Master of Science in Civil Engineering

Master Thesis

Assessing Public Transportation Accessibility Equity and Scoring Transit Lines Based on Their Contribution: An Open-Data Based Approach

Supervisor

Prof. Marco Diana

Dipartimento di Ingegneria dell'Ambiente, del Territorio e delle Infrastrutture - Politecnico di Torino

Co-Supervisor

Prof. Andrea Araldo

Département Réseaux et Services de Télécommunications - Institut Politechnique de Paris, Télécom SudParis

Candidate

Amirhesam Badeanlou

Academic Year 2021-2022

Abstract:

In order to encourage people to use alternative transport solutions while reducing the disproportional use of cars, studies that could provide relevant information to stakeholders and policymakers are required. This is due to the excessive use of cars and its negative externalities, such as traffic congestion and the production of high levels of carbon dioxide (CO2). Therefore, one strategy is to improve public transport by providing a fair level of access to it for the people to encourage them to have a multimodal behavior regarding transport.

This thesis proposes a methodology to assess public transportation accessibility inequity of land parcels served by transit systems in metropolitan areas. The methodology is based on the classic analysis tools of Lorenz curves and Gini indices, but the novelty resides in the fact that it can be easily applied in an automated way to several cities around the world, with no need for customized data treatment. Indeed, our equity metrics can be computed solely relying on open data, publicly available in a standardized form. We showcase our method by studying public transportation territorial equity in Paris, Madrid, Sydney, and Boston, and compare our findings with another recently proposed approach.

The above issue is related to the configuration of the transit offer in urban areas. Current transit suffers from an evident inequity: the level of service of transit in the suburbs is much less satisfying than in city centers, especially when budget constraints are tight. Consequently, private cars are still the dominant transportation mode for suburban people, which results in congestion and pollution. To achieve sustainability goals, transit should be (re)designed, placing equity among the main optimization objectives. To this aim, it is necessary to (i) quantify the "level of equity" of transit and (ii) identify the transit lines that are the most important to equity, which would then be the ones where the operator should invest the most in increasing service level (frequency or coverage) in order to reduce inequity in the system.

To the best of our knowledge, we are the first to tackle (ii). We propose efficient computational methods that rely solely on open data, allowing us to analyze the whole transit networks in Aachen, Berlin, Budapest, Helsinki, Manchester, Turin, and Vienna. Our method can be used to guide large-scale iterative optimization algorithms toward the goal of improving accessibility equity.

Abstract (Italiano):

Al fine di incoraggiare le persone a utilizzare soluzioni di trasporto alternative riducendo l'uso sproporzionato delle automobili, sono necessari studi che potrebbero fornire informazioni pertinenti alle parti interessate e ai responsabili politici. Ciò è dovuto all'uso eccessivo delle automobili e alle sue esternalità negative, come la congestione del traffico e la produzione di alti livelli di anidride carbonica (CO2). Pertanto, una strategia consiste nel migliorare il trasporto pubblico fornendo un livello equo di accesso ad esso per le persone per incoraggiarle ad avere un comportamento multimodale riguardo ai trasporti.

Questa tesi propone una metodologia per valutare la disuguaglianza di accessibilità al trasporto pubblico nelle diverse zone delle aree metropolitane. La metodologia si basa sui classici strumenti di analisi delle curve di Lorenz e degli indici di Gini, ma la novità sta nel fatto che può essere facilmente applicata in modo automatizzato a diverse città del mondo, senza necessità di un trattamento personalizzato dei dati. In effetti, le nostre metriche di equità possono essere calcolate esclusivamente basandosi su dati aperti, pubblicamente disponibili in una forma standardizzata. Mostriamo il nostro metodo, studiando l'equità territoriale nell'accesso al trasporto pubblico a Parigi, Madrid, Sydney e Boston e confrontiamo i nostri risultati con un altro approccio recentemente proposto.

Questa problematica è collegata alla configurazione dell'offerta di trasporto pubblico nelle aree urbane, Gli attuali sistemi di trasporto pubblico soffrono di un'evidente diseguaglianza: il livello del servizio nelle periferie è molto meno soddisfacente rispetto ai centri urbani, specialmente quando i vincoli di bilancio sono stringenti. Di conseguenza, le auto private sono ancora il mezzo di trasporto dominante per le persone che vivono nelle aree suburbane, il che si traduce in congestione e inquinamento. Per raggiungere gli obiettivi di sostenibilità, il trasporto pubblico dovrebbe essere (ri)progettato, ponendo l'equità tra i principali obiettivi di ottimizzazione. A tal fine è necessario (i) quantificare il "livello di equità" del trasporto pubblico e (ii) identificare le linee più importanti per l'equità, che sarebbero poi quelle in cui l'operatore dovrebbe investire di più in aumentare il livello di servizio (frequenza o copertura) al fine di ridurre le disuguaglianze nel sistema.

A quanto ci risulta, siamo i primi ad affrontare (ii). Proponiamo metodi computazionali efficienti che si basano esclusivamente su dati aperti, consentendoci di analizzare l'intera rete di trasporto pubblico ad Aquisgrana, Berlino, Budapest, Helsinki, Manchester, Torino e Vienna. Il nostro metodo può essere utilizzato per guidare algoritmi di ottimizzazione iterativa su larga scala verso l'obiettivo di migliorare l'equità dell'accessibilità.

List of Figures

List of Tables

1 Table of Contents

1. Chapter 1: Introduction

Public transport is one of the most important services in the cities, and it plays an undeniable role in people's travel from one place to another. Nowadays, because of the movements of the countries toward sustainability, authorities are trying to encourage and persuade people to use this system, considering that using private cars, especially by the suburban population, is among the leading causes of pollution in cities (Grelier, 2018). Therefore, reducing this car dependency is an essential purpose for the sustainability of urban transportation. Thus, authorities should enhance this system so that cars are not favored over public transport (Turcotte, 2008).

Transportation demand in the suburbs is too low, unlike in the city center, therefore it is not efficient to have a public transport service with a high frequency and a high number of stops, even if public transport is still needed in those areas also for social reasons. Consequently, suburban public transport users experience walking times and waiting times much higher than the city center's population. This inequality (Calabrò et al., 2021) is structural in public transport and increases car dependency in suburban areas (Welch and Mishra, 2013).

Accessibility can be described as the capacity of cities to allow people to move efficiently by guaranteeing equity and equal access to personal and professional opportunities. According to the previous discussion and as noted by (Biazzo et al., 2019), it is important to study equity. Transportation equity can be assessed by the geographical distribution of the chosen accessibility metric; if there is a big difference between accessibility in the city center and the suburb, this indicates high inequality.

Due to the rich datasets needed to study this accessibility inequality in the area under the study, e.g., households income and employment, one needs to contact the responsible authority. The lack of a standard format is another issue, because any country reports these data in their own specific format and their own language. Therefore, these data also need some data processing which is a time-consuming process or even infeasible. For this reason, most of the work on this matter focuses on one or two scenarios. Consequently, we based our method solely on open source and available data to be capable of replicating the method on multiple cities.

1.1. Objectives

The objectives of this thesis are presented in the following list:

a) To measure the inequality in the transportation level of service in the spatial distribution of the level of service of public transport

b) To find the lines of public transport that are affecting the most the level of inequality in different cities

c) To keep our method solely based on open data, like in the original (Biazzo et al., 2019) work, to have our methodology easily replicable in any city

d) To provide some suggestions on the better implementation of our method and proposal of future work.

1.2. Thesis structure

Chapter 1 gives a brief introduction about our motivations to follow this topic and inequality in public transport. Then, the main objectives of this research are shown. Finally, the thesis structure has been laid out with a brief description of each chapter in the following.

Chapter 2 shows the background on the accessibility and different ways of measuring it alongside with the advantages and disadvantages of each method. The last part is devoted to the background on equity in transport especially in public transport accessibility.

Chapter 3 explains the methodology of this research and database used. First, the input open data which are GTFS (General Transit Feed Specification) and population grid are introduced. And then for the methodology part we explain the work of (Biazzo et al., 2019) and how we replicate their method in our work to find the spatial accessibility distribution in different cities. Additionally, we explain the Lorenz curve and Gini coefficient and how we used them to catch the level of inequality. Finally, the last part of this chapter devoted to the methodology to find the most important transit lines for equity.

Chapter 4 shows all results of accessibility scores in different cities. And then talk about their general description and inequality in their geographical distribution. Additionally, we will trace the Lorenz curve, compute Gini indices, and compare our findings with (Biazzo et al., 2019). Finally, we will show the results of transit line scoring and discuss the methodology used.

Chapter 5 provides conclusions for the methodologies that we have adopted with explanation of the results. In addition, some suggestions are also provided for future implementation of methodology and future research.

2. Chapter 2: Background

2.1.Accessibility definitions

Accessibility is one of the main important subjects in transport planning and specifically in Public transport. It was first defined as 'the potential of opportunities for interaction' by (Hansen, 1959), which can also see the role of transport systems and land use in the mobility of people (Geurs and Van Wee, 2004). There are other definitions of accessibility in the literature but, in general, accessibility is labelled as the physical access to goods, services, and destinations. In the context of urban economics and geography, accessibility, which is one of the most important outcomes of the transportation system, is characterized as the facilitation in accessing a specific area or location (Mavoa et al., 2012). It is a measure of the advantage of the location of a zone or area compared to the other zones and areas (Biosca et al., 2013).Moreover, accessibility can be viewed from two perspectives: location-based accessibility, which includes mobility and land use, and individual-based accessibility, which focuses on the individual level (Miller, 2005). In addition, there are other perspectives like infrastructure-based accessibility, which focuses on mobility, and utility-based, which sees from an economic view (Geurs and Van Wee, 2004). According to the definitions above and considering a location based definition which is not completely depends on the specific area or location to be accessed and just considers the whole city to be accessed, accessibility can be described as the capacity of cities to allow people to move efficiently by guaranteeing equity and equal access to personal and professional opportunities (Biazzo et al., 2019).

Urban public transport system (PT) has drawn more attention recently due to its potential to enhance sustainability and urban life quality. If the transport infrastructure is not capable of meeting the demands, this causes an increase in waiting times and congestion in public transport and streets because people are more interested to use private cars (Lodovici and Torchio, 2015). So, the goodness of accessibility of public transport can improve the accessibility of other transport systems like private cars. Because, if we provide an accessible public transport system, the number of people who use the public system because it is beneficial for them increases. And, it causes less congested streets and provide faster paths for those who have to use their private cars (Abreha, 2007).

2.2.Accessibility measures

As we are going to focus on location-based accessibility, different metrics have been proposed in the literature to quantify location-based accessibility for various application scenarios. Although there are many different accessibility metrics, they may generally be divided into four groups (Handy and Niemeier, 1997, Kwan, 1998, Miller, 2020):

- Distance to the nearest location which could be a subway station, shopping center, medical center, etc.
- Opportunities that can be reached cumulatively within a certain access distance or travel time threshold (isochrones method)
- Gravity or entropy methods

- Random utility based measures

In order to describe these measures, first we have to define some terms which are used in these measures. The first term is the *travel impedance*. The impedance metric used in the most basic accessibility measurements is distance. This distance could be calculated as the simple straight-line (Euclidean) distance between two points, the right-angled (Manhattan) distance traveled along a rectangular grid network from point to point, or the network distance, typically computed as the shortest-path from point to point through the actual network, depending on the application.

However, travel time is a much better indicator of the effort or expense associated with traveling by car, public transportation, or even a bicycle. It doesn't really matter if a trip is 10 km or 15 km long as long as it takes to get there; if both trips take 30 minutes to complete, then access to both destinations should be the same (assuming they are equally desirable).

Additionally, travel time is a variable that is sensitive to policy (i.e., improvements in the performance of the transport system will result in shorter travel times, improving accessibility, and vice versa), whereas distance is depending on land use patterns that are less under the control of public authorities which in any case can change them only in the long run and less easily. Because they allow for the computation of accessibility implications of various policies and alternatives, travel-time based measurements are often considerably more valuable variables to include in planning analysis and evaluation.

The second needed term to define is *location attractiveness*. The most straightforward and typical approach to determining a location's attractiveness is to use a size variable of some kind, such as the number of jobs in a given zone for employment accessibility or the number of stores (or retail floor space or retail employees) in a zone for shopping accessibility. It is conceivable to use more sophisticated representations that include additional activity location characteristics that influence their suitability for prospective interaction (such as the quality and price of goods for shopping accessibility). However, in reality, these more in-depth descriptions of attractiveness are rarely applied, most likely as a result of data shortages and the resulting increase in analytical complexity.

Noting that factors like store opening/closing hours and levels of facility congestion (e.g., a popular restaurant may be fully booked and not available without a prior reservation) may also affect how appealing and/or accessible a given activity place is during the course of the day.

The third expression is *person level heterogeneity in accessibility*. Because everyone has different tastes and preferences, as well as different resources and constraints, it is reasonable to expect that the (dis)utility of travel to a particular location and the evaluation of the location's attractiveness as a potential destination will vary subjectively from person to person. Due to their larger spending power, those with higher incomes typically have access to a considerably wider variety of goods, services, and activities than those with lower incomes. People who have access to vehicles will be able to engage in a larger variety of activities than those who do not possess cars or who are unable to drive.

And, the last term is *location choice set*. The question of what locations are to be included in a given accessibility calculation is inherent in all accessibility measures.

In case of distance to the nearest location method, the measure can be expressed mathematically as:

$$a^{ip} = \underset{j \in L^p}{MIN}(d_{ij})$$

Where:

- a^{ip} is accessibility of zone i to the location of type p.
- L^p is set of locations of type p.
- d_{ij} is distance or travel time for a given mode from i to a location j in set L^p.

This measure has two limits:

a) it does not take into account the size or attractiveness of the closest place, thereby regarding all locations as equally attractive.
b) It does not take into account the cumulative effect of several accessible locations (for instance is a zone that is within 2.1 Km of two subway stations is lesser than he one within 2 Km of one single station?).

This metric is in line with a very straightforward location model where the closest place is always selected with probability 1.0. Which is:

$$P_j^{ip} = 1 \; if \; d_{ij} = \underset{j' \in L^p}{MIN} d_{ij'}; = 0 \; otherwise$$

where P_j^{ip} is the probability of choosing location j for purpose p given that one is located in zone i. Except in special cases, this is not a realistic choice model. Subway stations or highway junctions (only used if they are on the best route for a particular trip to the actual destination) or activity locations (for which there generally will be many competing locations of varying attractiveness). Therefore, this measure is more suitable as an explanatory variable in models for transportation or residence choices than as an independent measure of accessibility.

The second mentioned accessibility metrics mentioned above is isochrone, or cumulative count measures which is probably the most used accessibility measures in practice (Miller, 2020) which is defined by the following equation:

$$a^{ip} = \sum_{j \in L_{T|i}^p} X_j^p$$

Where:

- $L_{T|i}^p$ is set of locations of type p that are within a maximum distance or travel time T of zone i.
- X_j^p is size of activity type p (number of stores, jobs, etc.) at location j.

This measure is consistent with the location choice model of the form:

$$P_j^{ip} = \frac{X_j^p}{\sum_{j' \in L_{T|i}^p} X_{j'}^p} \; if\ j \in L_{T|i}^p; = 0\ otherwise$$

The main advantage of this measure is it is easy to compute and understand especially using the Geographic Information System (GIS) software and databases. However, there are some serious theoretical and methodological problems that undoubtedly severely limit its general use, as detailed below.

First, the measure assumes that users are unconcerned about travel distances/times to competing places as long as they all fall under the threshold T. While it is debatable if individuals are indifferent to slight changes in short distances/times (e.g., whether one needs to travel 20 or 22 minutes to rival places), the notion that people would rather spend less time/effort traveling than more is key to travel behavior theory. Consider two extreme but still important edge cases: one in which a particular collection of places is exactly 30 minutes from a specific zone (for a situation where T = 30 minutes) and one in which these same places are 5-10 minutes distant from the zone. Clearly, the latter instance provides a higher level of accessibility than the former.

Second, and similarly, the measure implies that any sites placed outside the threshold (even by an infinitesimal amount ε) are irrelevant to location i's accessibility and that the chance of visiting these locations is zero. There are no places. This is definitely incompatible with travel behavior theory. Consider the edge situation (again with T = 30 minutes) in which one scenario consists of a group of places all located 29.9 minutes from zone i and another in which this identical set of locations is placed 30.1 minutes distant. Is zone i's accessibility truly zero in the second example (or is it significantly different than in case one)? The obvious response is no.

Consider the scenario when all locations within a certain distance are the same size, X. The chance of selecting a location in this situation is simply 1/N for any location within the threshold (where N is the number of locations) and zero otherwise. A far more behaviorally realistic (and empirically verifiable) model is one in which the chance of choosing a place decreases as distance/travel time increases. This drop may begin slowly (relative indifference across places with similar, short distances/time), then accelerate in the vicinity of some threshold (at which point the travel impedance becomes progressively and discernibly onerous), and finally become vanishingly tiny beyond some point. Clearly, an accessibility metric that is compatible with this behavior should be favored over the unrealistic isochrone behavioral assumption.

The third mentioned measure was gravity or entropy measures which is linked back to (Hansen, 1959). In their simplest form the can be expressed as:

$$a^{ip} = \sum_{j \in L^{ip}} X_j^p \, f(d_{ij})$$

Where:

- L^{ip} is set of locations of type p in the choice set for zone i.

- $f(d_{ij})$ is the impedance function; $\frac{\partial f}{\partial d_{ij}} < 0$

Two key differences between isochrone and gravity approaches are: The gravity technique weights the attraction of places by the impedance function: nearby locations are weighted more strongly than farther distant locations. Rather than an arbitrary cut-off criterion, the choice set, L^{ip} determines the set of locations evaluated in the accessibility computation. Thus, the isochrone measure is a special case of gravity measure in which $f\left(d_{ij}\right) = 1$ and $L^{ip} = L^{p}_{T|i}$

So, the gravity measure equation is consistent with a location choice model of the form:

$$P_j^{ip} = \frac{X_j^p f(d_{ij})}{\sum_{j' \in L^{ip}} X_{j'}^p f(d_{ij'})}$$

From a behavioral standpoint, this equation generates a constantly declining choice probability with increasing distance/time and so provides a major improvement over the isochrone technique.

The last important category of accessibility measure taken into consideration comes from random utility theory. The modeling of discrete choices is addressed by the extension of neo-classical microeconomic theory known as random utility theory, which takes into account the probabilistic character of these choices from the modeler's perspective. Predicting a person's selection of one alternative from a group of plausible discrete options is the main challenge. Classic instances of this issue in travel demand models include selecting a particular travel mode and/or destination.

The multinomial logit model (MNL), which has the following general form for the case of a destination choice model, is by far the most prevalent type of random utility model.

$$P_j^{ip} = \frac{e^{V_j}}{\sum_{j' \in L^{ip}} e^{V_{j'}}} = \frac{e^{\beta Z_j}}{\sum_{j' \in L^{ip}} e^{\beta Z_{j'}}}$$

Where:

- $V_j = \beta Z_j$ is the systemaic utility of alternative j
- Z_j is the vector of explanatory variables
- β is the (row) vector of parameters

The actual perceived utility by a decision maker is:

$$U_j = V_j + \varepsilon_j$$

Where ε_j is the individual's idiosyncratic deviation in terms of how they perceive the utility of alternative j relative to the population average utility V_j. The person chooses the alternative that generates the maximum perceived utility U_j.

This actual perceived maximum utility is unobservable, but, for the case of the MNL model, it can be shown (Ben-Akiva et al., 1985) that the expected maximum utility (I^{ip}) associated with this choice is given by:

$$I^{ip} = E\left[MAX_j\left(U_j\right)\right] = \ln(\sum_{j\in L^{ip}} e^{\beta Z_j})$$

That is, it is the natural logarithm of the denominator of the logit choice model (commonly referred to by the term “logsum”). Further, it can also be shown that this expected maximum utility is the consumer’s surplus for this choice. Thus, it is a standard measure of economic benefit. Given this, (Ben-Akiva et al., 1985) argue that it also provides a behaviorally and economically sound definition of accessibility: “accessibility for a given activity is the expected utility that would be derived from participation in this activity, which is also the consumer surplus associated with this participation”. That is:

$$a^{ip} = \ln(\sum_{j\in L^{ip}} e^{\beta Z_j})$$

(Geurs and Van Wee, 2004) show how an accessibility measure should contain numerous essential factors to provide a complete evaluation, such as the layout of the transportation system, the distribution of houses and activities, and the time and economic constraints on journeys, in their accessibility review. They also show variations in accessibility throughout peak and off-peak hours, rivalry for activities (for example, if accessibility is assessed for more job-seeking people than available employment), and variance in the population's capacity to use the transportation system. Such a comprehensive evaluation might be impossible to compute and express. They conclude that it is necessary to be aware of the diverse nature of accessibility. However, select a less sophisticated measure to assess challenges when one feels the chosen measure represents the significant or knowable difference while understanding its limits.

2.3.Connection between accessibility and equality

Accessibility equity can be seen in two terms: horizontal and vertical equity. They are extensively used in the literature. Horizontal equity is when everyone gets their fair portion of the pie. Vertical equity refers to equality among unequal, in which equity is measured between groups based on socio-economic factors that may influence group members' need for or usage of public transportation, such as income, automobile ownership, age, and activity status. Horizontal equity, for example, might refer to inhabitants (of any socio-economic position) living within a reasonable walking distance of their first public transportation station. Vertical equity, on the other hand, assesses whether low-income households have the same walking distance as high-income inhabitants.

There are examples of horizontal and vertical equity in the literature on accessibility equity. (Lucas, 2012, Banister, 2018) persuasively argue that when analyzing the situation for the most disadvantaged, it is critical to examine the compounding impacts of variables such as poverty, disability, economic and social exclusion, disposable time, and unemployment. This suggests that sophisticated metrics or combinations of measures to represent accessibility are desirable when capturing vertical equity. Horizontal equity, on the other hand, might be measured using more easy accessibility indicators.

Public transport accessibility is one of the main issues in urban planning. In the design phase of the transit network, researchers are often more focused on minimizing the operator and user costs and less on access and equity (Murray and Berwick, 2003). Availability of infrastructure, simplicity of information, and cost and time minimization are the critical factors in designing attractive public transport and door-to-door access (Yatskiv et al., 2017).

Because the lack of access to public transit might imply social exclusion, land use and transport policies focus on accessibility to enable people to reach their destinations at a reasonable time and cost (Hawas et al., 2016). So, performing good, efficient, and accessible public transport is one of the main aims of policymakers and planners worldwide (Saghapour et al., 2016).

In our work we are going to focus on horizontal equity. We use Lorenz curve (Lorenz, 1905) to examine the distribution of accessibility and the Gini (Gini, 1912) scalar to obtain a single value reflecting the unevenness of the distribution which is completely described in 3.2.2.1 and 3.2.2.2. However, there is a large body of literature on Gini index, conceptualized by (Gini, 1912) with derivations by .

There are also tools to assess vertical inequity like the Suits coefficient. Suits (Suits, 1977) showed that by ordering the population on the x-axis by increasing income as in Figure 2-1, rather than by the amount of share of accessibility received as in Figure 2-2, one can construct a measure of how income-discriminatory a distribution is, i.e., if the distribution is benefitting low-income members more than high-income members of the population. In Figure 2-1 shows an example of how a Lorenz curve can look like when the population is ordered as Suits suggests. Note that in this setting the Lorenz curve can go above the diagonal OB since members of the population can receive high values of share of accessibiliy but be ordered to the left due to their low incomes. If K is the area of the triangle OAB, and P is the area under Lorenz curve contained by OAB, then the Suit index can be formulated as:

$$S = \frac{(P - K)}{K} = -1 + \frac{P}{K}$$

This formula takes values between − 1 and 1. Positive Suits values indicate a progressive distribution of accessibility, while negative values indicate a regressive distribution. In the progressive extreme case, all accessibility is received by the member with the lowest income, making L trace a line through OCB, yielding Suits = − 1 + 2 = 1. Alternatively, in the extreme regressive case, the member with the highest income receive all the share of accessibility and the Lorenz curve trace the line through OAB with Suits = − 1 + 0 = − 1. When L traces the diagonal OB, P is equal to K and Suits = − 1 + 1 = 0 which is the proportional case, i. e., the amount of accessibility that a group has is proportional to the groups share of the population and different for different income levels.

Figure 2-1 Lorenz curve for computing the Suits coefficient (Rubensson et al., 2020)

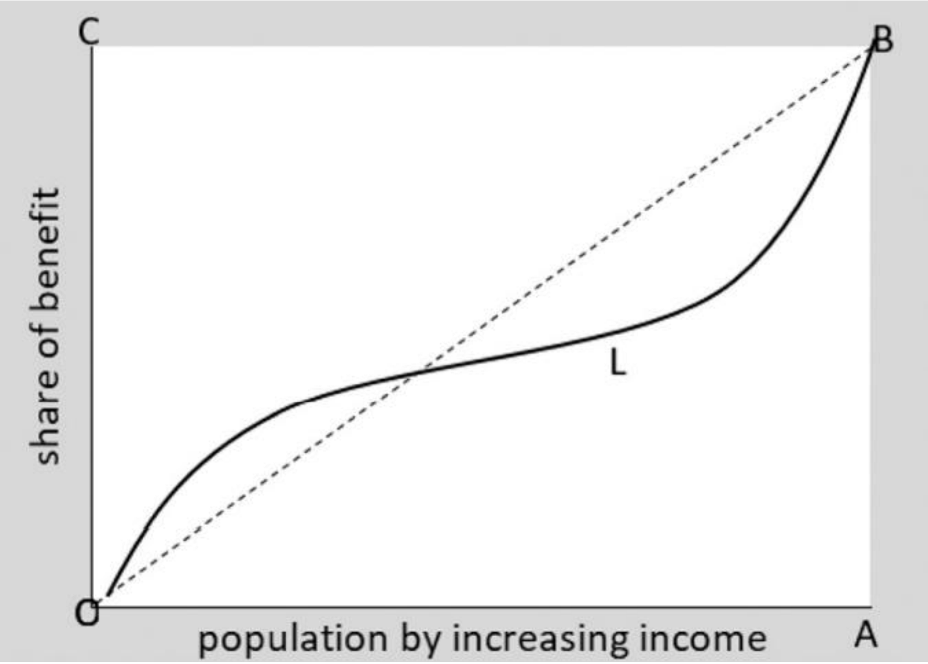
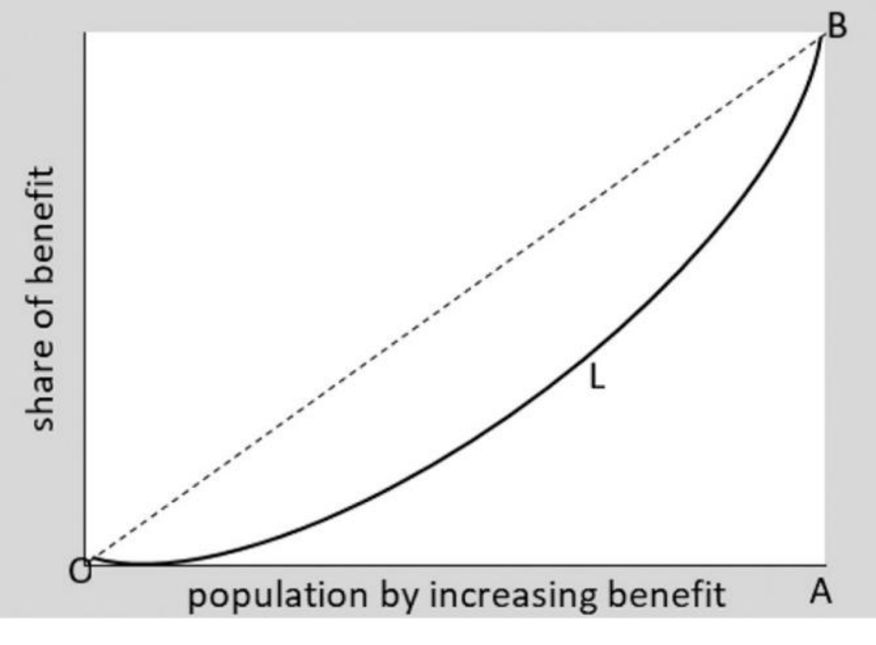


Figure 2-2 Lorenz curve for computing the Gini coefficient (Rubensson et al., 2020)



However, equality is not the only priority for policymakers. In the supply of accessibility and distribution, there are many other quality and performance considerations. What should the public transportation system look like and its goals? Aside from horizontal and vertical equality, the transportation system should run smoothly with a reasonable fare box recovery. It should also provide a dependable alternative for vehicle traffic during peak hours in the most congested areas. Proposed policy changes should be acceptable from a cost-benefit analysis (CBA) and an equality standpoint (Niehaus et al., 2016). (Golub and Martens, 2014), as previously mentioned, present an example of how to evaluate a policy proposal in the latter meaning. Another approach is proposed by (Wei et al., 2017), which addresses the problem of maximizing operational efficiency while also increasing public transportation stop coverage (accessibility) for several disadvantaged groups (elders, children, carless households, unemployed, disabled, poor, nonwhite).

3. Chapter 3: Database and Methodology

3.1. Open-access input data

3.1.1. General Transit Feed Specification (GTFS)

The General Transit Feed Specification (GTFS) is a data standard that enables public transportation providers to publish transit data in a format many software applications can ingest. As a result, thousands of public transportation operators now employ the GTFS data standard. The GTFS system is divided into two parts: a schedule component with timetable, fare, and geographic transit information and a real-time component with arrival estimates, vehicle locations, and service alerts (Google Transit Feed, 2022).

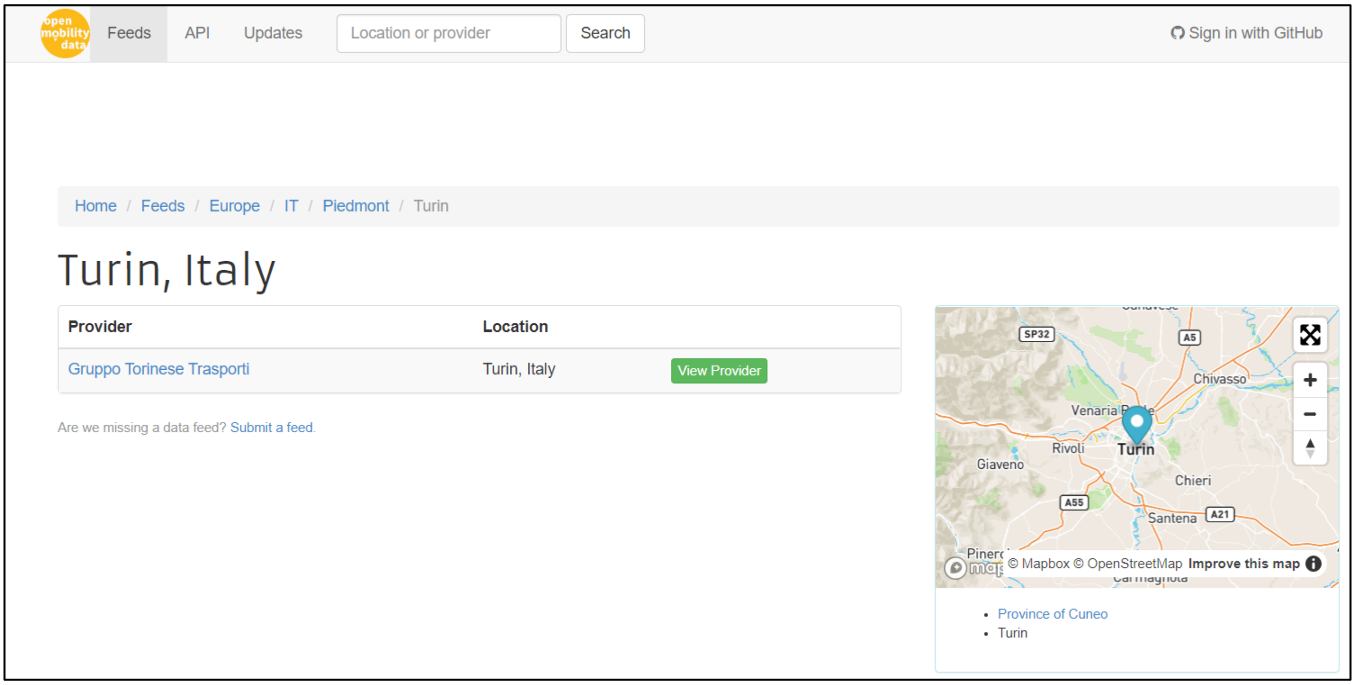
GTFS file of one city is a ZIP file containing comma-delimited text files that reflect fixed-route schedule, route, and bus stop data. The necessary text files list and what they show are in Table 3-1 (Google Transit Feed, 2022). GTFS data for cities of our interest can be downloaded from OpenMobilityData[1] by only searching the city of interest and choosing the respective authority of transport as an example in Figure 3-1 for Turin.

Table 3-1 GTFS Files Definition

Filename	Required or not?	Short description
agency.txt	Required	Transit agencies with service represented in this dataset.
stops.txt	Required	Stops where vehicles pick up or drop off riders. Also defines stations and station entrances.
routes.txt	Required	Transit routes. A route is a group of trips that are displayed to riders as a single service.
trips.txt	Required	Trips for each route. A trip is a sequence of two or more stops that occur during a specific time period.
stop_times.txt	Required	Times that a vehicle arrives at and departs from stops for each trip.
calendar.txt	Conditionally required	Service dates specified using a weekly schedule with start and end dates. This file is required unless all dates of service are defined in calendar_dates.txt.
calendar_dates.txt	Conditionally required	Exceptions for the services defined in the calendar.txt. If calendar.txt is omitted, then calendar_dates.txt is required and must contain all dates of service.
feed_info.txt	Conditionally required	Dataset metadata, including publisher, version, and expiration information.

[1] https://transitfeeds.com/

Figure 3-1 How to Download GTFS data for Turin



Home / Feeds / Europe / IT / Piedmont / Turin

Turin, Italy

Provider	Location	
Gruppo Torinese Trasporti	Turin, Italy	View Provider

Are we missing a data feed? Submit a feed.

- Province of Cuneo
- Turin

As it can be seen, the GTFS data of a particular urban area consists of several files, among which "stops.txt" shows the exact location of each transit stop (either bus stop or metro/train station) and "stop_times.txt" shows which line is serving each stop and at which time, which allows reconstructing the trajectory of each transit vehicle

3.1.2. Population Grid

3.1.2.1. Population density of European cities

The JRC-GEOSTAT 2018 is a regular grid map of 1 × 1 km pixels that shows the number of people living in Europe in 2018 in each square of the grid. It was created in the second part of 2020 by the European Commission Joint Research Centre, at the request of DG REGIO, with the help of Eurostat, as a follow-up to Eurostat's previous GEOSTAT editions from 2006 and 2011. The need to update population distribution information at high spatial resolution and insight into recent population changes at the local level was a significant motivator for the JRC-GEOSTAT 2018 [2] (Eurostat, 2018).

