobility patterns. Summarising, for each hourly and demographic combination, an aggregated number of trips and total travelled kilometers (km) are provided. As a result, the following variables are used in this study:

- Total amount of trips performed by hour and demographic.
- Total amount of km travelled by hour and demographic.
- Origin of the trip.
- Destination of the trip.

For the scope of this project, I rely on a representative temporal and spatial subset of the data: home-origin trips during March 2022, within the twenty-one districts of Madrid. The reason to use a 2022 mobility sample and not a more recent one is that the latest available district-level income data provided by the National Statistics Institute at the time of this research dates from 2021 (Instituto Nacional de Estadística, 2021a). Hence, I am trying to approximate the mobility sample as much as possible to the income data. The reason to retrieve March trips is because it is the first 31-day month in the year that does not have many festivities, as the beginning of January is still a holiday period in Spain.

The trips dataset is filtered to only contain trips in which individuals travel from home. The reason for this filtering is that the dataset does not contain information on the district of residence of each individual, but it does contain the district where aggregated trips start, and

whether the origin of the aggregated trips is home. By filtering only trips with home origin, it is possible to extrapolate the district of residence, and, hence, the socioeconomic status of the traveller. As a result, during the whole month of March 2022, I retrieve a dataset of 4,839,108 hourly and demographic combinations, adding up to around 3,3 quadrillion trips. Figure 3 illustrates the filtering process, and Table 2 displays a subsample of the raw dataset.

Date	Hour	Origin	Dest.	Distance	Activity Origin	...	Age	Sex	Trips	Trips Km
20220301	0	Centro	Centro	0.5-2	home	...	0-25	man	29,337	30,222
20220301	0	Centro	Centro	0.5-2	home	...	0-25	woman	34,143	27,522
20220301	0	Centro	Centro	0.5-2	home	...	25-45	man	92,799	63,331
...	...	...	...	...	...	...	...	...	...	...
20220331	23	Barajas	Barajas	2-10	home	...	25-45	woman	3,400	7,835
20220331	23	Barajas	Barajas	2-10	home	...	45-65	man	2,005	9,351
20220331	23	Barajas	Barajas	2-10	home	...	25-45	man	4,940	12,918

Table 2: Sample of the raw dataset containing demographic information and the total trips and total km travelled by each hourly OD and demographic combination. Some demographic columns have been removed to ease the visualisation.

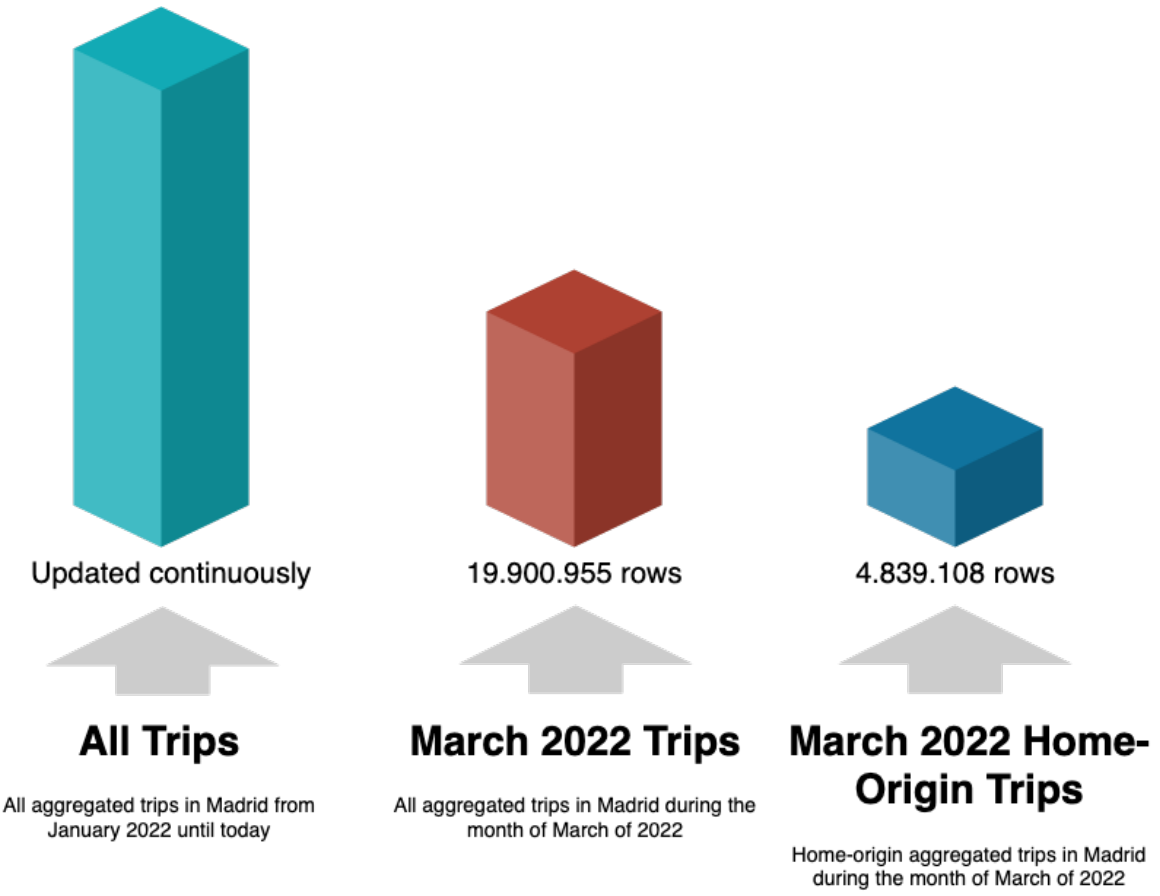

Figure 3: Filtering process of the trips dataset. For the purpose of the study, only March of 2022 home-origin trips have been used. Using home trips allows to infer the socioeconomic background of the individuals performing the trips.

Geometry

In addition, a set of shapefiles are included along the mobility data (Ministerio de Transportes y Movilidad Sostenible (MITMA), 2022), containing the geometries (polygons and centroids) of the different districts appearing in the study. These shapefiles allow the user to plot the different districts of choice. There are extra files containing the mapping between district IDs, their names and their population.

3.2 Income Data

I select the median income per unit of consumption and the Gini index for further quantitative analysis as income and segregation indicators respectively. In this section, I explain the reasoning behind such selection among different income variables.

Income Estimators Selection

To infer the economic background of each district, I initially retrieve the following income variables for the year of 2021 (Instituto Nacional de Estadística, 2021a) from the National Statistics Institute:

- Average net income per person.
- Average net income per household.
- Average income per unit of consumption.
- Median income per unit of consumption.
- Average gross income per person.
- Average gross income per household.

Figure 4 and Table 3 show a distribution of the income variables. Based on a preliminary analysis described in depth in the *Methods* section, I statistically prove that all income variables show significant clustering patterns. Despite its smaller Global Moran's coefficient, I select the median income per unit of consumption for further analysis, as median measurements reduce the influence of outliers in comparison to averages. The distribution of this variable throughout the districts is shown in Figure 5.

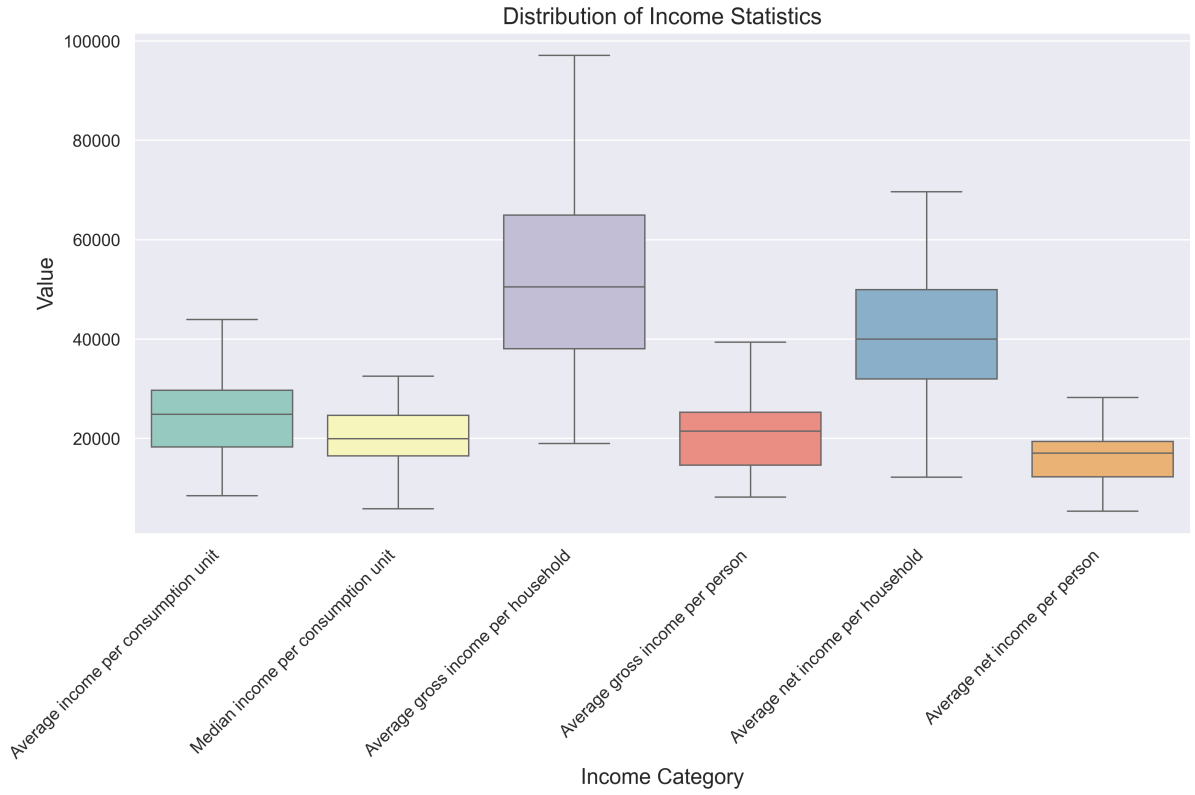

Figure 4: Distribution of income variables for all districts in 2021

	Average income per consumption unit	Median income per consumption unit	Average gross income per household	Average gross income per person	Average net income per household	Average net income per person
Count	21.0	21.0	21.0	21.0	21.0	21.0
Mean	27070.3	22350.0	57604.9	23275.4	44733.4	18045.0
Std Dev	8454.0	5829.7	18974.1	8209.7	12225.6	5363.2
Min	16116.0	14350.0	33395.0	12678.0	28681.0	10797.0
25%	20469.0	18550.0	42779.0	16498.0	35278.0	13719.0
50% (Median)	24874.0	19950.0	50487.0	21495.0	39991.0	17026.0
75%	32322.0	26950.0	72362.0	27311.0	55125.0	20671.0
Max	43930.0	32550.0	97093.0	39346.0	69670.0	28233.0

Table 3: Statistical summary of income data by category.

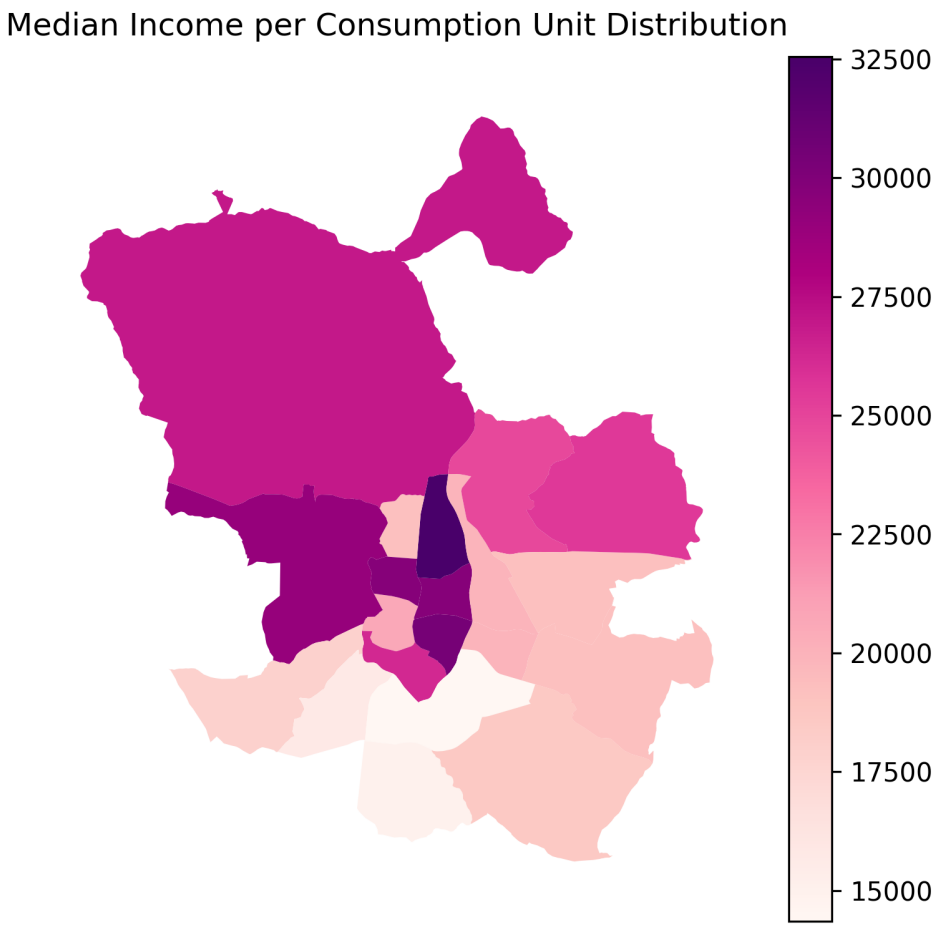

Figure 5: Median income per consumption unit by district

Gini index

In addition to the median income per consumption unit, I retrieve the Gini index by district (Instituto Nacional de Estadística, 2021b). The Gini index is a great estimator of the segregation within each of the districts of interest. It 'is a summary statistic that measures how equitably a resource is distributed in a population' (Farris, 2010). The index ranges from 0 (perfect equality) to 1 (perfect inequality), although it is often expressed as a percentage. In this case, the index is calculated on an income variable, 'the income per unit of consumption in the population, which is an income concept used internationally for a better comparison of individual incomes according to different types of households' (Instituto Nacional de Estadística, 2023). This makes it a great choice to represent income inequality within districts.

The different Gini indices by district are shown on Table 4, whereas Figure 6 shows a visual representation of the values on the map. As mentioned above, it is important to keep in mind that the Gini index is a measure of inequatily and not of wealth, and can enrich the analysis in terms of understanding segregation patterns, as a single measure cannot fully summarize a distribution. As Liu and Gastwirth explain, researchers can benefit from combining the Gini index with another measure which suits the study to overcome these limitations.