3.1.2.2. Population density of non-European cities

The Gridded Population of the World, Version 4 (GPWv4): Demographic Density, Revision 11 contains estimates of human population density (number of people per square kilometer) for the years 2000, 2005, 2010, 2015, and 2020, based on counts congruent with national censuses and population registries. The population counts were assigned to 30 arc-second grid cells (i.e., squares of about 1 km at the equator)

[2] https://ec.europa.eu/eurostat/web/gisco/geodata/reference-data/population-distribution-demography/geostat

using a proportionate allocation gridding technique that utilized about 13.5 million national and sub-national administrative units. The population density raster was constructed by dividing the population count raster for a specific target year by the land area raster. The data files were created as a global raster with a resolution of 30 arc seconds. The 30 arc-second count data were aggregated to 2.5 arc-minute, 15 arc-minute, 30 arc-minute, and 1-degree resolutions to generate a density raster to enable quicker worldwide processing and help research communities (SEDAC, 2020)[3]

3.1.3. Employment Data

3.1.3.1. Italy

The Italian National Institute of Statistics (Istat; Italian: Istituto Nazionale di Statistica) is the country's primary official data source. Population censuses, economic censuses, and a variety of social, economic, and environmental surveys and analyses are among its many operations. Istat is Italy's major generator of statistical data and a participant in the European Statistical System, which Eurostat oversees. The data of the general population and housing censuses and the industry and services censuses are published, which can be associated, through connection codes, with the partitions of the system of territorial bases (ISTAT, 2011)[4]. The description of the data used is in Table 3-2.

[3] https://sedac.ciesin.columbia.edu/data/set/gpw-v4-population-density-rev11/data-download

[4] https://www.istat.it/it/archivio/104317 → $Variabili\ censuarie$

Table 3-2 Metadata of industry and service

Field name	Short description
TIPO_SOGGETTO	NP ': non-profit institution -' IP ': public institution -' IM ': companies
CODREG	Numeric code that uniquely identifies the region within the national territory
REGIONE	Denomination of the region.
CODPRO	Numeric code that uniquely identifies the province within the national territory.
PROVINCIA	Name of the province.
CODCOM	Numeric code that uniquely identifies the municipality within the provincial territory.
COMUNE	Name of the municipality.
PROCOM	"Numeric code that uniquely identifies the municipality within the national territory. The value is obtained by concatenating the CODPRO field with the three-digit CODCOM field. "
SEZ2011	"Numeric code uniquely identifies the 2011 census section within the national territory. The value is obtained by concatenating the PROCOM field with the 7-digit NSEZ field. "
NSEZ	A number that uniquely identifies the 2011 census section within the municipal area.
ACE	A number that uniquely identifies the census area within the municipal area.
CODLOC	"Numeric code that identifies the 2011 location within the municipal area.
CODASC	Numeric code that uniquely identifies the sub-municipal area, where present, within the municipal area.
NUM_UNITA	The number of local units.
ADDETTI	Number of employees
ALTRI_RETRIB	Number of other paid workers
VOLONTARI	Number of volunteers

3.1.3.2. Île-de-France

Employment data about Île-de-France can be directly downloaded from the open data portal website of the region[5]. Within the definition of the population census, the employed working population comprises those who have a job and therefore fall into one of the following categories (Île-de-France, 2017):

- exercise a profession (salaried or not) even on a part-time basis
- help a family member with their work (even without pay)
- be an apprentice, a paid intern
- be a student or retired but working
- contingent soldiers (as long as this situation existed).

Unemployed people looking for a job are therefore excluded.

3.2. Methodology

3.2.1. Methodology of accessibility assessment from (Biazzo et al., 2019)

3.2.1.1. Isochronic maps

According to the definition given in 2.1, accessibility of a particular place measures how easy it is for a passenger to travel by using transit, starting from that place and going to a random place (Note that the destination in this definition is not the opportunities in the city and all of the places can be considered as a destination and they could be accessed). The more connected a place to public transport services, the higher its accessibility. The overall accessibility of a place is the key to defining its cohesion and viability. Indeed, high accessibility allows for a better socio-economic organization at the local level, with a significant impact on urban life, including employment opportunities, training, and education. In this sense, accessibility is closely linked to populations' social cohesion and well-being. A connected place, close to stations where several lines pass and with high service frequency, will benefit more than others. To quantify the accessibility, we resort to (Biazzo et al., 2019).

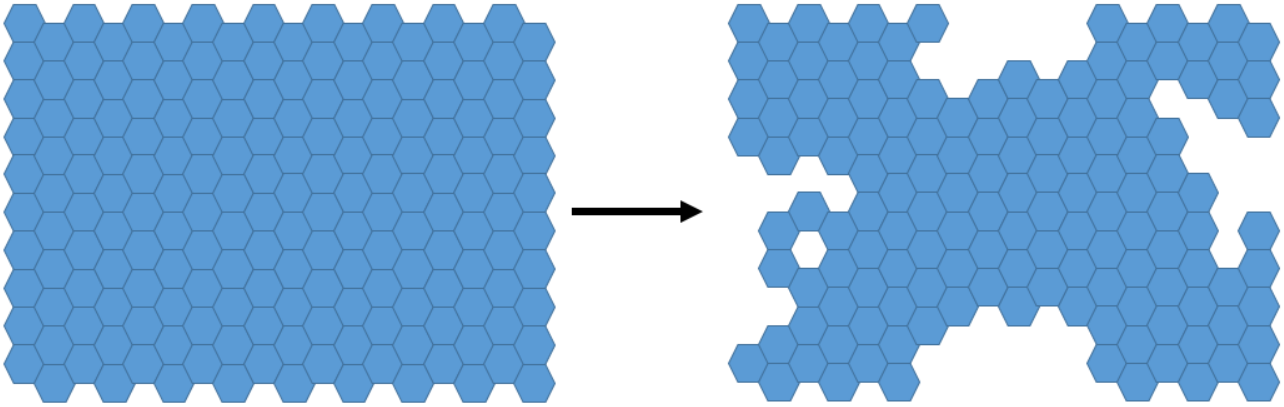
The accessibility measure is given in (Biazzo et al., 2019) is based on isochronic maps. The same authors have also developed a python script named ”public-transport-analysis” which is openly available on Github[6] we also used it to develop the following analyses. We partition the area under the study with a hexagonal tessellation $\lambda \in \Lambda$ each of 1 Km per side different from the original work of (Biazzo et al., 2019) that is 0.2 Km and this task is because by increasing the side from 0.2 to 1 Km, we could decrease the computation time drastically. The isochrone $I(\tau, (\lambda, t_0))$ is the area reachable from the center of the hexagon λ with the maximum travel time τ and the departure time t_0. It is important to note that hexagons do not cover the entire territory of the city. They cover all sections of a city with at least one public transportation station and all places reachable from any public transportation stop with a walking path less

[5] https://data.iledefrance.fr/explore/dataset/population-active-occupee-des-communes-dile-de-france-donnee-insee0/information/?location=11,48.93242,2.3909&basemap=jawg.streets

[6] https://github.com/CityChrone/public-transport-analysis

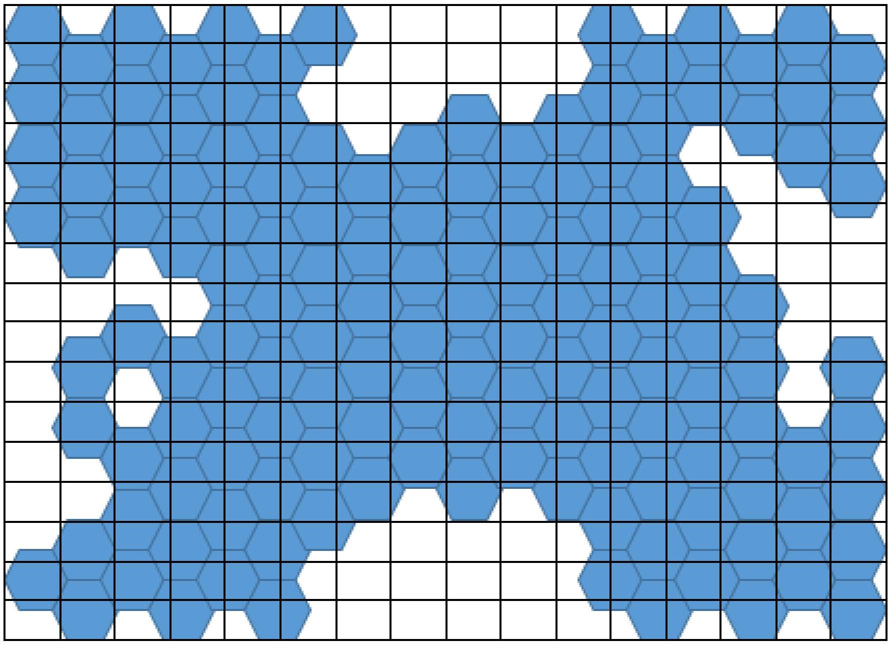
than 15 minutes, as in Figure 3-2. In order to compute the walking path between the center of hexagons and the public transportation stop, we use the backend version of the open-source routing machine (OSRM) (the description of implementing the tool in Python is available on Github[7])(Luxen and Vetter, 2011). The OSRM uses the OpenStreetMap network of the city where the analysis is being carried out to calculate the shortest walking pathways on each city's urban networks. In order to find the population of hexagons, we entered the population density data described in section 3.1.2 into the above introduced ”public-transport-analysis” script. These population density data applied to coarse-grained squares having a surface area of $1Km^2$. To match the size of the hexagons (about 2.6 Km^2) with the square's population density size, we split each square's population proportionately to the fraction of overlapping surface among the overlapping hexagons. As in Figure 3-3, we can see how hexagons and the population grid overlap. This process is done automatically by the above-mentioned python script from (Biazzo et al., 2019).

Figure 3-2 Process of Tessellation of the city of interest



[7] https://github.com/Project-OSRM/osrm-backend

Figure 3-3 Population grid and hexagons overlap

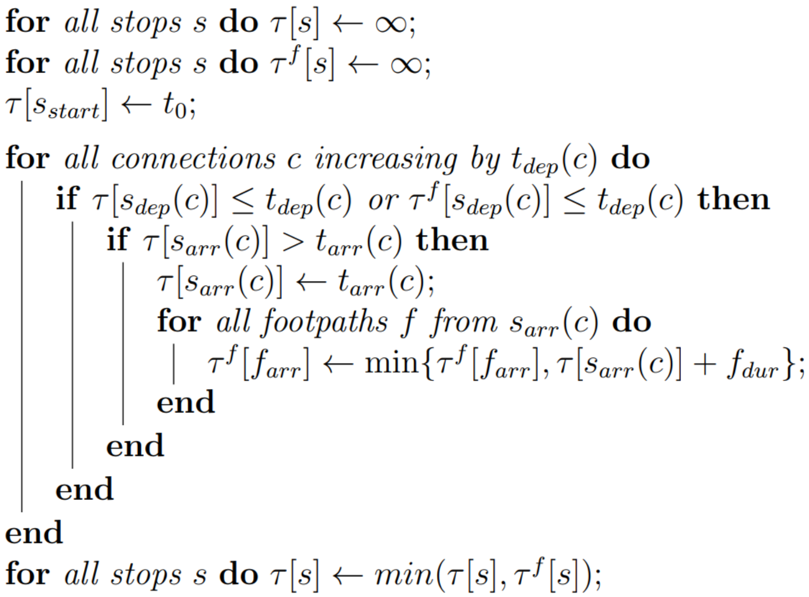


The final stage in producing isochronic maps is to combine the coarse-grained depiction of a city with the schedule of its public transportation system and compute journey times between any two hexagons of the tessellation at different times of day and/or different days of the week. The algorithm that has been used here is a modified version of the connection scan algorithm (CSA)(Dibbelt et al., 2013) which is called the Intransitive connection scan algorithm (ICSA). This algorithm is fully implemented in (Biazzo et al., 2019) work and the related pseudo-code implementation scheme is shown in Figure 3-4. Thanks to this modified algorithm, all the shortest-time paths linking the centers of every pair of hexagons in the tessellation at various starting times for a typical weekday can be computed. Each of these shortest-time paths will consider using public transit between two hexagons and the possibility of walking to neighboring hexagons to reach public transportation service locations within a particular area (Biazzo et al., 2019). The algorithm shown in Figure 3-4 works like the below:

- For each stop, we have a $1 \times n$ matrix, and n is the number of stops in the city; at first, we assign infinite to all arrays of this matrix
- Then, sort the connection file that comes directly from the "stop_times.txt" of the GTFS file that shows which line is serving each stop and at which time with the departure time.
- By a loop, we check that starting from our stop (a) which transit line we can take as soon as possible, and we take it, and we assign arrival time to the next stop (b) in the matrix of stop (a).
- Now, we check is it possible to go to stop (b) by walking with less time than using public transport. If yes, we assign the arrival time to reach stop (b) using walking in the matrix of stop (a)

- Again, with the loop, we check starting from stop (b) which connection we can take as soon as possible, and then we take it and arrive to stop (c) and assign the arrival time to the matrix of stop (a) and also check the walking for shorter time path and so on.
- Finally, we have a matrix for stop (a), which shows how much it takes to reach from this stop to all other stops.
- We repeat the same procedure for other stops of public transit network.

Figure 3-4 Intransitive Connection Scan Algorithm (Biazzo et al., 2019)



```
for all stops $s$ do $\tau[s] \leftarrow \infty$;
for all stops $s$ do $\tau^f[s] \leftarrow \infty$;
$\tau[s_{start}] \leftarrow t_0$;
for all connections $c$ increasing by $t_{dep}(c)$ do
    if $\tau[s_{dep}(c)] \leq t_{dep}(c)$ or $\tau^f[s_{dep}(c)] \leq t_{dep}(c)$ then
        if $\tau[s_{arr}(c)] > t_{arr}(c)$ then
            $\tau[s_{arr}(c)] \leftarrow t_{arr}(c)$;
            for all footpaths $f$ from $s_{arr}(c)$ do
                $\tau^f[f_{arr}] \leftarrow \min\{\tau^f[f_{arr}], \tau[s_{arr}(c)] + f_{dur}\}$;
            end
        end
    end
end
for all stops $s$ do $\tau[s] \leftarrow min(\tau[s], \tau^f[s])$;
```

3.2.1.2. Accessibility scores

As in (Biazzo et al., 2019), we have two types of accessibility metrics that make comparing different parts of the city or the cities easier, namely velocity score and sociality score. In addition, we have defined another score also to consider the opportunities in the cities called attraction score. Accessibility scores aim to quantify the operation of public transit to connect places and people.

3.2.1.2.1. Velocity score

Velocity score $\nu(\lambda)$ can be defined as the speed of expansion, i.e., the average speed at which it is possible to move from the center of the hexagon and go toward a random direction using public transportation. The movement of people can be described by an origin-destination matrix (ODM). The computation of the velocity score assumes a uniform ODM. In order to compute the velocity score, consider the isochrone $I(\tau, (\lambda, t_0))$ which is centered at hexagon λ at the time t_0 corresponding to travel time τ. The covered area $A(\tau, (\lambda, t_0))$ of isochrone at time τ will be the area within the $I(\tau, (\lambda, t_0))$. In order to have the average traveled distance $\bar{r}$ from the center of hexagon λ toward a random direction, we approximate the area of isochrone $A(\tau,(\lambda,t_0))$ by a circle (Figure 3-5).

$$\bar{r}\big(\tau,(\lambda,t_0)\big) = \sqrt{\frac{A(\tau,(\lambda,t_0))}{\pi}}$$

And then, by dividing the average travel distance $\bar{r}$ by the travel time τ, we can compute the average expansion speed at time τ of a circular isochrone with the same area as the real one. This quantity can also be considered the average trip velocity of duration τ toward a random direction from the center of the hexagon λ. This quantity can be computed for any departure time t_0, travel time τ and for any hexagons of the tesselation. Therefore, the velocity score can be computed by averaging over departure time t_0. Additionally, it is important to define a new variable T as the maximum possible travel time in order to consider all the possible travel time between 0 and T and using integration and not only consider one specific travel time.

$$v(\lambda) = \frac{\sum_{t_0=t_{start}}^{t_{end}} \int_0^T v\big(\tau,(\lambda,t_0)\big)d\tau}{number\ of\ departure\ instants}$$

Here velocity score indicates a measure of the average speed at which an individual can move away from a hexagon λ, in a randomly chosen direction starting with departure time t_0.

Figure 3-6 and Figure 3-7 show an example of the spatial distribution of the Velocity score in Île-de-France for illustrative purposes, whereas the full results will be presented entirely later.

Figure 3-5 Isochrones with hexagonal tessellation at different times(Biazzo et al., 2019)

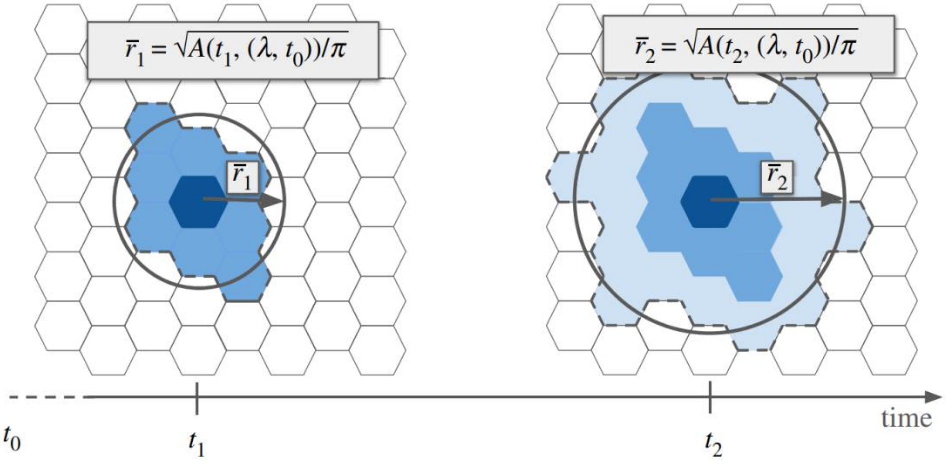

$\bar{r}_1 = \sqrt{A(t_1,(\lambda,t_0))/\pi}$
$\bar{r}_2 = \sqrt{A(t_2,(\lambda,t_0))/\pi}$
$\bar{r}_1$
$\bar{r}_2$
time
t_0
t_1
t_2

Figure 3-6 Velocity Score of Île-de-France

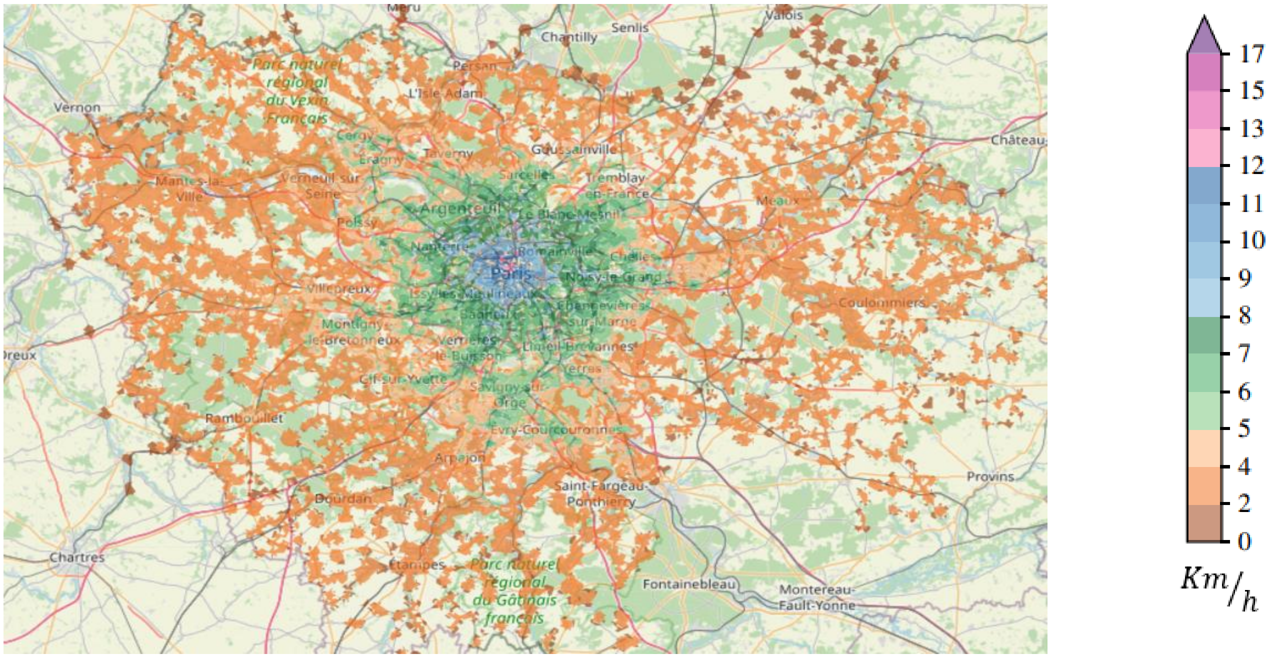
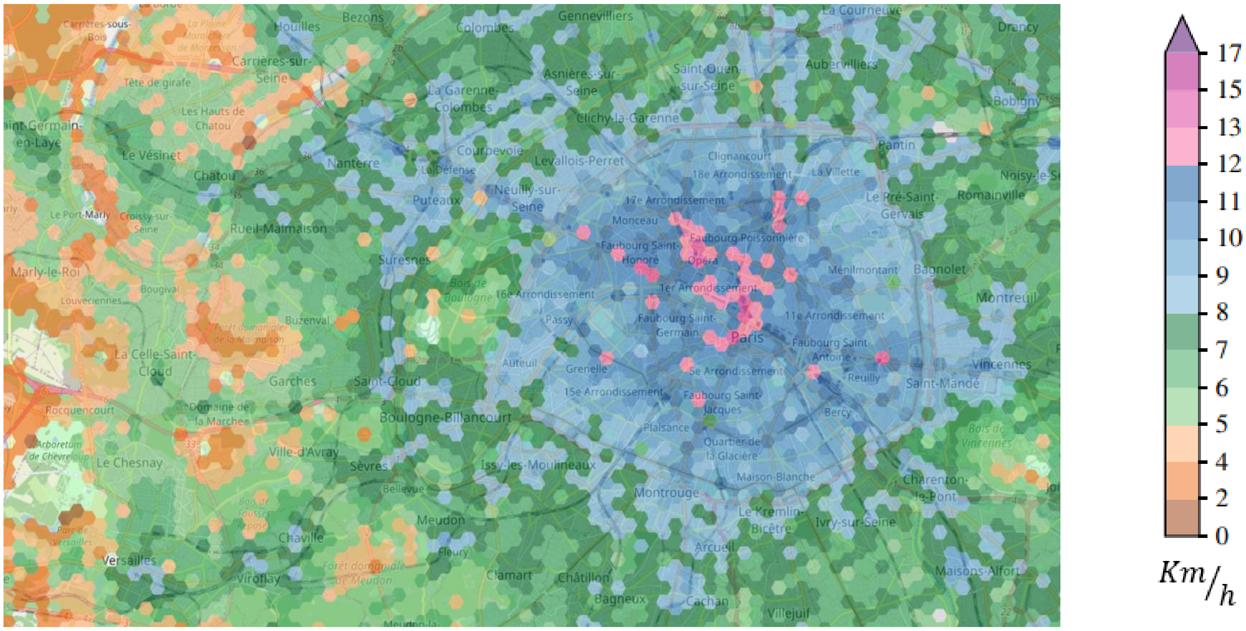


Figure 3-7 Zoomed Version of Velocity Score of Île-de-France



3.2.1.2.2. Sociality score

The above-mentioned velocity score measures how well the public service facilitates the rapid exploration of urban space. However, there is a strong interplay between the population density and the efficiency of public transport. Therefore, while improving service in densely populated areas is common, low-density locations run the danger of being underserved by public transit. The second score has been introduced to take into account this interplay, which quantifies the goodness of public transit in connecting people. Here we consider $P(\tau,(\lambda,t_0))$ as the population residing in the isochrone $I(\tau,(\lambda,t_0))$. Consequently, like what we have done for the velocity score, the sociality score obtain as

$$s(\lambda) = \frac{\sum_{t_0=t_{start}}^{t_{end}} \int_0^T P\big(\tau,(\lambda,t_0)\big)d\tau}{number\ of\ departure\ instants}$$

Here the sociality score indicates the number of individuals that may be reached in a single trip starting from hexagon λ at departure time t_0 and with the duration of τ. for example, in Figure 3-8, an individual start his/her trip from the hexagon highlighted in green in t_0 and with the travel duration of τ he/she can reach the hexagons highlighted in orange, so the number of individuals that is possible to reach is the sum of the population of the green hexagon and orange hexagons.

Figure 3-9 and Figure 3-10 show an example of the spatial distribution of the sociality score in Île-de-France, and the full results will be presented entirely later.

Figure 3-8 presentations of one possible trip

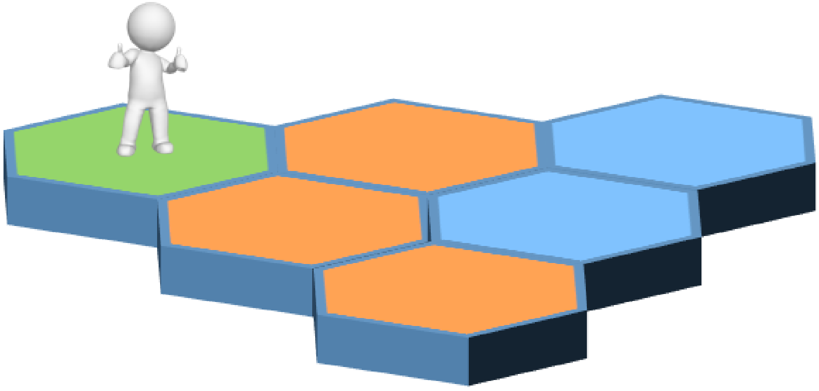


Figure 3-9 Sociality Score of Île-de-France

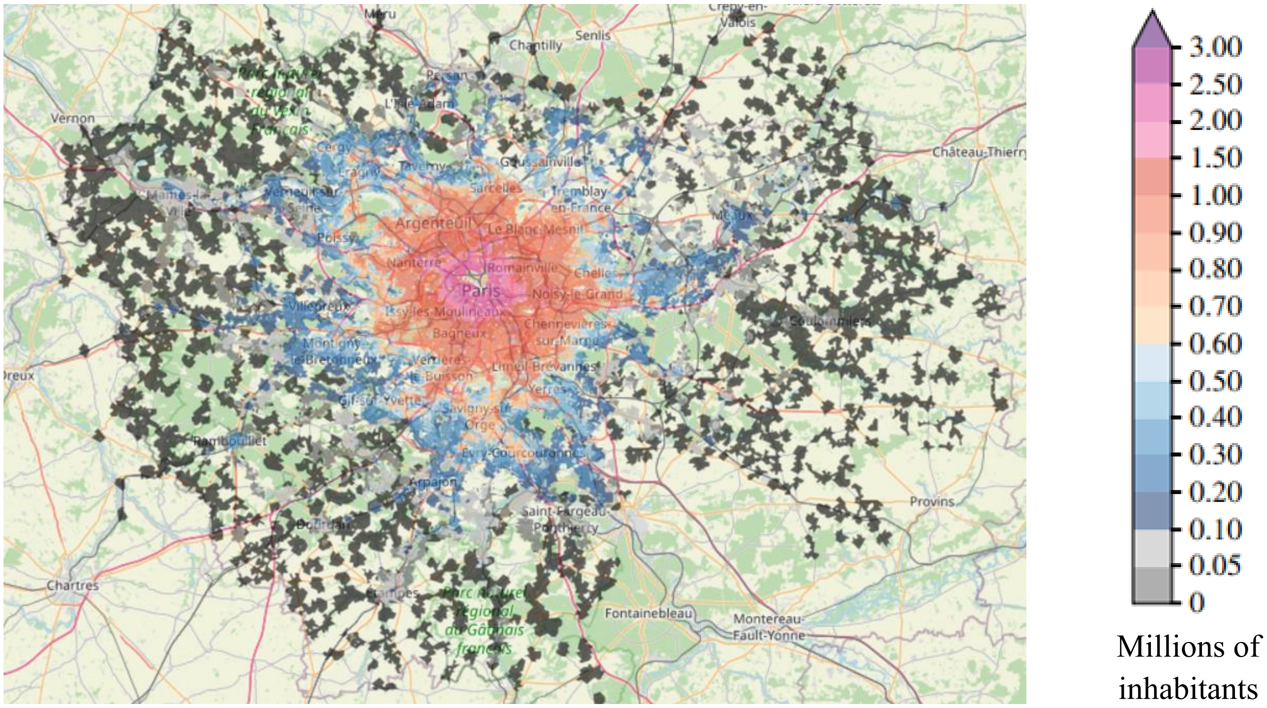
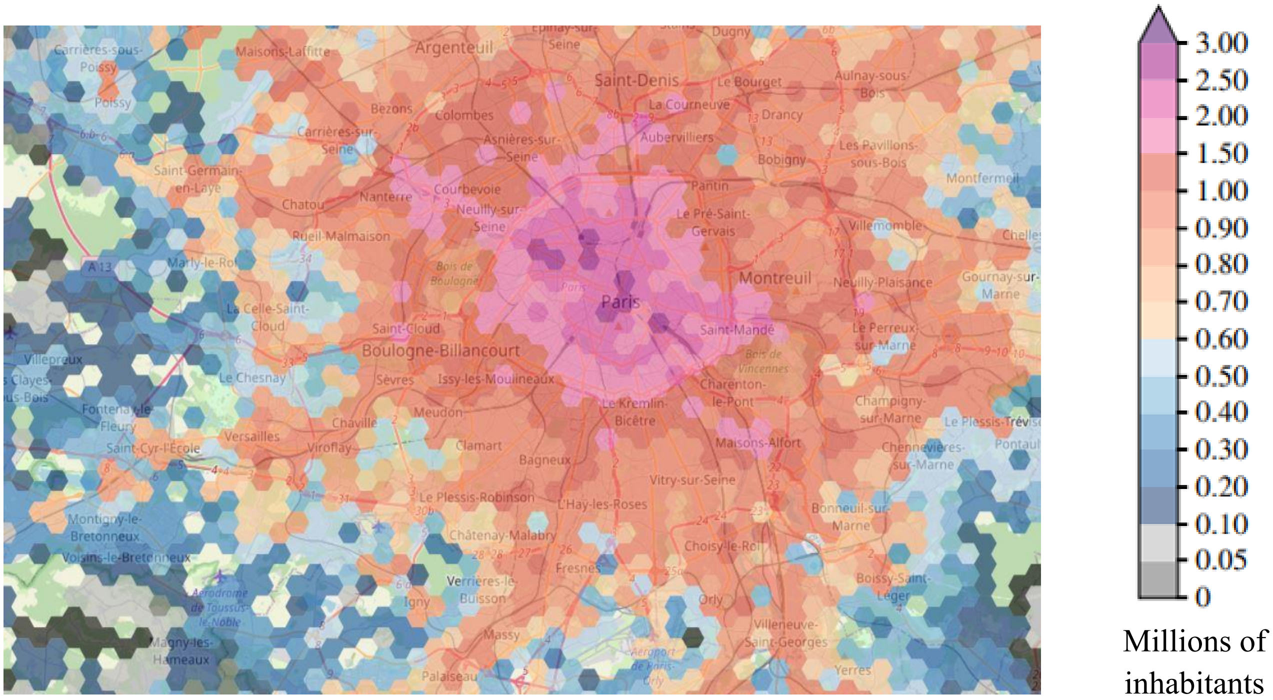


Figure 3-10 Zoomed version of Sociality Score in Île-de-France



3.2.1.2.3. Attraction score

We have also added another score to the work of (Biazzo et al., 2019) called the attraction score. By considering the population in sociality score, we consider the individuals that may start a trip from their place of residence but do not consider the opportunities in the city that may be interesting for those individuals to start their trip. Because, according to 2.1 most of the definitions of accessibility depends also on the type of the opportunities. So, in order to also consider these opportunities in our computation, we also introduced the attraction score. Here we consider $D(\tau, (\lambda, t_0))$ as the number of jobs in the isochrone $I(\tau, (\lambda, t_0))$. Consequently, like what we have done for the other two scores, the attraction score obtain as

$$\mathcal{R}(\lambda) = \frac{\sum_{t_0=t_{start}}^{t_{end}} \int_0^T D\big(\tau, (\lambda, t_0)\big) d\tau}{number\ of\ departure\ instants}$$

Therefore, the attraction score indicates the number of workplaces that may be reached in a single trip starting from hexagon λ at departure time t_0 and with a travel time of τ.

Figure 3-11 and Figure 3-12 shows an example of the spatial distribution of the attraction score in Île-de-France, and the full results will be presented entirely later. Trip definition is same here as 3.2.1.2.2, but instead of population, we sum the number of jobs.

Figure 3-11 Attraction Score of Île-de-France

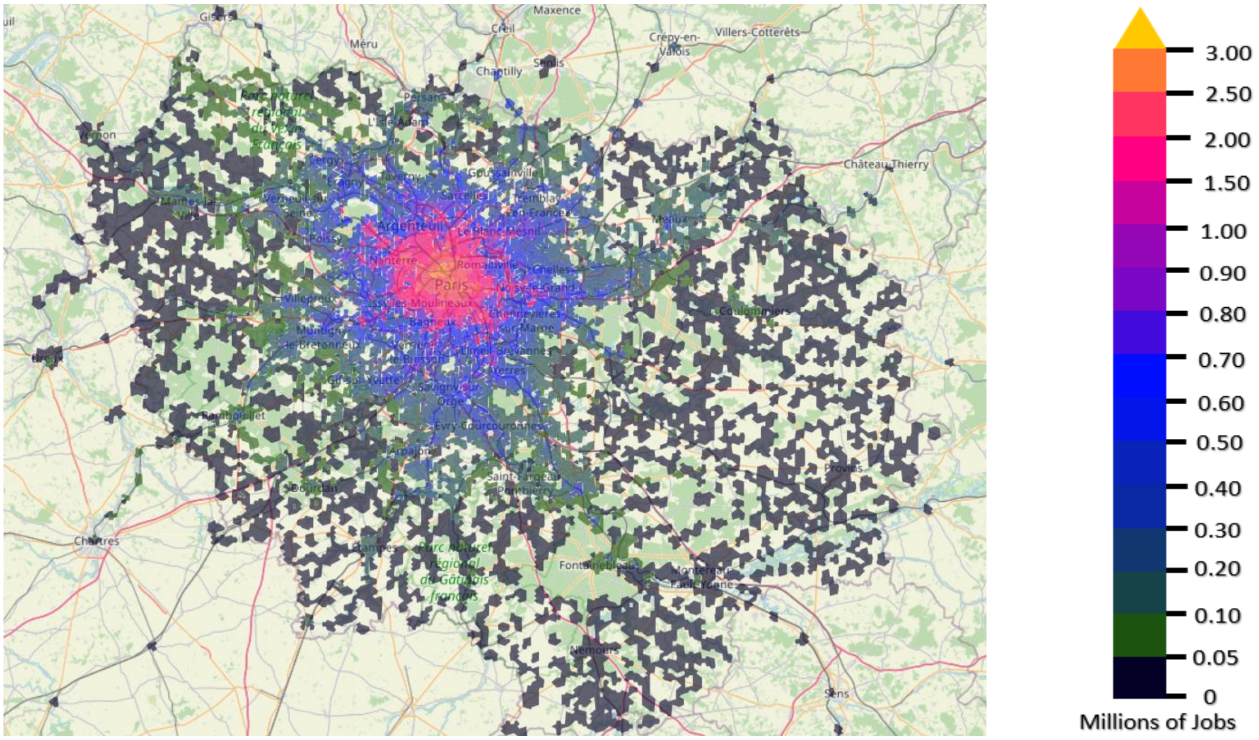
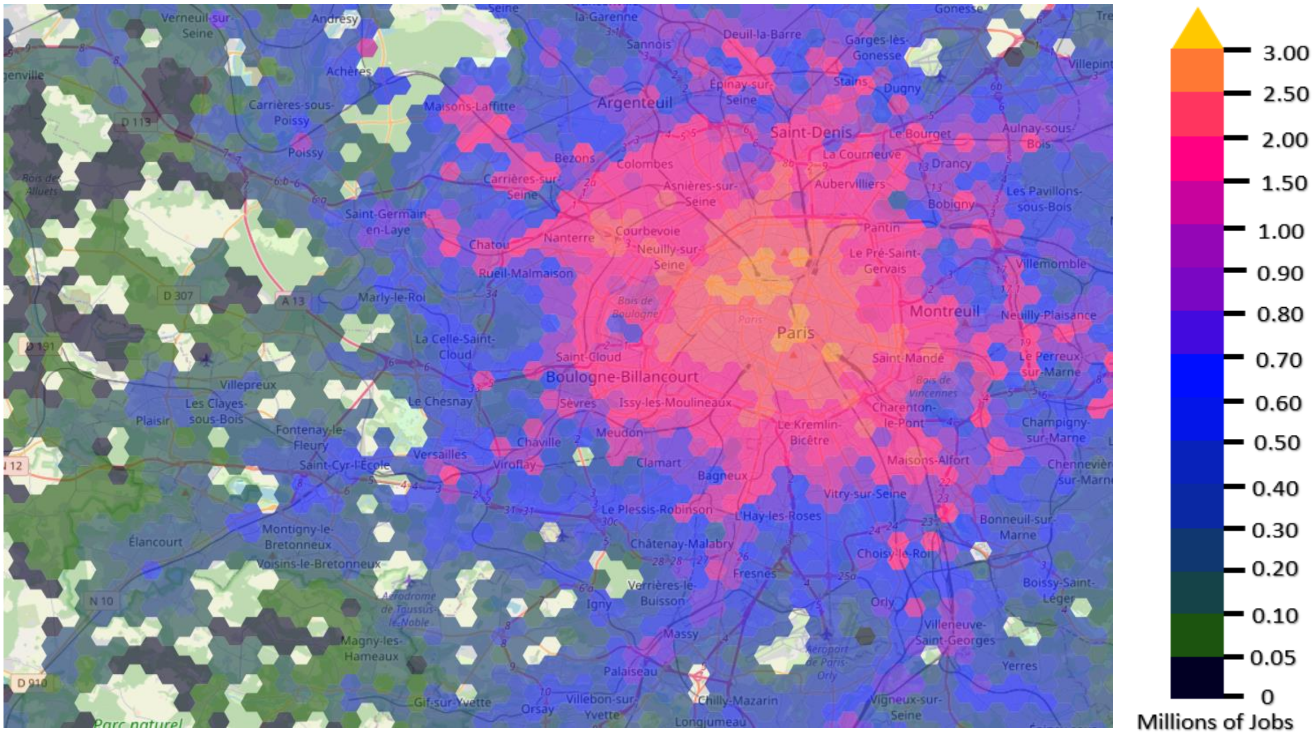


Figure 3-12 Zoomed version of Attraction Score of Île-de-France

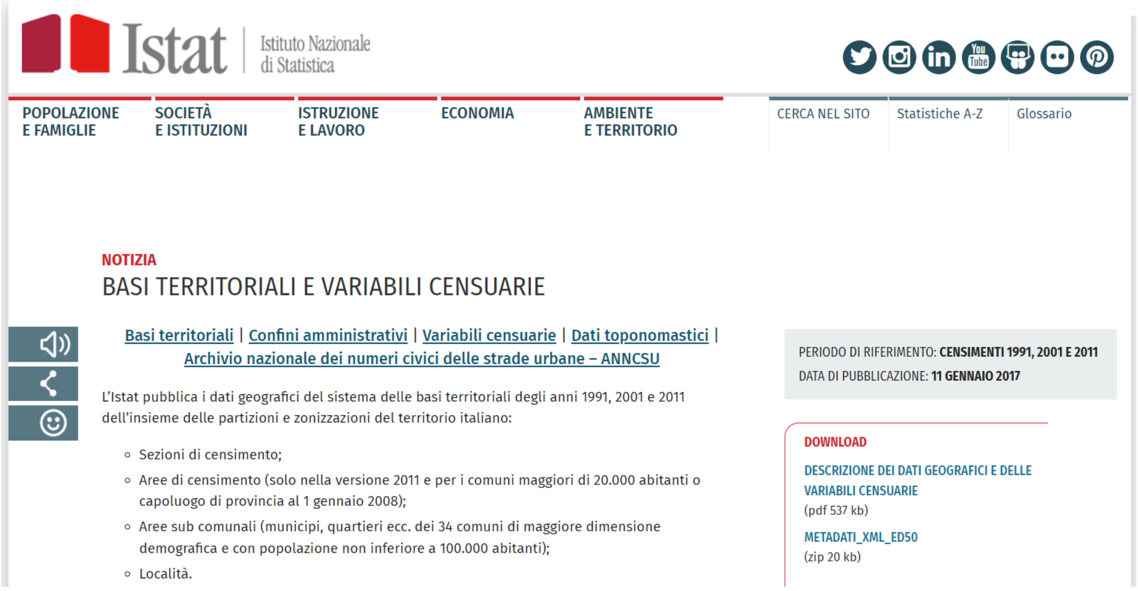


3.2.1.2.3.1. Data processing for Attraction score

Here, we show the procedure for Turin. As described in 3.1.3.1, we can download the data from (ISTAT, 2011). One problem is that the website is in Italian (at least the page for the census data). Like in Figure 3-13, we should download the data about industry and services (Censimento dell'industria e dei servizi[8]).

[8] https://www.istat.it/it/archivio/104317

Figure 3-13 ISTAT Census data webpage



NOTIZIA

BASI TERRITORIALI E VARIABILI CENSUARIE

Basi territoriali | Confini amministrativi | Variabili censuarie | Dati toponomastici | Archivio nazionale dei numeri civici delle strade urbane – ANNCSU

L'Istat pubblica i dati geografici del sistema delle basi territoriali degli anni 1991, 2001 e 2011 dell'insieme delle partizioni e zonizzazioni del territorio italiano:

- Sezioni di censimento;
- Aree di censimento (solo nella versione 2011 e per i comuni maggiori di 20.000 abitanti o capoluogo di provincia al 1 gennaio 2008);
- Aree sub comunali (municipi, quartieri ecc. dei 34 comuni di maggiore dimensione demografica e con popolazione non inferiore a 100.000 abitanti);
- Località.

PERIODO DI RIFERIMENTO: **CENSIMENTI 1991, 2001 E 2011**
DATA DI PUBBLICAZIONE: **11 GENNAIO 2017**

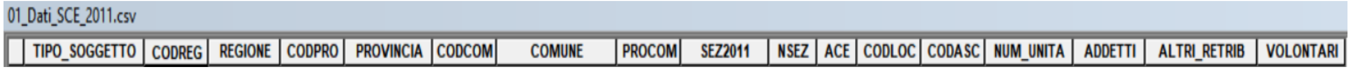
DOWNLOAD

DESCRIZIONE DEI DATI GEOGRAFICI E DELLE VARIABILI CENSUARIE
(pdf 537 kb)

METADATI_XML_ED50
(zip 20 kb)

The downloaded data is in old Excel format, which is not working in ArcGIS software. So, we have to convert this data to CSV format and extract the part about Piemonte. Then, we can add this file to the ArcMap software. Unfortunately, this file does not have any accurate geographic data about the Census sections as in Figure 3-14.

Figure 3-14 Head row of the industry and services in Piemonte


01_Dati_SCE_2011.csv

TIPO_SOGGETTO	CODREG	REGIONE	CODPRO	PROVINCIA	CODCOM	COMUNE	PROCOM	SEZ2011	NSEZ	ACE	CODLOC	CODASC	NUM_UNITA	ADDETTI	ALTRI_RETRIB	VOLONTARI

Therefore, we use another data from ISTAT: territorial bases census sections of Piemonte[9]. There are two similar columns in these two data: "SEZ2011" and by using this column, we can join these two layers and have the job places data for Turin, and at the end, we can add this finalized shapefile to the "public-transport-analysis" (Biazzo et al., 2019).

3.2.2. Assessing accessibility inequality

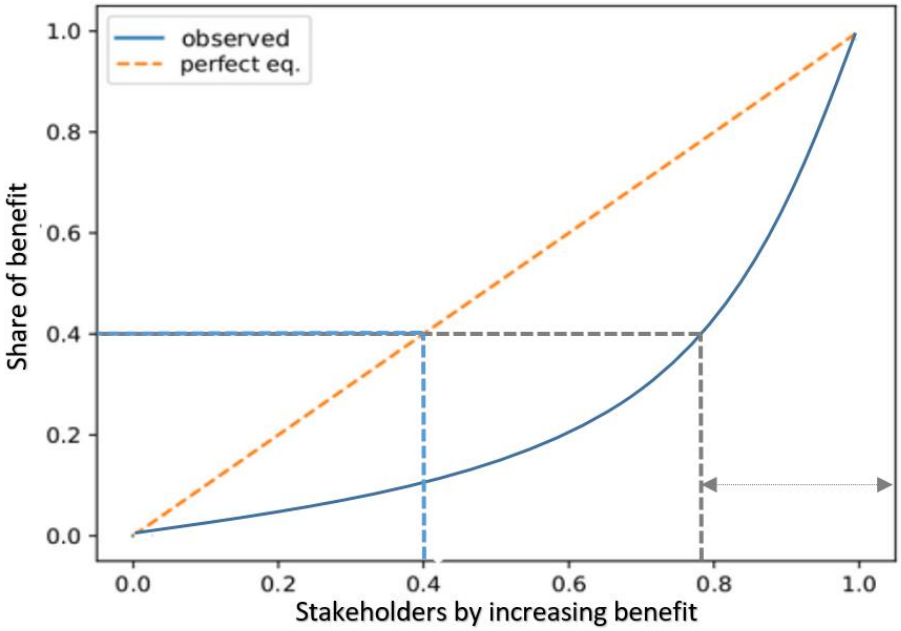
3.2.2.1. Lorenz curve

Lorenz curves are a graphical depiction of the population's cumulative distribution function of wealth in economics (Lorenz, 1905). Figure 3-15 is an example of the Lorenz curve. The dashed line depicts a population with fully equal benefit distribution; the solid curved line depicts an inequitable wealth distribution (for example, approximately 80% of the population owns around 40% of the total benefits,

[9] https://www.istat.it/it/archivio/104317

and the other 20% owns the remaining 60% of benefits). The smaller the distance between the Lorenz curve and the perfect-equality curve, the more the equity of the benefit distribution.

Figure 3-15 Example of Lorenz curve



Lorenz curves are also used for any quantity that can be aggregated over a population, not simply income. And in our case, we will use it for accessibility distribution among the population, and for this aim, we define two types of Lorenz curves below.

3.2.2.1.1. Hexagon-based Lorenz curve

Let us denote $a(\lambda)$ the accessibility of the hexagon λ. Here we consider each hexagon as a stakeholder; we order the hexagons from the worst to the best in terms of accessibility. Therefore, $a(\lambda_i) \leq a(\lambda_{i+1})$. We now build a plot where we put such hexagons $\lambda_1, \lambda_2, \dots, \lambda_{|\Lambda|}$ on the x-axis and the corresponding cumulative values of the velocity score of each hexagon, according to the above defined order, the y-axis (Figure 3-16). It is then possible to derive from such plot a Lorenz curve, which simply represents the normalization between 0 and 1 of the values on both axes of (Figure 3-16 For each hexagon, the corresponding value of the Lorenz curve is

$$L_{a^{10}}^{hex}(\lambda_i) = \frac{1}{K} \cdot \sum_{j=1}^{i} a(\lambda_j)$$

Constant K is the normalization factor; therefore, the Lorenz curve goes from 0 to 1 . i.e.

[10] Please consider that the corresponding value for the Lorenz curve which is in terms of accessibility score can be velocity or sociality score. As you see in Figure 3-17, we showed velocity score just as an example. However, the exact same graph can be also produce for sociality score and you will see in results and discussion chapter.

$$K = \sum_{j=1}^{|\Lambda|} a(\lambda_j)$$

We finally normalize the x-axis, so it goes from 0 to 1. Here the normalization factor is the number of hexagons. The final result is shown in (Figure 3-17).

Figure 3-16 Hexagon based on its Velocity score in a cumulative way for Paris

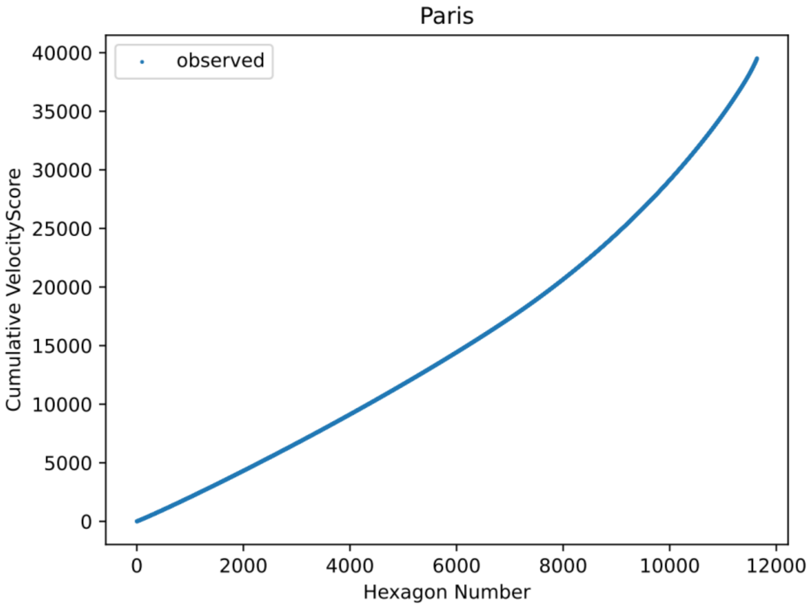
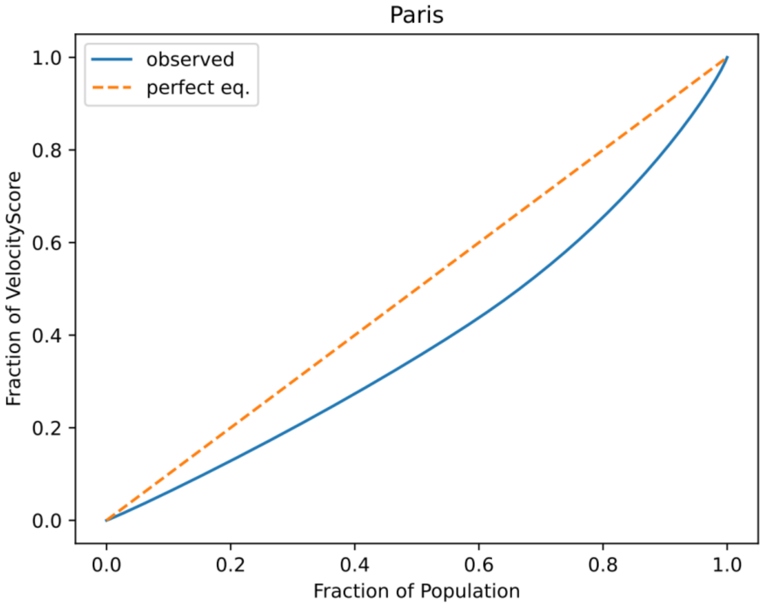


Figure 3-17 Hexagon-based Lorenz curve for Paris



3.2.2.1.2. Individual-based Lorenz curve

In order to solve the limitation of the hexagon-based Lorenz curve, that is, we give the same weight to all of the hexagons no matter the amount of their population, we also compute the second type of Lorenz curve, which is the Individual-based Lorenz curve.

This time we consider an individual as a stakeholder. Let us consider an individual p living in hexagon λ (we indicate this with $p \in \lambda$). We assume that all the individuals living in any hexagon $\lambda \in \Lambda$ enjoy the accessibility of that hexagon, i.e., $a(p) = a(\lambda)\ \forall p \in \lambda$. Let us denote P the set of all individuals. We order the individuals $p_1, p_2, \dots, p_{|P|}$ such that $a(p_i) \leq a(p_{i+1})$. If we browse individuals in such an order, we will first encounter all individuals in the worst hexagon (the one with the smallest accessibility), then the second worst and so on. We put the individuals $p_1, p_2, \dots, p_{|P|}$ on the x-axis, and the corresponding value for each individual p_i the Lorenz curve is

$$L_{a^{11}}^{ind}(\lambda_i) = \frac{1}{K'} \cdot \sum_{j=1}^{i} a(p_j)$$

And, similarly as before K' is the normalization factor to have the values of the Lorenz curve from 0 to 1. So, the normalization factor is

$$K' = \sum\nolimits_{j=1}^{|\mathrm{P}|} a(p_j)$$

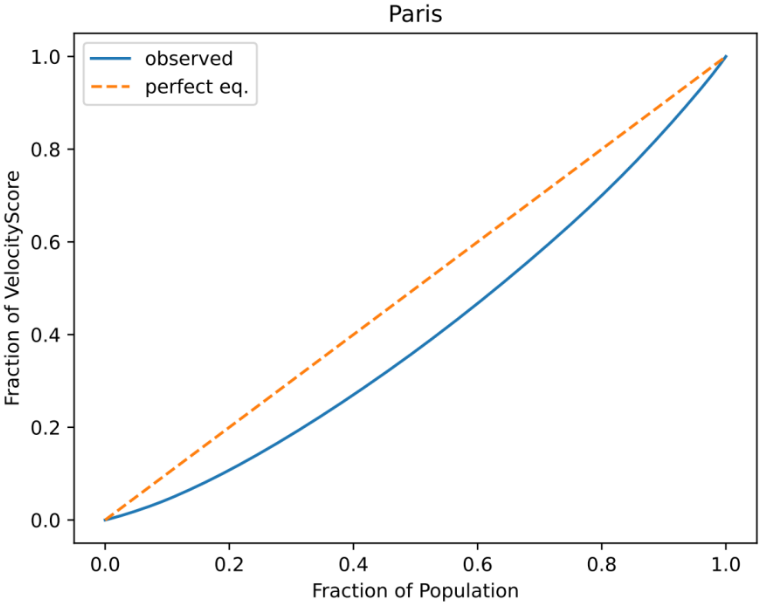
We also normalize the x-axis by the study area's population as in Figure 3-18.

In principle, Lorenz curves could be derived for any of the three scores introduced in subsection 2.2.1.1. However, given the lack of data for all considered cities, we computed hexagon-based and individual-based curves only for velocity and sociality score, thus four types of Lorenz curves in total (section 4.2)

As we know Lorenz curve is a curve used to see inequality visually. However, to have a mathematically correct number to be capable of comparing the inequality in different cities in the next section, we are going to define the Gini coefficient.

[11] Please consider that the corresponding value for the Lorenz curve which is in terms of accessibility score can be velocity or sociality score. As you see in Figure 3-18, we showed velocity score just as an example. However, the exact same graph can be also produce for sociality score and you will see in results and discussion chapter.

Figure 3-18 Individual-based Lorenz curve for Paris



3.2.2.2. Gini index

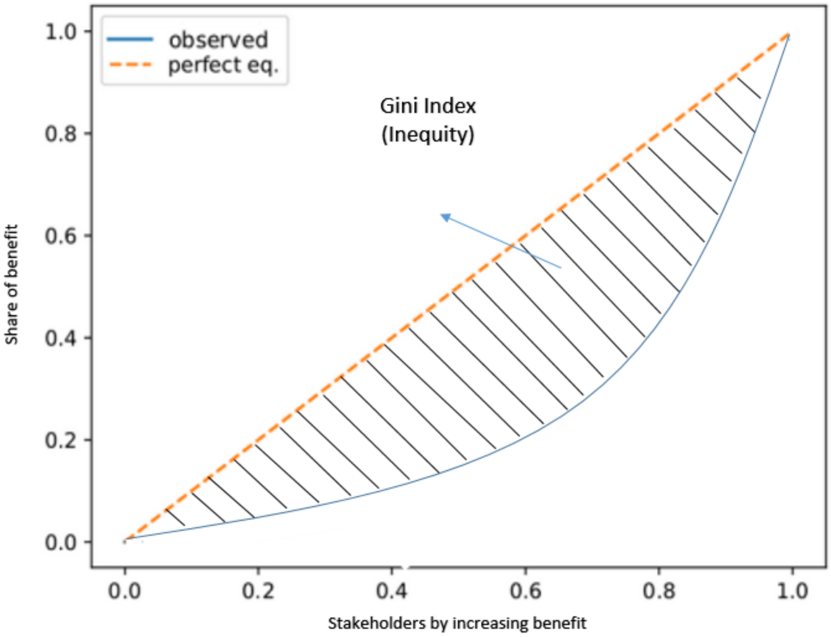
Gini coefficient is a single simple mathematical metric that represents the overall degree of inequity. The area between the line of equality and the observed curve is divided by the entire area under the line of equality to get this ratio. With this index, the distributions of two separate Lorenz curves can be compared numerically (Gini, 1912). A Gini coefficient of 0 expresses perfect equality, where all values are the same (i.e., everyone has the same share of benefits). A Gini coefficient of 1 (or 100 %) expresses maximal inequality among values (i.e., for many people where only one person has all the benefits and all others have none, the Gini coefficient will be nearly one). The Gini coefficient is difficult to calculate mathematically; however, it may be approximated using the following formula(Delbosc and Currie, 2011)

$$G = 1 - \sum_{k=1}^{n} (X_k - X_{k-1})(Y_k - Y_{k-1})$$

Where X_k is the cumulated proportion of the population variable, for $k = 0, \dots, n$ with $X_0 = 0, X_n = 1$ and Y_k is the cumulated proportion of the benefits that here are represented by the accessibility score, for $k = 0, \dots, n$ with $Y_0 = 0, Y_n = 1$. Figure 3-19 is an example of the Lorenz curve and Gini index.

Finally, by having the accessibility scores for each hexagon in the city, we can derive the Lorenz curve and also compute the Gini coefficient. The reason behind using the Lorenz curve and Gini coefficient together is because with the Lorenz curve, we have just a visual presentation of the distribution of accessibility among the population and this cannot help us to compare different cities inequity. However, by computing the Gini index, we arrive at a specific number for each city which can be compared in different cities.

Figure 3-19 Example of Gini Index



3.2.3. Public transportation system improvement toward equity

3.2.3.1. Motivation and terminology

The calculation of accessibility scores explained in 3.2.1.2 is based on General Transit Feed Specification (GTFS) explained in detail in 3.1.1.

Let us denote with $\mathcal{L}$ the set of transit lines. Thanks to this information, starting from the center of any hexagon $\lambda \in \Lambda$, with the departure time of t, we can compute all the hexagons that can be reached within the time range T. An isochrone is then calculated to include the reached hexagons. The accessibility score $a(\lambda, t, \mathcal{L})$ could be both velocity and sociality scores according to 3.2.1.2 . This methodology uses the "public-transport-analysis" script in 3.2.1 as a base to be capable to studying the concept of lines in the accessibility scores, to which other scripts were developed and added as we are going to describe in following sections.

It is easy to show that the more lines available, the more the accessibility. Fixing any λ and t, $a(\lambda, t, \mathcal{L})$ is an increasing set function, i.e., $a(\lambda, t, \mathcal{L}') \leq a(\lambda, t, \mathcal{L})$ if $\mathcal{L}' \subset \mathcal{L}$.

3.2.3.2. Definition of a score for equity

Let us consider the set of lines $l \in \mathcal{L}$, either bus, metro, or commuter rail. We want to associate an equity score to any line l, to measure how significant its contribution is to the overall transit equity.

To this aim, we compute the contribution of line l to the Gini index. As we previously mentioned, passing from $\mathcal{L}\backslash\{l\}$ to $\mathcal{L}$ undoubtedly improves the accessibility of all hexagons. However, it is essential to understand how such an improvement is distributed across hexagons. i.e., for which hexagons this improvement is considerable and for which others it is instead negligible.

Suppose that, when passing from $\mathcal{L}\backslash\{l\}$ to $\mathcal{L}$, the most considerable accessibility improvement is observed in the "unfortunate" hexagons, i.e., those suffering from low accessibility. Therefore, this means that l has a beneficial impact on equity. Observe that the unfortunate hexagons correspond to the population on the left of the Lorenz curve (Figure 3-15). So, in this case, the left part of the Lorenz curve "inflates" while the right part deflates due to its scaling to 1. As a result, the Lorenz curve approaches the perfect equity curve as geometrical evidence, and the Gini index decreases. Therefore, in this case, we can say that l is positively contributing to equity.

If, instead, the most considerable improvement is observed in the "fortunate" hexagons. i.e., the ones enjoying high accessibility, the opposite consideration holds that in this case, line l is not beneficial to equity, and the Gini index increases. In this case, it would not mean that it does not help improve equity (although it may improve average accessibility or overall travel time). This would not mean that l has to be eliminated but that, if a transit authority has a limited budget to improve its offer, it should invest in improving other lines rather than l, e.g., by increasing frequency or introducing advanced technology like automation.

We have thus shown that the change in the Gini index when passing from $\mathcal{L}\backslash\{l\}$ to $\mathcal{L}$ indicates its contribution to equity. This means that if the Gini index increases a lot, l worsens inequity. Otherwise, l positively contributes to equity. More formally, let us denote with $G(\mathcal{L},t)$ the Gini index when all lines are active, considering departures at time t. Let us denote with $G(\mathcal{L}\backslash\{l\},t)$ the Gini index without line l. We define the equity score as

$$\Delta G(l\,,t) = G(\mathcal{L}\backslash\{l\},t) - G(\mathcal{L},t)$$

If $\Delta G(l\,,t)$ is largely positive, then l is important for equity at time t, while if $\Delta G(l\,,t)$ is small or negative, that line is irrelevant in terms of equity. The equity score, as defined before, might be difficult to compute as it requires a relatively large amount of computation, which may become prohibitive in large cities.

3.2.3.2.1. Proof of computational complexity of equity score

Let us fix a specific departure time t. In order to compute the Gini index $G(\mathcal{L},t)$, we need to compute the Lorenz curve, which in turn requires computing the set of all accessibility values $\{a(\lambda,t,\mathcal{L})|\lambda \in \Lambda\}$. In order to get these values, we have to compute the earliest arrival path from hexagon λ to λ' with departure time t is the sequence of movements that allow arriving at λ' As soon as possible. A movement can be (i) to walk from one hexagon to another, (ii) to wait at a stop, (iii) to board a vehicle (bus, train, or metro) up to a specific other stop, (iv) to alight. Observe that one can alight a vehicle and board another on another line. We thus need to compute $|\Lambda| \cdot |\Lambda|$ earliest arrival paths. We take the computation of the shortest arrival times from (Biazzo et al., 2019) and the algorithm described in 3.2.1.1.

To compute $\{\Delta G(l\,,t)|l \in \mathcal{L}\}$, we need to compute $G(\mathcal{L}\backslash\{l\},t)$ for all lines $l \in \mathcal{L}$, in addition to $G(l\,,t)$. Therefore, the computation of the earliest arrival paths must be repeated $\mathcal{L} + 1$ times.

This computational complexity is very impractical if the computation of equity scores is used within an optimization loop, in which lines are iteratively modified to get, at the end of the loop, a transit structure with better equity. We thus propose in the following subsection a method to compute an alternative approximate score for each line $e(l\,,t), \forall l \in \mathcal{L}$, more computationally efficient than $\Delta G(l\,,t)$, while preserving its meaning. i.e., indicating the contribution of each transit line toward equity.

3.2.3.3. Line importance function

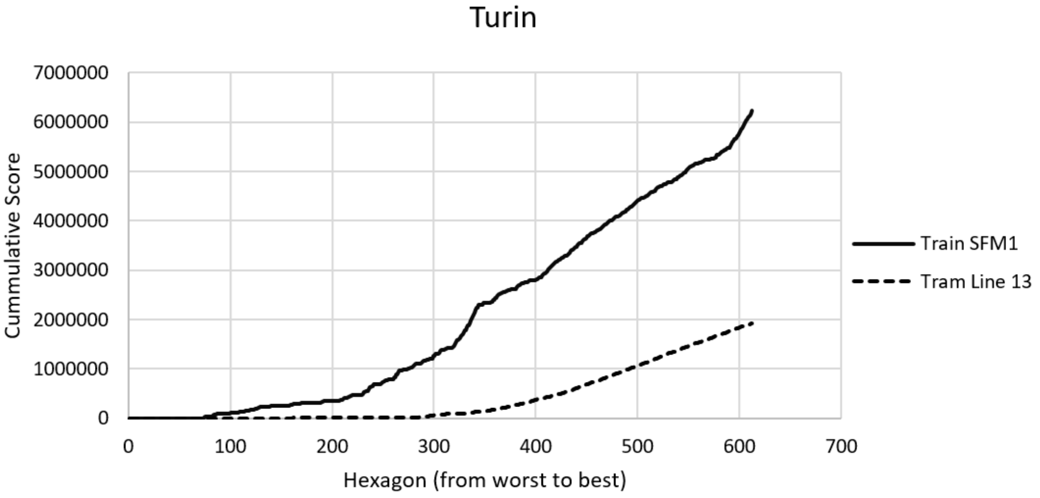
We instrument the earliest arrival path algorithm with an importance function $i(\lambda, l\,,t)$ which we initialize to $i(\lambda, l\,,t) = 0, \forall\lambda \in \Lambda, l \in \mathcal{L}$. While constructing the earliest arrival time paths from λ to any other hexagons, departing at time t and within travel time T, we increment $i(\lambda, l\,,t)$ by the distance traveled via line l (0 in case l is never used). In this way, we give more weight to the lines that take a user departing from hexagon λ as furthest as possible. Indeed, such lines are likely to be those that contribute the most to the accessibility of λ. Therefore, at the end of the earliest arrival path computation, we obtain, with practically no additional computational cost, the importance function $i(\lambda, l\,,t)$, telling how significant the contribution of each line l is for the accessibility of any hexagon $\lambda \in \Lambda$, considering departure time t.

We now order hexagons $\lambda = 1, \dots, |\Lambda|$ from the "worst" (the one with the lowest accessibility) to the "best" (the one with the highest accessibility). We define the cumulative importance function as

$$I(\lambda, l\,,t) = \sum_{\lambda'=1}^{\lambda} i(\lambda', l\,,t)$$

It expresses the importance of line l for the λ worst hexagons, which are $\lambda' = 1, \dots, \lambda$, following the order above. We depict an example in Figure 3-20, which used the new approach for computing the equity score of two public transport lines in Turin, namely Train SFM1 and Tram line 13.

Figure 3-20 Cumulative importance function for two lines $I(\lambda, l_1, t)$ and $I(\lambda, l_2, t)$ as a function of λ in Turin



It is evident that a line whose cumulative importance is very high for the worst hexagons (l_1 of Figure 3-20) is beneficial for equity, as it allows people in disadvantaged hexagons to go relatively far. On the contrary, a line whose cumulative importance function remains low for the worst hexagons and only increases for the best hexagons does not contribute to equity. For instance, in Figure 3-20 line l_2 does not contribute at all to the accessibility of the worst 300 hexagons. Usually cumulative importance functions are all convex like the above depicted two, which eases their comparison in order to unequivocally understand which line is mostly contributing to equity.

3.2.3.4. Extraction of the worst hexagons

According to 3.2.3.3, we have arrived at the formula that computes the line importance value for the worst hexagons, and those worst hexagons are the ones located in the suburbs. To be capable of selecting and extracting the hexagons that are both relevant to analyse and located in the suburbs, we need to define two different criteria. Population criterion

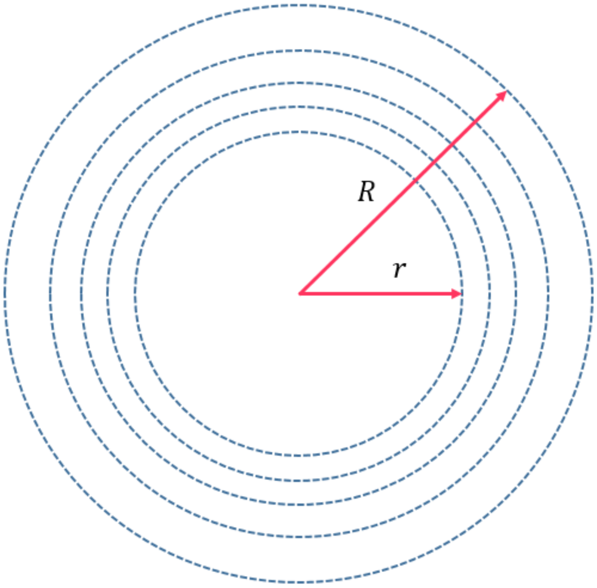
As we know, we may have computed the importance function value for a hexagon that, in reality, no one lives there. So, in order to remove these types of errors, we remove those hexagons whose population density is not significant (less than 100 people per Km^2).