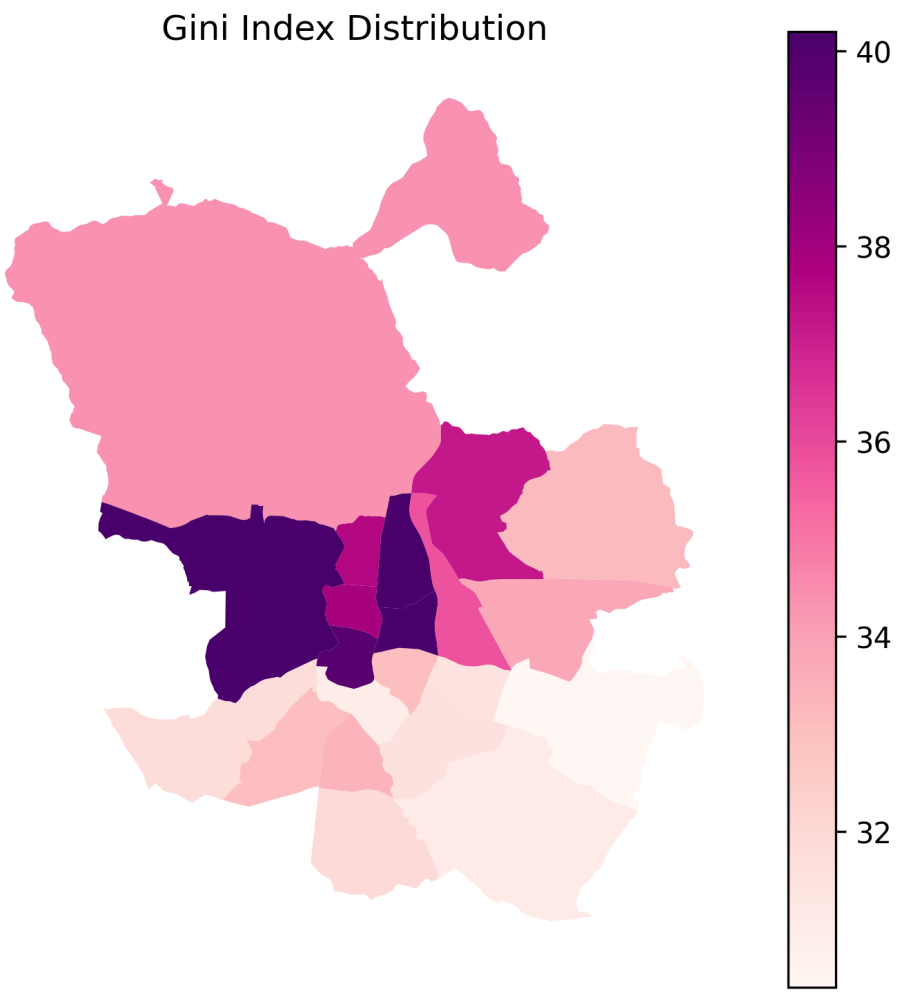

Figure 6: Gini index by District

District	Gini Index
Centro	39.8
Arganzuela	31.0
Retiro	33.1
Salamanca	40.2
Chamartín	40.2
Tetuán	37.6
Chamberí	37.9
Fuencarral-El Pardo	34.4
Moncloa-Aravaca	40.2
Latina	31.8
Carabanchel	33.1
Usera	33.4
Puente de Vallecas	31.6
Moratalaz	31.5
Ciudad Lineal	35.8
Hortaleza	37.2
Villaverde	31.9
Villa de Vallecas	31.1
Vicálvaro	30.4
San Blas-Canillejas	33.8
Barajas	33.2

Table 4: Gini index by district in Madrid.

4 Methodology

The following section first describes the different methods used to test the state of income segregation in the year 2021 and to select an adequate income indicator. Then, delves into the methods used to test hypotheses 1, 2 and 3.

To test hypothesis 1, I build an OD dataset containing normalised trip counts within and between districts. These counts are used to build a mobility network and a consequent adjacency matrix. The construction of a network allows to quantitatively analyse the in and out-weights of the nodes. Lastly, I build complementary assortativity matrices based on trips between income and Gini index deciles to study whether there are any mobility patterns across socio-economic groups.

To test hypothesis 2, I use the previously created trip counts dataset. The districts of origin and destination are classified into four quantiles based on the median income and also on the Gini index values, and I compare the differences among the furthest quantiles. This allows to further analyse the total number of trips between and within districts based the income background of the individuals.

To test hypothesis 3, I analyse the average distance of the trips based on the four aforementioned quantiles by variable. I retrieve the average trip distance by OD pair, and compare these based on the top and lowest quantile.

4.1 Measuring Inequality

Spatial autocorrelation is used to describe the extent to which a variable is correlated with itself through space. This concept is closely related to Tobler's First Law of Geography, which states that 'everything is related to everything else, but near things are more related than distant

things' (Tobler, 1970). To understand the state of income distribution and inequality in Madrid in 2021, I use the traditional spatial autocorrelation Global Moran's I on each of the income variables. I also recreate and compare the results against a null model, where the socioeconomic variable of interest is swapped randomly across districts.

Global Moran's I

Global Moran's I (Moran, 1950) is a spatial autocorrelation measure used to evaluate whether a specific pattern is clustered, dispersed, or random across spatial units. For this measure, the null hypothesis states that the variable being analysed is randomly distributed among the features in the area of study. To test the significance of the results, the Global Moran's I is often calculated along with a z-score and a p-value. To obtain these, I perform Monte Carlo randomisation, a typical method to simulate spatial randomness by reassigning the observed median income per consumption unit values among districts and calculating a randomised distribution for the Moran's I. This randomised distribution aids on testing the significance of the results.

A higher Global Moran's I indicates higher spatial autocorrelation, whereas a lower Global Moran's I indicates lower spatial autocorrelation. Z-score values below -1.96 indicate negative spatial correlation, whereas z-score values over 1.96 indicate positive spatial autocorrelation. I calculate the Global Moran's I estimators for each variable of interest, as well as the corresponding p-values and z-scores. Once these p-values are obtained, I reject the null hypothesis when the p-value is lower than 0.05 ($\alpha < 0.05$). In the context of this estimator, it is needed to define what a neighbour is. To do so, I apply row-standardised Queen contiguity weights.

Local Moran's I

It is important to note that the Global Moran's statistic describes a complete spatial pattern, which could indicate clustering, but it does not capture any nuances in the location of the clusters. To overcome this limitation, I calculate one of the many Local Indicators of Spatial Autocorrelation (LISA), the Local Moran's I values, with their respective local p-values and local z-scores, one for each district of interest for the median income per consumption unit. The mechanisms behind the Local Moran's I are similar to those used for calculating the global Moran's I, but they are applied to each of the twenty-one observations, resulting in twenty-one statistics, instead of just one. To obtain these statistics, I previously needed to calculate spatial weights based on district adjacency. To keep the results consistent with the global estimators, I again apply the Queen contiguity weights.

Moran's Plot

The Moran's Plot (Anselin, 1996) is a graphical device that aids on visualising the strength of spatial autocorrelation of a spatial variable. The variable is displayed against its spatial lag, which are the values of the neighbouring districts, in this case calculated using again Queen contiguity weights. The values are standardized, meaning that they are centred on the mean in each axis, and the units represent standard deviations from the mean. The Moran's Plot is often displayed with a line of best fit for the dots in it. The slope of this line is interestingly the value of the Global Moran's I obtained for the variable of choice.

To ease the interpretation of the obtained Moran's values, I build the Moran's Plot for the median income per consumption unit.

4.2 Construction of OD Trip Counts and Mobility Network

In the previous section, I address the state of income distribution and segregation in 2021 in the different districts of Madrid. In the following sections, I describe the methods used to study the interaction between individual mobility patterns and income within the districts of interest for

the month of March of 2022. These methods mainly help testing hypothesis 1. To do so (1) I create an OD trip counts dataset, followed by (2) a mobility network and an inherent adjacency matrix to visually and quantitatively delve on potential patterns of mobility based on nodes' weights and trips. Lastly, (3) I build several assortativity matrices following closely the study of Duran-Sala et al..

OD Trip Count Dataset

To further inspect mobility interactions between districts, I build a dataset that contains the normalised trip counts T_{ab} for March of 2022 between and within districts. I normalise the total trip count by the total number of trips starting from the district of origin. The reason behind this normalisation is that the number of trips originating from each district differs widely. For example, Puente de Vallecas or Carabanchel are the districts where most trips start from in the original dataset, but Barajas shows the lowest amount of trips. These disparities are closely related to the population of the districts. See Figure 20 in Appendix A for a visualisation of these disparities.

Equation 1 formalises the normalisation of trips, where T_{ab} is the normalised trip count from district a to district b, $\delta_k(a,b)$ is an indicator function that equals 1 if the k-th trip originates in district a and ends in district b, and n is the total number of trips in the dataset.

Tables 5 and 6 how a sample of the final OD trip counts dataset and their statistics, respectively. These counts comprise the base to construct the weights of the edges in the mobility network. Figure 7 shows the distribution of the trip counts after normalisation.

$$T_{ab} = \frac{\sum_{k=1}^{n} \delta_k(a,b)}{\sum_b \sum_{k=1}^{n} \delta_k(a,b)} \tag{1}$$

Trip Count	Origin	Destination	Total Trips from Origin	Normalised Trip Count
24,324,971,550,694	Centro	Centro	136,712,522,521,830	0.18
18,484,663,675,589	Centro	Arganzuela	136,712,522,521,830	0.14
...	...	...	...	...
14,590,854,025,006	Vicálvaro	Moratalaz	119,163,153,410,873	0.12
...	...	...	...	...
3,455,182	Barajas	Villaverde	78,342,805,567,739	0.00000004410337
31,152,927,901,561	Barajas	Barajas	78,342,805,567,739	0.40

Table 5: Sample of the trip counts dataset. The normalised trip count is obtained by dividing the total trip count between two districts by the total number of trips from the district of origin.

Statistic	Value
Mean	0.0476
Standard Deviation	0.0675
Min	0.0000000366
25%	0.0038
50%	0.0129
75%	0.0740
Max	0.3976

Table 6: Statistics of the normalised trips between OD pairs.

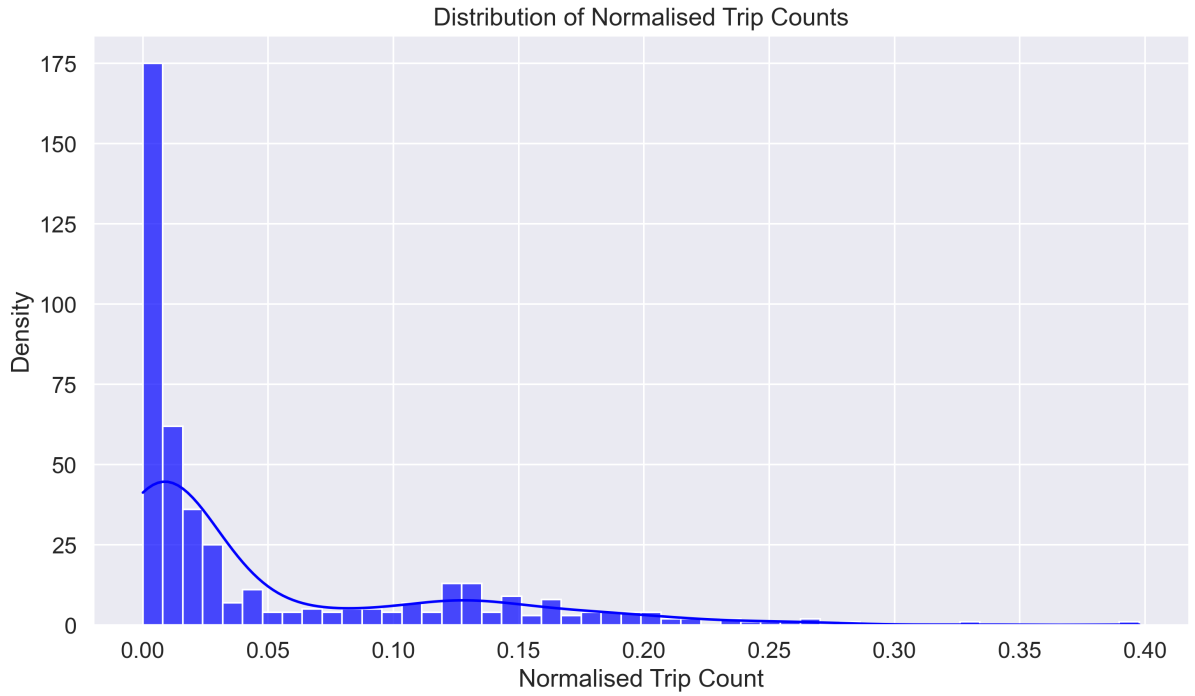

Figure 7: Distribution of the normalised trip counts between OD pairs. The minimum trip count is 0.0000000366, whereas the maximum trip count is 0.398. The mean trip count is 0.0476.

Mobility Network

To further quantify the interaction between mobility and income, I construct a mobility network of the city for the trips of choice. Real networks often 'display a large heterogeneity in the capacity and intensity of the connections' (Barrat et al., 2003). To capture this heterogeneity, weights can be assigned to the edges of a network to represent the closeness (or, in contrast, the farness) of connections between the nodes. Closeness can have many definitions based on the purpose of the study. In this case, it represents a higher number of trips from one district to another. The higher number of trips from one district to another, the closer these districts are. As a result, I build a weighted, directed network $G = (V, E, W)$ where:

- V is the set of nodes, representing the districts, with $n = |V| = 21$.
- E is the set of edges, representing trips between districts, with $m = |E| = 441$. The reason for obtaining 441 edges is that there is always at least one trip from each district to another in the dataset, so all nodes connect to each other.
- W is the set of possible weights.
- Each edge in E is an ordered OD pair (a, b), indicating a directed trip from district a to district b.
- Each edge (a, b) is assigned a weight w_{ab}, representing the normalised number of trips from district a (origin) to district b (destination). These weights are retrieved from the OD trip count dataset described earlier. To ease the analysis and visualisation of the network, the weights are scaled to fall within the range $[0, 1]$.
- The graph is, as a result, weighted and directed, with a weight function $\omega : E \to R$, where $\omega(a, b) = w_{ab}$ is the weight of the edge (a, b).

An adjacency matrix is 'the basic representation of a graph as a matrix. Each row/column corresponds to a node' (Coscia, 2021). A weighted, directed network 'can be represented mathematically by an adjacency matrix with entries that are not simply zero or one, but are equal instead to the weights of the edges' (Newman, 2004):

$$A_{ab} = \text{weight of connection from } a \text{ to } b. \tag{2}$$

I retrieve and display the adjacency matrix A of the mobility network as a heat map. This provides a visual understanding of the weights of the OD pairs of trips between districts. These weights are key to study interconnectivity between districts.

The construction of the network mainly aids on capturing the in-weight and out-weight of the districts. However, when displayed visually, networks can help understand patterns better. Refer to Figure 21 in Appendix A for a detailed look at a pair of nodes and their weights. Self-loops have been kept throughout the analysis and visualisation as they represent a high share of trips.

4.3 Assortativity Matrix

The introduction of matrices has been widely used in the literature to study segregation structures and mobility patterns. For example, Kazmina et al. introduce what they call the *mixing matrix*, which 'represents the connectivity between different subgroups of the population as defined by their attributes'. These matrices often provide a visual representation of the segregation patterns within the population, and can also serve as an input for further levels of quantification, like the calculation of assortativity coefficients or correlations. In this case, they also group districts based on deciles, displaying agglomerated patterns that the adjacency matrix cannot provide.

Following Bokányi et al. approach, later used by Duran-Sala et al. in the context of a mobility network and income segregation, I calculate two assortativity matrices (X) between ten different income deciles (D) for all of the trips. The income deciles are calculated based on the two relevant variables: the median income per consumption unit and the Gini index. These assortativity matrices encode the probability C_{ij} of travellers starting a trip from districts of a given decile $D = i$ to travel to districts with income decile $D = j$, where u are the number of trips originating from districts of decile i and travelling to districts of j, as shown in Equation 3. The entries of these matrices are later normalised to fall between 0 and 1 ($\tilde{X}$).

$$C_{ij} = \frac{\sum_{\{u|D_{u,\text{home}=i},D_{u,\text{destination}=j}\}} 1}{\sum_{\{u|D_{u,\text{home}=i}\}} 1} \tag{3}$$

Equation retrieved from Duran-Sala et al.

Afterwards, I calculate the *assortativity index* p with the Pearson correlation coefficient of the normalised matrix entries $\tilde{X}$, following Equation 4. Assortativity can be described as the tendency of nodes to connect or 'attach' to nodes with similar properties in a graph, also referred to as homophily. Dissortativity describes exactly the opposite tendency. In other words, a perfect diagonal matrix where complete homophily is observed (no trips are performed across deciles) will have an *assortativity index* p of 1, whereas a matrix representing indifferent behaviour when it comes to the destination of the trips will show an *assortativity index* p close to 0. A p-value is calculated along each *assortativity index* p.

$$\rho = \frac{\sum_{i,j} ij\tilde{X}_{ij} - \sum_{i,j} i\tilde{X}_{ij} \sum_{i,j} j\tilde{X}_{ij}}{\sqrt{\sum_{i,j} i^2\tilde{X}_{ij} - \left(\sum_{i,j} i\tilde{X}_{ij}\right)^2}\sqrt{\sum_{i,j} j^2\tilde{X}_{ij} - \left(\sum_{i,j} j\tilde{X}_{ij}\right)^2}} \tag{4}$$

Equation retrieved from Duran-Sala et al.

4.4 Relationship Between In-Weight, Out-Weight and Income

I study the pull-push patterns by district and their relationship to income and inequality by calculating the total in-weight and out-weight of each node in the graph. To do so, I set the weighted in-degree to be the sum of the weights of all edges directed into a node, and the weighted out-degree to be the sum of the weights of all edges directed out of a node. To represent the out-weights with more granularity, I exclude self-loops, as these do not represent trips to another district, but trips within the district. I do include these self-loops when performing the in-weight calculation.

To test whether the variables of choice are related to the in-weight and out-weight of the nodes in the mobility network, I run the four following regression models:

Model 1: In-weights and the median income per consumption unit.

$$y = \beta_0 + \beta_{in} x + \epsilon \tag{5}$$

Where:

- y: Total in-weight of districts.
- x: Median income per consumption unit, scaled by dividing by 1,000.
- β_0: Intercept, the expected total in-weight when $x = 0$.
- β_{in}: Coefficient for x, measuring the change in y for a 1,000-unit increase in x.
- ϵ: Error term.

Hypotheses:

- **Null Hypothesis** (H_0): The total in-weight of districts is not related to the median income per consumption unit.

$$H_0 : \beta_{in} = 0 \tag{6}$$

- **Alternative Hypothesis** (H_1): The total in-weight of districts is positively related to the median income per consumption unit. Districts with higher median incomes tend to have higher total in-weights. A significant positive value for β_{in} would provide evidence supporting the alternative hypothesis.

$$H_1 : \beta_{in} > 0 \tag{7}$$

Model 2: Out-weights and the median income per consumption unit.

$$y = \beta_0 + \beta_{out} x + \epsilon \tag{8}$$

Where:

- y: Total out-weight of districts.
- x: Median income per consumption unit, scaled by dividing by 1,000.
- β_0: Intercept, the expected total out-weight when $x = 0$.
- β_{out}: Coefficient for x, measuring the change in y for a 1,000-unit increase in x.

- ϵ: Error term.

Hypotheses:

- **Null Hypothesis** (H_0): The total out-weight of districts is not related to the median income per consumption unit.

$$H_0 : \beta_{out} = 0 \tag{9}$$

- **Alternative Hypothesis** (H_1): The total out-weight of districts is negatively related to the median income per consumption unit. Districts with higher median incomes tend to have lower total out-weights. A significant negative value for β_{out} would provide evidence supporting the alternative hypothesis.

$$H_1 : \beta_{out} < 0 \tag{10}$$

Model 3: In-weights and the Gini index.

$$y = \beta_0 + \gamma_{in} x + \epsilon \tag{11}$$

Where:

- y: Total in-weight of districts.
- x: Gini index.
- β_0: Intercept, the expected total in-weight when $x = 0$.
- γ_{in}: Coefficient for x, measuring the change in y for a one-unit increase in x.
- ϵ: Error term.

Hypotheses:

- **Null Hypothesis** (H_0): The total in-weight of districts is not related to the Gini index.

$$H_0 : \gamma_{in} = 0 \tag{12}$$

- **Alternative Hypothesis** (H_1): The total in-weight of districts is negatively related to the Gini index. Districts with higher Gini index tend to have lower total in-weights. A significant negative value for γ_{in} would provide evidence supporting the alternative hypothesis.

$$H_1 : \gamma_{in} < 0 \tag{13}$$

Model 4: Out-weights and the Gini index.

$$y = \beta_0 + \gamma_{out} x + \epsilon \tag{14}$$

Where:

- y: Total out-weight of districts.
- x: Gini index.

- β_0: Intercept, the expected total out-weight when $x = 0$.
- γ_{out}: Coefficient for x, measuring the change in y for a one-unit increase in x.
- ϵ: Error term.

Hypotheses:

- **Null Hypothesis** (H_0): The total out-weight of districts is not related to the Gini index.

$$H_0 : \gamma_{out} = 0 \tag{15}$$

- **Alternative Hypothesis** (H_1): The total out-weight of districts is positively related to the Gini index. Districts with higher Gini index tend to have higher total out-weights. A significant positive value for γ_{out} would provide evidence supporting the alternative hypothesis.

$$H_1 : \gamma_{out} > 0 \tag{16}$$

Using this approach, β_{in}, β_{out}, γ_{in} and γ_{out} represent the slope of the relationship between the income variables and total in or out-weight of nodes in a linear regression model where the parameters are estimated by using Ordinary Least Squares (OLS).

4.5 Analysis of the Number of Trips Based on Quantiles

The graph and adjacency matrix built in the previous sections provide a visual understanding of the distribution of trips throughout the districts. Studying the relationship between the node weights and income helps understanding the pull-push mobility patterns. On the other hand, to get a quantitative understanding of the distribution of trips in depth, I further inspect the weights of the adjacency matrix. Hence, to answer hypothesis 2, I calculate four income quantiles based on the median income per consumption unit and the Gini index of the districts. These quantiles are used to classify individuals into low or high income and segregation districts of residence. Table 7 illustrates the breakdown of quantiles by income variable.

Quantiles	Median Income Districts	Gini Index Districts
q = 0 (Low)	Latina, Carabanchel Usera, Puente de Vallecas Villaverde, Villa de Vallecas	Arganzuela, Latina Puente de Vallecas, Moratalaz Villa de Vallecas, Vicálvaro
q = 3 (High)	Retiro, Salamanca Chamartín, Chamberí Moncloa-Aravaca	Centro, Salamanca, Chamartín Chamberí, Moncloa-Aravaca

Table 7: Classification of districts into quantiles based on median income per consumption unit and Gini index. Districts in quantiles 1 and 2 are excluded from this analysis.

To quantify the extent to which residents from low-income districts tend to travel to high-income districts (and vice versa) relative to all trips, I retrieve the normalised number of trips originating from low quantiles ($q = 0$) to high quantiles ($q = 3$) and vice versa. I then divide this sum by the total number of trips. Equation 17 is applied to each variable of interest and for every combination of quantiles (low to low, low to high, high to high, high to low):

$$p^v_{x \to y} = \frac{t^v_{x \to y}}{t^v_{\text{total}}} \times 100 \tag{17}$$

- $v \in \{\text{income}, \text{gini}\}$ refers to the variable of choice.
- $x, y \in \{l, h\}$ represent the quantile of origin and destination (l: low, h: high).
- $t^v_{x \to y}$ are the total normalised trips between quantiles x and y for variable v.
- t^v_{total} represent the total normalised trips for variable v, defined as:

$$t^v_{\text{total}} = \sum_{x,y} t^v_{x \to y} \tag{18}$$

4.6 Analysis of the Distance of Trips Based on Quantiles

To answer hypothesis 3, I am particularly interested in quantifying the extent to which residents travel different distances based on the median income per consumption unit or the Gini index values of their district of residence.

To quantify these potential differences, (1) I separately classify districts into the aforementioned quantiles: lowest median income quantile ($q = 0$), lowest Gini index quantile ($q = 0$) and highest median income ($q = 3$) and highest Gini index quantile ($q = 3$). Then, (2) I classify the trips based on their quantile of origin, both for the median income and the Gini index. Lastly, (3) I calculate the average distance per trip in km for each OD pair for trips. This is done by dividing the total distance of the trips in km by the total number of trips for each OD pair. The total distance of the trips and the total number of trips are two variables available in the original dataset. This process is shown in Equation 19, where a indicates the district of origin, and b is the destination district. (4) I compare the resulting average distances based on the quantile of origin to quantify whether there are substantial differences based on the median income or Gini index values.

The average distance per trip for a given OD pair is calculated as:

$$D_{ab} = \frac{\text{Total km}_{ab}}{\text{Total Number of Trips}_{ab}} \tag{19}$$

5 Results

In this research, I delved into the state of income segregation in conjunction with the study of mobility patterns based on income in Madrid during March of 2022.

5.1 Initial Analysis: Measuring Inequality and Income Variable Selection

First, I analysed the state of income segregation by comparing different global and local Moran's I values for several income variables to test whether they were viable to study income segregation.

Table 8 shows the Global Moran's coefficients obtained for each income variable in the year 2021. All p-values show statistical significance ($\alpha < 0.05$), meaning that the null hypothesis stating that income is dispersed randomly across districts is rejected at least with 95% confidence. All of the obtained z-scores are higher than 1.96, indicating positive spatial autocorrelation. Positive Moran's I values indicate a tendency towards clustering, while negative Moran's I values indicate a tendency towards dispersion. In this case, all income variables show a tendency towards clustering, with the strongest tendency for the average net income per person. Figure 8 displays a visual comparison of the strength of the resulting Global Moran's coefficients among all the income variables, and their significance.

Based on this initial analysis, and despite having the lowest clustering coefficient, the median income per consumption unit is the selected variable to study the interaction between mobility

due to its insensibility to outliers. This variable is further complemented throughout the study with the Gini index per district.

Variable	Global Moran's I	P-value	Z-Score
Average income per consumption unit	0.417	0.001***	3.814
Median income per consumption unit	0.315	0.006**	2.755
Average gross income per household	0.370	0.001***	3.441
Average gross income per person	0.453	0.003**	3.956
Average net income per household	0.358	0.004**	3.299
Average net income per person	0.459	0.001***	4.018

Table 8: Global Moran's I, p-values, and z-scores for income variables.

Notes: The results indicate statistical significance for models where $p < 0.05$. Significance levels: * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$.

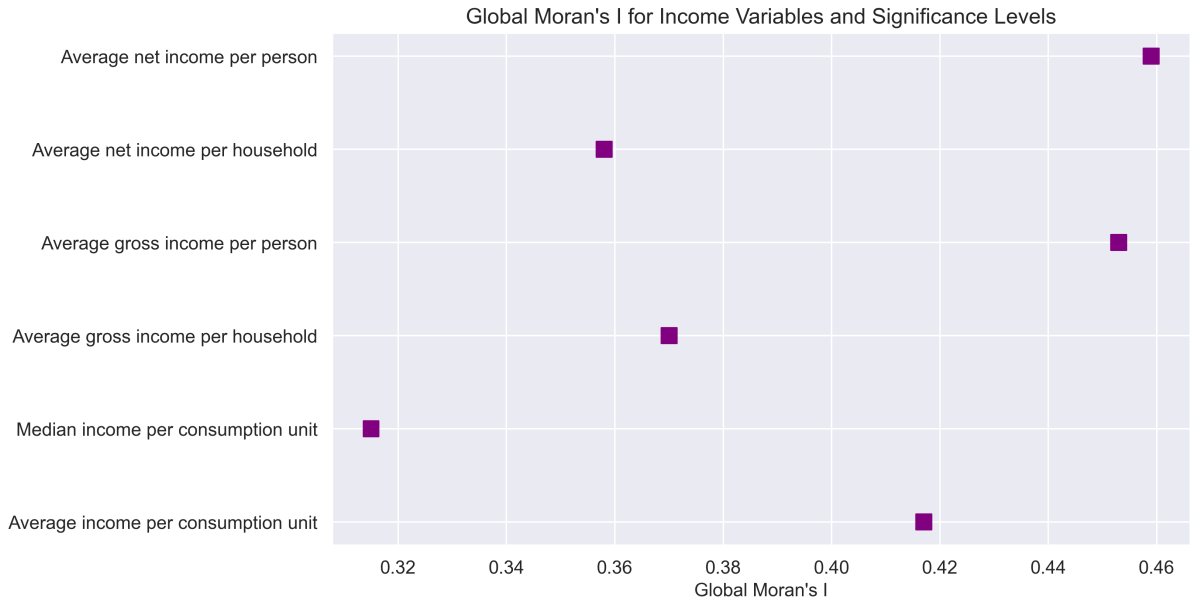

Figure 8: Global Moran's coefficients for each variable of interest and their significance. Purple dots indicate significant coefficients ($p < 0.05$) based on the Monte Carlo test after running 999 simulations.

Figure 9 shows the Global Moran's I simulated reference distribution on the left, obtained after running 999 random simulations, and the Moran's Plot for the median income per consumption unit on the right.

The reference distribution helps compare the obtained statistic (marked with a vertical red line) to the randomised results (the mean is marked with a blue vertical line). The Moran's scatter plot displays the relationship between the median income value of each district and the median income values of the neighbouring districts. The upper-right and lower-left quadrant indicate positive spatial autocorrelation, whereas the lower-right and upper-left indicate negative spatial autocorrelation. Fourteen out of the twenty-one districts are located in the upper-right and lower-left quadrants, indicating that they tend to follow some clustering pattern. This essentially means that low median income districts tend to be placed close to other low median income districts, and high median income districts tend to be located around other high median income districts.

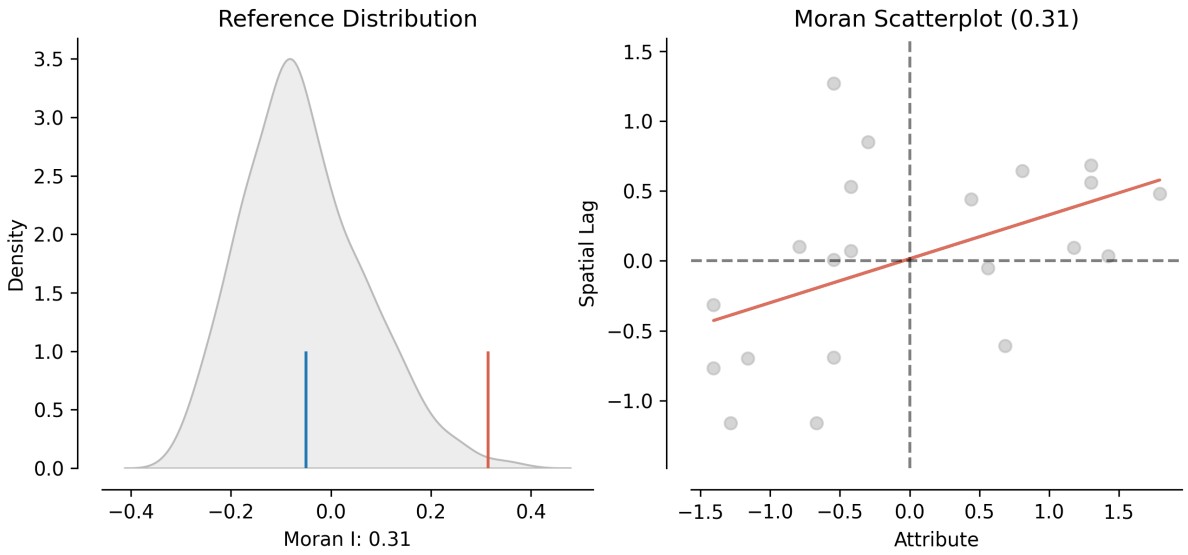

Figure 9: Reference distribution and Moran's Plot for the median income per consumption unit. In the reference distribution, the red vertical line marks the observed Moran's I value in relation to the simulated sample, whereas the blue vertical line indicates the mean of the reference distribution. In the Moran's Plot, each dot represents a district, and the red line shows the line of best fit, whose slope is the value of the Global Moran's I (0.31).

The obtained Local Moran's I estimators for the median income are displayed in Figure 10. For each district, a Local Moran's coefficient, a p-value and a z-score were obtained. Based on the individual p-values, 48% of the districts have degrees of local spatial association strong enough to reject the idea of pure chance, meaning that their assigned p-values are below 0.05 ($\alpha < 0.05$). In other words, following this analysis, around 48% of the districts are considered to significantly form a spatial cluster in terms of the median income per consumption unit, meaning that the median income is fairly clustered across districts. These results are visually reflected in Figure 11, where the districts are coloured based on their assigned income cluster, High-High (HH), Low-High (LH), Low-Low (LL) and Not Significant (ns).