3.2.3.4.1. Suburb criterion

In order to select those hexagons located in the suburbs, we experimentally define the following criteria. We denote R as the maximum distance from the center of the city (the distance between the farthest hexagon and the center of the city) and r as the distance of a certain hexagon from the center. The considered criteria are as follows:

i. Consider hexagons with $\frac{r}{R} \geq \frac{1}{3}$
ii. Consider hexagons with $\frac{r}{R} \geq \frac{1}{6}$
iii. Consider hexagons with $\frac{r}{R} \geq \frac{1}{12}$
iv. Sort the hexagons from worst to the best based on the sociality score and take the first 40% of the hexagons.
v. Sort the hexagons from worst to the best based on the sociality score and take the first 65% of the hexagons.
vi. Sort the hexagons from worst to the best based on the sociality score and take the first 90% of the hexagons.
vii. Consider hexagons with a sociality score of less than 25% of the maximum sociality score.
viii. Consider hexagons with a sociality score of less than 50% of the maximum sociality score.
ix. Consider hexagons with a sociality score of less than 75% of the maximum sociality score.

Figure 3-21 Suburb criterion - Radius



As in Figure 3-21, for instance, if we choose criterion (i), we will get those hexagons located at a distance from the center $r \geq 1/3 \cdot R$.

Figure 3-22 An example for suburb criterion using sociality score

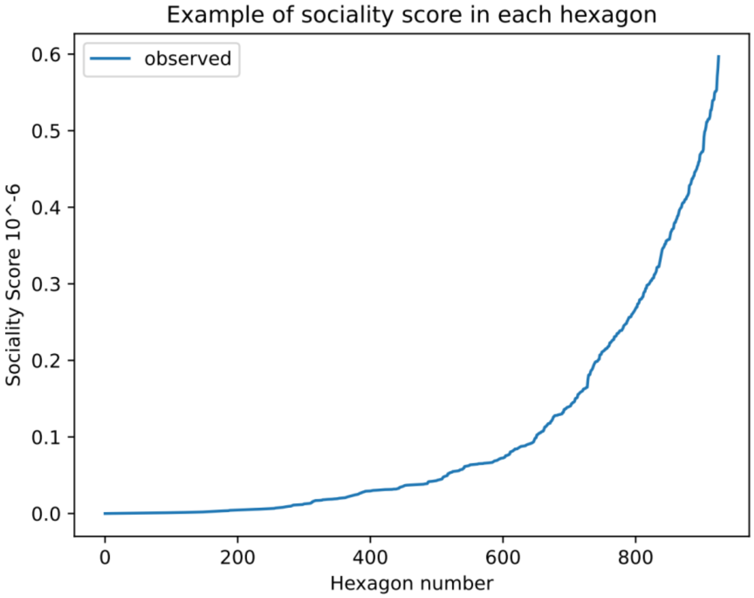
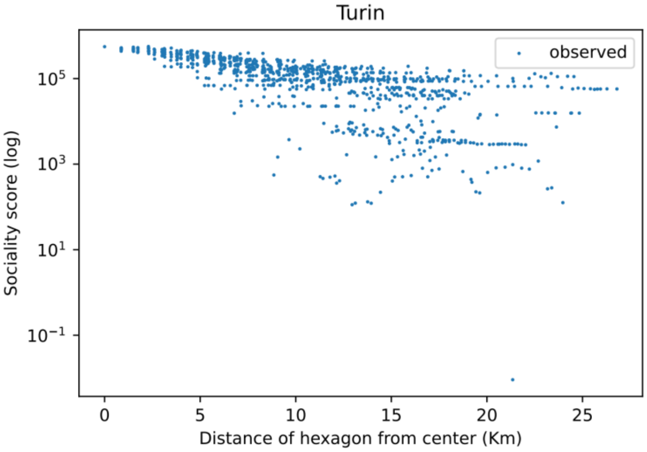


Figure 3-23 Sociality score and distance from the center dependence for Turin



As in Figure 3-22, the hexagons are sorted based on their sociality score, and for instance, if we choose the criterion (iv), we will get the first 40% of hexagons, roughly the first 400 hexagons in this example that also have low sociality score and according to the example in Figure 3-23 as much as we go from the center to the periphery the sociality score decreases. So, those hexagons with low sociality score are the ones located in the periphery.

Finally, for the last three criteria, we catch the maximum sociality score in the city, and we select those hexagons with a sociality score lower than, for example, 25% of the maximum sociality score as in criterion (vii).

The choice of the best criterion is case-specific and it will therefore be discussed in when presenting the results.

3.2.4. Considered cities

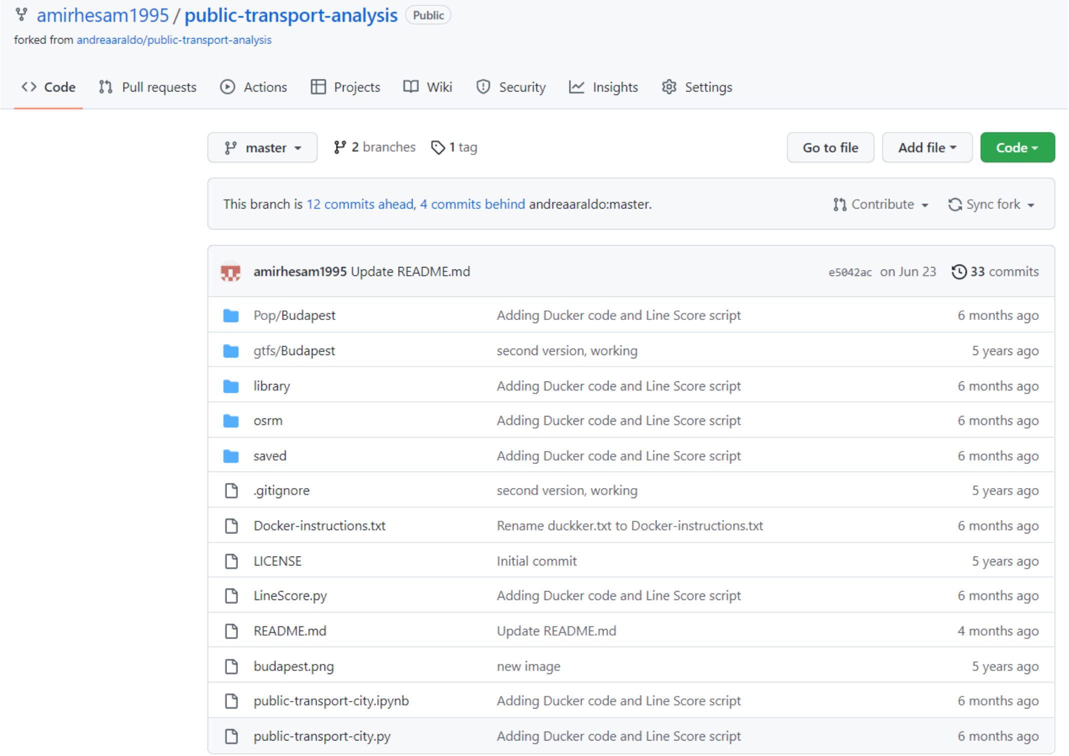
We consider two distinct sets of cities to run experiments. The first part, which is the comparison of the results of accessibility equity between our method based on section 3.2.2 and the results from (Biazzo et al., 2019) considers Paris, Boston, Sydney, and Madrid. This selection is because in the (Biazzo et al., 2019) report, the accessibility equity results for Paris, New York, Madrid, Montreal, Sydney, and Boston and between those, we selected the four cities above.

In the second part, which is the finding of the most important lines of the transit system for equity, we considered Turin, Aachen, Manchester, Helsinki, Vienna, Budapest, and Berlin. The reason behind this selection and the difference between the cities in the first part and second part is that the four cities in the first part have high number of stops in their transit network, increasing the computation time supernumerary. Finally, because each city has to do the computation as many times as the number of lines, we have to consider cities with a reasonable number of stops.

3.2.5. Scripts for computation of line importance function

The script for computing the line importance function reported in Appendix D. However, if the reader wants to run the code on his/her machine the code for computing the line importance function and also the accessibility scores with description is available on Github [12] (Figure 3-24).

Figure 3-24 Github Webpage containing the needed scripts



This branch is 12 commits ahead, 4 commits behind andreaaraldo:master.

amirhesam1995 Update README.md e5042ac on Jun 23 33 commits

Pop/Budapest	Adding Ducker code and Line Score script	6 months ago
gtfs/Budapest	second version, working	5 years ago
library	Adding Ducker code and Line Score script	6 months ago
osrm	Adding Ducker code and Line Score script	6 months ago
saved	Adding Ducker code and Line Score script	6 months ago
.gitignore	second version, working	5 years ago
Docker-instructions.txt	Rename duckker.txt to Docker-instructions.txt	6 months ago
LICENSE	Initial commit	5 years ago
LineScore.py	Adding Ducker code and Line Score script	6 months ago
README.md	Update README.md	4 months ago
budapest.png	new image	5 years ago
public-transport-city.ipynb	Adding Ducker code and Line Score script	6 months ago
public-transport-city.py	Adding Ducker code and Line Score script	6 months ago

[12] https://github.com/amirhesam1995/public-transport-analysis

4. Chapter 4: Results and discussion

The development of this chapter will carry out in two parts. In the first part, we will show the results of accessibility scores of ,Paris, Boston, Sydney, and Madrid. And then talk about their general description and inequality in their geographical distribution; finally, we will trace the Lorenz curve, compute Gini indices, and compare our findings with (Biazzo et al., 2019). In the second part, we will show the results of transit line scoring and discuss the methodology used.

According to what we have described in 3.2.1.2 for the two defined scores, which are velocity and sociality scores, we have arrived at the following formulas:

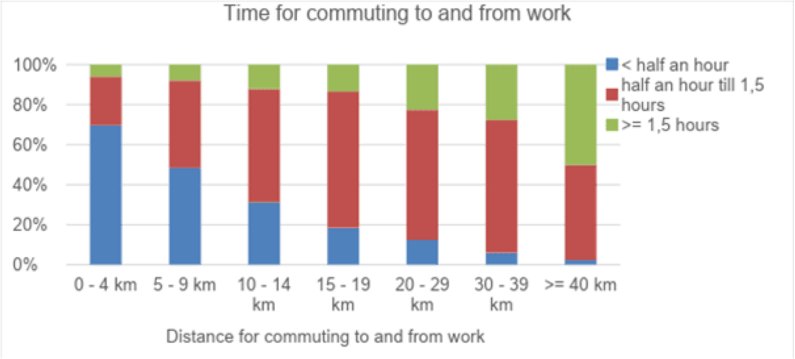
$$v(\lambda) = \frac{\sum_{t_0=t_{start}}^{t_{end}} \int_0^T v\big(\tau, (\lambda, t_0)\big)d\tau}{number\ of\ departure\ instants}$$

$$s(\lambda) = \frac{\sum_{t_0=t_{start}}^{t_{end}} \int_0^T P\big(\tau, (\lambda, t_0)\big)d\tau}{number\ of\ departure\ instants}$$

We have two variables here which are the departure time t_0 and maximum travel time T. These two variables have to be set as an experimental value, and to do so, for the departure time, we considered 6 AM until 10 PM every two hours in order to include all the different departure times in a day that public transit works more or less at a reasonable frequency and also we can include the different peak times of a day in this interval. So now, according to the formula, by averaging over different departure times, we can reach one specific number for accessibility of each hexagon that more or less shows the accessibility in one specific day.

The second experimental variable that has to be set is the maximum travel time T. According to (Worx, 2017) and Figure 4-1, more than half of European respondents (56.8%) say that, on average, they take less than an hour to get to and from work each day. One-third of them (29.8%) takes less than half an hour a day, while 27% take between 30 minutes and an hour. Of the countries surveyed, the British spend the most time traveling, with 28.8% claiming to take 90 minutes or more on their commute. So as an average and also to consider the confidence interval here, we consider 1 hour.

Figure 4-1 Time for commuting to and from work (Worx 2017)



4.1. General description of accessibility scores

As described in 3.2.1.2, we have computed the accessibility scores using the python script "public-transport-analysis" of the work (Biazzo et al., 2019). According to the visualization in Figure 3-6 and Figure 3-9, we can see the unequal distribution of accessibility among the population residing in the center of the city and the periphery population. However, to have a mathematically correct number and visualization beside our eyes to quantify this inequity, we resort to the Lorenz curve and Gini index described by the methodology in 3.2.2.1 and 3.2.2.2.

4.2. Quantify the level of inequity in accessibility distribution

We compared Paris, Madrid, Boston, and Sydney, among the analyzed cities from (Biazzo et al., 2019). They quantify the inequality by two metrics:

- The ratio between the average accessibility of the best 1% hexagons and the average of the other 99% of hexagons
- The ratio between the average accessibility of the best 1% population and the average of the other 99% of the population

The above-mentioned inequality metrics, we feel, are unreliable. Indeed, there is no compelling reason to select the top 1% of scores. In fact, if a different proportion is utilized (say, 5%), inequality results may be drastically different. Therefore, we have decided to use the Gini index on Lorenz curves to measure inequality. The Gini index is more general than the two metrics mentioned above. Moreover, it does not require arbitrary choosing of any parameter.

Table 4-1 Types of Lorenz curve or Gini indices

Notation Gini Index	**Notation** Lorenz Curve	**Description**	**Unit of analysis**	**Accessibility score**
$\boldsymbol{G_v^{hex}}$	$L_v^{hex}(\lambda_i)$	Hexagon-based Lorenz curve or Gini index of the velocity score	Hexagons	Velocity
$\boldsymbol{G_s^{hex}}$	$L_s^{hex}(\lambda_i)$	Hexagon-based Lorenz curve or Gini index of the sociality score	Hexagons	Sociality
$\boldsymbol{G_v^{ind}}$	$L_v^{ind}(\lambda_i)$	Individual-based Lorenz curve or Gini index of the velocity score	Individuals	Velocity
$\boldsymbol{G_s^{ind}}$	$L_s^{ind}(\lambda_i)$	Individual-based Lorenz curve or Gini index of the sociality score	Individuals	Sociality

We denote the Gini indexes computed on the respective Lorenz curves as G_v^{hex}, G_s^{hex},G_v^{ind}, G_s^{ind} .

As in Table 4-1, We have four types of Lorenz curve and Gini Index as well. In what is reported in Figure 3-4, Figure 3-5, Figure 3-6 and Figure 3-7 each type of Lorenz curve is depicted for all four considered cities.

The Gini indexes we calculated are reported in Table 4-2. Recall that the lower the Gini index, the better the equity. To allow easier comparison between cities, we normalize each column via mean normalization by the formula:

$$x' = \frac{x - \text{average}(x)}{\max(x) - \min(x)}$$

Moreover, we report the results in Table 4-3. For a more straightforward interpretation, we plot in Figure 4-6. If the corresponding Gini index in each column was below the average, the normalization number is negative, which is highlighted in green, and if the Gini index is higher than average, the normalized value will be positive and highlighted in red.

Figure 4-2 Hexagon-based Lorenz curves of the velocity score $L_v^{hex}(\lambda_i)$

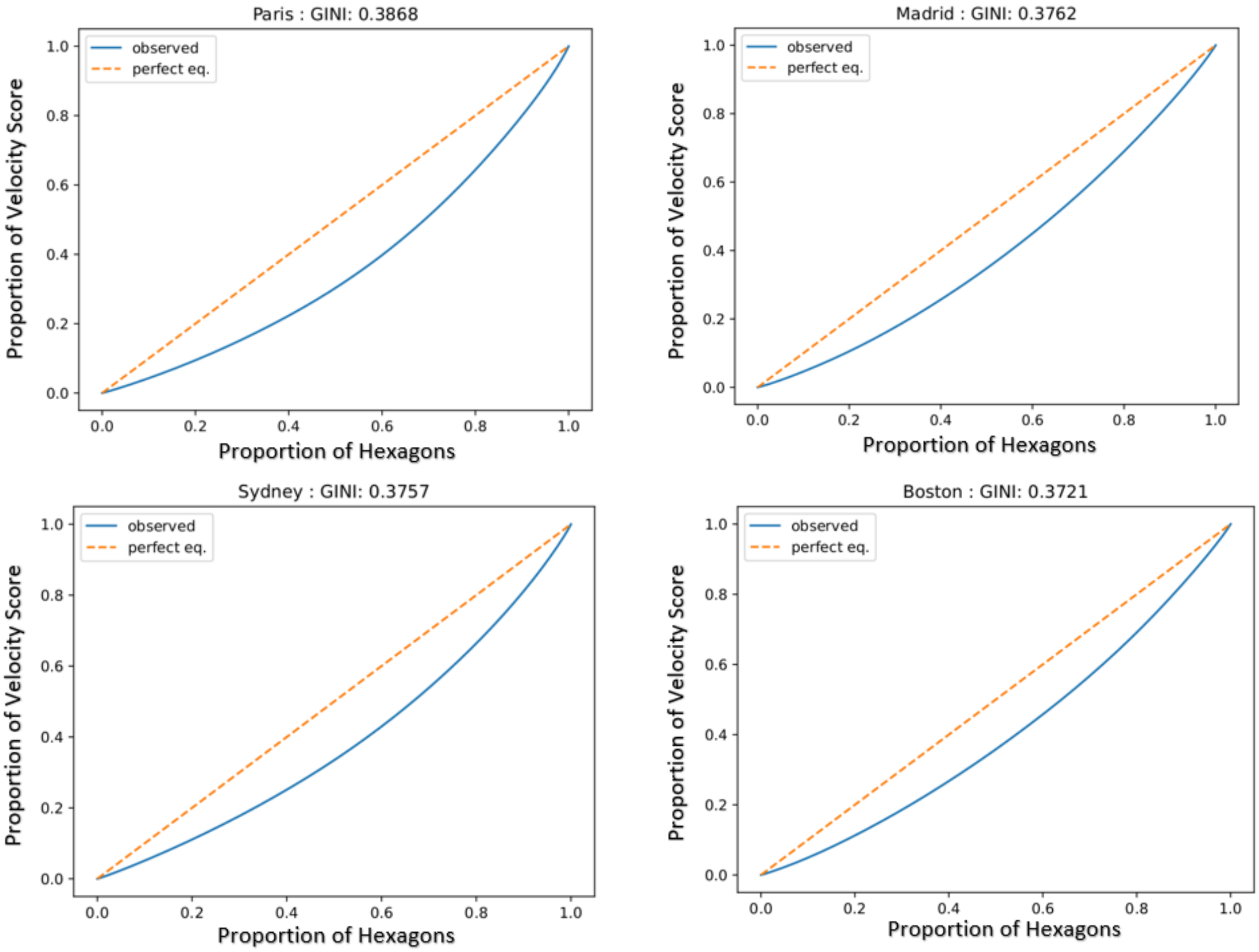


Figure 4-3 Hexagon-based Lorenz curves of the sociality score $L_s^{hex}(\lambda_i)$

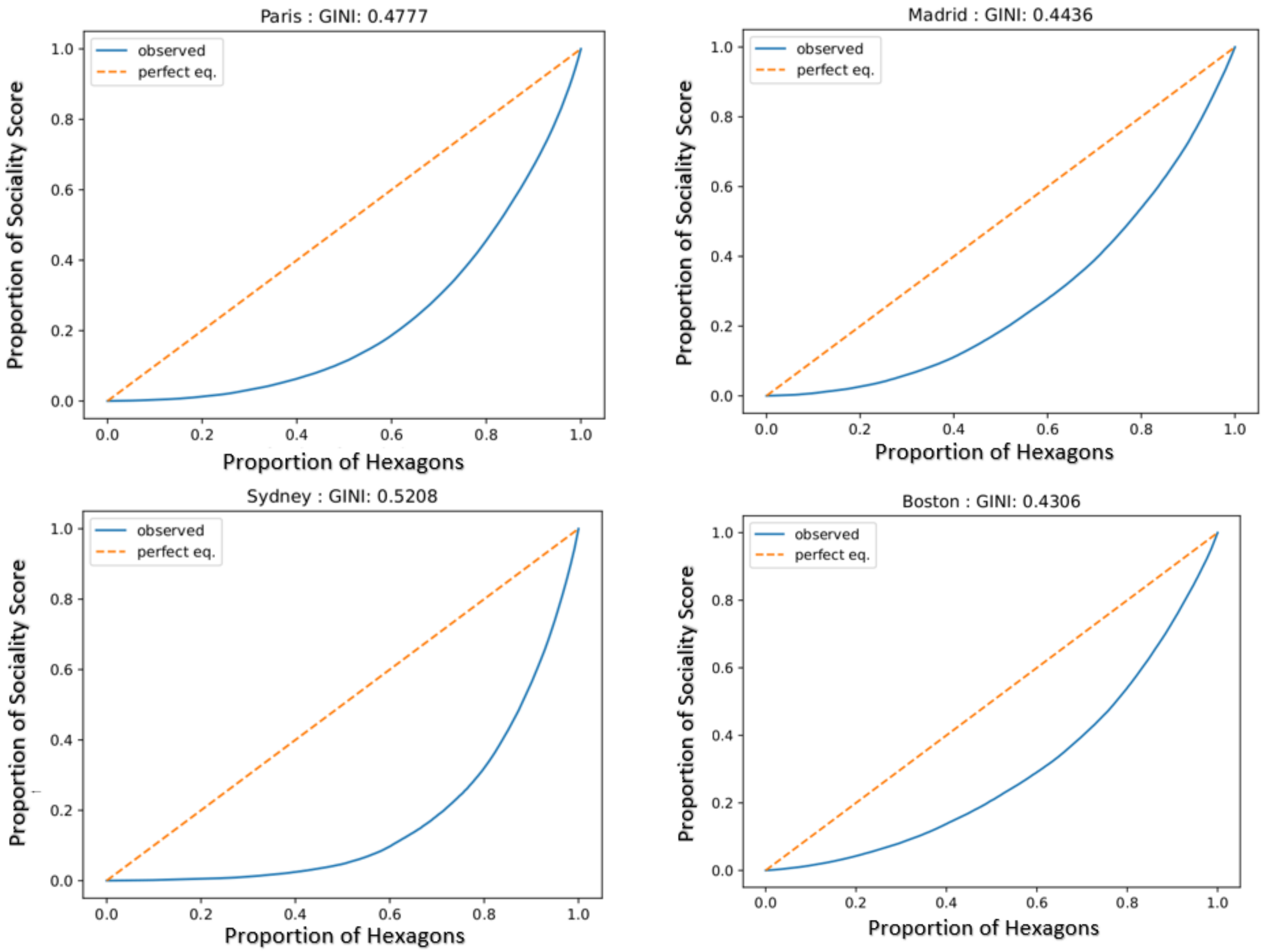
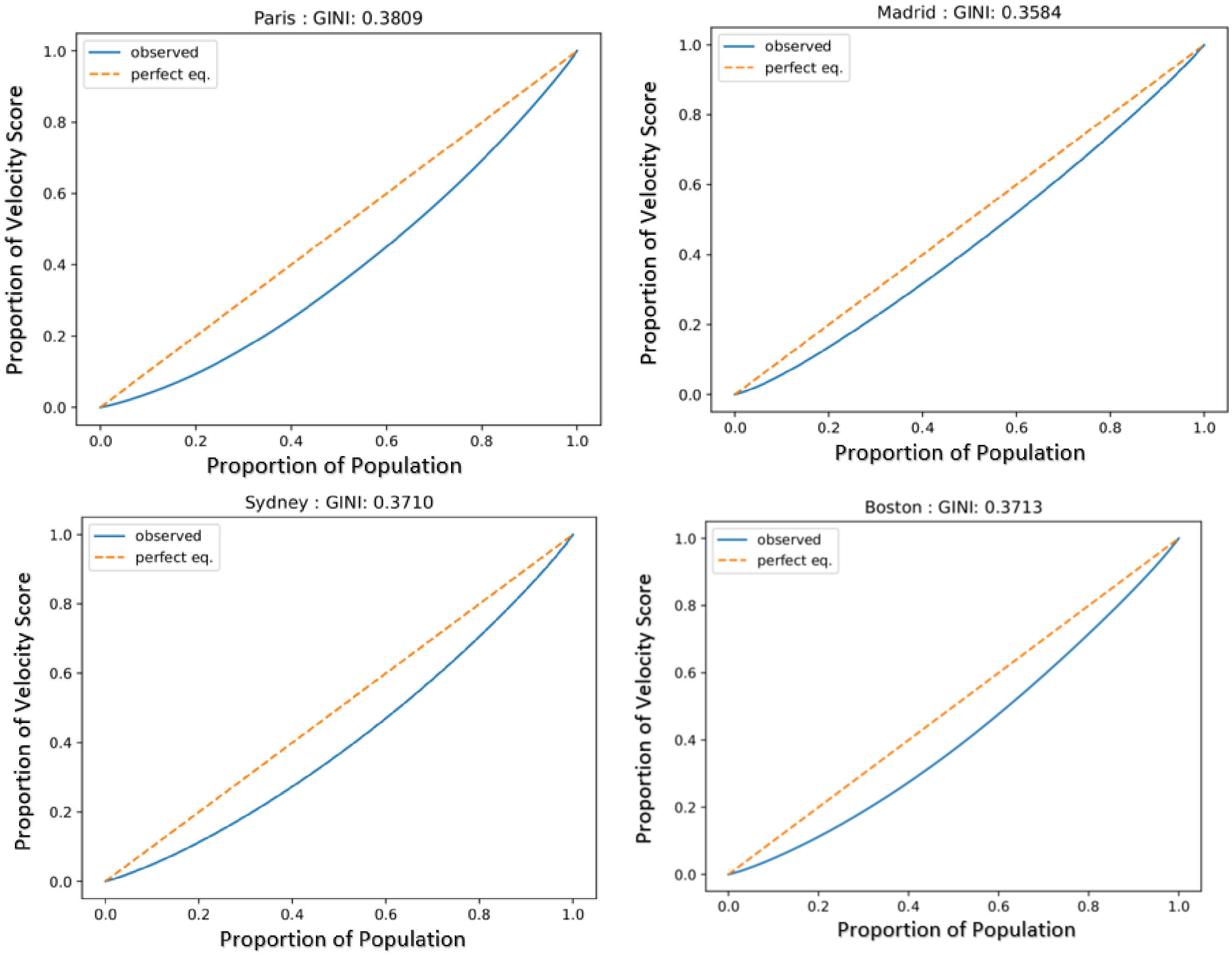


Figure 4-4 Individual-based Lorenz curves of the velocity score $L_v^{ind}(\lambda_i)$



Paris : GINI: 0.3809

Madrid : GINI: 0.3584

Sydney : GINI: 0.3710

Boston : GINI: 0.3713

Figure 4-5 Individual-based Lorenz curves of the sociality score $L_s^{ind}(\lambda_i)$

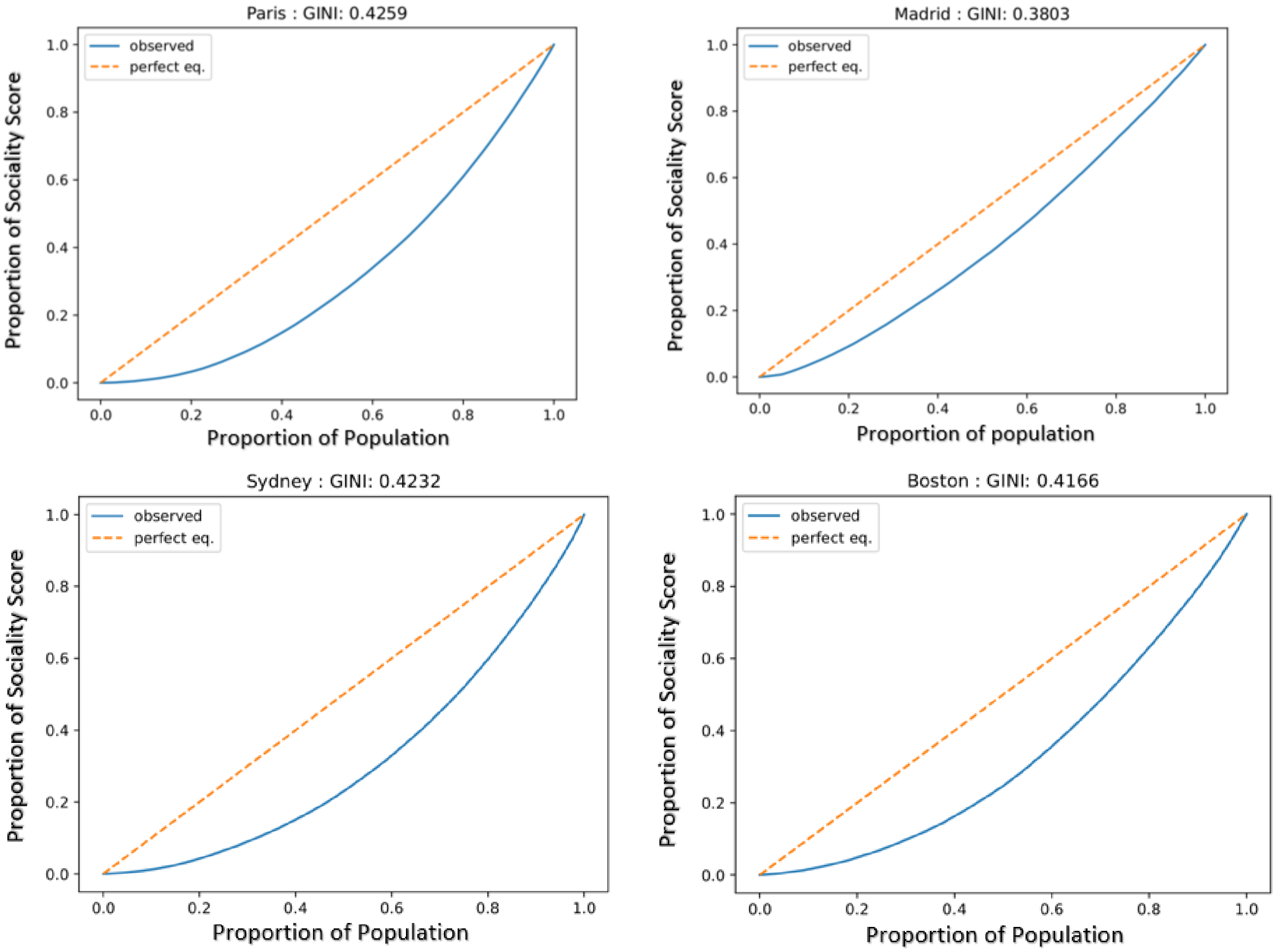


Table 4-2 Gini indices of four mentioned cities

City	$\boldsymbol{G_v^{hex}}$	$\boldsymbol{G_s^{hex}}$	$\boldsymbol{G_v^{ind}}$	$\boldsymbol{G_s^{ind}}$
Paris	0.3868	0.4777	0.3809	0.4259
Madrid	0.3762	0.4436	0.3584	0.3803
Boston	0.3721	0.4306	0.3713	0.4166
Sydney	0.3757	0.5208	0.371	0.4232

Table 4-3 Normalized Gini indices of four considered cities

City	G_v^{hex}	G_s^{hex}	G_v^{ind}	G_s^{ind}
Paris	0.62	0.11	0.47	0.32
Madrid	-0.10	-0.27	-0.53	-0.68
Boston	-0.38	-0.42	0.04	0.11
Sydney	-0.14	0.58	0.03	0.26

Figure 4-6 Normalized Gini indices (the larger, the higher inequity)

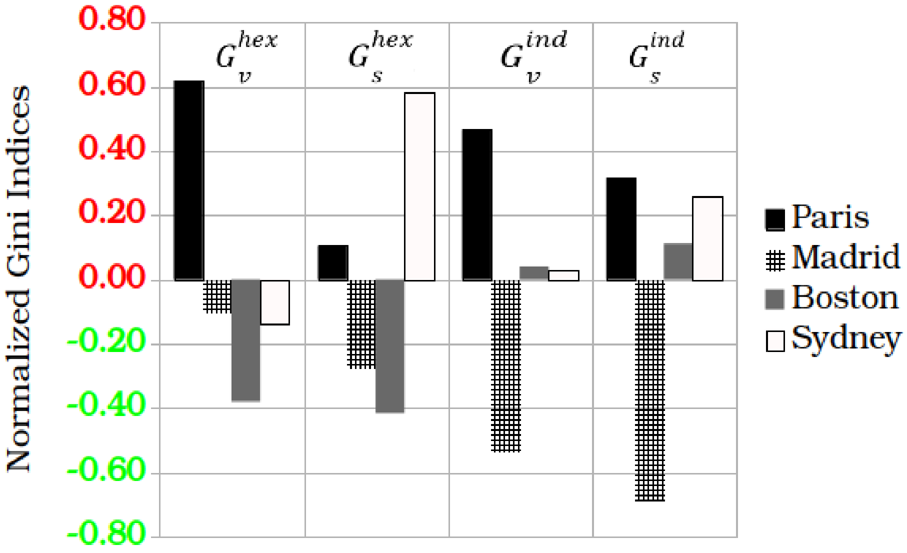
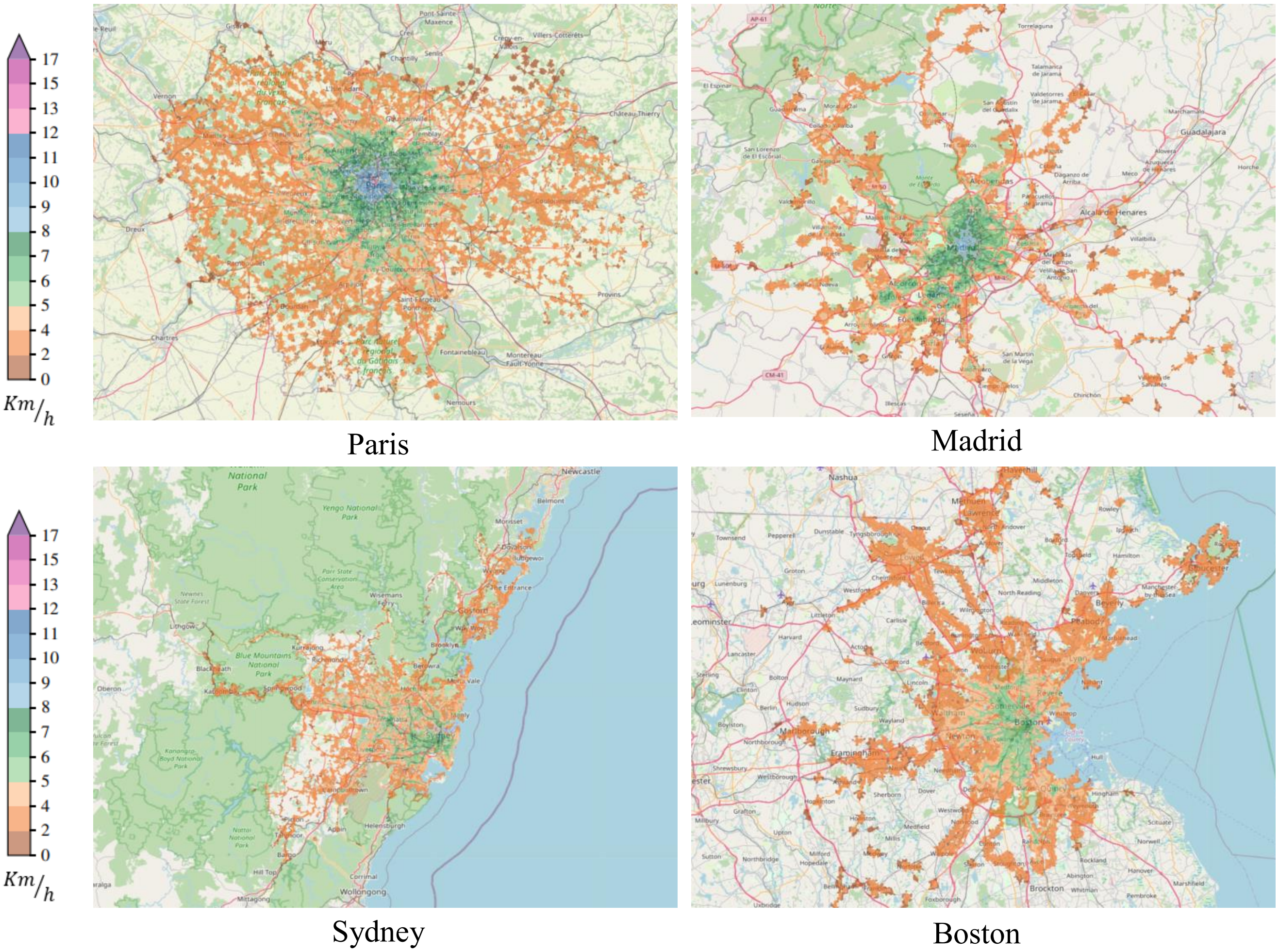


Figure 4-7 Velocity Score of four considered cities



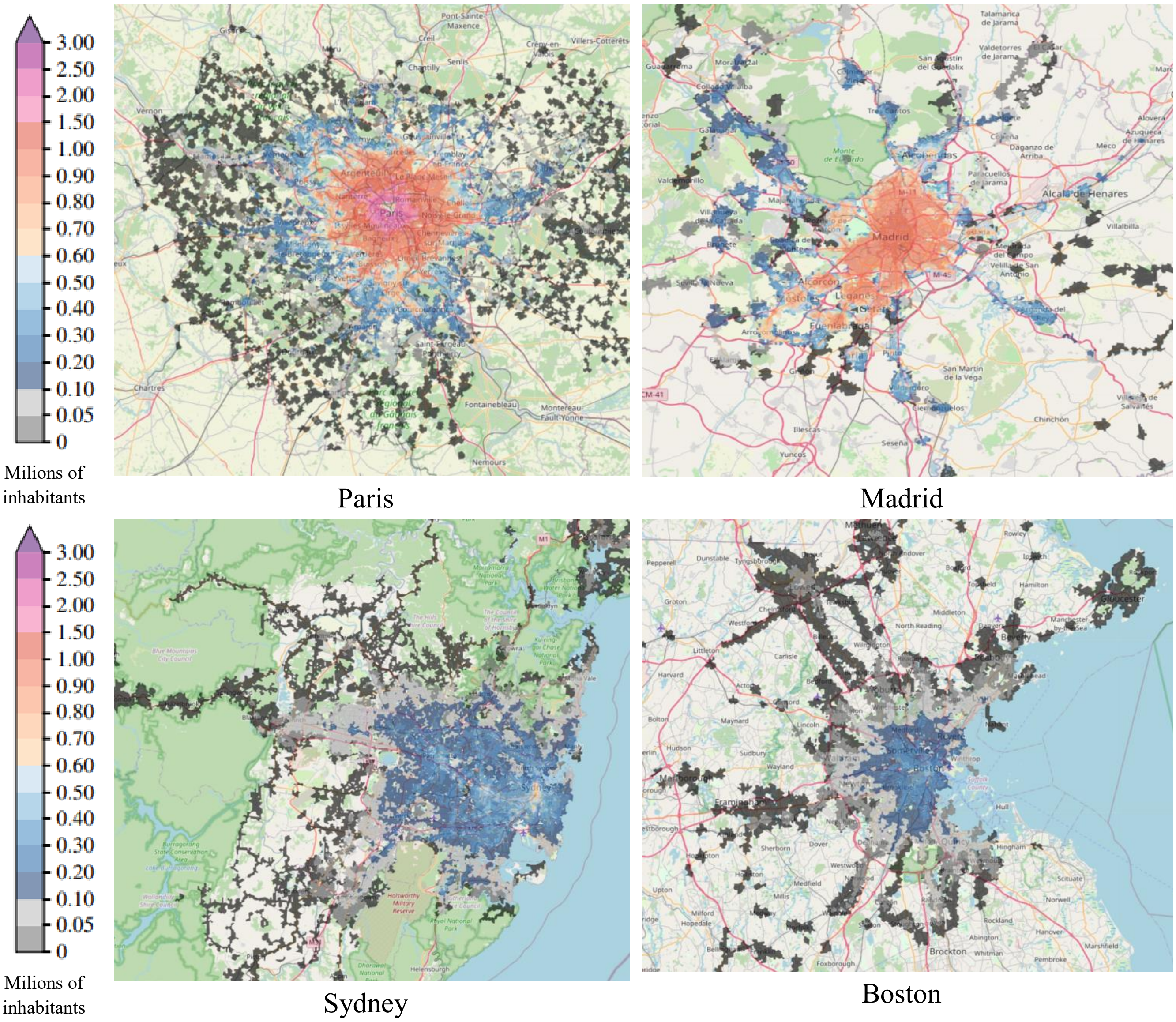
Paris

Madrid

Sydney

Boston

Figure 4-8 Sociality Score of four considered cities (Millions of Inhabitants)



Paris

Madrid

Sydney

Boston

We first notice that Paris suffers the most from high inequity, with most of the four different types of the Gini coefficient used. Because, as in Table 4-3 in the first column, Paris is the only one that has a value higher than the average that shows the most inequitable city, and then in the second column, which is hexagon-based velocity score Gini index Paris and Sydney have the value higher than average. However, the value of Paris is lower than Sydney, which shows Sydney is the most inequitable city. for the third and fourth column, Paris, Madrid, and Sydney has value higher than average, but the value for Paris is higher than others, so it is the most inequitable city. Overall, we can say that in three of the four metrics used, Paris is the most inequitable city. On the contrary, Madrid enjoys the best equity. This is confirmed by visually comparing the geographical distribution of the accessibility scores from the two cities in Figure

4-7 and Figure 4-8. It is evident that the accessibility gap between the center and suburbs is way more significant in Paris, while in Madrid, accessibility is more evenly distributed.

As you may see in Figure 4-7 for the velocity score we have hexagons with velocity score higher than even 10 Km/h in Paris and Madrid but we have not hexagons with these high numbers in Boston and Sydney. This difference is because the overall level of public transport and speed of moving with it is higher in Paris and Madrid. Similarly in Figure 4-8 for the sociality score again because the overall level of public transport is better so it can moves much higher number of people in Paris and Madrid. Consequently, we expect higher numbers in these cities.

We can order cities from the worst (high inequity) to the best (high equity) based on the four metrics defined in (Biazzo et al., 2019) and based on our four Gini-index-based metrics. We do this in Table 4-4. Note that all metrics confirm the general trends, but there are differences. For example, if we focus on the hexagon-based inequity of (Biazzo et al., 2019) velocity scores (first two rows of Table 4-4), based on our metric, Madrid is better than Sydney, in contrast to (Biazzo et al., 2019), that claim the opposite. However, visual inspection (Figure 4-7 shows that our claim is correct. Instead of sociality scores (Figure 4-8), our Gini-index-based metrics do not seem to be more accurate than (Biazzo et al., 2019).

Table 4-4 Comparison of the ranking of cities from (Biazzo et al., 2019) and the ranking based on our computation (Figure 4-6)

Metric	**Worst city**	**2nd worst**	**2nd best**	**Best**
Velocity score (top 1% hexagons)	Paris	Madrid	Sydney	Boston
G_v^{hex}	*Paris*	*Sydney*	*Madrid*	*Boston*
Sociality score (top 1%hexagons)	Paris	Boston	Sydney	Madrid
G_s^{hex}	*Sydney*	*Paris*	*Madrid*	*Boston*
Velocity score (top 1% individuals)	Paris	Boston	Sydney	Madrid
G_v^{ind}	*Paris*	*Boston*	*Sydney*	*Madrid*
Sociality score (top 1% individuals)	Boston	Sydney	Paris	Madrid
G_s^{ind}	*Paris*	*Sydney*	*Boston*	*Madrid*

Our results reported in italic. Each ranking orders the cities from the one that suffers the highest inequity, to the one that enjoys the highest equity. Green background indicates that the two rankings correspond.

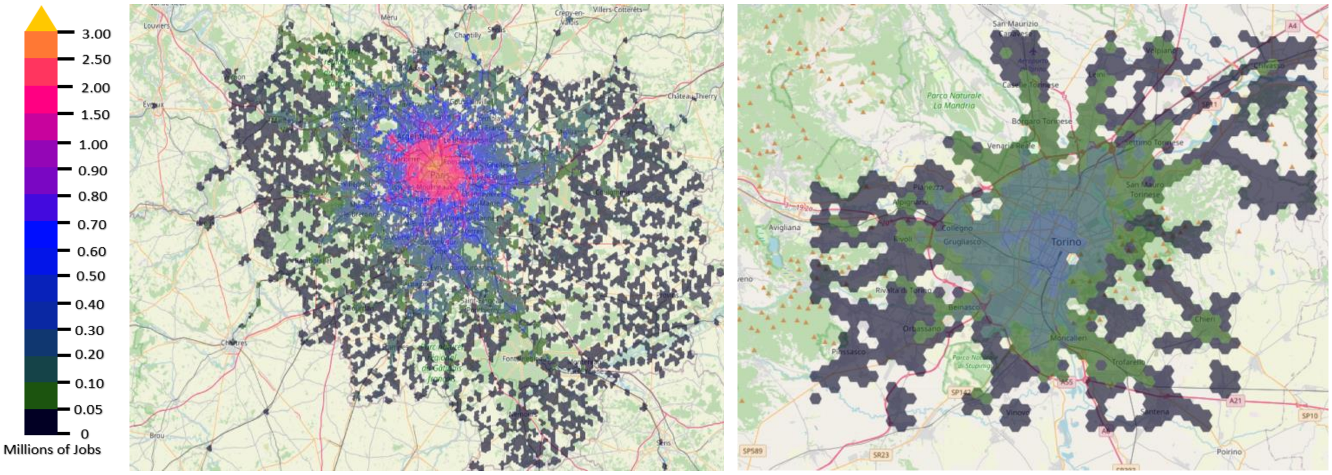
Overall, one reassuring finding emerges: all metrics manage to capture the evident differences in Accessibility equity from one city to another. This encourages the possibility of automating equity analysis across cities, with no need for visual inspection, only based on open data.

4.3. A brief consideration of Attraction Score

The attraction score was the third score we showed to also consider the opportunities in the city as described in 3.1.3 and 3.2.1.2.3. However, here we want to show this score for just two cities among the considered cities in 3.2.4, Paris and Turin, because, as we described in 3.2.1.2.3.1, it is hard to get the data about workplaces for different cities. First, we show the spatial distribution of the attraction score. Then, by the Lorenz curve and computing the Gini index compares the level of inequity in these two cities.

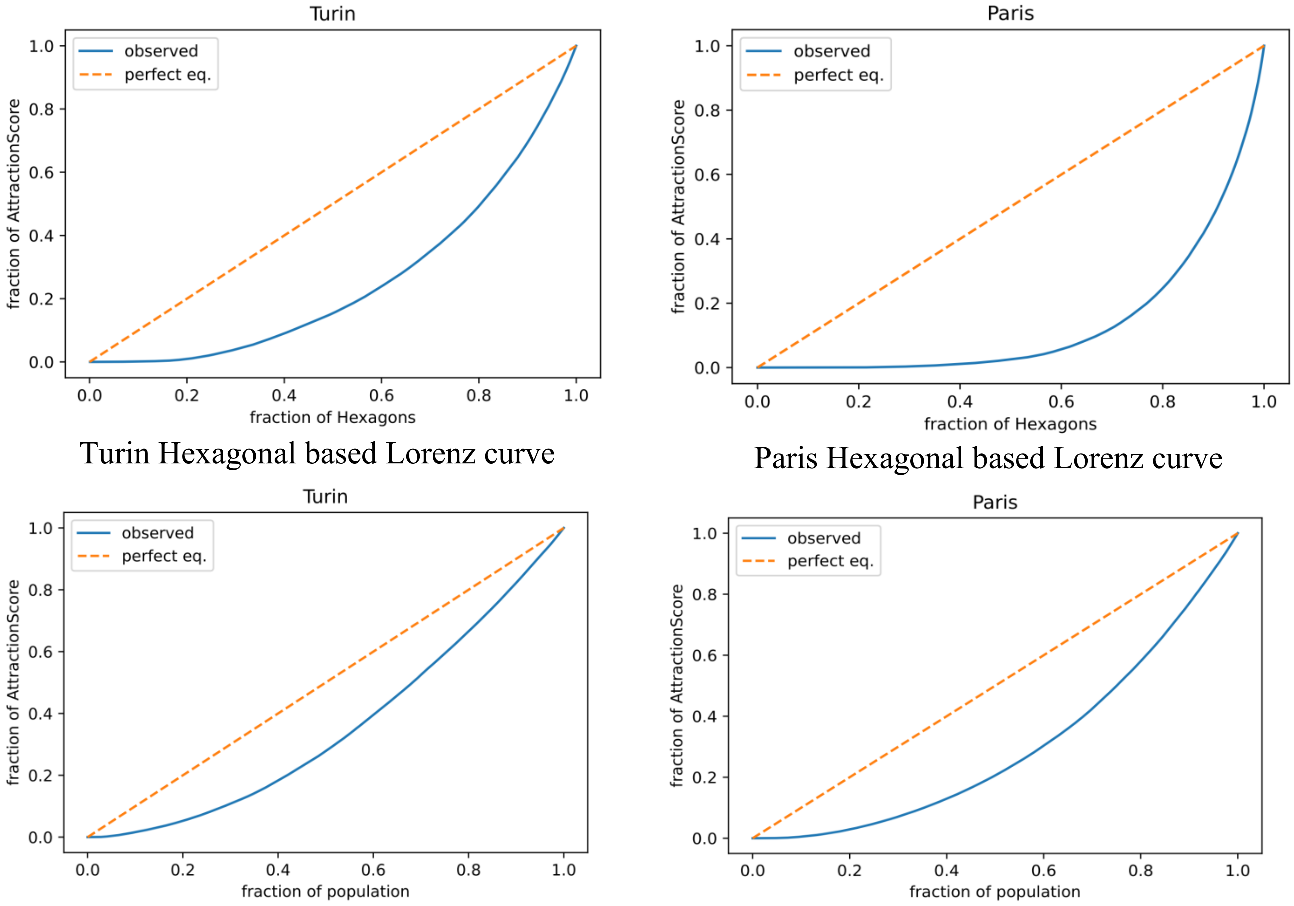
As we can see in Figure 4-9, the distribution of attraction score is entirely different in these two cities, and this makes sense because the number of jobs in Paris is far higher than in Turin. Consequently, we see the highest attraction score in Turin is around 0.3 million of jobs but in Paris is around 3 million of jobs.

Figure 4-9 Attraction Score of Paris (left) and Turin (right)



The next step is to depict the Lorenz curve and also compute the Gini index and compare the results (recall that according to 3.2.2.1, we have two Lorenz curve types: Hexagonal-based Lorenz curve and Individual-based Lorenz curve) (Figure 4-10).

Figure 4-10 Lorenz curve of the two considered cities (Paris and Turin)



Turin Hexagonal based Lorenz curve

Paris Hexagonal based Lorenz curve

Turin Individual based Lorenz curve

Paris Individual based Lorenz curve

Table 4-5 Gini indices of two considered cities (Turin and Paris)

City	Hexagonal Attraction Score	Individual Attraction Score
Turin	0.49	0.29
Paris	0.73	0.40

According to the Gini numbers that we can see in Table 4-5, no matter the metrics used, Paris is more inequitable than Turin, and this could also be shown in section 3.2 that Paris was the most inequitable among the considered cities.

The concept of attraction score is presented to show that it is possible to compute accessibility with employment data like the sociality score with the population grid. A single index reflecting the two ends of a journey would be required for a proper accessibility measurement in the transportation sector. However, as we noted in the introduction, these data are challenging to get, and we do not have employment data for all considered cities in 3.2.4. As a result, we exclusively used available data in our methodology and based it on the population grid to find the lines contributing to accessibility.

4.4. Finding transit lines that benefit the equity in cities

According to 3.2.3.2, we defined the equity score, but the computation of this score requires a very high computation time as in 3.2.3.2.1. So, because of this computational complexity, we use the line importance function according to 3.2.3.3. However, to be capable of comparing the values for both equity score and line importance function, we will first show the results for both of them and then talk about their correlation and how we can move from one to the other.

4.4.1. Results for the equity score and line importance function

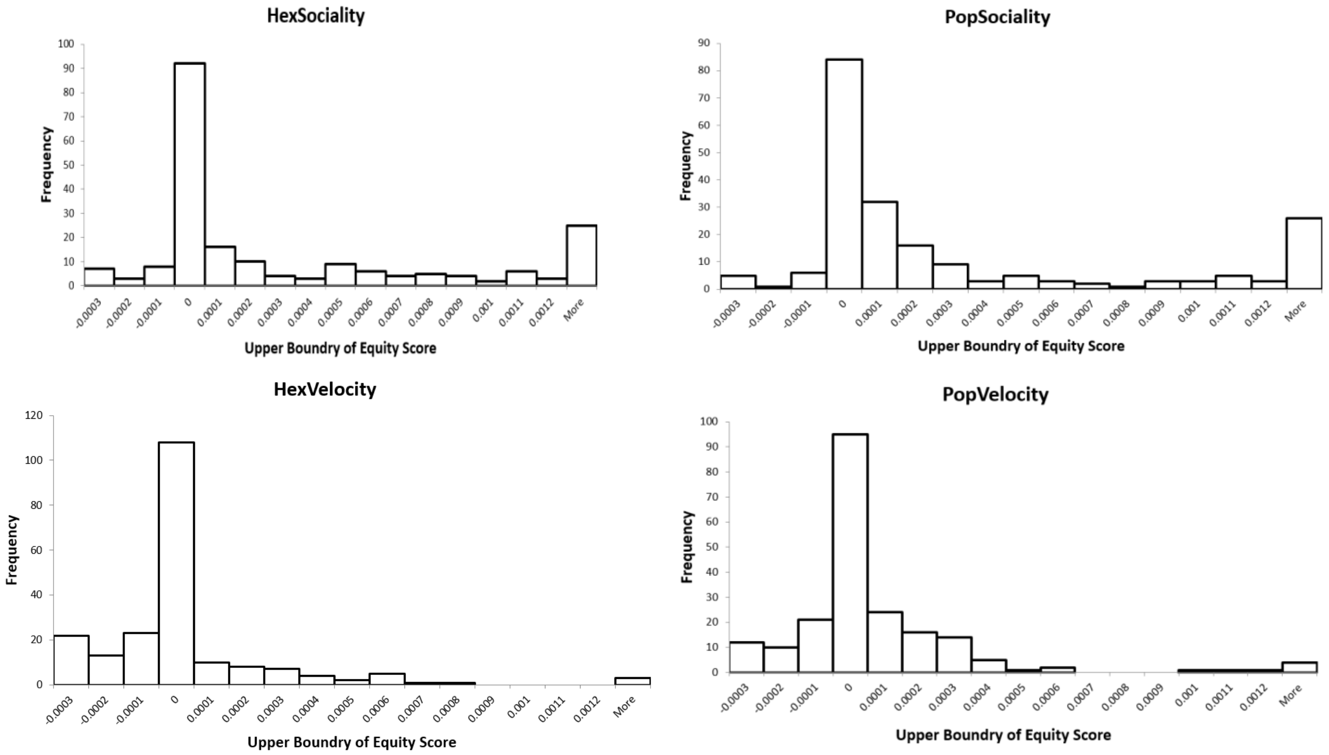
, When computing the accessibility score for different cities in 3.2.1, we took different departure times from 6AM up to 10PM. On the other hand, here we consider the departure time equal to 8AM for the computation of equity score and importance function, because we have to compute the Gini index, which is the basis of the equity score, each time we remove a line. Consequently, considering different departure times makes our computation time much longer, and it was impossible to do it in our case. So, we have to consider just one departure time, and between all those possible departure times, 8AM can more or less capture the peak in the morning. Moreover, also for the importance function we have to consider also 8AM in order to be capable to correlate the results between equity score and line importance function.

As in 3.2.3.2, equity score for each line of the transit system is computed using the Gini index. To better understand this way of computing, we depict the result of equity score for the best and worst five lines of public transit in terms of equity for Turin in Table 4-6, and the same results for other cities will report in the Appendix B. The Transit lines highlighted in green are the ones with the lowest score, and those are lines that are less important for equity, and the ones highlighted in red are those which are the most important for equity.

Table 4-6 Four types of equity score for the five best and five worst lines in Turin

Rank	Transit Line	Hex Sociality	Transit Line	Pop Sociality	Transit Line	Hex Velocity	Transit Line	Pop Velocity
1st	**Metro Line**	-0.0081	**Tram Line 10**	-0.0008	**Metro Line**	-0.0044	**Metro Line**	-0.0071
2nd	**Bus Line 2**	-0.0015	**Metro Line**	-0.0007	**Bus Line 62**	-0.0015	**Bus Line 2**	-0.0011
3rd	**Tram Line 10**	-0.0006	**Bus Line 57**	-0.0005	**Bus Line 2**	-0.0009	**Train Line SFM1**	-0.0007
4th	**Bus Line 60**	-0.0004	**Train Line SFM6**	-0.0004	**Tram Line 4**	-0.0008	**Tram Line 10**	-0.0007
5th	**Bus Line 58**	-0.0004	**Bus Line 58**	-0.0004	**Bus Line 5**	-0.0008	**Tram Line 4**	-0.0005
5th to last	**Bus Line 59**	0.0055	**Bus Line 1091**	0.0025	**Bus Line 70**	0.0006	**Bus Line 1510**	0.0012
4th to last	**Bus Line 1511**	0.0064	**Train Line SFM2**	0.0030	**Bus Line 1511**	0.0008	**Bus Line 30**	0.0016
3rd to last	**Bus Line 30**	0.0070	**Bus Line 1432**	0.0032	**Bus Line 3096**	0.0012	**Bus Line 1511**	0.0017
2nd to last	**Bus Line 1432**	0.0082	**Bus Line 3107**	0.0034	**Bus Line 30**	0.0016	**Bus Line 1432**	0.0020
Last	**Bus Line 3107**	0.0131	**Bus Line 62**	0.0046	**Bus Line 3107**	0.0026	**Bus Line 3107**	0.0028

Figure 4-11 Probability distribution function of four types of equity score in Turin



As in Table 4-6 for Turin, we can see that depending on the metrics that we use, the equity score $\Delta G(l\,,t)$ and the ranking of the lines that we get could be different. The same data for other considered cities reported in Appendix B.

As in Figure 4-11, we depict the histogram of different equity scores for Turin. In hexagonal and individual velocity equity scores, most of the scores gathered around the 0, which are the lines with a very low impact on accessibility equity. On the other hand, for the hexagonal and individual sociality, most scores gathered a little higher than zero, due to a much longer right tail of the distribution compared to the two velocity scores. To sum up, depending on the metric, the results could be different. The same data as in Figure 4-11, for other considered cities reported in Appendix B.

Similarly in Table 4-7 for hexagonal velocity equity score the 2nd and 3rd quartile are both equal to zero which means 50% of the scores are equal to zero. Likely for individual velocity and hexagonal sociality equity scores the 1st and 2nd quartile are both equal to zero which means 50% of the scores are equal to zero. But, for the individual sociality equity score only the 1st quartile is equal to zero which shows 25% of the scores are equal to zero. We can say by choosing the individual sociality equity score we probably can better catch the lines that have impact on accessibility equity.

Table 4-7 Average and quartile values of the equity scores

	Hex Sociality	Pop Sociality	Hex Velocity	Pop Velocity
Average	0.00052	0.00037	-4.5E-05	1.09E-05
Quartile 1	0	0	-0.00013	0
Quartile 2	0	0.00001	0	0
Quartile 3	0.00051	0.00032	0	0.0001
Quartile 4	0.00645	0.00303	0.00076	0.0016

As for the line importance function, the results for the different combinations mentioned in 3.2.3.4 is reported in Table 4-8. As much as the importance function value is higher, the line is more important for equity. And, as in Table 4-8 the ranking of the lines depending on the combination can be different. For instance, for combination (i) the best line is bus line 3107 but for combination (ii) the best line is Train SFM1. In Table 4-8, the bottom five lines are highlighted in green, and the top five lines are highlighted in red.

Table 4-8 Top and bottom five Importance function value for different combinations for Turin

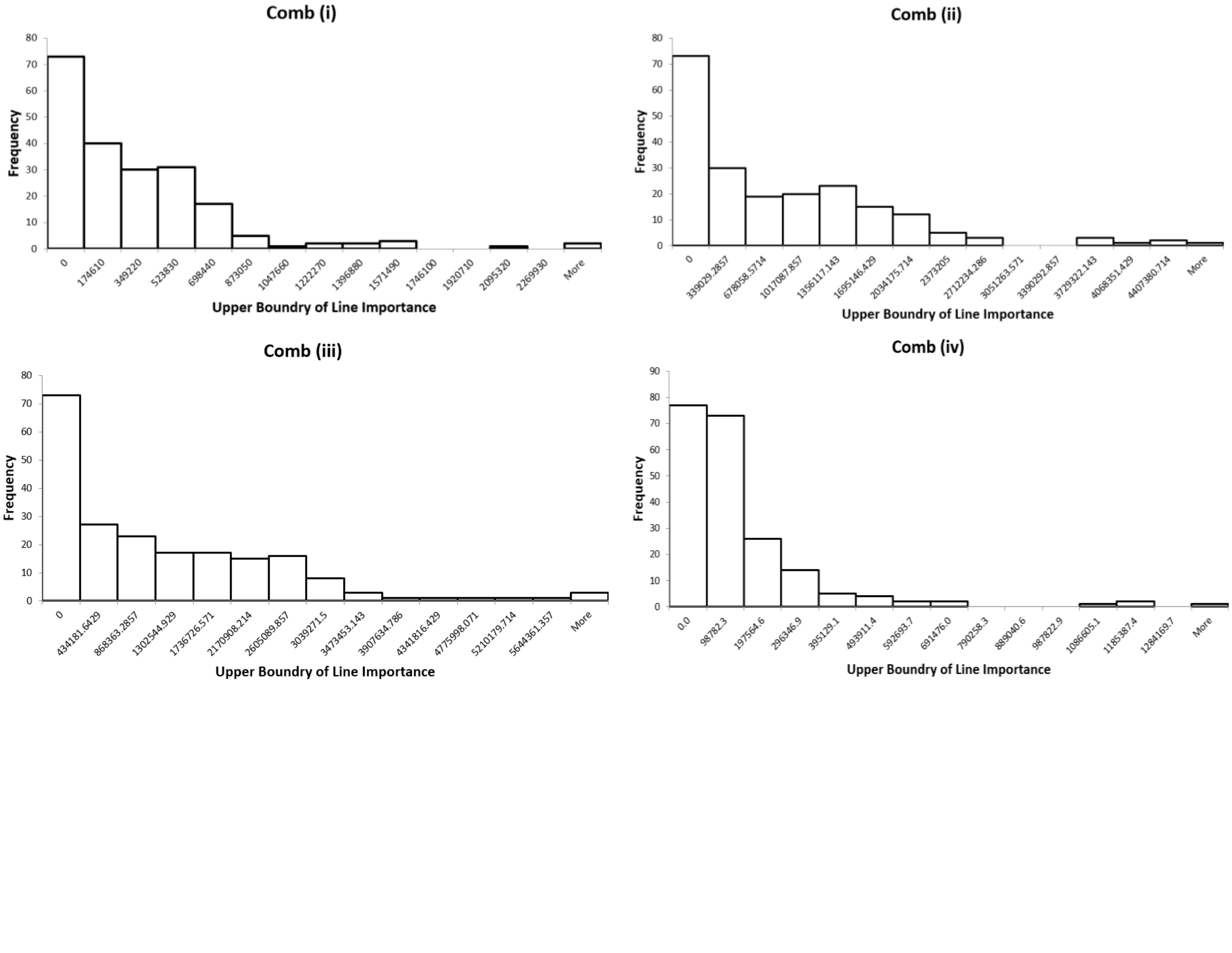
Rank	Transit Line	Comb (i)	Transit Line	Comb (ii)	Transit Line	Comb (iii)	Transit Line	Comb (iv)	Transit Line	Comb (v)	Transit Line	Comb (vi)	Transit Line	Comb (vii)	Transit Line	Comb (viii)	Transit Line	Comb (ix)
1st	Bus Line 80	76592	**Bus Line 80**	286281	**Bus Line 53**	379656	**Bus Line 47**	14100	**Bus Line 6**	123798	**SF2 Bus**	301533	**Bus Line 3006**	126712	**Bus Line 80**	255679	**Bus Line 53**	344676
2nd	Bus Line 84	78717	**Bus Line 19N**	286637	**Bus Line 1087**	395509	**Bus Line 42**	14816	**Bus Line 65**	125189	**Bus Line 19N**	312173	**Bus Line 58/**	127484	**Bus Line 13**	256614	**Bus Line 19**	359061
3rd	Bus Line 35	80802	**Bus Line 84**	306646	**Bus Line 19**	399284	**Bus Line 71**	15238	**Bus Line 3006**	126712	**Bus Line 3316**	326023	**Bus Line 19**	131899	**Bus Line 6**	261505	**SF2 Bus**	361908
4th	Bus Line 70	80962	**Bus Line 13**	308511	**Bus Line 80**	402338	**Bus Line 64**	15711	**Bus Line 19**	131899	**Bus Line 80**	344561	**Bus Line 3133**	136615	**Bus Line 3133**	264633	**Bus Line 80**	380933
5th	Bus Line 58/	82214	**SF2 Bus**	309018	**Bus Line 1085**	409702	**Bus Line 72**	16162	**Bus Line 3133**	135309	**Bus Line 6**	344767	**Bus Line 65**	141599	**Bus Line 1086**	265014	**Bus Line 1087**	392198
5th to last	Train SFM2	1464948	**Bus Line 3107**	3698718	**Bus Line 62**	5173196	**Train SFM1**	688243	**Tram Line 4**	1859343	**Bus Line 62**	4143255	**Tram Line 4**	1905732	**Bus Line 2**	3462352	**Bus Line 62**	4767457
4th to last	Bus Line 1510	1570575	**Bus Line 2**	3979053	**Train SFM2**	5642450	**Bus Line 1511**	1038776	**Train SFM2**	1991141	**Train SFM2**	4552055	**Train SFM2**	2061657	**Bus Line 3107**	3509461	**Train SFM2**	5165753
3rd to last	Bus Line 1432	2046003	**Train SFM2**	4255338	**Bus Line 2**	5713483	**Bus Line 3107**	1158882	**Bus Line 1432**	2057232	**Bus Line 2**	4572215	**Bus Line 1432**	2087111	**Train SFM2**	3541582	**Bus Line 2**	5270282
2nd to last	Train SFM1	2345637	**Tram Line 4**	4403214	**Train SFM1**	5762736	**Bus Line 1510**	1174192	**Bus Line 3107**	2695789	**Tram Line 4**	5014303	**Bus Line 3107**	2701136	**Tram Line 4**	3917820	**Train SFM1**	5287677
Last	Bus Line 3107	2444540	**Train SFM1**	4746410	**Tram Line 4**	6078543	**Bus Line 1432**	1382952	**Train SFM1**	2805728	**Train SFM1**	5078344	**Train SFM1**	2819266	**Train SFM1**	4469991	**Tram Line 4**	5662419

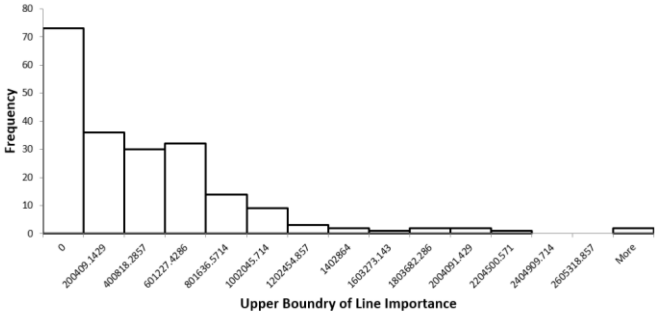
Repeating the same procedure for equity score here for the line importance function leads us to nine histograms as in Figure 4-12. The same data as in Figure 4-12, for other considered cities reported in Appendix C.

A concrete example in Turin is given in Figure 4-13, where we have on the left a line positively contributing to equity. Note that both $\Delta G(l\,,t)$ and $e(l\,,t)$ are higher for line on the left than on the right. The same example for other considered cities shown in Appendix A.

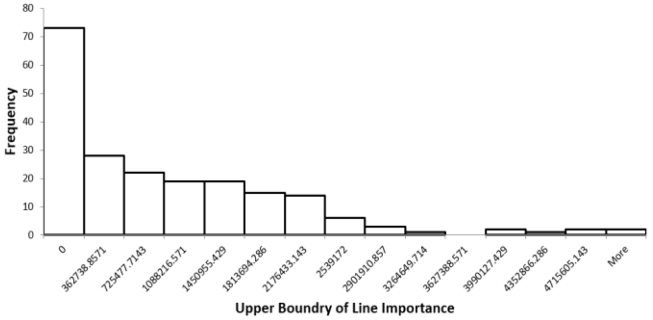
Note also that the $\Delta G(l\,,t)$ numbers show by removing a specific line from the city public transport network how much the value of the Gini index changes. If the line is important for equity, the value of the Gini index goes up, which shows that by removing that line, the area between the Lorenz curve and the perfect equity line has been increased and vice versa. For the line importance, the value of $e(l\,,t)$ increases every time that individual use that line. In other words, the distance an individual travels between two consecutive stops of that specific line will be added to the line importance of that line.