Based on this analysis, I conclude that there is spatial autocorrelation in terms of the median income per consumption unit in the city of Madrid in 2021, with significant clustering patterns in 48% of the districts.

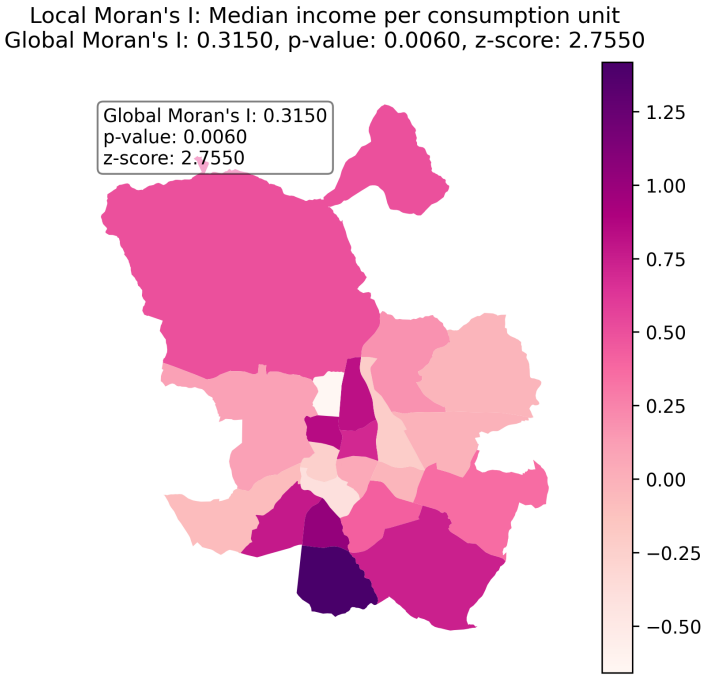

Figure 10: Local Moran's choropleth map for the median income per consumption unit.

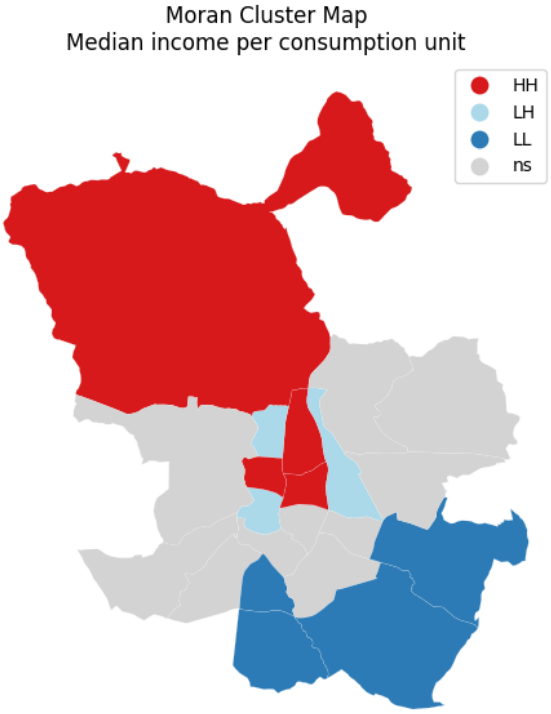

Figure 11: Clusters obtained after calculating the Local Moran's estimates. The obtained income clusters reflect a north-south divide.

5.2 Hypothesis 1: Pull-Push Patterns In Madrid

Hypothesis 1 refers to the pull-push patterns introduced by Ravenstein and whether they are replicated, in a smaller scale, in Madrid. To understand mixing patterns in the city, I constructed a mobility network containing normalised trip counts and a consequent adjacency matrix. I built four assortativity matrices to quantify the interaction between mobility across districts and individual-level socio-economic status (SES). Lastly, I built four simple regression models to explore the in-weights and out-weights of each district in detail.

Adjacency Matrix

The adjacency matrix (see Figure 12) provides a more nuanced display of the weights of OD pairs that a network visualisation cannot provide. Rows represents the origin of the trips, while columns display the destination of these. By observing a specific row of the matrix, one can develop an understanding on where most trips are directed to for each origin. When observing a column, one can grasp which districts travel the most to that specific district.

Based on the high weights on the diagonal, it reveals a tendency of travellers to stay within their own district, or those nearby. Barajas is the district with the highest number of self-loops, or intra-district trips, followed by Villaverde. This essentially means that individuals travel short distances overall, preferring to stay in their district of residence, or those nearby. This matrix reveals some patterns, but it is the assortativity matrices which contain more income and inequality information.

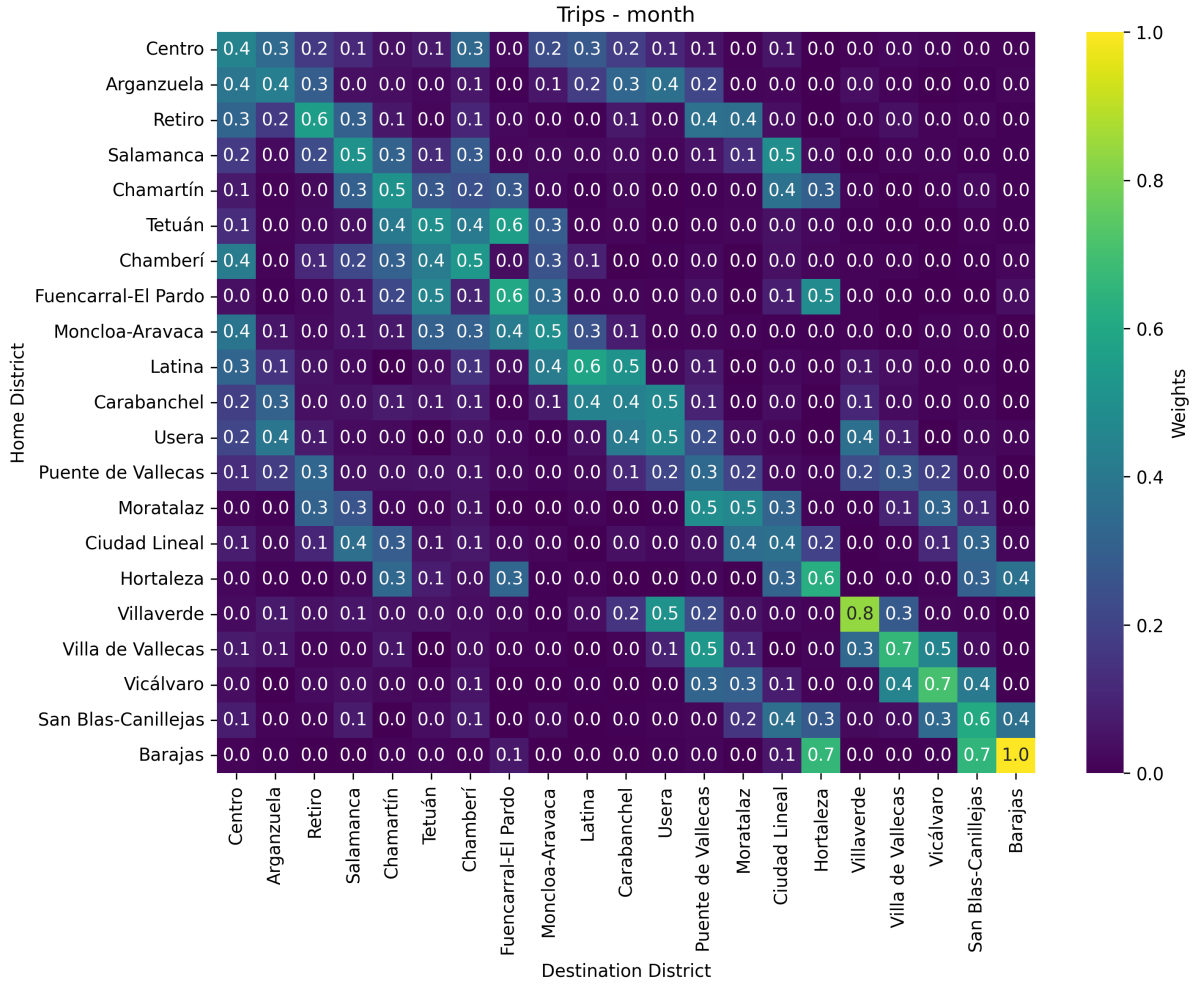

Figure 12: Adjacency matrix displayed as a heat map of normalised trips during March of 2022. The weights have been normalised to fall between 0 and 1, and reduced to one decimal point only to ease the visualisation.

Assortativity Matrices

To study mixing patterns based on income and segregation, I built assortativity matrices. These helped quantifying to what extent residents from lower income districts tend to travel to higher income districts in comparison to the rest of residents. Figures 13 and 14 show the unnormalised resulting matrices.

Assortativity is also known as homophily, hence lower assortativity indices indicate lower homophily, and vice versa. The *assortativity index* p is significant only when taking into account the Gini index (see Figure 15). With a resulting $p = 0.27$, the results indicate certain mixing between deciles, and is thus far from perfect assortativity (or homophily), indicated by $p = 1$, but closer to an indifferent behaviour, in which $p = 0$. In simple terms, the results suggest that residents from Madrid lack any specific preference for travelling to other districts, despite their segregation background. There is a highly weighted square comprised by deciles 7 to 9, indicating relatively higher assortativity for highly segregated deciles. This means that residents from the top three highest segregated districts tend to make a relatively high amount of trips within themselves. Based on the relatively strong weights on the diagonal entries, individuals tend to travel slightly more to similar Gini index deciles.

Even a lower mixing pattern is observed when taking into account the median income per consumption unit (see Figure 16), but these results are not significant based on the high p-value of the resulting index. Despite it all, the lowest income decile (decile 0) shows the highest number of intra-district trips, followed by decile 7. This essentially means that residents from the lowest median income districts show the highest homophily, or preference to travel within districts with similar median income values. Based on the results, individuals also tend to travel slightly more to similar median income deciles. Yet, further research is needed to investigate the significance of these particular results.

Based on the assortativity matrices and the adjacency matrix, I conclude that there is a moderate tendency of individuals to stay within their own Gini index deciles, performing a relatively high number of intra-district trips, but still showing certain inter-quantile interaction. A lower assortativity pattern was found in the case of median income deciles, indicating higher mixing, but the results were not found to be statistically significant.

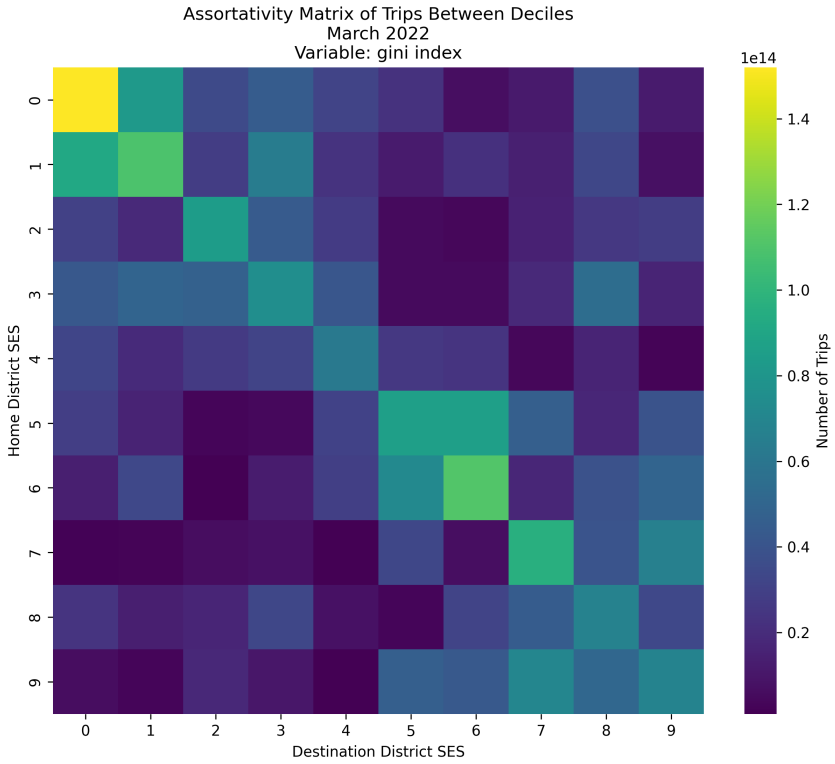

Figure 13: Assortativity matrix between Gini index deciles. Higher district SES values indicate a higher segregation index.

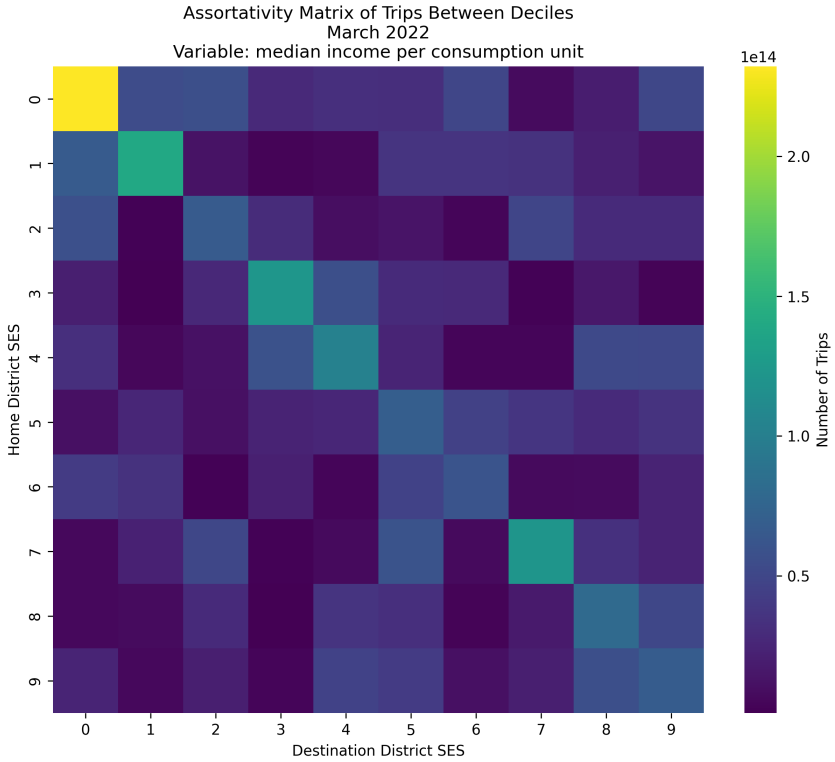

Figure 14: Assortativity matrix between median income per consumption unit deciles. Higher district SES values indicate higher median income per consumption unit.