Figure 4-12 Probability distribution function of nine types of Line importance in Turin

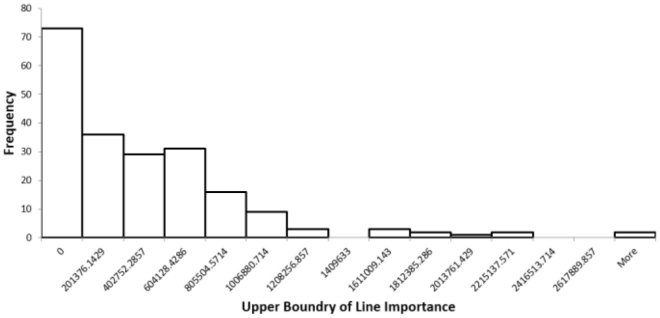




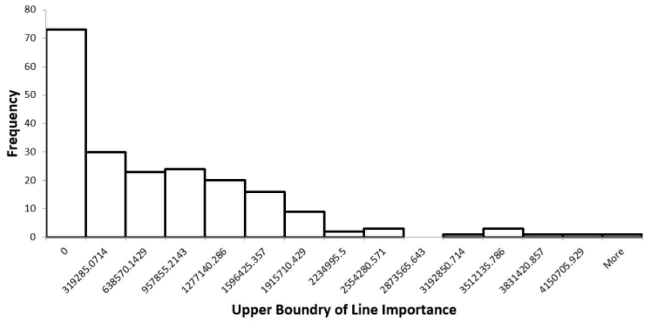
Comb (v)



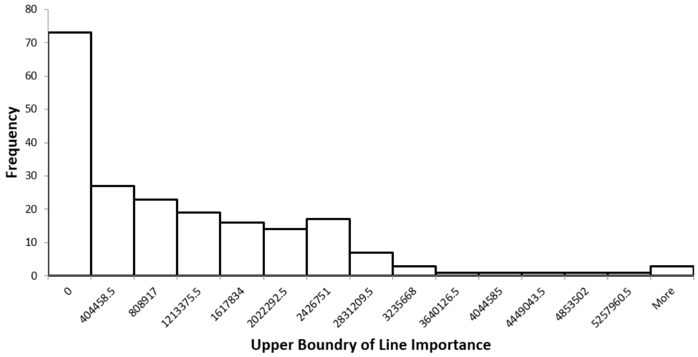
Comb (vi)



Comb (vii)

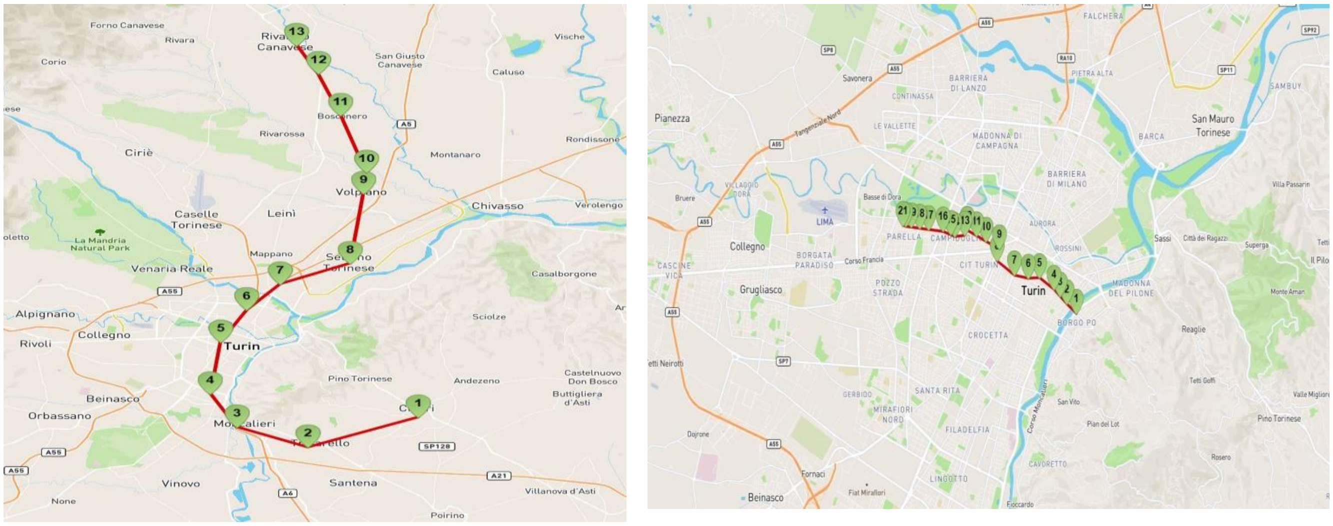


Comb (viii)



Comb (ix)

Figure 4-13 Example of equity score and line importance function in Turin



Turin, line important for equity.
(Train SFM1)
$\Delta G(l\,,t) = 1.3 \cdot 10^{-3}$, $e(l\,,t) = 2.8 \cdot 10^{6}$

Turin, line not contributing to equity.
(Tram line 13)
$\Delta G(l\,,t) = -2.2 \cdot 10^{-4}$, $e(l\,,t) = 3.6 \cdot 10^{5}$

4.4.2. Correlation between the results from equity score $\Delta G(l\,,t)$ and the line importance function $e(l\,,t)$

We select the hexagons of our interest which are those in suburbs based on the nine combinations described in 3.2.3.4.1 and the results shown in 4.4.1.In order to find the correlation between the results from equity score and line importance function, we have to define a new combination that considers the nine combinations defined in 2.2.3.4.2 and four different types of Gini index for equity score. The obtained combinations are reported in Table 4-9. In order to use these combinations to quantify the correlation, we resort to the Pearson coefficient and p-values. To better understand, for example, in combination number 8, we correlate the results of equity score of hexagons based velocity score of the city and the results of line importance function considering the combination (ii) originated from 3.2.3.4.1.

Table 4-9 Different combinations between equity score and line importance function

Combination number	Definition of combination
1	Considering hexagon based sociality score and combination (i) from 3.2.3.4.1
2	Considering hexagon based sociality score and combination (ii) from 3.2.3.4.1
3	Considering hexagon based sociality score and combination (iii) from 3.2.3.4.1
4	Considering individual based sociality score and combination (i) from 3.2.3.4.1
5	Considering individual based sociality score and combination (ii) from 3.2.3.4.1
6	Considering individual based sociality score and combination (iii) from 3.2.3.4.1
7	Considering hexagon based velocity score and combination (i) from 3.2.3.4.1
8	Considering hexagon based velocity score and combination (ii) from 3.2.3.4.1
9	Considering hexagon based velocity score and combination (iii) from 3.2.3.4.1
10	Considering individual based velocity score and combination (i) from 3.2.3.4.1
11	Considering individual based velocity score and combination (ii) from 3.2.3.4.1
12	Considering individual based velocity score and combination (iii) from 3.2.3.4.1
13	Considering hexagon based sociality score and combination (iv) from 3.2.3.4.1
14	Considering hexagon based sociality score and combination (v) from 3.2.3.4.1
15	Considering hexagon based sociality score and combination (vi) from 3.2.3.4.1
16	Considering hexagon based sociality score and combination (vii) from 3.2.3.4.1
17	Considering hexagon based sociality score and combination (viii) from 3.2.3.4.1
18	Considering hexagon based sociality score and combination (ix) from 3.2.3.4.1
19	Considering individual based sociality score and combination (iv) from 3.2.3.4.1
20	Considering individual based sociality score and combination (v) from 3.2.3.4.1
21	Considering individual based sociality score and combination (vi) from 3.2.3.4.1
22	Considering individual based sociality score and combination (vii) from 3.2.3.4.1
23	Considering individual based sociality score and combination (viii) from 3.2.3.4.1
24	Considering individual based sociality score and combination (ix) from 3.2.3.4.1
25	Considering hexagon based velocity score and combination (iv) from 3.2.3.4.1
26	Considering hexagon based velocity score and combination (v) from 3.2.3.4.1
27	Considering hexagon based velocity score and combination (vi) from 3.2.3.4.1
28	Considering hexagon based velocity score and combination (vii) from 3.2.3.4.1
29	Considering hexagon based velocity score and combination (viii) from 3.2.3.4.1
30	Considering hexagon based velocity score and combination (ix) from 3.2.3.4.1
31	Considering individual based velocity score and combination (iv) from 3.2.3.4.1
32	Considering individual based velocity score and combination (v) from 3.2.3.4.1
33	Considering individual based velocity score and combination (vi) from 3.2.3.4.1
34	Considering individual based velocity score and combination (vii) from 3.2.3.4.1
35	Considering individual based velocity score and combination (viii) from 3.2.3.4.1
36	Considering individual based velocity score and combination (ix) from 3.2.3.4.1

We can show the correlation result for all 36 combinations for the seven considered cities that were listed in subsection 3.2.4 in Table 4-10 and Table 4-11. In these tables, we can see the value of the Pearson coefficient is from -1 to +1. When we have a negative Pearson coefficient value, this number is saying the equity score does not increase with the increase of line importance function and this is something against our assumption. As in Table 4-10 and Table 4-11 we do not have a negative Pearson coefficient for Vienna but for the rest of the cities, for example combination 7, for Helsinki and Berlin we have a very high negative value of Pearson coefficient and also, the p-value is significant. So, we can say although this combination is not beneficial for us but it still shows a good correlation between two scores.

Table 4-10 Different combination Pearson coefficient and p-value for all of the considered cities (part1)

		Turin	Aachen	Manchester	Helsinki	Vienna	Budapest	Berlin
Comb 1	**Pearson Coeff**	0.5363	0.3359	0.5862	0.3687	0.3432	0.3838	0.3827
	***p*-value**	8.2E-17	8.7E-11	4.9E-60	1.5E-16	1.4E-08	2.7E-11	1.1E-42
Comb 2	**Pearson Coeff**	0.2871	0.2346	0.1836	-0.0835	0.3668	0.1451	0.2922
	***p*-value**	2.7E-05	8.2E-06	3.1E-06	2.3E-01	1.1E-09	1.5E-02	8.4E-25
Comb 3	**Pearson Coeff**	0.2316	0.1428	-0.0708	-0.1120	0.3804	0.1273	0.2916
	***p*-value**	7.9E-04	7.1E-03	7.4E-02	1.1E-01	2.4E-10	3.3E-02	1.1E-24
Comb 4	**Pearson Coeff**	0.6291	0.1531	0.5529	0.6351	0.4369	0.4775	0.6314
	***p*-value**	3.3E-24	3.9E-03	2.8E-52	8.9E-25	1.7E-13	2.1E-17	4.8E-133
Comb 5	**Pearson Coeff**	0.5917	0.3500	0.2833	0.5825	0.4700	0.4021	0.6383
	***p*-value**	6.2E-21	1.2E-11	3.2E-13	3.3E-20	1.2E-15	2.4E-12	7.6E-137
Comb 6	**Pearson Coeff**	0.5754	0.3555	0.0198	0.5681	0.4852	0.4015	0.6367
	***p*-value**	1.2E-19	5.6E-12	6.2E-01	4.3E-19	1.1E-16	2.6E-12	6.2E-136
Comb 7	**Pearson Coeff**	-0.1084	-0.0981	-0.0467	-0.4707	0.2141	-0.1237	-0.6148
	***p*-value**	1.2E-01	6.5E-02	2.4E-01	3.1E-27	5.2E-04	3.8E-02	2.5E-124
Comb 8	**Pearson Coeff**	-0.3417	-0.4281	-0.5279	-0.6052	0.2348	-0.3801	-0.6819
	***p*-value**	4.7E-07	3.3E-17	5.3E-47	3.3E-48	1.4E-04	4.3E-11	3.8E-163
Comb 9	**Pearson Coeff**	-0.3735	-0.5861	-0.6234	-0.6252	0.2456	-0.3801	-0.6757
	***p*-value**	3.0E-08	4.9E-34	7.3E-70	3.1E-52	6.5E-05	4.4E-11	3.8E-159
Comb 10	**Pearson Coeff**	0.0679	-0.0289	0.1697	0.0978	0.2676	-0.0581	-0.3716
	***p*-value**	3.3E-01	5.9E-01	1.7E-05	3.4E-02	1.3E-05	3.3E-01	3.6E-40
Comb 11	**Pearson Coeff**	-0.1393	-0.1691	-0.3310	-0.1149	0.3032	-0.3231	-0.4321
	***p*-value**	4.5E-02	1.4E-03	9.5E-18	1.3E-02	6.6E-07	3.0E-08	3.5E-55
Comb 12	**Pearson Coeff**	-0.1633	-0.2977	-0.4870	-0.1534	0.3197	-0.3153	-0.4193
	***p*-value**	1.9E-02	1.1E-08	3.0E-39	8.6E-04	1.4E-07	6.7E-08	9.8E-52
Comb 13	**Pearson Coeff**	0.6765	0.3254	0.6239	0.3309	0.3554	0.3850	0.6039
	***p*-value**	4.8E-29	3.6E-10	5.3E-70	1.9E-13	4.0E-09	2.3E-11	7.1E-119
Comb 14	**Pearson Coeff**	0.4806	0.2709	0.4411	0.2449	0.3434	0.1885	0.3755
	***p*-value**	2.3E-13	2.3E-07	1.0E-31	7.8E-08	1.4E-08	1.5E-03	4.7E-41
Comb 15	**Pearson Coeff**	0.2581	0.1356	0.0239	0.2089	0.3691	0.1218	0.2860
	***p*-value**	1.7E-04	1.1E-02	5.5E-01	5.1E-06	8.9E-10	4.1E-02	8.8E-24
Comb 16	**Pearson Coeff**	0.4727	0.2609	0.2195	0.3632	0.4039	0.2203	0.3644
	***p*-value**	6.4E-13	6.4E-07	2.2E-08	4.5E-16	1.4E-11	2.0E-04	1.4E-38
Comb 17	**Pearson Coeff**	0.3056	0.1650	-0.1115	0.2307	0.3398	0.1299	0.2874
	***p*-value**	7.5E-06	1.8E-03	4.8E-03	4.4E-07	2.0E-08	2.9E-02	5.3E-24
Comb 18	**Pearson Coeff**	0.2344	0.0986	-0.1577	0.2069	0.3646	0.1303	0.2907
	***p*-value**	6.8E-04	6.4E-02	6.4E-05	6.3E-06	1.5E-09	2.9E-02	1.5E-24

Table 4-11 Different combination Pearson coefficient and p-value for all of the considered cities (part2)

		Turin	Aachen	Manchester	Helsinki	Vienna	Budapest	Berlin
Comb 19	**Pearson Coeff**	0.5523	0.3472	0.5176	0.6112	0.4226	0.5168	0.5900
	***p*-value**	6.3E-18	1.8E-11	6.1E-45	1.4E-22	1.2E-12	1.4E-20	3.2E-112
Comb 20	**Pearson Coeff**	0.6237	0.3400	0.4727	0.5817	0.4392	0.4162	0.6566
	***p*-value**	1.0E-23	5.0E-11	9.1E-37	3.9E-20	1.2E-13	3.4E-13	2.6E-147
Comb 21	**Pearson Coeff**	0.5776	0.3494	0.1200	0.5677	0.4755	0.3951	0.6394
	***p*-value**	8.2E-20	1.3E-11	2.4E-03	4.7E-19	5.2E-16	6.1E-12	2.0E-137
Comb 22	**Pearson Coeff**	0.6260	0.3412	0.3054	0.6178	0.4460	0.4324	0.6553
	***p*-value**	6.4E-24	4.2E-11	3.2E-15	3.5E-23	4.6E-14	3.1E-14	1.5E-146
Comb 23	**Pearson Coeff**	0.5879	0.3509	-0.0267	0.5747	0.4315	0.3937	0.6407
	***p*-value**	1.2E-20	1.1E-11	5.0E-01	1.4E-19	3.6E-13	7.5E-12	3.6E-138
Comb 24	**Pearson Coeff**	0.5744	0.3414	-0.0768	0.5672	0.4705	0.4048	0.6378
	***p*-value**	1.4E-19	4.1E-11	5.3E-02	5.1E-19	1.1E-15	1.7E-12	1.6E-136
Comb 25	**Pearson Coeff**	0.1889	-0.2822	0.1099	-0.5530	0.2603	-0.1157	-0.3683
	***p*-value**	6.4E-03	6.6E-08	5.5E-03	6.5E-39	2.2E-05	5.3E-02	1.9E-39
Comb 26	**Pearson Coeff**	-0.1738	-0.4041	-0.2886	-0.6037	0.2260	-0.3488	-0.6476
	***p*-value**	1.2E-02	2.5E-15	1.1E-13	6.6E-48	2.4E-04	1.8E-09	4.4E-142
Comb 27	**Pearson Coeff**	-0.3591	-0.6009	-0.5946	-0.6254	0.2371	-0.3874	-0.6855
	***p*-value**	1.1E-07	4.1E-36	3.7E-62	2.9E-52	1.2E-04	1.7E-11	1.6E-165
Comb 28	**Pearson Coeff**	-0.1835	-0.4127	-0.4890	-0.5318	0.3234	-0.3231	-0.6544
	***p*-value**	8.1E-03	5.4E-16	1.4E-39	1.3E-35	1.0E-07	3.0E-08	4.8E-146
Comb 29	**Pearson Coeff**	-0.3299	-0.5765	-0.6309	-0.6119	0.2265	-0.3887	-0.6876
	***p*-value**	1.2E-06	9.8E-33	5.2E-72	1.6E-49	2.4E-04	1.4E-11	6.7E-167
Comb 30	**Pearson Coeff**	-0.3725	-0.6255	-0.6376	-0.6277	0.2348	-0.3755	-0.6753
	***p*-value**	3.3E-08	7.9E-40	5.6E-74	9.5E-53	1.4E-04	7.7E-11	7.0E-159
Comb 31	**Pearson Coeff**	0.3324	-0.0244	0.2744	-0.0270	0.2886	-0.1185	-0.0977
	***p*-value**	9.9E-07	6.5E-01	1.8E-12	5.6E-01	2.3E-06	4.7E-02	7.5E-04
Comb 32	**Pearson Coeff**	0.0018	-0.1479	-0.0486	-0.1170	0.2790	-0.3093	-0.3872
	***p*-value**	9.8E-01	5.3E-03	2.2E-01	1.1E-02	5.1E-06	1.2E-07	9.4E-44
Comb 33	**Pearson Coeff**	-0.1561	-0.3181	-0.4314	-0.1558	0.3088	-0.3268	-0.4370
	***p*-value**	2.5E-02	9.1E-10	2.9E-30	7.1E-04	4.0E-07	2.0E-08	1.5E-56
Comb 34	**Pearson Coeff**	-0.0050	-0.1536	-0.2782	0.0038	0.3393	-0.2881	-0.3976
	***p*-value**	9.4E-01	3.8E-03	8.7E-13	9.3E-01	2.1E-08	9.0E-07	3.0E-46
Comb 35	**Pearson Coeff**	-0.1337	-0.2943	-0.5079	-0.1349	0.2747	-0.3341	-0.4408
	***p*-value**	5.5E-02	1.7E-08	4.4E-43	3.4E-03	7.3E-06	9.4E-09	1.3E-57
Comb 36	**Pearson Coeff**	-0.1636	-0.3485	-0.5290	-0.1579	0.3044	-0.3094	-0.4174
	***p*-value**	1.9E-02	1.5E-11	3.3E-47	6.0E-04	5.9E-07	1.2E-07	3.1E-51

p-value can be seen as the probability of observing the measured correlation "by chance", under the hypothesis that there is actually no dependency between $e(l\,,t)$ and $\Delta G(l\,,t)$. As an experimental setting, we consider the cut-off p-value equal to 0.05, which means when p-value is higher than 0.05, the correlation is considered totally "by chance".As in Table 4-10 and Table 4-11, p-values with the value below 0.05 highlighted in red, and with a value higher than 0.05 are highlighted in green. Consequently, the red values shows that the probability that our Pearson correlation coefficient result be random are insignificant. Which means our Pearson coefficient shows the correlation which is not random.

In order to find the best combination, one methodological choice could be to exclude those combinations with negative Pearson coefficient and p-value higher than the cut-off value, which is 0.05 in at least one city. by doing so, we reach combinations 1, 4, 5, 13, 14, 16, 19, 20, 21 and 22.

Table 4-12 Remained combination after exclusion those with negative Pearson coefficient or high p-value in at least one city

Combination number	Definition of combination
1	Considering hexagon based sociality score and combination (i) from 3.2.3.4.1
4	Considering individual based sociality score and combination (i) from 3.2.3.4.1
5	Considering individual based sociality score and combination (ii) from 3.2.3.4.1
13	Considering hexagon based sociality score and combination (iv) from 3.2.3.4.1
14	Considering hexagon based sociality score and combination (v) from 3.2.3.4.1
16	Considering hexagon based sociality score and combination (vii) from 3.2.3.4.1
19	Considering individual based sociality score and combination (iv) from 3.2.3.4.1
20	Considering individual based sociality score and combination (v) from 3.2.3.4.1
21	Considering individual based sociality score and combination (vi) from 3.2.3.4.1
22	Considering individual based sociality score and combination (vii) from 3.2.3.4.1

In an effort to select one combination among the above mentioned ones, we have to exclude the combinations one by one to reach to the best. In the first step, between the combination 1 and 4 we exclude combination 1. Because according to 3.2.2.1.1 and 3.2.2.1.2, the individual-based sociality score is scaled by the population of each hexagon but the hexagon-based is not scaled by the population of each hexagon. In the next step, between combinations 13 and 19 based on the same reason as before we exclude combination 13. In the third step, between combination 14 and 20 we exclude combination 14. In the last step, between 16 and 22, we exclude combination 16.

Between the remaining combinations, we also exclude combinations 4 and 5. Because, they are based on $\frac{r}{R}$ which is a measure not fit for cities that are not mono centric. Finally, we have combinations 19, 20, 21, and 22 which between them is hard to select just one combination.

Another way to select best combination is to compute the average for each combination among different cities as in Table 4-13. The two maximum averages in the table are for combinations 4 and 20.

Table 4-13 Different combination's Pearson coefficient and average over all cities

Combination Number	Average
1	0.4195
2	0.2037
3	0.1416
4	0.5023
5	0.4740
6	0.4346
7	-0.1783
8	-0.3900
9	-0.4312
10	0.0206
11	-0.1723
12	-0.2166
13	0.4716
14	0.3350
15	0.2005
16	0.3293
17	0.1924
18	0.1668
19	0.4983
20	0.5043
21	0.4464
22	0.4892
23	0.4218
24	0.4170
25	-0.1086
26	-0.3201
27	-0.4308
28	-0.3244
29	-0.4284
30	-0.4399
31	0.0897
32	-0.1042
33	-0.2166
34	-0.1113
35	-0.2244
36	-0.2316

Combination 4 is the correlation between the equity score of the individual-based sociality score with the line importance function value of the hexagons with $\frac{r}{R} \geq \frac{1}{3}$, whereas combination 20 is the correlation between the equity score of the individual-based sociality score with the line importance function value of the first 65% of hexagons sorted by increasing sociality score.

Here we will choose combination 20 because it has the highest average Pearson correlation and very small p-values. Moreover, it does not depend on $\frac{r}{R}$, so it can be applied in cities that are not mono-centric, it also considers the population which is hidden in the computation of sociality score. So, now we use the combination 20 to replicate our method in different cities, which means the formula from 4.4.2 updates as follows:

$$e(l\,,t) = I(\lambda^{65th}, l\,, t) = \sum_{\lambda'=1}^{\lambda^{65th}} i(\lambda', l\,, t)$$

We now empirically confirm that $e(l)$ captures the same information of $\Delta G(l)$ in a much more computationally efficient way. The correlation between the two equity scores $\Delta G(l\,,t)$ and $e(l\,,t)$ appears in the scatterplots of Figure 4-14, where each point corresponds to a line l, whose x coordinate is $\Delta G(l\,,t)$ and the y coordinate is $e(l\,,t)$. To generalize this observation, Table 4-14 lists the selected combination Pearson's correlation coefficient between the values $\Delta G(l\,,t)$ and $e(l\,,t)$, for all lines $l \in \mathcal{L}$ in all seven considered cities, we observe a relatively strong correlation. The fact that the two scores carry very similar information is confirmed by extremely low p-values.

Figure 4-14 Correlation between two equity scores for all of the considered cities

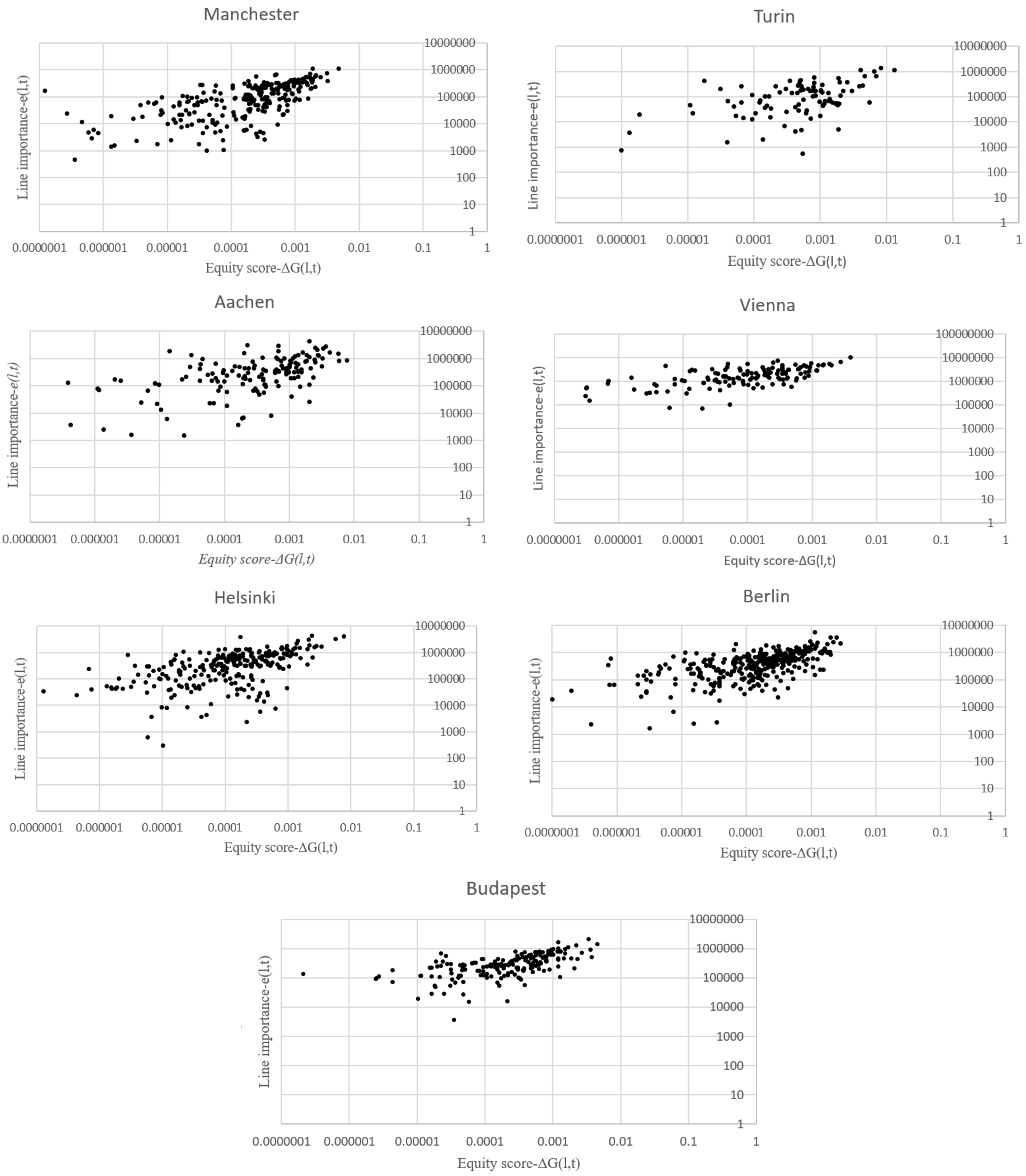


Table 4-14 Correlation between $e(l\,,t)$ and $\Delta G(l\,,t)$

City	Pearson Coeff	*p*-value
Manchester	0.62	6E-18
Turin	0.34	2E-11
Aachen	0.47	6E-45
Vienna	0.58	1E-22
Helsinki	0.43	1E-12
Berlin	0.42	1E-20
Budapest	0.66	3E-112

4.4.2.1. Computational gain

Computing $e(l\,,t)$ is several hundred times faster than $\Delta G(l\,,t)$. To give an idea of the order of magnitude of computation time, we report in Table 4-15 the time needed to compute the Gini index $G(\mathcal{L}\,,t)$, the time needed to compute the set of all equity scores $\Delta G(l\,,t)$ and to compute the set of all line importance function $e(l\,,t)$. The results are obtained on a virtual machine using 32 AMD EPYC 7532 CPUs and 64GB of memory. Note that the time to compute $\{\Delta G(l\,,t)|l \in \mathcal{L}\}$ is estimated (multiplying by $|\mathcal{L}|$ the time for computing $G(\mathcal{L}\,,t)$). It is evident that for bigger cities like Paris, computing $\{\Delta G(l\,,t)|l \in \mathcal{L}\}$ into an optimization loop would be impractical. In smaller cities, one should wait some hours or days to get $\{\Delta G(l\,,t)|l \in \mathcal{L}\}$, which might be acceptable in some cases. However, it would be impossible to tolerate such high times if one wants to use the information about equity score in some optimization loops for network design. In this case, if metaheuristics, e.g., generic algorithms, or artificial intelligence, e.g., reinforcement learning, are used, these optimization iterations could be hundreds or thousands. In this case, it would be impossible to compute $\{\Delta G(l\,,t)|l \in \mathcal{L}\}$ and we could resort to $\{e(l\,,t)|l \in \mathcal{L}\}$, which is several orders of magnitude faster to compute.

Table 4-15 Computation time

	$\boldsymbol{G(t)}$	$\boldsymbol{\{\Delta G(l,t)|l \in \mathcal{L}\}}$	$\boldsymbol{\{e(l,t)|l \in \mathcal{L}\}}$
Manchester	300 sec	2.2 days	~5 min
Turin	60 sec	3.5 hours	~1 min
Aachen	90 sec	8-9 hours	~90 sec
Vienna	40 sec	2.9 hours	~40 sec
Helsinki	95 sec	12.3 hours	~ 95 sec
Berlin	210 sec	2.8 days	~4 min
Budapest	50 sec	3.9 hours	~ 50 sec
Paris	720 sec	15 days	~12 min

5. Chapter 5: Conclusions

This work has developed a method to assess public transportation accessibility equity and to evaluate the contribution of individual transit lines to achieve better equity through an open data based approach. Thus, in the first part of the work, to capture the inequality in the city, we have proposed a methodology to compute equity in the distribution of transportation quality, which can be fully automated and easily performed for several cities worldwide. This is guaranteed by the fact that the computation only relies on open data in a standardized form. We performed a comparative analysis of four cities, and our results confirm previous work findings, in some cases better capturing the accessibility equity differences between cities.

In the second part, we have proposed a methodology to find the most important lines to improve equity for the population in suburbs. The proposed line importance function can be used to guide the investment choices of transit operators. Indeed, if the budget were infinite, achieving a good accessibility distribution would be easy: massive investment could be dedicated to increasing all lines' frequency (and thus the fleet). However, due to the limitedness of the budget, operators need to make choices and prioritize certain lines over others. We advocate that such prioritization choices should not be (solely) made to improve average or total accessibility or social welfare. We believe instead that it is fairer to prioritize those lines that favor accessibility.

By looking at the results in 4.4.1, and especially for Turin in Figure 4-13, if we want to find an observational reason to say a line is more important than another for equity, most of the time, the lines that touch the periphery of the city has more influence on the equity because that line could be used for the mobility of the residence people in a suburb like train SFM1. On the other hand, Tram line 13, since it does not touch the periphery, does not have a high line importance function value and is less important. However, please note that this observational procedure can be used just for the lines whose scores have a high difference and one of them touches the periphery. If we want to compare two lines that both serve the periphery, we should rely just on the proposed method.

According to the above discussion, the limitations of the methodologies are the following:

1. Considering the center of the hexagon as the destination point, that can influence the accuracy of results
2. The accessibility measures do not take into account the demand. We proxy the demand by considering the population, which is not the actual demand, although it surely influences it.
3. We face the edge effect by considering the city as our area of study. Because the extension of hexagons ends somewhere more or less at the city's border, however, there could be some individuals that live out of the area of study but work in the city, and they should be considered in the computation.
4. This methodology would not be helpful in cities where just some part of the city is serve by public transport like the center and the other parts are not served or serve by scattered stations. According to the methodology in 3.2.1.1, we remove those hexagons with a walking path higher than 15

minutes. Consequently, we eliminate those hexagons that are far from any public transit stop and some accessible hexagons remains, and we get a very equitable city because, the remaining hexagons are those with high accessibility score and parts of the city without public transport service are out of the computation. However, this limitation makes us realize that the walking path is another experimental setting that could be different in different cities to capture all the individuals, even those far from any public transit stop.

5. It is good to consider also vertical equity, which is equality between unequal based on socioeconomic characteristics like using Suits coefficient according to what we described in 2.3.

Despite this study's limitation, the results can be a step forward in deepening our knowledge about inequity in the provision of transit services. Furthermore, finding the lines most contributing to the accessibility equity of the network can be a guide to distinguishing them and choosing the best countermeasures to improve the equity.

Nevertheless, this study is just an onset to provide a methodology to assess the transit network to find the most important lines for inequity. According to the literature, there were no previous studies about this matter. However, further research could study the optimal design of future transit, in which demand-responsive buses co-exist with classic fixed lines. The proposed method will guide selecting which fixed lines to keep and which could be replaced by demand-responsive buses. When available, it would also be possible to consider how to enrich accessibility computation with additional (no-public) data, like employment, business locations, and types.

References

1. ABREHA, D. A. Analysing public transport performance using efficiency measures and spatial analysis: The case of Addis Ababa, Ethiopia. 2007. ITC Enschede, The Netherlands.

2. BANISTER, D. 2018. *Inequality in transport*.

3. BEN-AKIVA, M. E., LERMAN, S. R. & LERMAN, S. R. 1985. *Discrete choice analysis: theory and application to travel demand*, MIT press.

4. BIAZZO, I., MONECHI, B. & LORETO, V. 2019. General scores for accessibility and inequality measures in urban areas. *Royal Society open science,* 6**,** 190979.

5. BIOSCA, O., SPIEKERMANN, K. & STĘPNIAK, M. 2013. Transport accessibility at regional scale. *Europa XXI,* 24**,** 5-17.

6. CALABRÒ, G., ARALDO, A., OH, S., SESHADRI, R., INTURRI, G. & BEN-AKIVA, M. Integrating fixed and demand-responsive transportation for flexible transit network design. TRB 2021: 100th Annual Meeting of the Transportation Research Board, 2021. Transportation Research Board, TRBAM-21-02493.

7. DELBOSC, A. & CURRIE, G. 2011. Using Lorenz curves to assess public transport equity. *Journal of Transport Geography,* 19**,** 1252-1259.

8. DIBBELT, J., PAJOR, T., STRASSER, B. & WAGNER, D. Intriguingly simple and fast transit routing. International Symposium on Experimental Algorithms, 2013. Springer, 43-54.

9. EUROSTAT. 2018. *Eurostat 2018 Population Grid* [Online]. https://ec.europa.eu/eurostat/web/gisco/geodata/reference-data/population-distribution-demography/geostat.

10. GEURS, K. T. & VAN WEE, B. 2004. Accessibility evaluation of land-use and transport strategies: review and research directions. *Journal of Transport geography,* 12**,** 127-140.

11. GINI, C. 1912. *Variabilità e mutabilità: contributo allo studio delle distribuzioni e delle relazioni statistiche.[Fasc. I.]*, Tipogr. di P. Cuppini.

12. GOLUB, A. & MARTENS, K. 2014. Using principles of justice to assess the modal equity of regional transportation plans. *Journal of Transport Geography,* 41**,** 10-20.

13. GOOGLE TRANSIT FEED. 2022. *GTFS Data Specification* [Online]. https://developers.google.com/transit/gtfs.

14. GRELIER, F. 2018. CO 2 EMISSIONS FROM CARS: the facts A report by Transport & Environment Acknowledgements.

15. HANDY, S. L. & NIEMEIER, D. A. 1997. Measuring accessibility: an exploration of issues and alternatives. *Environment and planning A,* 29**,** 1175-1194.

16. HANSEN, W. G. 1959. How accessibility shapes land use. *Journal of the American Institute of planners,* 25**,** 73-76.

17. HAWAS, Y. E., HASSAN, M. N. & ABULIBDEH, A. 2016. A multi-criteria approach of assessing public transport accessibility at a strategic level. *Journal of Transport Geography,* 57**,** 19-34.

18. ÎLE-DE-FRANCE, R. 2017. *Population active occupée des communes d'Île-de-France (données Insee)* [Online]. Available: https://data.iledefrance.fr/explore/dataset/population-active-occupee-des-communes-dile-de-france-donnee-insee0/information/?location=11,48.93242,2.3909&basemap=jawg.streets [Accessed].

19. ISTAT. 2011. *BASI TERRITORIALI E VARIABILI CENSUARIE* [Online]. Available: https://www.istat.it/it/archivio/104317 [Accessed].

20. KWAN, M. P. 1998. Space-time and integral measures of individual accessibility: a comparative analysis using a point-based framework. *Geographical analysis,* 30**,** 191-216.

21. LODOVICI, M. S. & TORCHIO, N. 2015. Social inclusion in EU public transport.

22. LORENZ, M. O. 1905. Methods of measuring the concentration of wealth. *Publications of the American statistical association,* 9**,** 209-219.

23. LUCAS, K. 2012. Transport and social exclusion: Where are we now? *Transport policy,* 20**,** 105-113.

24. LUXEN, D. & VETTER, C. Real-time routing with OpenStreetMap data. Proceedings of the 19th ACM SIGSPATIAL international conference on advances in geographic information systems, 2011. 513-516.

25. MAVOA, S., WITTEN, K., MCCREANOR, T. & O'SULLIVAN, D. 2012. GIS based destination accessibility via public transit and walking in Auckland, New Zealand. *Journal of transport geography,* 20**,** 15-22.

26. MILLER, E. J. Measuring accessibility: Methods and issues. 2020. International Transport Forum Discussion Paper.

27. MILLER, H. J. 2005. Place-based versus people-based accessibility. *Access to destinations.* Emerald Group Publishing Limited.

28. MURRAY, M. & BERWICK, D. M. 2003. Advanced access: reducing waiting and delays in primary care. *Jama,* 289**,** 1035-1040.

29. NIEHAUS, M., GALILEA, P. & HURTUBIA, R. 2016. Accessibility and equity: An approach for wider transport project assessment in Chile. *Research in Transportation Economics,* 59**,** 412-422.

30. RUBENSSON, I., SUSILO, Y. & CATS, O. 2020. Fair accessibility–Operationalizing the Distributional effects of policy interventions. *Journal of Transport Geography,* 89**,** 102890.

31. SAGHAPOUR, T., MORIDPOUR, S. & THOMPSON, R. G. 2016. Public transport accessibility in metropolitan areas: A new approach incorporating population density. *Journal of Transport Geography,* 54**,** 273-285.

32. SEDAC. 2020. *Gridded Population of the World (GPW), v4* [Online]. Available: https://sedac.ciesin.columbia.edu/data/set/gpw-v4-population-density-rev11 [Accessed].

33. SUITS, D. B. 1977. Measurement of tax progressivity. *The American Economic Review,* 67**,** 747-752.

34. TURCOTTE, M. 2008. Dependence on cars in urban neighbourhoods: Life in metropolitan areas. *Canadian Social Trends, Statistics Canada (www. statcan. ca)*.

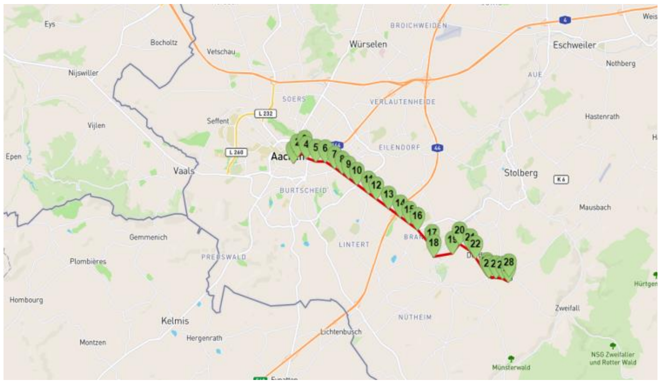
35. WEI, R., LIU, X., MU, Y., WANG, L., GOLUB, A. & FARBER, S. 2017. Evaluating public transit services for operational efficiency and access equity. *Journal of transport geography,* 65**,** 70-79.

36. WELCH, T. F. & MISHRA, S. 2013. A measure of equity for public transit connectivity. *Journal of Transport Geography,* 33**,** 29-41.

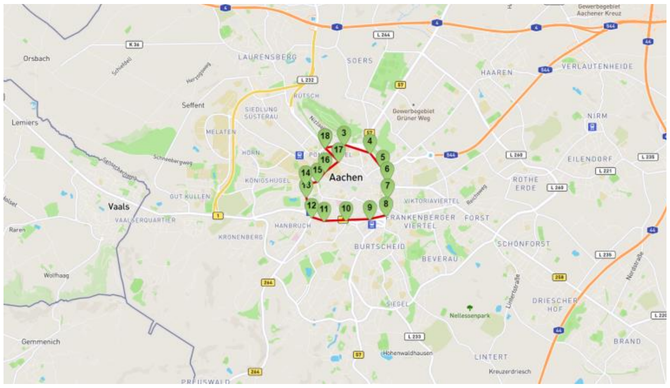
37. WORX, P. H. S. P. S. 2017. *Time for commuting to and from work* [Online]. Available: https://www.sdworx.com/about-sd-worx/press/2018-09-20-more-20-europeans-commute-least-90-minutes-daily [Accessed].

38. YATSKIV, I., BUDILOVICH, E. & GROMULE, V. 2017. Accessibility to Riga public transport services for transit passengers. *Procedia Engineering,* 187**,** 82-88.

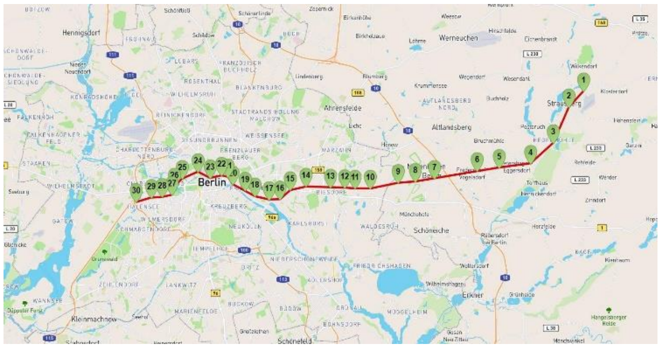
Appendix A: lines mostly and lastly contributing to equity in the considered cities



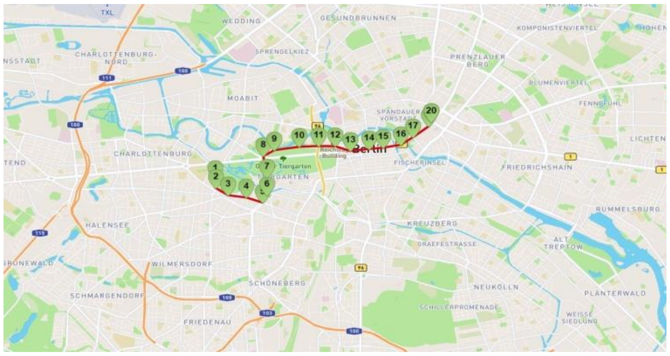
Aachen, line important for equity.
(Local Bus Line 15)
$\Delta G(l,t) = 3.3 \cdot 10^{-5}$, $e(l,t) = 1.6 \cdot 10^{4}$



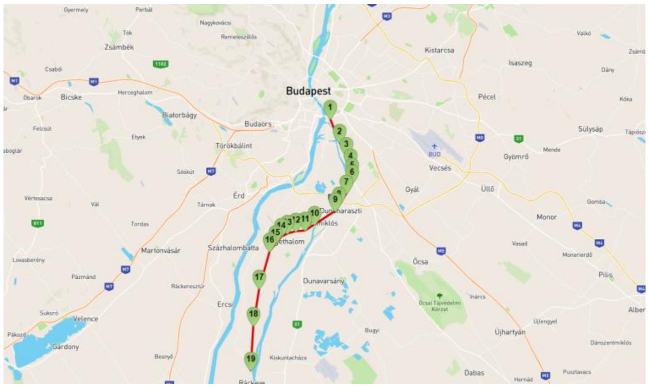
Aachen, line important for equity.
(Local Bus Line 13B)
$\Delta G(l,t) = -2.2 \cdot 10^{-4}$, $e(l,t) = 4 \cdot 10^{3}$



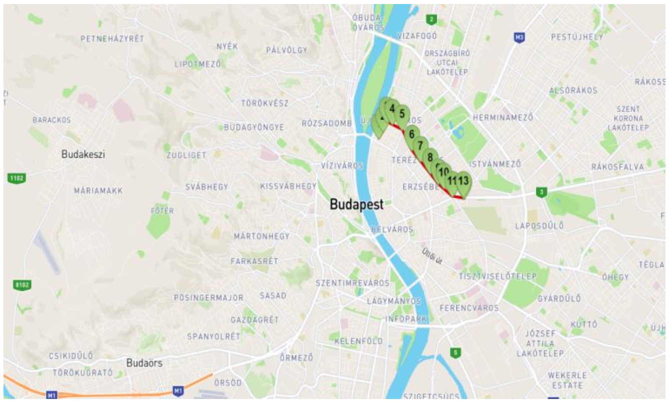
Berlin, line important for equity.
(Rail Line S5)
$\Delta G(l,t) = 7.6 \cdot 10^{-4}$, $e(l,t) = 8.0 \cdot 10^{6}$



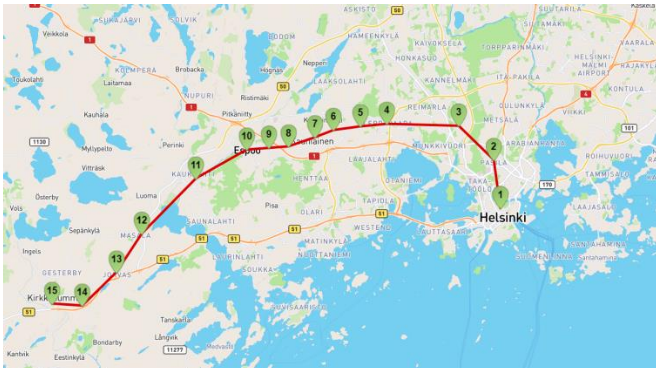
Berlin, line important for equity.
(Bus Line 100)
$\Delta G(l,t) = 7.3 \cdot 10^{-7}$, $e(l,t) = 2.0 \cdot 10^{6}$



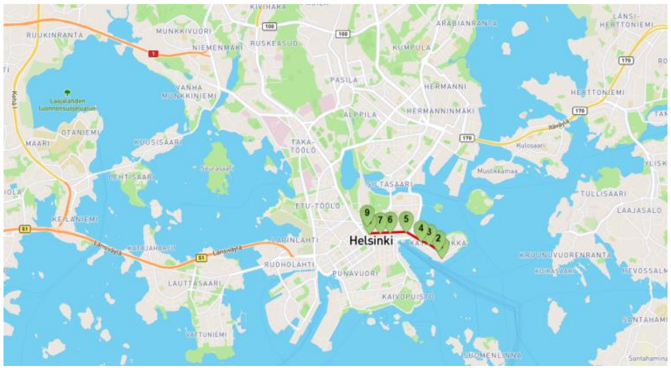
Budapest, line important for equity.
(Train Line H6)
$\Delta G(l\,,t) = 3 \cdot 10^{-3}$, $e(l\,,t) = 1.5 \cdot 10^{6}$



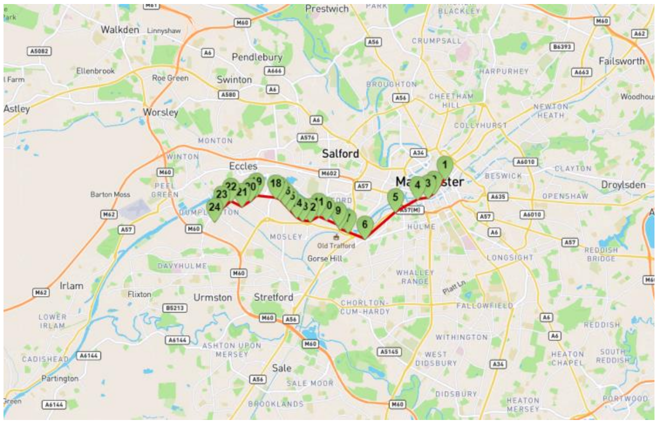
Budapest, line important for equity.
(Bus Line 76)
$\Delta G(l\,,t) = -5.6 \cdot 10^{-5}$, $e(l\,,t) = 1.3 \cdot 10^{5}$



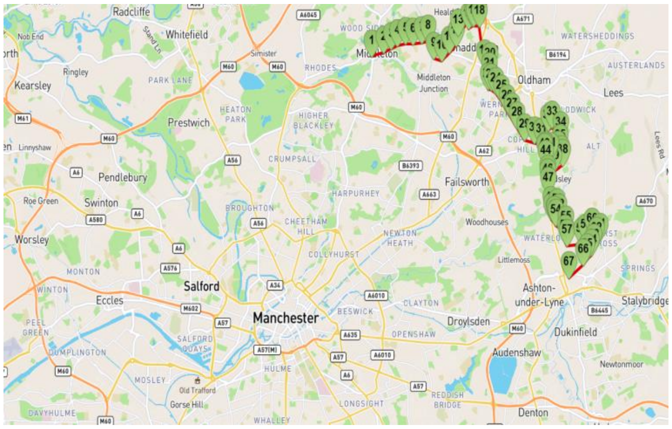
Helsinki, line important for equity.
(Train Line U)
$\Delta G(l\,,t) = 1.3 \cdot 10^{-3}$, $e(l\,,t) = 4.3 \cdot 10^{6}$



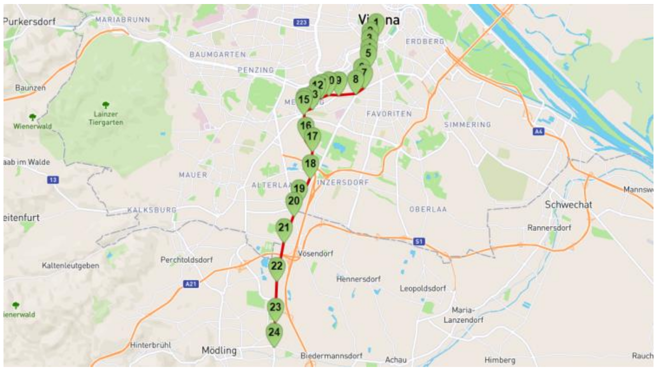
Helsinki, line important for equity.
(Tram Line 5)
$\Delta G(l\,,t) = 4.6 \cdot 10^{-5}$, $e(l\,,t) = 1.7 \cdot 10^{4}$



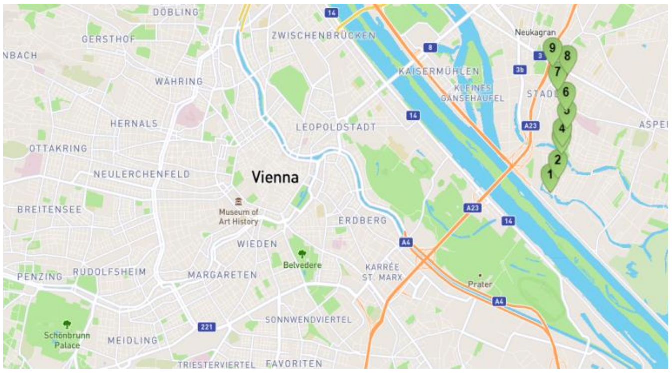
Manchester, line important for equity.
(Bus Line X50)
$\Delta G(l\,,t) = 1.3 \cdot 10^{-3}$, $e(l\,,t) = 9.6 \cdot 10^{4}$



Manchester, line important for equity.
(Bus Line X50)
$\Delta G(l\,,t) = -1.2 \cdot 10^{-4}$, $e(l\,,t) = 6.9 \cdot 10^{3}$



Vienna, line important for equity.
(Tram WLB)
$\Delta G(l\,,t) = 1.5 \cdot 10^{-2}$, $e(l\,,t) = 3.5 \cdot 10^{6}$



Vienna, line important for equity.
(Bus Line 96A)
$\Delta G(l\,,t) = 2.8 \cdot 10^{-6}$, $e(l\,,t) = 1.1 \cdot 10^{5}$

Appendix B: Histograms of four types of equity score and the value for the five best and worst lines and in the considered cities

Figure 0-1 Vienna

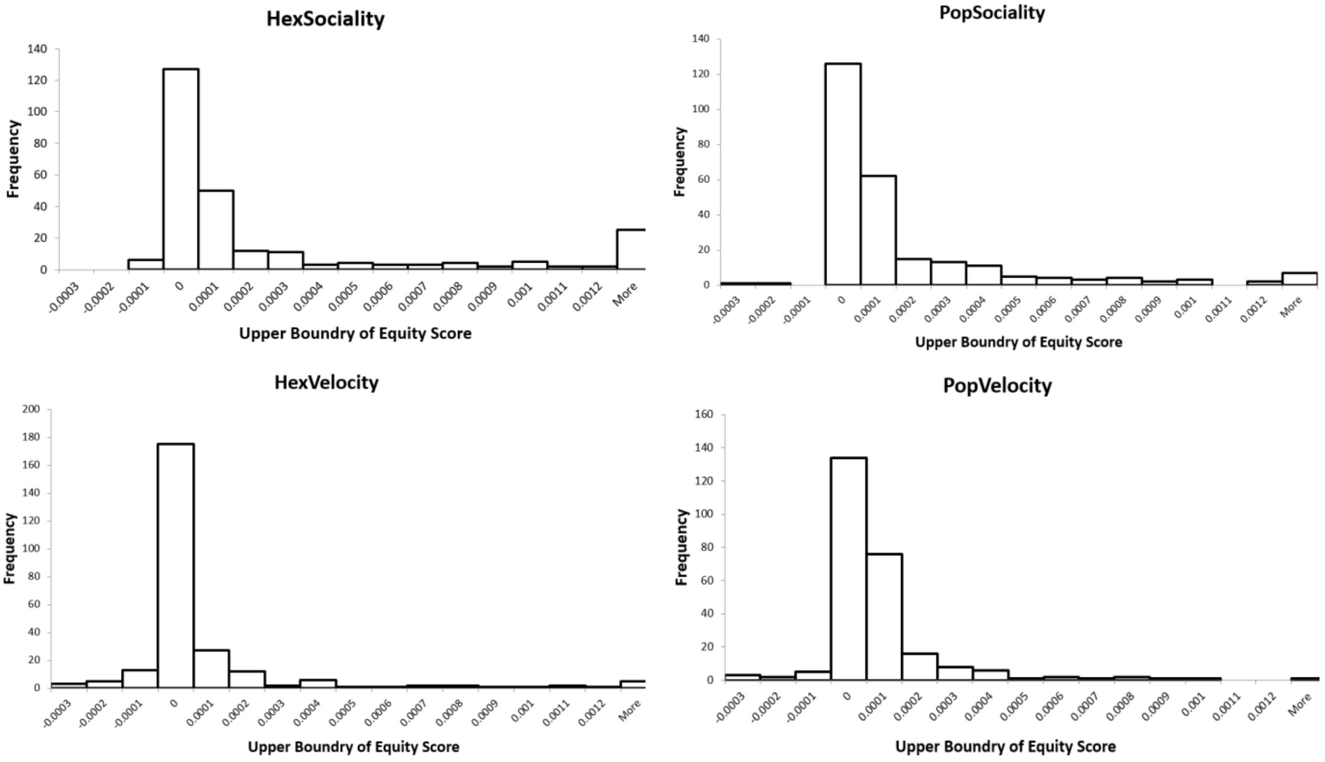


Table 0-1 Vienna

Rank	Transit Line	HexSociality	Transit Line	PopSociality	Transit Line	HexVelocity	Transit Line	PopVelocity
1st	Bus Line 20B	-0.0002	Bus Line 57A	-0.0004	Metro Line U6	-0.0008	Metro Line U6	-0.0009
2nd	Tram Line 18	-0.0002	Tram Line 5	-0.0003	Tram Line 26	-0.0004	Metro Line U3	-0.0003
3rd	Bus Line 11B	-0.0001	Bus Line 13A	-0.0001	Tram Line 2	-0.0003	Bus Line 57A	-0.0003
4th	Tram Line 5	-0.0001	Bus Line 14A	-0.0001	Tram Line 25	-0.0003	Tram Line 5	-0.0003
5th	Bus Line 57A	-0.0001	Bus Line 11B	-0.0001	Metro Line U3	-0.0003	Tram Line O	-0.0002
5th to last	Bus Line 79B	0.0032	Tram Line 11	0.0015	Bus Line 89A	0.0012	Bus Line 66A	0.0008
4th to last	Bus Line 16B	0.0035	Bus Line 66A	0.0018	Bus Line 60A	0.0013	Bus Line 60A	0.0008
3rd to last	Bus Line 50A	0.0036	Bus Line 15A	0.0020	Bus Line 32A	0.0017	Bus Line 15A	0.0009
2nd to last	Bus Line 32A	0.0043	Tram Line 26	0.0027	Bus Line 25A	0.0018	Bus Line 32A	0.0010
Last	Tram Line WLB	0.0155	Tram Line WLB	0.0039	Tram Line WLB	0.0103	Tram Line WLB	0.0040

Figure 0-2 Manchester

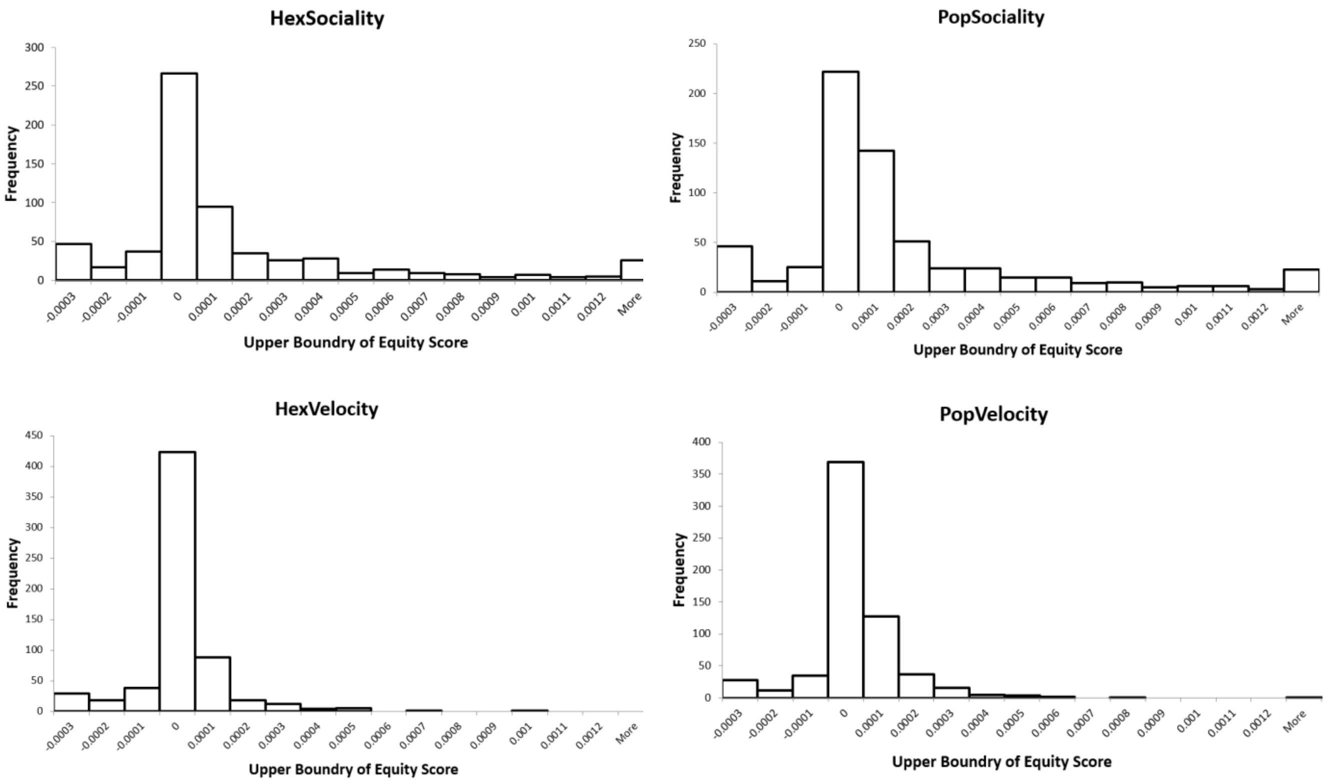


Table 0-2 Manchester

Rank	Transit Line	HexSociality	Transit Line	PopSociality	Transit Line	HexVelocity	Transit Line	PopVelocity
1st	Tram Line 5i	-0.0025	Tram Line 2o	-0.0034	Bus Line 471	-0.0010	Tram Line 5i	-0.0010
2nd	Tram Line 2o	-0.0024	Tram Line 4o	-0.0029	Tram Line 5i	-0.0009	Tram Line 4o	-0.0008
3rd	Tram Line 4o	-0.0020	Bus Line 192	-0.0026	Bus Line 471	-0.0008	Tram Line 5o	-0.0008
4th	Bus Line 192	-0.0018	Tram Line 5i	-0.0024	Bus line 17	-0.0007	Bus Line 8	-0.0007
5th	Bus Line 59	-0.0016	Bus Line 53	-0.0016	Tram Line 5o	-0.0007	Bus Line 582	-0.0007
5th to last	Bus Line 199	0.0021	Bus Line 330	0.0018	Bus Line 88	0.0004	Bus Line 375	0.0005
4th to last	Bus Line 352	0.0024	Bus Line 10	0.0020	Bus Line 358	0.0004	Bus Line 352	0.0005
3rd to last	Bus Line 125	0.0031	Bus Line 125	0.0023	Bus Line 375	0.0005	Bus Line 125	0.0006
2nd to last	Bus Line 1	0.0032	Bus Line 464	0.0025	Bus line 1	0.0006	Bus Line 464	0.0008
Last	Bus Line 464	0.0048	Bus Line 1	0.0028	Bus Line 199	0.0009	Bus Line 1	0.0013

Figure 0-3 Helsinki

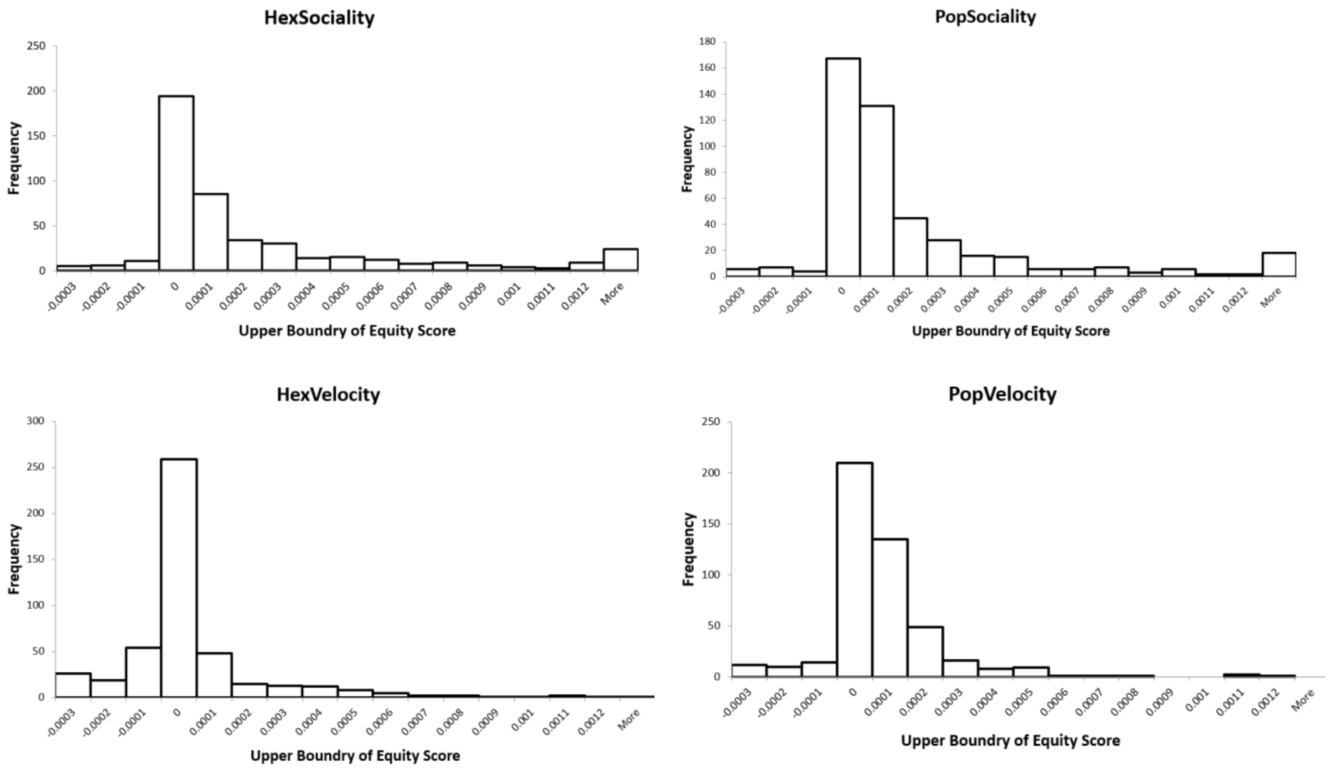


Table 0-3 Helsinki

Rank	Transit Line	HexSociality	Transit Line	PopSociality	Transit Line	HexVelocity	Transit Line	PopVelocity
1st	Bus Line 54	-0.0016	Bus Line 57	-0.0005	Train Line K	-0.0017	Bus Line 54	-0.0010
2nd	Metro Line 2	-0.0009	Bus Line 70	-0.0005	Bus Line 560	-0.0014	Train Line I	-0.0010
3rd	Bus Line 57	-0.0007	Bus Line 506	-0.0004	Train Line P	-0.0014	Metro Line 2	-0.0010
4th	Bus Line 550	-0.0007	Bus Line 280	-0.0004	Bus Line 550	-0.0013	Train Line P	-0.0008
5th	Train Line A	-0.0003	Bus Line 500	-0.0003	Metro Line 1	-0.0013	Bus Line 550	-0.0008
5th to last	Bus Line 841	0.0025	Bus Line 544	0.0026	Bus Line 907	0.0009	Bus Line 171	0.0007
4th to last	Bus Line 246T	0.0025	Train Line E	0.0028	Bus Line 961	0.0011	Bus Line 21	0.0008
3rd to last	Train Line K	0.0031	Bus Line 560	0.0034	Bus Line 246T	0.0011	Train Line R	0.0010
2nd to last	Bus Line 345	0.0032	Metro Line 1	0.0057	Bus Line 987A	0.0012	Metro Line 1	0.0011
Last	Bus Line 844	0.0037	Train Line K	0.0076	Bus Line 788K	0.0019	Bus Line 641	0.0012