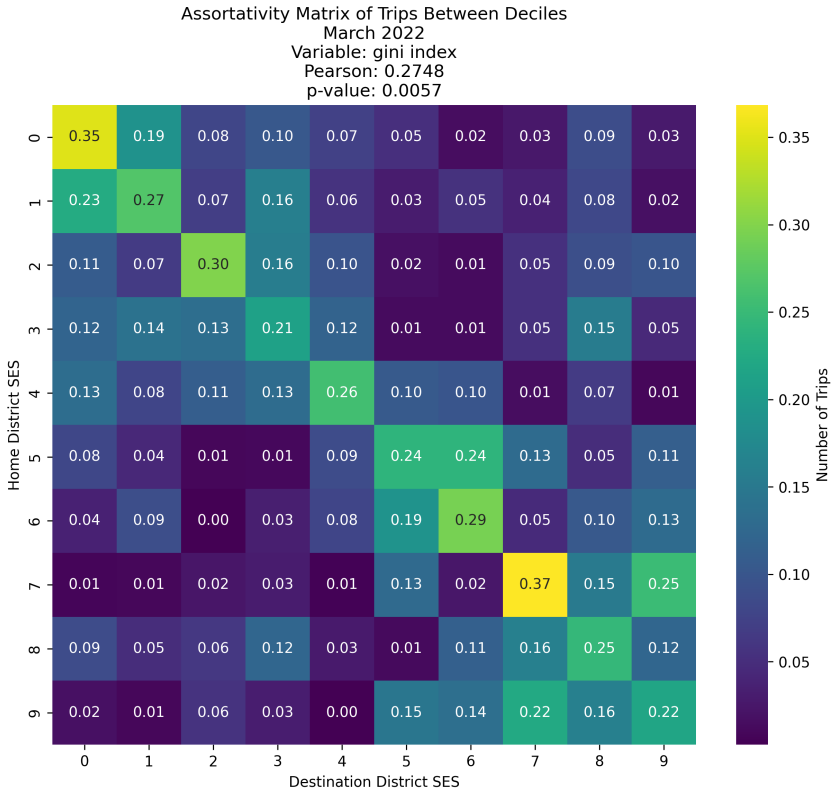

Figure 15: Normalised assortativity matrix between Gini index deciles, where $p = 0.27$. Higher district SES values indicate a higher segregation index.

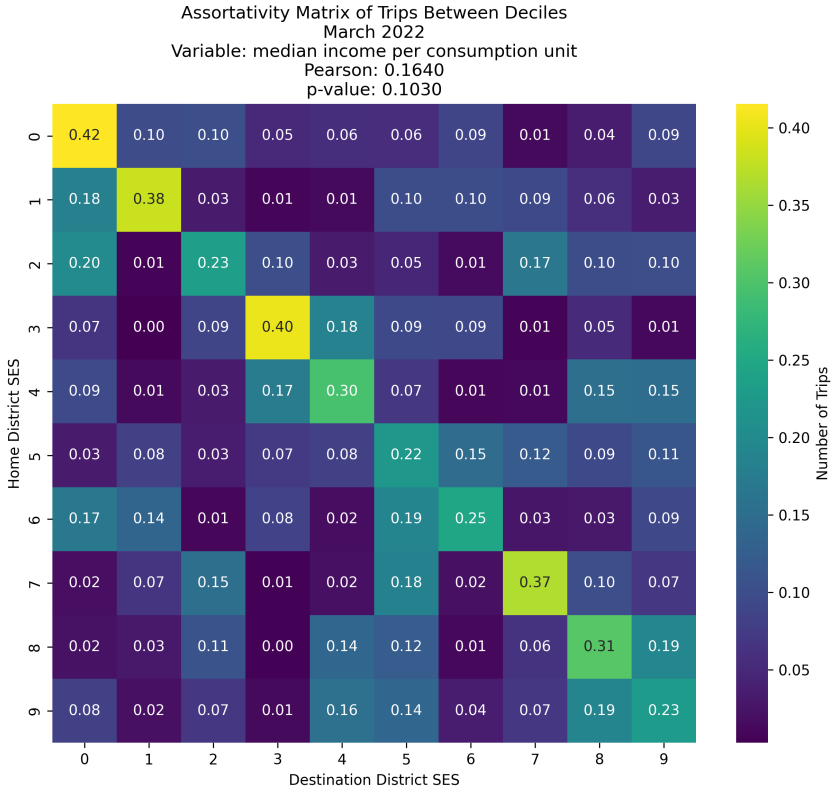

Figure 16: Normalised assortativity matrix between median income per consumption unit deciles, where the obtained $p = 0.16$ is not statistically significant. Higher district SES values indicate higher median income per consumption unit.

In-Weights and Out-Weights

Table 9 contains a sum of all the in-weights and out-weights for each node in the network. Figures 17 shows the values on the Madrid map. These weights per node help understand which districts push and pull more travellers in the city for the whole month of study.

The districts with the highest in-weight are Centro, Puente de Vallecas and Chamberí, indicating that they receive a relatively high number of trips. On the other spectrum stand the districts of Barajas, Latina and Villa de Vallecas, with the lowest in-weight, indicating low travel preference. Puente de Vallecas, Ciudad Lineal and Carabanchel display the highest out-weight values, indicating a relatively high number of trips outside of these districts, whereas the lowest out-weight values are located in Barajas, Villaverde and Vicálvaro, showing a relatively low number of trips originating from these districts to other areas in the city.

Table 10 displays the coefficients and significance of the four regression models. Based on the p-values, only Model 3 shows statistically significant results, where $p < 0.05$. The model reflects a statistically significant positive effect of the Gini index on the in-weight of the districts, meaning that districts showing high Gini index values pull residents in. This result does not align with the previous hypothesis that districts showing higher Gini index values push citizens out of them.

The rest of the models' coefficients did not reach statistical significance, and the resulting coefficients are small. Despite this lack of significance, in every case, results align with the alternative hypothesis. The results suggest a small positive relationship between the median income on the in-weight of the districts, and a potential negative relationship between the median income and the out-weight of the districts. Model 4 shows a modest potential positive

relationship between the out-weight and the Gini index, though further research with a larger sample is needed to obtain robustest results.

Refer to Figures 22, 23, 24, 25 in Appendix A for a visualisation of the results.

District	Total In-weight	Total Out-weight
Centro	3.54	2.07
Arganzuela	2.26	2.09
Retiro	2.48	1.96
Salamanca	2.61	2.01
Chamartín	2.90	2.00
Tetuán	2.56	2.02
Chamberí	3.06	1.98
Fuencarral-El Pardo	2.49	1.92
Moncloa-Aravaca	2.29	2.02
Latina	2.01	1.93
Carabanchel	2.41	2.08
Usera	2.35	2.05
Puente de Vallecas	3.19	2.18
Moratalaz	2.23	2.05
Ciudad Lineal	2.77	2.15
Hortaleza	2.70	1.89
Villaverde	2.09	1.68
Villa de Vallecas	2.07	1.85
Vicálvaro	2.24	1.81
San Blas-Canillejas	2.65	1.91
Barajas	1.89	1.51

Table 9: Total in-weight and out-weight by district for the home-origin trips of March of 2022. Self-loops have been excluded for the calculation of out-weights.

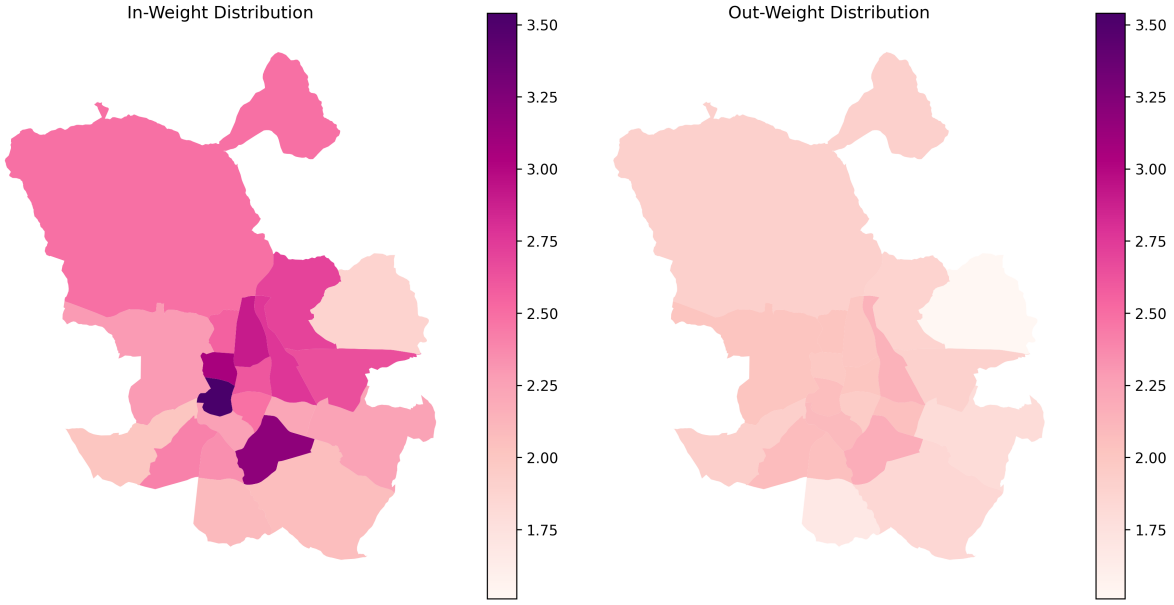

Figure 17: Distribution of the obtained in-weights per district. It can be observed that the in-weight values are higher overall than the out-weight distribution. This is in part because self-loops add up to a big amount of trips.

Model	Dependent Variable	F-statistic	R-Squared	Coefficients	P-value
1	In-weight	0.3551	0.018	const = 2.2992	0.000
				Median income (β_{in}) = 0.0096	0.596
2	Out-weight	0.0659	0.003	const = 1.9950	0.000
				Median income (β_{out}) = -0.0016	0.800
3	In-weight	8.484	0.309	const = -4.7103	0.823
				Gini Index (γ_{in}) = 0.2534	0.009***
4	Out-weight	1.156	0.057	const = 1.5829	0.000
				Gini Index (γ_{out}) = 0.0109	0.296

Table 10: OLS regression results.

Notes: The results indicate statistical significance for models where $p < 0.05$. Significance levels: * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$. For the Median income variable, the coefficient represents the effect per unit increase in income, while for the Gini Index, it shows the effect of each additional point.

Based on these findings, I conclude that the total in-weight of districts is positively related to the Gini index. I also conclude that the previous analysis does not provide sufficient evidence to reject the null hypothesis that the total out-weight of the districts is not related to the median income per consumption unit and the Gini index, and neither to reject the null hypothesis that the total in-weight of districts is not related to the median income per consumption unit.

Summarising, the findings failed to substantiate hypothesis 1, indicating that no significant relationship between low median income, highly segregated districts and out-weight was found, nor a relationship between high median income, low segregated districts and in-weight. In contrast, a significant, positive relationship between highly segregated districts and the in-weight was found.

5.3 Hypothesis 2: Number of Trips

Hypothesis 2 is related to what extent residents from lower income districts tend to travel to higher income districts in terms of number of trips. To answer this hypothesis, the number of trips were analysed based on income and Gini index quantiles. Table 11 displays the different share of trips for the variables of interest. Residents from the lowest median income quantile travel 1.33% more to the highest income quantile than vice versa. Lowest income quantile residents performed 6.57% more intra-quantile trips (low to low) than those residents from the highest income quantile (high to high).

In terms of the Gini index, it is key to notice that there is a higher share of trips from low segregated districts to high segregated districts than the other way around. The results show that the lowest Gini index quantile residents perform up to 1.85% more trips to the highest Gini index quantile than vice versa. In addition, residents from the least segregated quantile perform up to 2.3% more intra-quantile districts (low to low) than those residents from the most segregated districts.

	Median Income Per Consumption Unit	**Gini Index**
$p_{l \to h}$	3.65%	4.70%
$p_{h \to l}$	2.32%	2.85%
$p_{l \to l}$	17.67%	14.90%
$p_{h \to h}$	11.10%	12.60%

Table 11: Percentage of trips from home to any other destination by income deciles.

I conclude that residents from lower median income per consumption unit districts perform 1.33% more trips to higher income quantiles than vice versa. I also conclude that residents from low segregated districts also tend to perform around 1.85% more trips to higher segregated districts than vice versa. The differences in intra-quantile share of trips are 6.57% and 2.3% for the median income and Gini index respectively.

Based on these findings, I suggest that low median income per consumption unit districts are indeed associated with a higher travel share to high median income per consumption unit districts, consistent with hypothesis 2. However, the expected trip share from highly segregated districts to low segregated districts was not supported. In fact, the opposite relationship was found, where there are a higher number of trips from low segregated districts to highly segregated ones.

5.4 Hypothesis 3: Distance of Trips

Hypothesis 3 supports a relationship between income and segregation and the average trip distance. Figures 18 and 19 represent the distribution of the average distance per trip, stratified by median income and Gini index quantiles. Each dot in the plot shows the average distance of a specific OD pair. The longest average distance is highlighted for each variable and group. As expected, the longest average trips are directed towards Barajas, one of the peripheral, industrial districts.

Residents from the lowest median income quantile travel, on average, 2.09 km per trip, whereas residents from the highest median income quantile perform slightly shorter trips, with an average of 1.68 km per trip. This is a difference in the average trip of 0.41 km. This difference is slightly higher when taking into account the Gini index. Residents from the most segregated districts perform trips of 1.70 km on average, whereas those living in the least segregated district perform trips of 2.33 km on average. This is a difference of 0.63 km.

Despite the almost insignificant differences in the results, it is key to notice that the distribution of distances reflects a heavy concentration around shorter trips (between 0.5 km and 2.5

km) for all residents, despite their background. In addition, low-income and low-Gini quantiles present a slightly longer tail, indicating that these subgroups are more likely to take longer trips.

I conclude that residents from the lowest median income quantile travel slightly longer distances on average than residents from the highest median income quantile. Residents from the highest Gini index quantile travel shorter distances on average than residents from the lowest Gini index quantile. I also conclude that the average distance of most trips falls under a 0.5 km - 2.5 km window, and that those performing the longest average trips belong to the low-income and low-Gini quantiles.

Hence, the observed results support that residents from low median income per consumption unit districts tend to perform, on average, longer distance trips than those living in high-income districts, supporting partly hypothesis 3. However, the results indicate a negative relationship between the Gini index and the average distance of the trips, indicating that those living in highly segregated districts perform, on average, shorter trips.

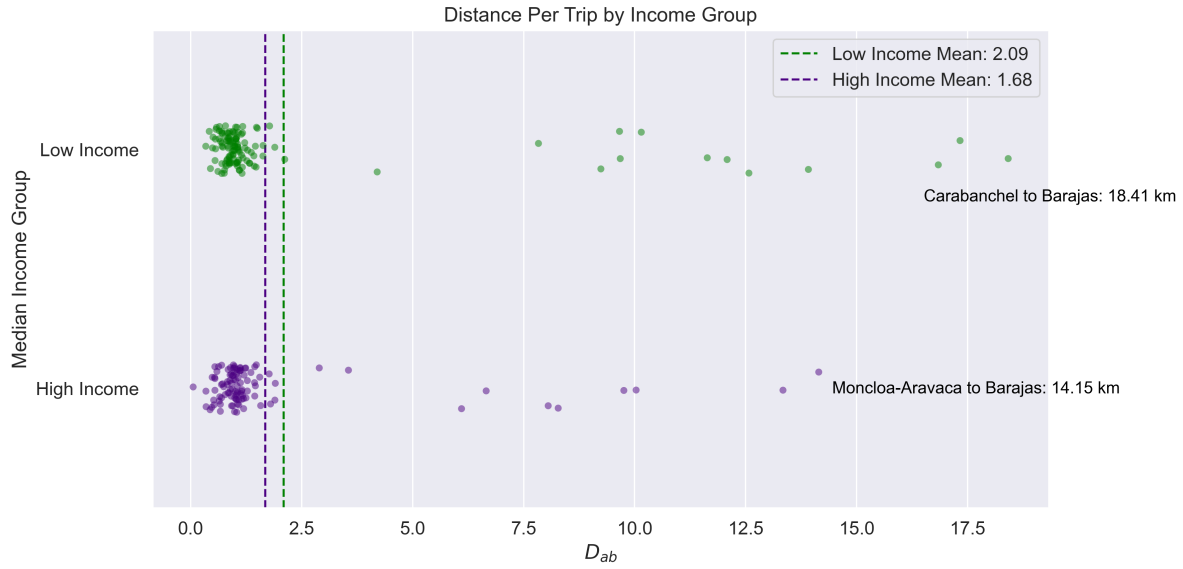

Figure 18: Distribution of the average distance of trips in km for low and high income deciles. The deciles were obtained using the median income per consumption unit. Each dot represents an OD pair, placed on the plot based on the average distance of the trips between the OD pair. The longest OD trips for each quantile group are labelled.

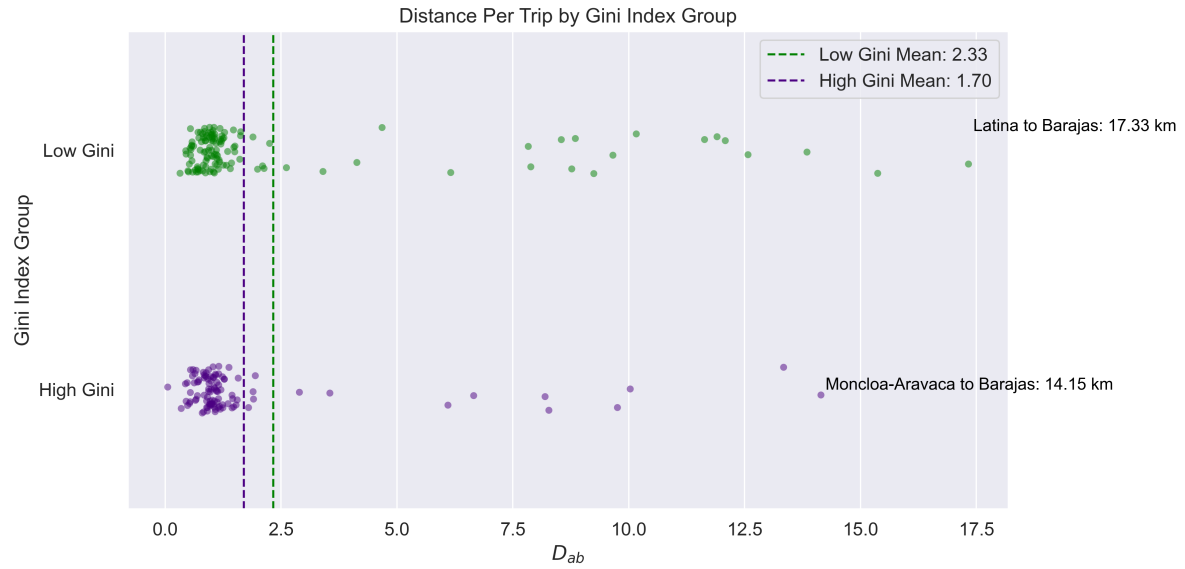

Figure 19: Distribution of the average distance of trips in km for low and high income deciles. The deciles were obtained using the Gini index. Each dot represents an OD pair, placed on the plot based on the average distance of the trips between the OD pair. The longest OD trips for each quantile group are labelled.

6 Discussion

RQ: To what extent do residents from economically segregated districts in Madrid move differently than those in other districts?

6.1 Exploratory Analysis: Is Madrid Really Segregated?

Existing literature shows that Madrid presents a social and economic clustering pattern in which higher income districts tend to be located around the North-West part of the city, whereas lower income districts tend to cluster around the South-East (González-García et al., 2024). I prove that Madrid indeed shows clustering patterns in the year 2021 based on six income indicators by calculating Global and Local Moran's I values. 48% of the districts have degrees of local spatial association strong enough to reject the idea of pure chance.

The districts of Fuencarral-El Pardo, Chamartín, Salamanca and Chamberí are part of the HH cluster, whereas Usera, Villaverde, Villa de Vallecas and Vicálvaro form the LL cluster. Tetuán, Centro and Ciudad Lineal form the LH cluster. These results correlate with the clear spatial manifestation divided by a *poverty line* previously observed by González-García et al..

In addition, proving statistically that there is spatial autocorrelation based on the median income per consumption unit made it more sensible to use this variable to study segregation for the rest of the study. The use of the median income was complemented with the Gini index to add an extra dimension to the study, as the Gini index is a great indicator of intra-district segregation. In this regard, the districts with higher median income values correlate with high Gini index values, and vice versa. The posterior results are hence a reflection of this correlation.

6.2 Pull-Push Patterns in Districts: Who Travels Where?

I hypothesised that low-income and highly segregated districts push residents out for daily mobility purposes, like work or recreational, whereas high-income and low segregated districts

pull residents, as these districts tend to offer higher amenities and work possibilities.

Adjacency Matrix

By looking at the relatively high set of weights located at the diagonal of the matrix (see Figure 12), it is observable that most travellers tend to stay within their own districts of origin or those that are proximate. These results align with the work of Alessandretti et al., who propose a container model in which human mobility is organized based on a hierarchical structure of spatial containers, or places, and with previous research showing that mobility patterns tend to follow a heavy-tailed distribution (González et al., 2008), where individuals travel short distances often, but travel long distances with less frequency. This *short trip tendency* is later reflected in Figures 18 and 19, where the mean of most trips is around two km.

Residents of Barajas show the highest tendency of travelling within their own district. The Barajas residents' most travelled destinations after their own district are Hortaleza and San-Blas Canillejas, which are precisely the only two contiguous districts. Barajas falls under the relatively high median income quantile, and the relatively low Gini index. Even though further research is needed to draw any conclusions, I hypothesise that Barajas offers a mix of good employment opportunities, high living standards and it is a big, peripheral district, located relatively far from the city centre. These conditions make the district become a city within the city, in which most residents can find all they need without needing to leave the district. This is related to the emergence of *polycentric cities*, metropolitan areas where urban functions are distributed among a series of subcenters (Van Criekingen et al., 2007). Gordon, Richardson and Wong concluded that a polycentric urban layout in Los Angeles between 1970 and 1980 has been associated with shorter work trips, particularly intra-county trips, in the most peripheral counties in cities like Los Angeles (Gordon et al., 1986). In a smaller scale, this could be the case of Barajas and other peripheral areas of Madrid, where residents do not need to leave their subcenters as often, especially if these offer job opportunities. The implications of polycentric cities on segregation are conflicting and the literature is extense on whether they reduce or increase inequality overall.

Summarising, the adjacency matrix reveals a strong tendency of travellers to stay within their own district of residence for most trips, and of travelling short distances, but further quantification of the trips and income stratification is needed to draw conclusions. The development of assortativity matrices plus the in-weight, out-weight analysis help reveal the reasons behind these patterns.

Assortativity Matrices

When taking into account the Gini index to build an assortativity matrix, it is observable a highly weighted square comprised by deciles 7 to 9 (see Figure 15 again for reference). This indicates that these highly segregated deciles tend to have high assortativity, meaning that residents from the top three segregated districts show relatively higher homophilic patterns. This is potentially the case because the most segregated districts are as well the wealthiest, hence most residents do not really need to leave these districts as they offer a high amount of job opportunities and amenities. In terms of the median income matrix (see Figure 16), deciles 0 and 1 show some of the highest diagonal weights, particularly decile 0. This shows that there are relatively strong homophily patterns across the lowest income districts. One potential reason is that these districts are often located in the peripheral areas, and transportation might be less accessible than in other districts, making it difficult for residents to access the central, wealthier districts. Another reason could be the emergence of the aforementioned *polycentric cities*, where individuals do not need to need the area to access all they need, but this is unlikely in the case of the poorest districts, as they are often not equiped with such amenities. Further

research on several factors is needed to accurately understand the causes of such homophilic patterns.

These results are reflected in the district out-weights, where some of the districts classified as LL (Usera, Villaverde, Villa de Vallecas and Vicálvaro) also show some the lowest out-weights (particularly Villaverde and Vicálvaro), indicating that residents do not leave the district as much as other residents do.

Summarising, based on the a*ssortativity indices p*, individuals tend to slightly travel to similar median income per consumption unit and Gini index deciles, but these patterns are far from showing a strong homophilic behaviour. One of the main reasons for the relatively stronger diagonal weights is that lots of individuals tend to travel within their own districts (self-loops) or often travel short distances. As median income and Gini index values are fairly clustered in the city, travelling short distances often implies staying in similar deciles to the district of residence, increasing the resulting diagonal weights of the matrices. In addition, further research to understand the causes of the homophilic patterns is needed.

In-Weights and Out-Weights

The weights per node reported in Table 9 help understand which districts push and pull more travellers in the city for the whole month of study. The range of in-weights is slightly wider than the range of out-weights, partly because self-loops (trips within districts) have been excluded for the out-weight calculation, but kept for the in-weight calculation.

At first sight, higher income, higher Gini index and geographically central districts like Centro, Chamberí or Chamartín have a relatively higher in-weight. Puente de Vallecas, which shows the second highest in-weight, is an outlier, as it is not one of the richest districts, nor so central. The reason for such high in-weight is that it receives a fairly high share of trips from Villa de Vallecas, the nearby district that belongs to the low income cluster. This could mean that a lot of Villa de Vallecas residents travel to Puente de Vallecas, a contiguous, slightly richer district, potentially for work purposes. Villa de Vallecas shows the the third lowest in-weight, indicating that it does not receive many travellers.

Barajas shows the lowest in-weight value. This indicates low travel preference to it. As shown in the adjacency matrix, Barajas displays the highest number of intra-district trips, yet it does not receive many trips from the rest of the districts.

The highest out-weight values are observed in Puente de Vallecas, Ciudad Lineal and Carabanchel. Puente de Vallecas displays the lowest intra-district preference (a weight of 0.3), indicating that most residents travel to other districts instead of staying in their own. By referring to the adjacency matrix, Puente de Vallecas shows fairly distributed travelling patterns across the city. Something similar, in a smaller scale, happens with Ciudad Lineal and Carabanchel, yet the later seems to show stronger travelling patterns to two particular districts: Usera and Latina, which are contiguous districts.

To further quantify these observations, the analysis of the four regression models is key.

Regressions

The linear regression results show that (1) there is a statistically significant positive effect of the Gini index on the in-weight of the districts, and that there is no statistically significant evidence to support a relationship between (2) low median income and out-weight, (3) high median income and in-weight, or (4) a high Gini index and out-weight. These results do not align with hypothesis 2, stating that residents from highly segregated districts might travel more often to less segregated districts. In fact, the results indicate that travellers tend to perform trips towards districts with high Gini index values.

A explanation for this tendency might be that, in the original dataset, the higher income districts also show some of the highest Gini index values, as these richer districts tend to also

have an unequal distribution of income. These highly unequal districts are also mostly located in the central area of the city, fulfilled with job opportunities and amenities. These potential explanations lead to further research on why this high-segregation, high-income dichotomy happens, and whether it is a local or a global trend.

6.3 Trip Counts: Who Travels More?

Based on the percentage of trips per quantiles (see Table 11), it is observable that residents from lower median income quantiles perform slightly more trips to high income quantiles than vice versa. These results support the hypothesis that lower income residents might need to travel more to perform daily activities, like work or daily purchases, and some leisure activities. Figure 5 shows that some of the wealthiest districts are located in the most central areas of the city, often fulfilled with amenities. This could explain the slightly higher number of trips from low to high income districts, especially in a city like Madrid, were a lot of the economic activities evolve around tourism, and most touristic attractions are placed in the central areas.

In terms of the Gini index, it is key to notice that there is a higher share of trips from low segregated districts to high segregated districts than the other way around, showing a flow from low to high Gini index districts. This aligns with the previously discussed results: as higher income districts also score high in inequality (or Gini index), the results reflected this correlation throughout the study.

The results also sustain the previously observed homophilic tendency, where intra-district and intra-quantile trips compose a high share of trips, especially when considering low median income and low Gini index quantiles. Even though further research is needed to find the real causes behind these patterns, literature often shows that low income classes tend to have less access to transportation possibilities, making it difficult to travel (Park and Kwan, 2017). Yet, they often need to perform longer trips for work purposes (Ureta, 2008).

Even though the aforementioned differences in the share of trips across low and high deciles comprise only a few decimal points for both the median income and the Gini index, it does show a potential pattern, as this study is dealing with trillions of trips within the city. I suggest that a bigger sample, both in terms of the temporal and spatial variables, could reflect stronger patterns.

6.4 Trips Based on Distance: Who Travels the Longest?

The distribution of trips by distance is reflected in Figures 18 and 19. The differences in the proportion of trips based on distance is smaller than expected initially, but, in case of the median income, the results do align with the initial hypothesis that the average distance in km for low-income residents is higher than high-income residents. On the other hand, the results show an unexpected behaviour in terms of the Gini index, as low-Gini index residents tend to travel slightly longer average trips than high-Gini index residents, failing to substantiate hypothesis 3. Again, the most feasible explanation for these unexpected results in terms of the Gini index is that the wealthiest districts tend to score the highest in the Gini index.

The longest average trips are performed to Barajas, which, as explained above, falls under the medium-high median income quantile, and under the low-medium Gini index quantile. This district is located in the peripherals of the city, and it counts with a fairly strong industrial activity, including the international Madrid airport, potentially attracting a high number of trips.

6.5 Gini Index and Median Income per Consumption Unit

The Gini index is a typical measure of inequality, but it cannot capture all the economic aspects behind a district. For that reason, it has been combined with an income indicator, the

median income per consumption unit. Due to this combination, it was possible to observe that the wealthiest districts are often the most unequal ones. This enriches the study, as, despite showing high inequality patterns, most of the richest districts are fulfilled with amenities and touristic attractions, pulling travellers towards them and displaying high in-weight numbers. This research shows that, in Madrid, the intra-district inequality does not reduce the attraction of a district, particularly if the district offers amenities, job opportunities and touristic attractions.

The final purpose of this research is not to determine why this high-income, high-Gini index dichotomy takes place, but it does reveal interesting patterns based on these two variables that need further study.

7 Limitations

7.1 Complex Segregation Patterns

The nuances of segregation are multifaceted and dynamic, often impossible to identify and measure. This research aimed to add a dimension to the study of segregation and inequality by focusing on granular mobility patterns in combination with median income per consumption unit and Gini index values. But the forms of segregation 'do not exist in isolation; they intersect and evolve through diverse social, economic, political, spatial and technological processes' (Netto et al., 2024). By using the previously defined methods and analyses, I inherently reduce such complex problem to a set of classifications and boundaries that cannot grasp all of the multiple domains and scales in which segregation occurs.

Indices' Problems

Some of the main problems when calculating spatial indices are the measurement of spatial interaction, the selection of an adequate spatial scale and how to measure significance. Yao et al. discuss some of these common spatial representation problems in detail: 'It should be noted that there are numerous ways to represent spatial interaction, analyse spatial scale and derive statistical significance' (Yao et al., 2019). In this particular research, I use the Moran's I as the main segregation index, with the consequent limitation that it is sensitive to changes in the weight matrix of choice (Maruyama, 2015). Hence, the use of a different weight matrix or standardisation process can influence the resulting coefficients. Refer to Figure 26 on Appendix A for a comparison of Queen and Rook cardinalities.

In addition, I set the statistical thresholds of significance to follow what is established in the scientific community. On the other hand, especially when referring to the study of such multifactorial patterns, making a binary choice based on a threshold does not capture all the nuances behind the scenes.

Referring to the choice of spatial scale, I selected districts as units of study. The size of the districts considered in this research is diverse. Some of the most central districts are relatively small compared to the most peripheral ones. As a result, the number of trips from one district to another is susceptible to these size differences. For example, it only takes a ten minute car ride to go from one central district to another, but it can take up to forty minutes for inhabitants of a big peripheral district to reach a contiguous one. Due to these size disparities, inhabitants from big peripheral districts might show a strong intra-district weight, like it happens for example in the case of Barajas. This is not *per se* an inaccurate result, but it does hide a lot of valuable information of intra-district behaviour, as these intra-district trips can, in some cases, be longer than some district-to-district trips, but not be reflected.

Measuring Wealth and Inequality

My measure of wealth relies on the median income per consumption unit by district, which does not exhaust the concept of individual-level SES. In addition, due to the impossibility of obtaining the data on the individuals' districts of residence, I had to filter only home-origin trips to infer the socioeconomic background of the travellers. This reduces the richness of the mobility sample, potentially skewing the results. The reason a more granular approach could not be taken is due to ethical and privacy constrains, as retrieving mobility and income data at an individual level is often not possible and trespasses individual privacy limits. As an alternative, Duran-Sala et al. improve the approximation of the individual-level SES by incorporating information on individual's income and demographic status present in the mobility data source, on top of the residential income level. This addition increases the granularity of the individual profiles, adding an extra dimension to further understand segregation and inequality processes. Other authors further identify the usual home areas of the users by performing a thorough analysis of the overnight stays through several weeks (Ponce-de-Leon et al., 2021), available in other dataset from the same source.

Transportation and Amenities

Two of the many factors that influence how people move in urban areas are transportation and amenities. Due to their complexity, these variables have been left out for the study. Yet, they influence movement heavily and there is extensive research on how they do.

As Anderson and Galaskiewicz summarise, a specific ramification of the literature states that the higher spheres 'prioritize certain parts of the city, typically by race and class, to receive key public amenities, including public transit', what has been named as *transport disadvantage*. This form of social exclusion can heavily limit the mobility of the citizens. Something similar happens with amenities, where central, richer districts tend to allocate a higher number of them, influencing the way citizens travel and interact with the urban space.

Yet a general analysis of these variables is used to interpret the results of this research, further quantification of these two variables will benefit the methods used in this study to further understand the results.

7.2 Sampling

Temporal Sample

As a temporal subset, I used the month of March of 2022 as a representative sample for the study of the mobility patterns. Monthly mobility patterns can provide valuable insights in terms of mobility, but most mobility research focuses on periods longer than that, like González et al., who used a six-month sample, or Edsberg Møllgaard et al., who used a one-year sample. In addition, a better approach to represent a typical commuting period would have been to artificially construct a typical week and weekend by adding individual, non-festive days. This construction would allow to study the week and weekend differences better.

Spatial Sample

For this research, I do not take into account the several commuter towns that shape the metropolitan area of Madrid, which has been classified as the largest metropolitan area in Spain, and the second in the European Union as of 2021 (OECD, 2021). The definitions of the extension of the metropolitan area change based on the year of study and different sources. To get a rough idea, Gómez Giménez and Hernández Aja state that the Madrid metropolitan area is in continuous expansion, and between 2001 and 2011, up to 40% of the population decided to locate further than 25 km from the functional city centre. As a result, this metropolitan

area plays a big role in the influx of travellers in the city, making it important to include as part of the area of study. Due to the time frame available and the need for a detailed study of the characteristics of the towns and income data availability, this metropolitan area has not been included in the research. On the other hand, the mobility data source provides mobility patterns for the whole Spanish territory, including this metropolitan area. The methods used in this study can easily be extrapolated to include these towns by just filtering a different set of districts.

7.3 Trips as Edges

The edges in the mobility network are used to represent normalised trips across and within districts. Their weight is based on the number of trips, but, as it often happens with network representations, these weights do not capture all of the information that is provided in the dataset, becoming inherently a limitation, as there are many other factors that influence trips than just frequency. For example, the weights assigned to the edges do not contain information on the purpose of the trip, or the gender of the group performing the trip. Adding more information to the edges, particularly the purpose of the trip, which indicates whether individuals are travelling to work or to other locations, could enrich the analysis massively.

7.4 Mobile Phone Data

Despite being a great proxy for individual-level mobility patterns, phone data can have certain limitations that skew the analysis. For example, some populations are under represented, especially those demographic groups in which mobile phone penetration is incomplete, like elderly, kids, or homeless individuals. In addition, the use mobile phone data raises ethical concerns due to its sensibility, and potential privacy breaks. As Sieg et al. explore in their research, mobile phone data must be aggregated and not individualised. It is important to find a balance between granularity and privacy. The data used in this research has been made publicly available by The Ministry of Transport and Sustainable Mobility (Ministerio de Transportes y Movilidad Sostenible (MITMA), 2022). The available data is already aggregated and anonymised, making it impossible to follow individual trips, ensuring the privacy of the participants. All the steps throughout this research respect this format and ensure the privacy of the residents.

8 Conclusion

This research is relevant because it delves into the mobility patterns and segregation processes generated in a southern European metropolis, providing insights that can help apprehend segregation patterns and inequality processes. To answer the research question, the study reveals moderate homophilic tendencies within the city, where residents from the lowest median income districts travel slightly more, and longer, than those from the highest income districts. The results also show that residents from the most unequal districts travel the least, and the shortest.

Further research could complement this study by deepening further into the role of transportation and amenities on the results, and further filtering the trips by purpose. In addition, including the metropolitan area and developing a normalisation technique to account for the unequal size of the districts could reveal further insights into the number of trips.

The insights provided in this study can be used by urban planners and policymakers to reduce inequality in an already segregated city like Madrid. Reducing such inequalities can increase overall life satisfaction, increase safety in the city and build a more prosperous society. In addition, the methods defined in the following sections can be applied to any dataset containing OD trips in combination with relevant socio-demographic data. These methods can be used to determine segregation patterns and reducing inequality among different territories.

Bibliography

Alessandretti, L., Aslak, U., and Lehmann, S. (2020). The scales of human mobility. *Nature*, 587(7834):402–407.

Alessandretti, L., Natera Orozco, L. G., Saberi, M., Szell, M., and Battiston, F. (2023). Multimodal urban mobility and multilayer transport networks. *Environment and Planning B: Urban Analytics and City Science*, 50(8):2038–2070.

Anderson, K. F. and Galaskiewicz, J. (2021). Racial/Ethnic Residential Segregation, Socioeconomic Inequality, and Job Accessibility by Public Transportation Networks in the United States. *Spatial Demography*, 9(3):341–373.

Anselin, L. (1996). The Moran scatterplot as an ESDA tool to assess local instability in spatial association. In Fischer, M., Scholten, H. J., and Unwin, D., editors, *Spatial Analytical Perspectives on GIS*, pages 111–126. Routledge, 1 edition.

Ardura Urquiaga, A., Lorente-Riverola, I., and Ruiz Sanchez, J. (2020). Platform-mediated short-term rentals and gentrification in Madrid. *Urban Studies*, 57(15):3095–3115.

Barrat, A., Barthelemy, M., Pastor-Satorras, R., and Vespignani, A. (2003). The architecture of complex weighted networks.

Bassolas, A., Sousa, S., and Nicosia, V. (2021). Diffusion segregation and the disproportionate incidence of COVID-19 in African American communities. *Journal of The Royal Society Interface*, 18(174):20200961.

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

Bokányi, E., Juhász, S., Karsai, M., and Lengyel, B. (2021). Universal patterns of long-distance commuting and social assortativity in cities. *Scientific Reports*, 11(1):20829.

Borja, J. and Castells, M. (1997). *Local y global: la gestión de las ciudades en la era de la información*. UNCHS.

Candipan, J., Phillips, N. E., Sampson, R. J., and Small, M. (2021). From residence to movement: The nature of racial segregation in everyday urban mobility. *Urban Studies*, 58(15):3095–3117.

Coscia, M. (2021). The Atlas for the Aspiring Network Scientist.

Duncan, O. D. and Duncan, B. (1955). A Methodological Analysis of Segregation Indexes. *American Sociological Review*, 20(2):210–217.

Duran-Sala, M., Balachandran, A. K., Morandini, M., Naushirvanov, T., Prabhakaran, A., Renninger, A., and Mazzoli, M. (2024). Disentangling individual-level from location-based income uncovers socioeconomic preferential mobility and impacts segregation estimates.

Edsberg Møllgaard, P., Lehmann, S., and Alessandretti, L. (2022). Understanding components of mobility during the COVID-19 pandemic. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 380(2214):20210118.

Farris, F. A. (2010). The Gini Index and Measures of Inequality. *The American Mathematical Monthly*, 117(10):851–864.

Feitosa, F. F., Koschitzki, T., Silva, M. P. S., and Monteiro, A. M. V. (2007). Global and local spatial indices of urban segregation. *International journal of geographical information science : IJGIS*, Vol.21(3):p.299–323.

Gómez Giménez, J. M. and Hernández Aja, A. (2018). *Retos de Las Áreas Urbanas Funcionales Españolas: El Caso Madrileño, 1991-2011*, volume 6 of *Formas Urbanas y Territorio*. Universidad de Zaragoza, II Congreso Internacional ISUF-H Zaragoza 2018.

González, M. C., Hidalgo, C. A., and Barabási, A.-L. (2008). Understanding individual human mobility patterns. *Nature*, 453(7196):779–782.

González-García, I., Fernández-Ramírez, C., and Rodríguez-Alonso, R. (2024). The Construction of Residential Exclusion in Madrid and Its Metropolitan Area. In Lois-González, R. C. and Rio Fernandes, J. A., editors, *Urban Change in the Iberian Peninsula: A 2000–2030 Perspective*, pages 87–108. Springer Nature Switzerland, Cham.

Gordon, P., Richardson, H. W., and Wong, H. L. (1986). The Distribution of Population and Employment in a Polycentric City: The Case of Los Angeles. *Environment and Planning A: Economy and Space*, 18(2):161–173.

Graham, S. and Marvin, S. (2002). *Splintering Urbanism: Networked Infrastructures, Technological Mobilities and the Urban Condition*. Routledge, London.

Haandrikman, K., Costa, R., Malmberg, B., Rogne, A. F., and Sleutjes, B. (2023). Socio-economic segregation in European cities. A comparative study of Brussels, Copenhagen, Amsterdam, Oslo and Stockholm. *Urban Geography*, 44(1):1–36.

Instituto Nacional de Estadística (2021a). Censo de población y viviendas 2021: Tabla 31097. https://ine.es/jaxiT3/Datos.htm?t=31097.

Instituto Nacional de Estadística (2021b). Indicadores demográficos básicos. Tabla 37727. https://ine.es/jaxiT3/Datos.htm?t=37727#_tabs-tabla.

Instituto Nacional de Estadística (2023). IOE metodología. Technical report.

Jakubs, J. F. (1981). A distance-based segregation index. *Socio-Economic Planning Sciences*, 15(3):129–136.

Jargowsky, P. A. (1996). Take the money and run: Economic segregation in U.S. metropolitan areas. *Institute for Research on Poverty Discussion Papers*, (1056-95).

Kaufmann, T., Radaelli, L., Bettencourt, L. M. A., and Shmueli, E. (2022). Scaling of urban amenities: Generative statistics and implications for urban planning. *EPJ Data Science*, 11(1):1–19.

Kazmina, Y., Heemskerk, E. M., Bokányi, E., and Takes, F. W. (2024). Socio-economic segregation in a population-scale social network. *Social Networks*, 78:279–291.

Liu, Y. and Gastwirth, J. L. (2020). On the capacity of the Gini index to represent income distributions. *METRON*, 78(1):61–69.

Marcuse, H. (2022). La lucha contra el liberalismo en la concepción totalitaria del Estado. Constelaciones. *Constelaciones. Revista de Teoría Crítica*, 13:487–522.

Maruyama, Y. (2015). An alternative to Moran's I for spatial autocorrelation.

Massey, D. S. and Denton, N. A. (1988). The Dimensions of Residential Segregation*. *Social Forces*, 67(2):281–315.

Ministerio de Transportes y Movilidad Sostenible (MITMA) (2022). Estudios de movilidad con big data: Open data movilidad.

Moran, P. A. P. (1950). Notes on Continuous Stochastic Phenomena. *Biometrika*, 37(1/2):17–23.

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(1):4633.

Municipio en Cifras (2023). El municipio en cifras. https://portalestadistico.com/municipioencifras/?pn=madrid&pc=ZTV21.

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(7):1062–1083.

Müürisepp, K., Järv, O., Tammaru, T., and Toivonen, T. (2022). Activity Spaces and Big Data Sources in Segregation Research: A Methodological Review. *Frontiers in Sustainable Cities*, 4(861640).

Neira, M., Molinero, C., Marshall, S., and Arcaute, E. (2024). Urban segregation on multilayered transport networks: A random walk approach. *Scientific Reports*, 14(1):8370.

Netto, V., Krenz, K., Fiszon, M., Peres, O., and Rosalino, D. (2024). Decoding Segregation: Navigating a century of segregation research across disciplines and introducing a bottom-up ontology.

Newman, M. E. J. (2004). Analysis of weighted networks. *Physical Review. E, Statistical, Nonlinear, and Soft Matter Physics*, 70(5 Pt 2):056131.

Nilforoshan, H., Looi, W., Pierson, E., Villanueva, B., Fishman, N., Chen, Y., Sholar, J., Redbird, B., Grusky, D., and Leskovec, J. (2023). Human mobility networks reveal increased segregation in large cities. *Nature*, 624(7992):586–592.

OECD (2021). OECD Data Explorer - Archive • Metropolitan areas. https://data-explorer.oecd.org.

Park, Y. M. and Kwan, M.-P. (2017). Multi-Contextual Segregation and Environmental Justice Research: Toward Fine-Scale Spatiotemporal Approaches. *International Journal of Environmental Research and Public Health*, 14(10):1205.

Ponce-de-Leon, M., del Valle, J., Fernandez, J. M., Bernardo, M., Cirillo, D., Sanchez-Valle, J., Smith, M., Capella-Gutierrez, S., Gullón, T., and Valencia, A. (2021). COVID-19 Flow-Maps an open geographic information system on COVID-19 and human mobility for Spain. *Scientific Data*, 8(1):310.

Ravenstein, E. G. (1885). The Laws of Migration. *Journal of the Statistical Society of London*, 48(2):167–235.

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

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

Sieg, L., Gibbs, H., Gibin, M., and Cheshire, J. (2024). Ethical Challenges Arising from the Mapping of Mobile Phone Location Data. *The Cartographic Journal*, 0(0):1–14.

Tobler, W. R. (1970). A Computer Movie Simulating Urban Growth in the Detroit Region. *Economic Geography*, 46:234–240.

Ureta, S. (2008). To Move or Not to Move? Social Exclusion, Accessibility and Daily Mobility among the Low-income Population in Santiago, Chile. *Mobilities*, 3(2):269–289.

Van Criekingen, M., Bachmann, M., Guisset, C., and Lennert, M. (2007). Towards polycentric cities. An investigation into the restructuring of intra-metropolitan spatial configurations in Europe. *Belgeo. Revue belge de géographie*, (1):31–50.

Vaughan, L. and Arbaci, S. (2011). The Challenges of Understanding Urban Segregation. *Built Environment*, 37(2):128–138.

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

Wilson, W. J. (1987). *The Truly Disadvantaged: The Inner City, the Underclass, and Public Policy, Second Edition*. University of Chicago Press, Chicago, IL.

Wong, D. W. S. (1998). Measuring Multiethnic Spatial Segregation. *Urban Geography*, 19(1):77–87.

Yao, J., Wong, D. W., Bailey, N., and Minton, J. (2019). Spatial Segregation Measures: A Methodological Review. *Tijdschrift voor Economische en Sociale Geografie*, 110(3):235–250.

Zambon, I., Serra, P., Sauri, D., Carlucci, M., and Salvati, L. (2017). Beyond the 'Mediterranean city': Socioeconomic disparities and urban sprawl in three Southern European cities. *Geografiska Annaler: Series B, Human Geography*, 99(3):319–337.

Appendix A. Extra Figures

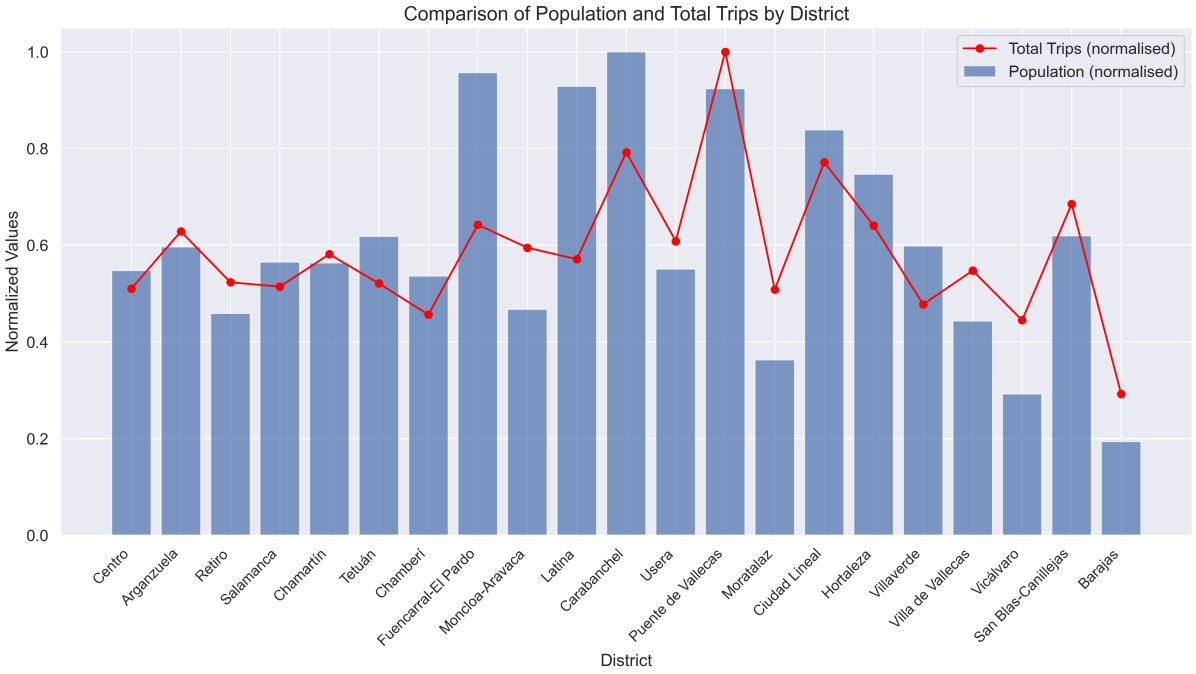

Figure 20: Standardised relationship between the population of the districts and the number of trips originating from each district in the original dataset.

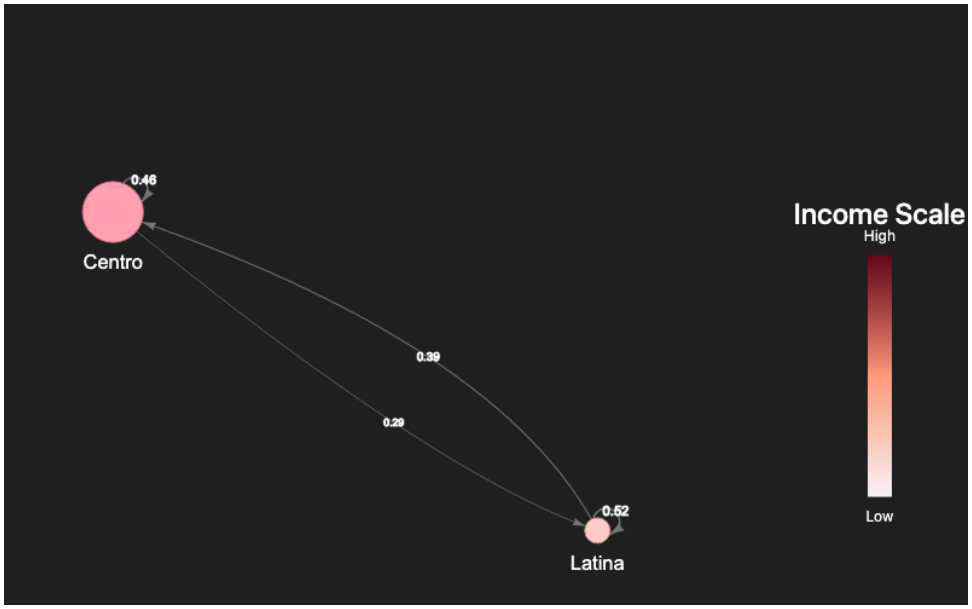

Figure 21: Sample of normalised trips during March of 2022 between two districts. Node sizes represent the in-weight, and the colour of the nodes represents the median income per consumption unit range.

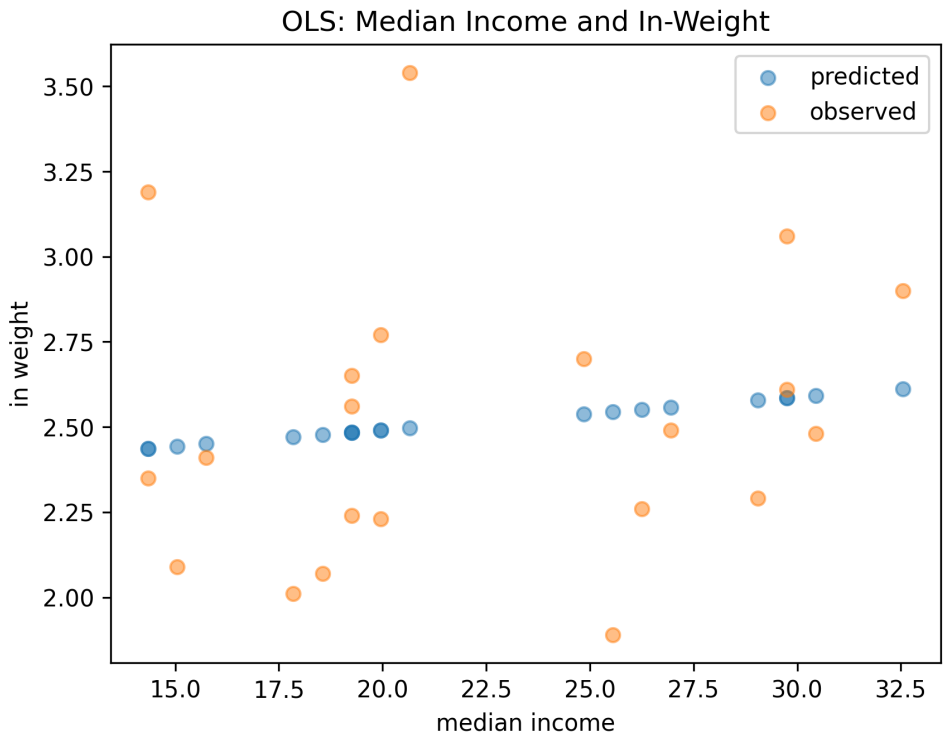

Figure 22: Model 1 results for the median income per consumption unit and in-weight.

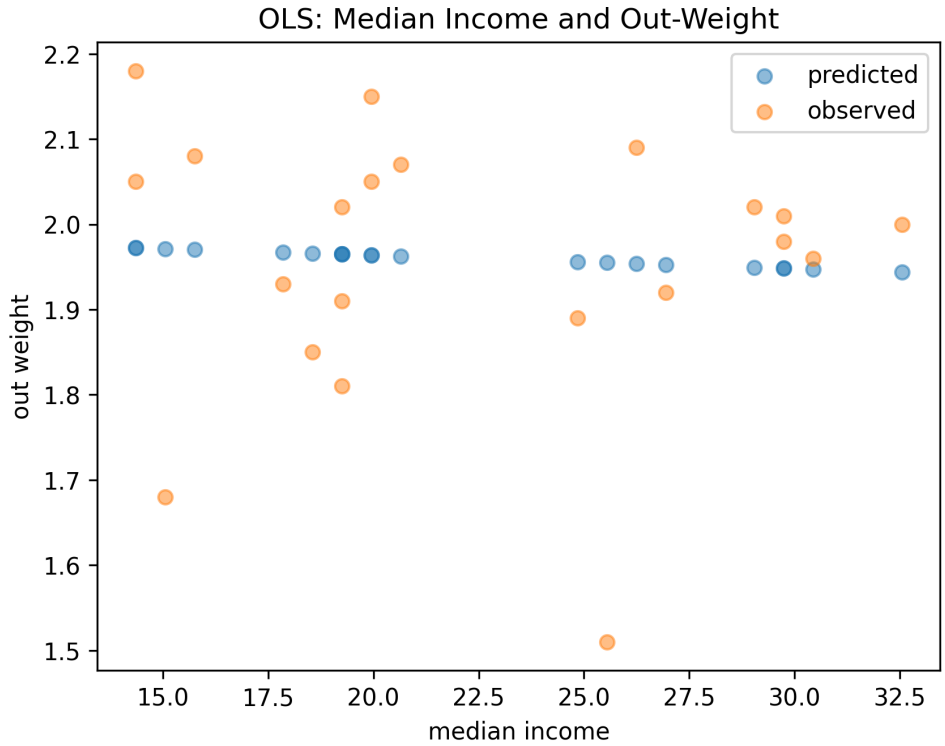

Figure 23: Model 2 results results for the median income per consumption unit and out-weight.

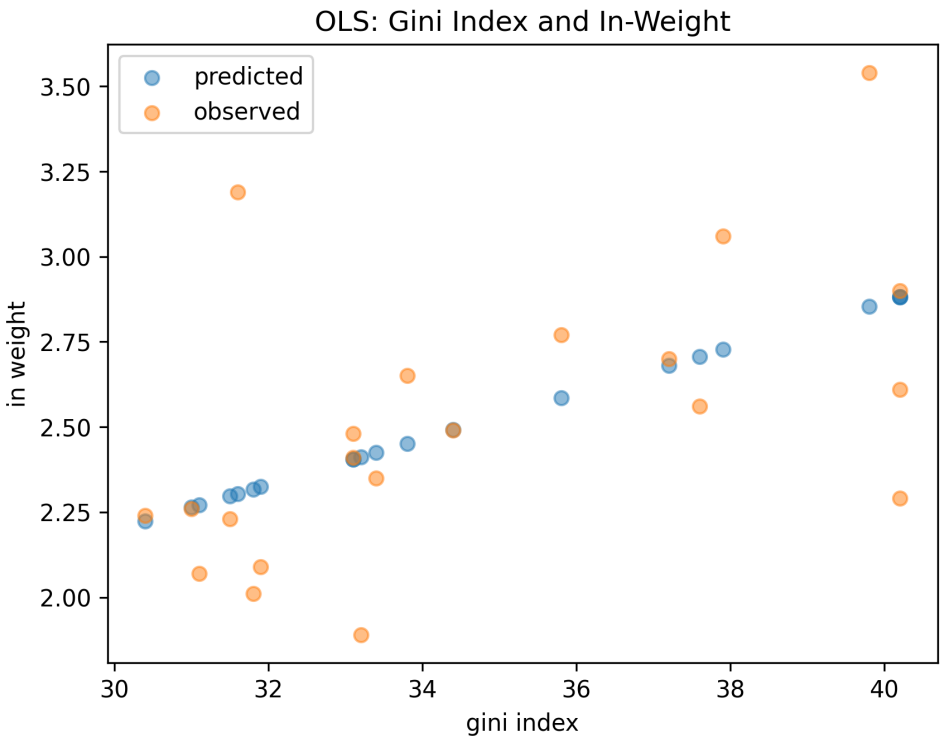

Figure 24: Model 3 results for the Gini index and in-weight.

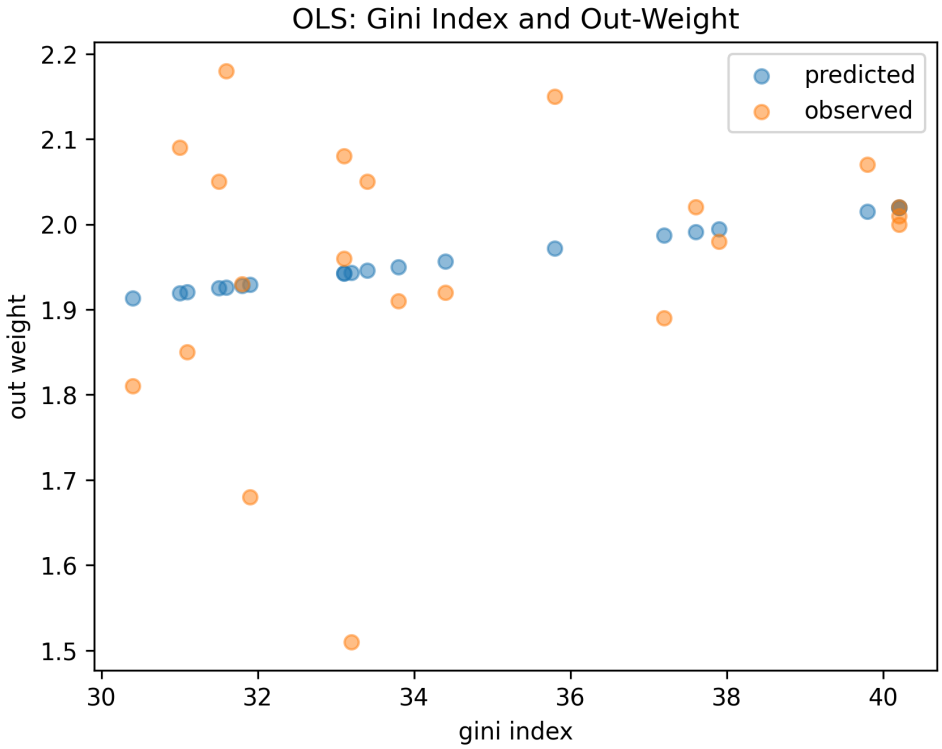

Figure 25: Model 4 results for the Gini index and out-weight.

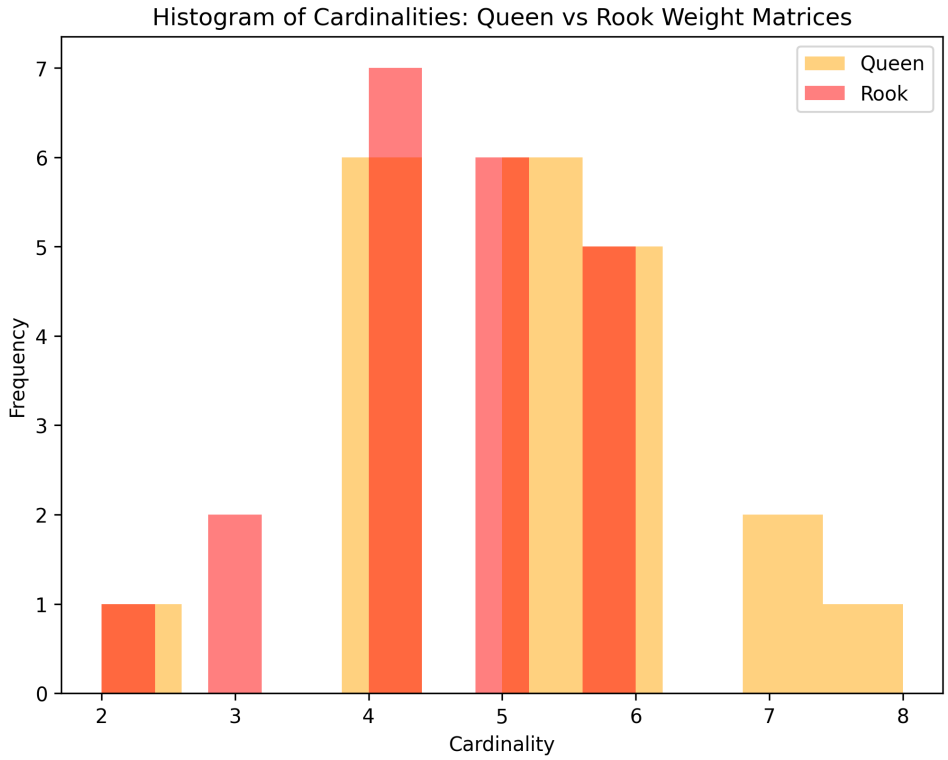

Figure 26: Comparison of Rook and Queen cardinalities to illustrate how weight matrix choices as measures of contiguity can influence the results.

Appendix B. Use of AI tools

The following AI tools have been utilized while developing code: ChatGPT-3.5. GPT-3.5 has been assisting in designing the LaTeX equations and tables, and helped develop the code used to perform the study.

Appendix C. Code Availability

All the code needed to replicate the study is available in the following Github repository: https://github.com/carobs9/segregation-madrid. An interactive visualisation of the resulting network of the districts and their weights is available in the following link: https://carobs9.neocities.org/no_threshold.