Figure 0-4 Budapest

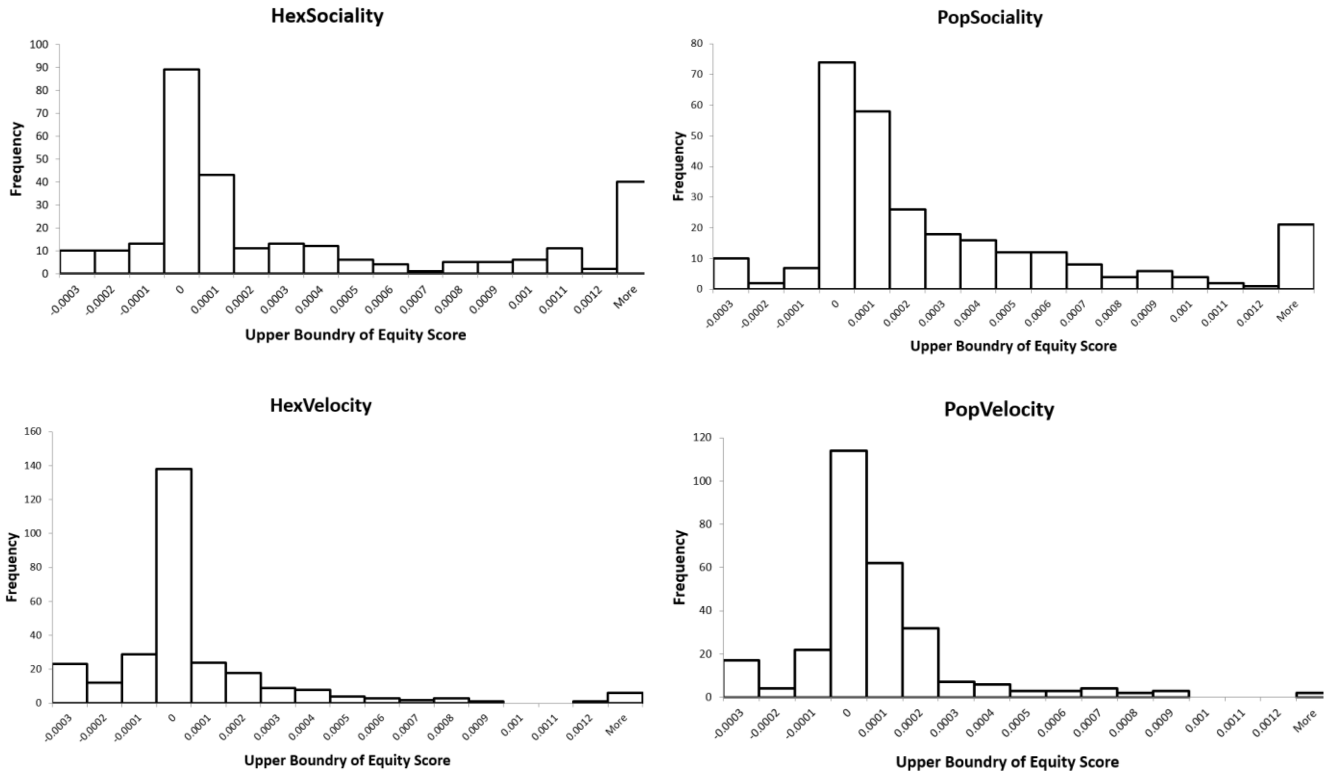


Table 0-4 Budapest

Rank	Transit Line	HexSociality	Transit Line	PopSociality	Transit Line	HexVelocity	Transit Line	PopVelocity
1st	Tram Line 1	-0.0023	Metro Line 2	-0.0028	Metro Line 2	-0.0032	Metro Line 2	-0.0061
2nd	Metro Line 2	-0.0019	Bus Line M3	-0.0009	Tram Line 1	-0.0029	Metro Line 4	-0.0024
3rd	Bus Line M3	-0.0015	Metro Line 1	-0.0008	Metro Line 3	-0.0014	Bus Line M3	-0.0020
4th	Bus Line 8E	-0.0007	Bus Line 72	-0.0006	Metro Line 4	-0.0013	Bus Line 99	-0.0013
5th	Bus Line 20E	-0.0006	Tram Line 6	-0.0006	Bus Line M3	-0.0010	Tram Line 6	-0.0012
5th to last	Bus Line 169E	0.0033	Bus Line 200E	0.0027	Bus Line 169E	0.0013	Bus Line 169E	0.0008
4th to last	Bus Line 138	0.0038	Tram Line 1	0.0034	Train Line H8	0.0017	Bus Line 26	0.0009
3rd to last	Train Line H7	0.0041	Bus Line 151	0.0036	Bus Line 63	0.0018	Train Line H8	0.0009
2nd to last	Train Line H5	0.0066	Train Line H7	0.0038	Train Line H5	0.0018	Train Line H6	0.0015
Last	Bus Line 200E	0.0069	Train Line H5	0.0046	Train Line H6	0.0032	Train Line H5	0.0018

Figure 0-5 Berlin

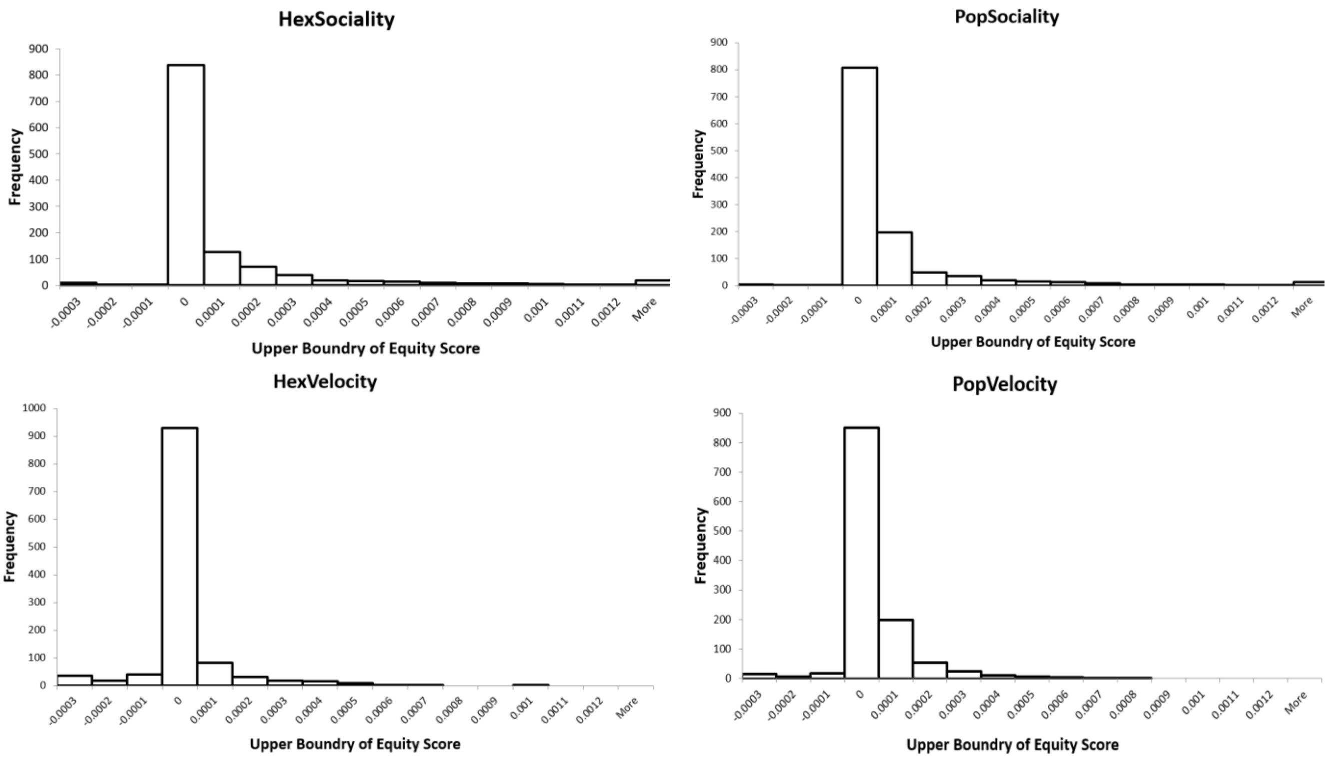


Table 0-5 Berlin

Rank	Transit Line	HexSociality	Transit Line	PopSociality	Transit Line	HexVelocity	Transit Line	PopVelocity
1st	Train Line U8	-0.0016	Train Line RB10	-0.0011	Train Line U7	-0.0032	Train Line U7	-0.0029
2nd	Train Line U9	-0.0013	Train Line S42	-0.0009	Train Line U6	-0.0023	Train Line U9	-0.0028
3rd	Train Line S41	-0.0009	Train Line U8	-0.0007	Train Line U9	-0.0015	Train Line U6	-0.0021
4th	Train Line S42	-0.0008	Train Line U2	-0.0005	Train Line U8	-0.0013	Train Line U8	-0.0019
5th	Train Line S45	-0.0007	Train Line S41	-0.0004	Train Line U5	-0.0010	Train Line S42	-0.0015
5th to last	Bus Line 222	0.0020	Train Line RB13	0.0020	Bus Line 614	0.0005	Bus Line 197	0.0005
4th to last	Train Line S1	0.0020	Train Line S1	0.0021	Bus Line 951	0.0006	Bus Line 136	0.0005
3rd to last	Train Line S8	0.0022	Train Line S25	0.0024	Bus Line 806	0.0006	Bus Line 222	0.0005
2nd to last	Train Line S5	0.0025	Train Line U5	0.0028	Bus Line 950	0.0007	Bus Line 806	0.0007
Last	Train Line S2	0.0029	Train Line U7	0.0049	Bus Line 671	0.0010	Bus Line 62	0.0007

Figure 0-6 Aachen

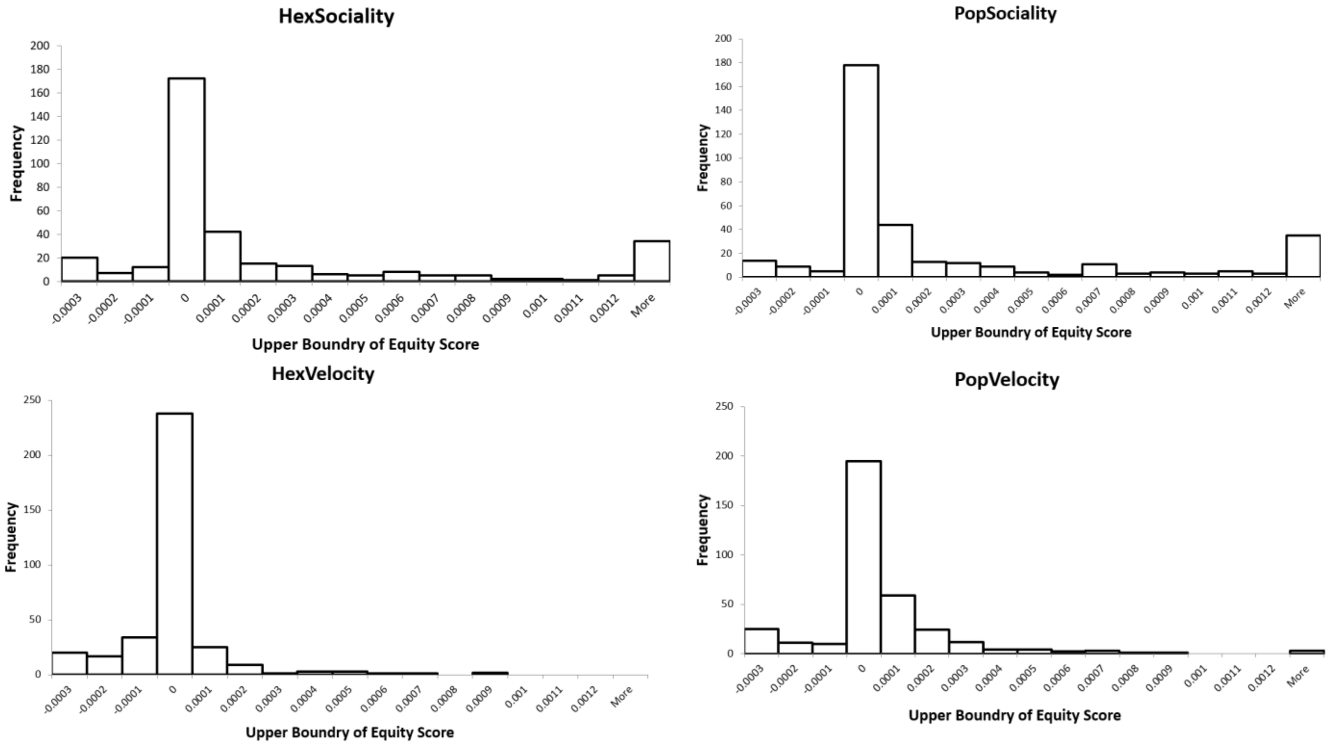
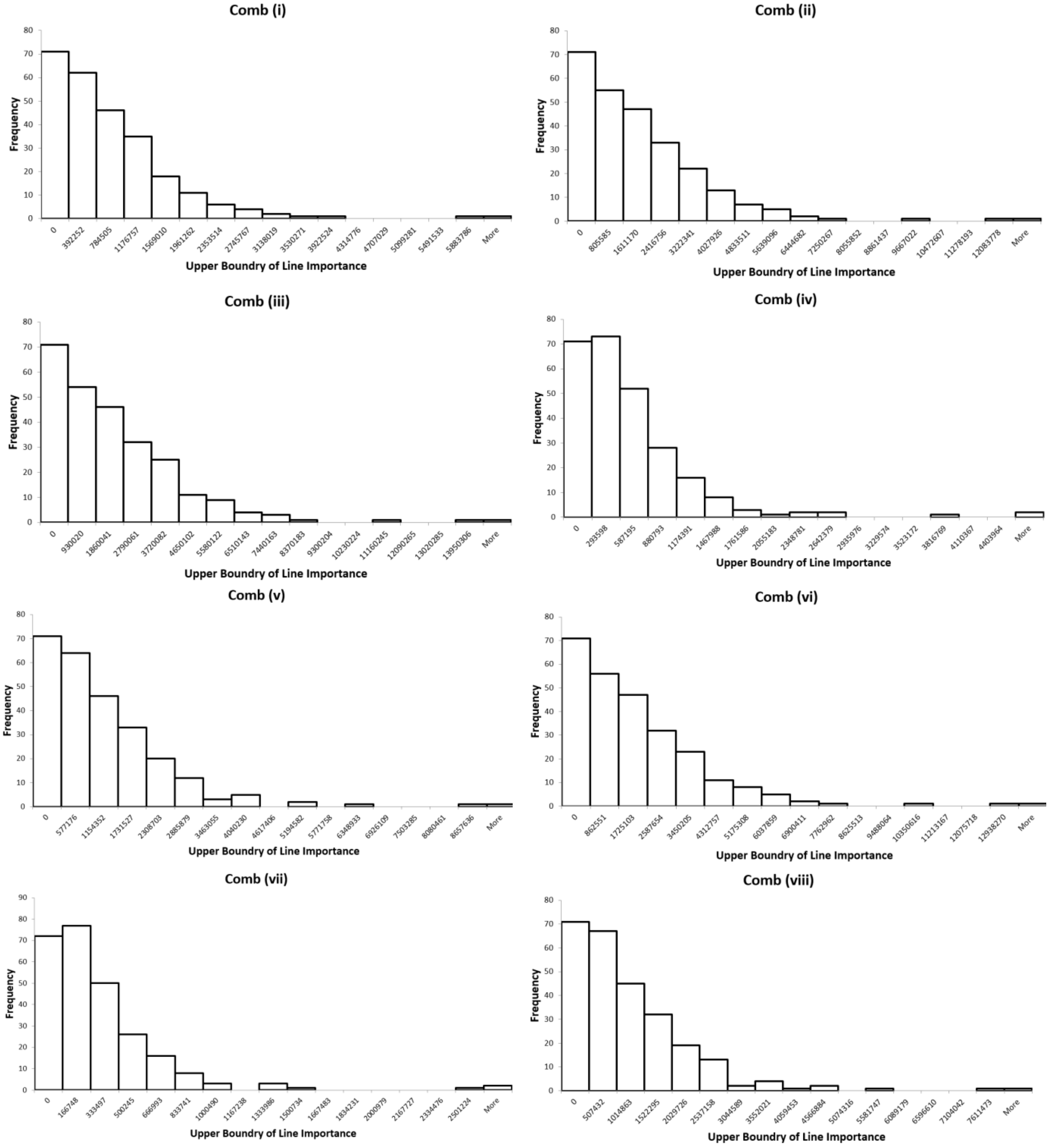


Table 0-6 Aachen

Rank	Transit Line	HexSociality	Transit Line	PopSociality	Transit Line	HexVelocity	Transit Line	PopVelocity
1st	Bus Line 54	-0.0023	Train Line RE6	-0.0023	Bus Line 51	-0.0013	Train Line RB20	-0.0018
2nd	Train Line RB20	-0.0019	Bus Line 45	-0.0016	Bus Line SB63	-0.0010	Netliner Bus	-0.0016
3rd	Bus Line 47	-0.0015	Bus Line 34	-0.0013	Train Line RB20	-0.0010	Bus Line 54	-0.0016
4th	Bus Line 25	-0.0012	Bus Line 2	-0.0012	Bus Line SB66	-0.0010	Bus Line 2	-0.0012
5th	Bus Line 34	-0.0012	Bus Line 3B	-0.0008	Netliner Bus	-0.0010	Bus Line SB63	-0.0012
5th to last	Bus Line SB63	0.0035	Bus Line 51	0.0036	Bus Line EK1	0.0004	Bus Line 216	0.0008
4th to last	Train Line RB21	0.0036	Bus Line 11	0.0042	Bus Line 298	0.0005	Bus Line 286	0.0009
3rd to last	NetLiner Bus	0.0039	Bus Line 28	0.0057	Netliner Bus	0.0007	Bus Line 28	0.0013
2nd to last	Bus Line 298	0.0061	Bus Line 296	0.0058	Bus Line 475	0.0008	Bus Line 296	0.0016
Last	Bus Line 14	0.0074	Bus Line 44	0.0076	NetLiner2 Bus	0.0009	Bus Line 44	0.0025

Appendix C: Histograms of nine types of line importance in the considered cities

Figure 0-1 Vienna



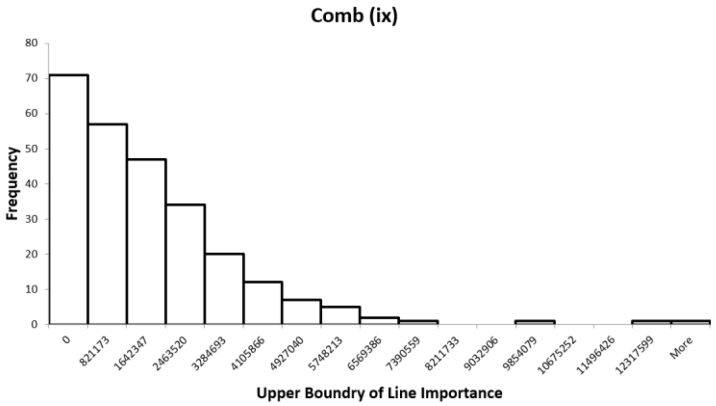
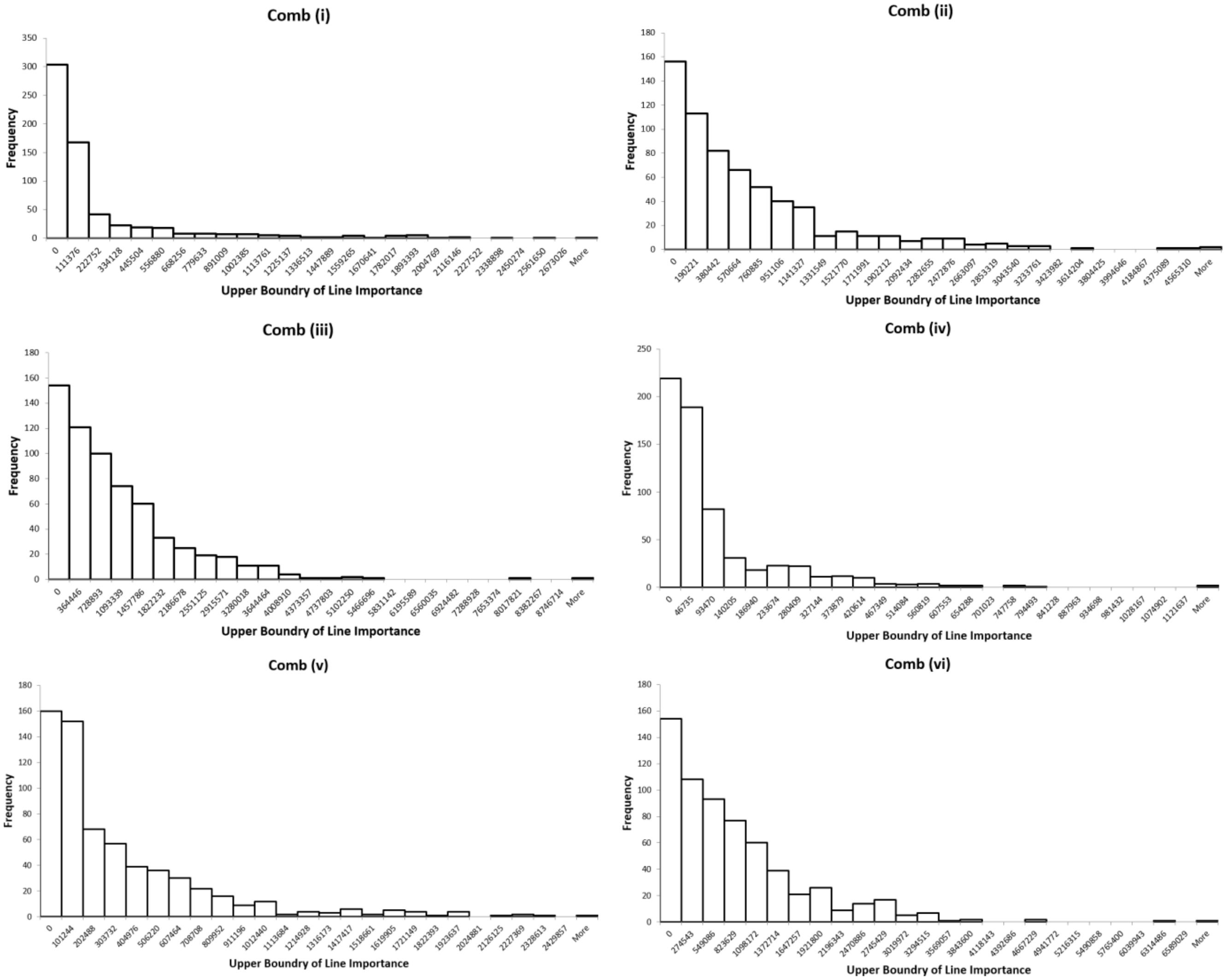


Figure 0-2 Manchester



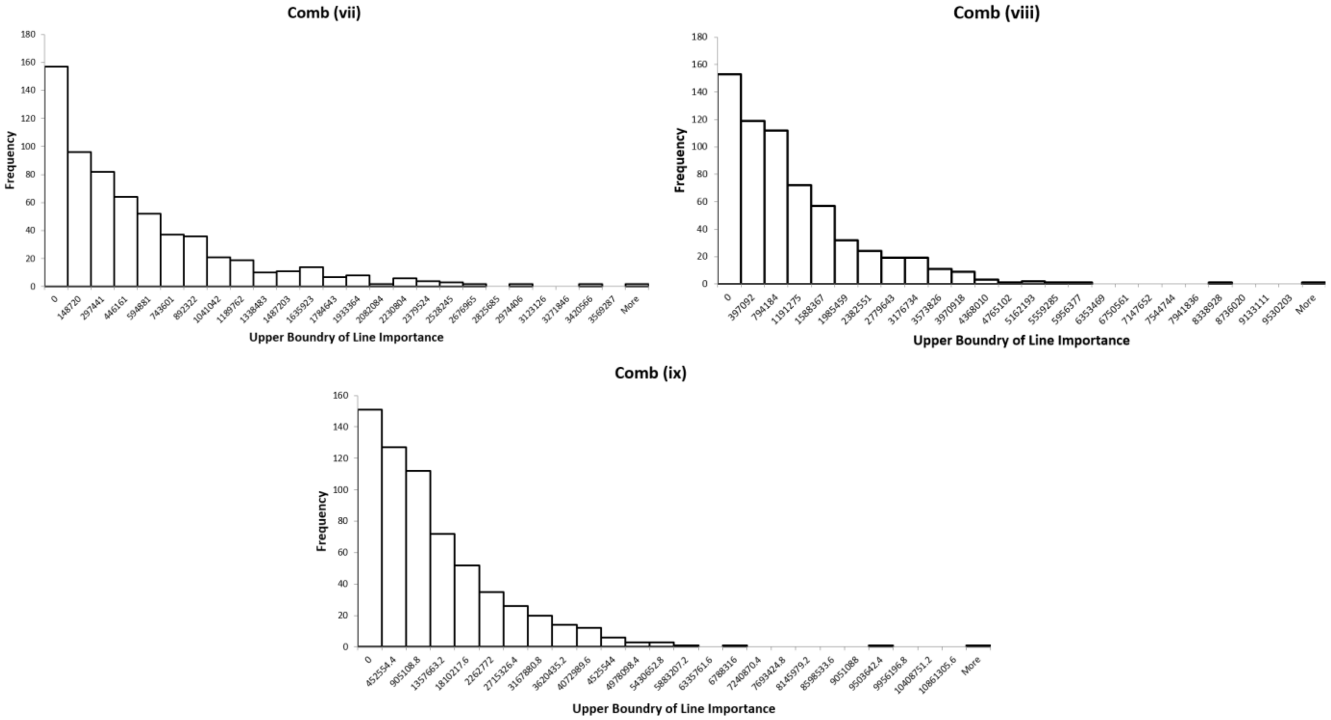
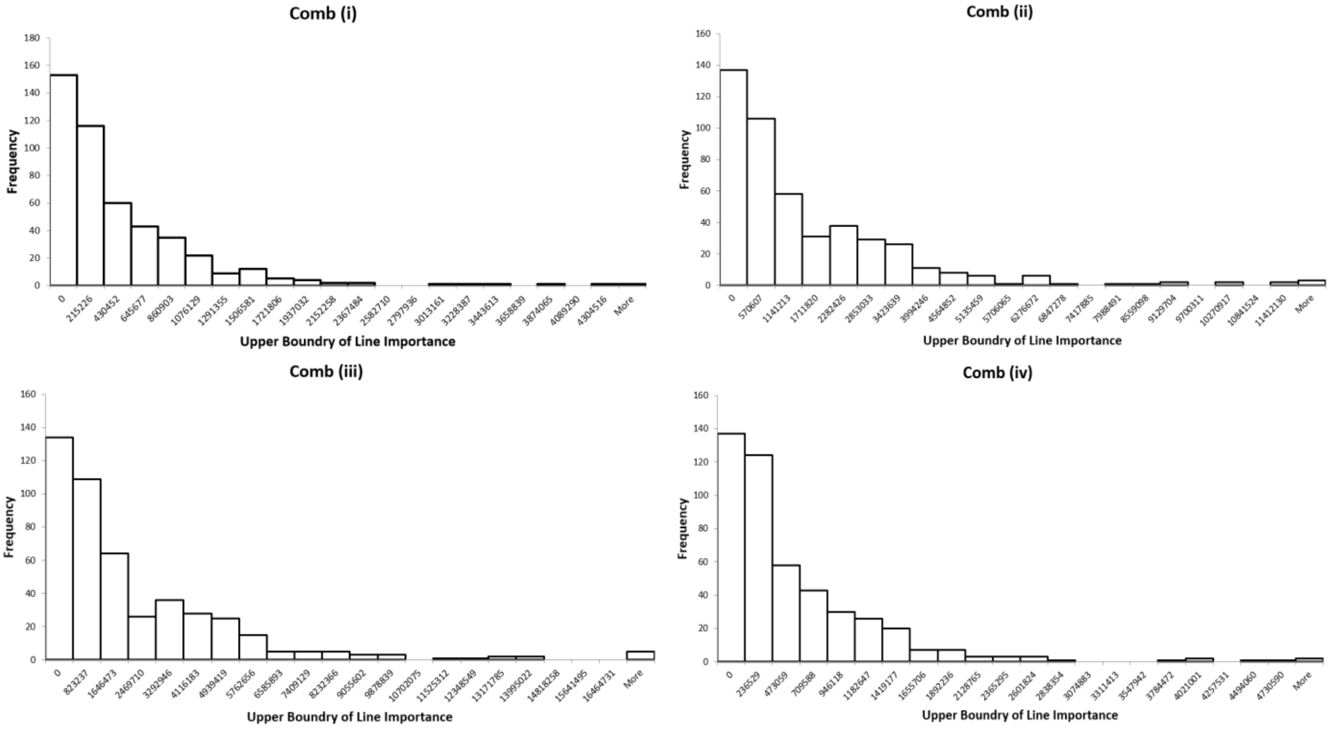


Figure 0-3 Helsinki



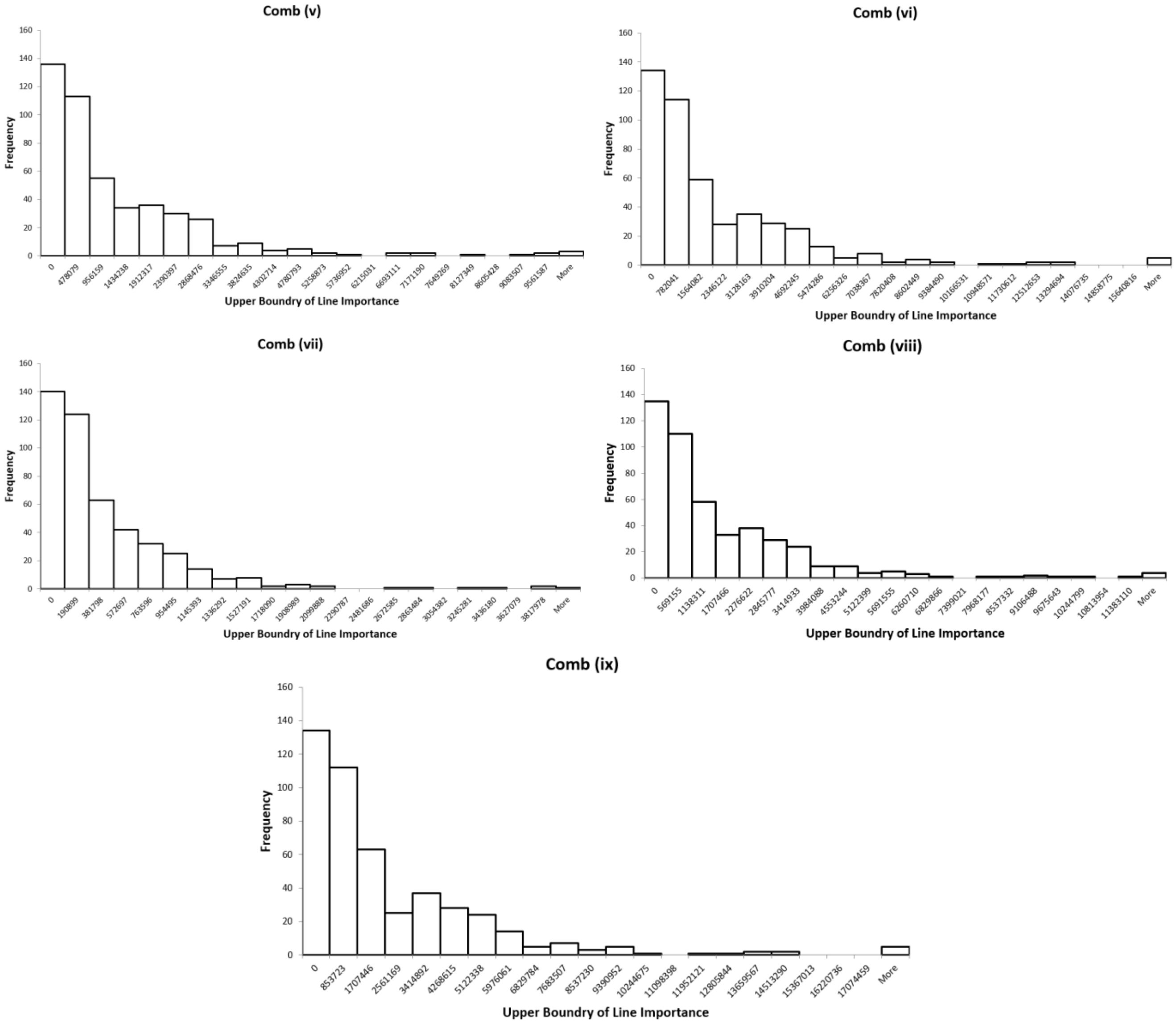
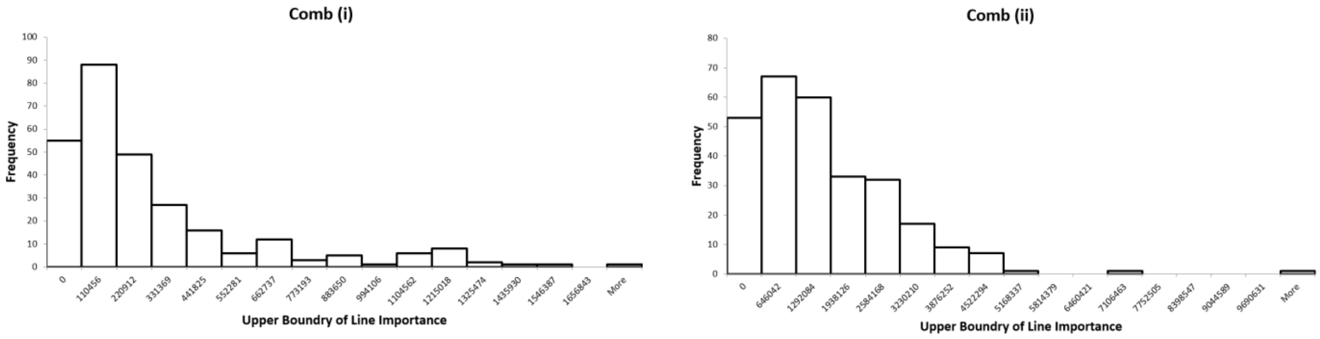
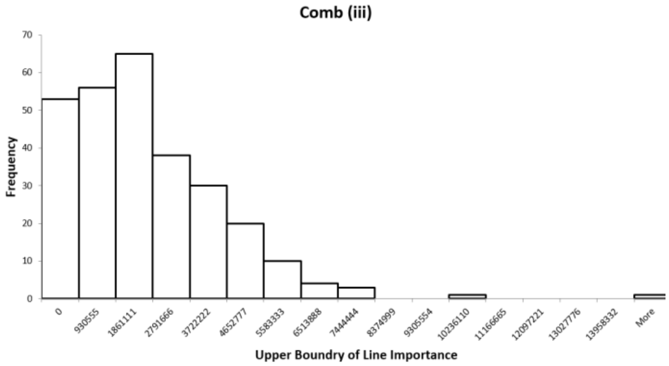
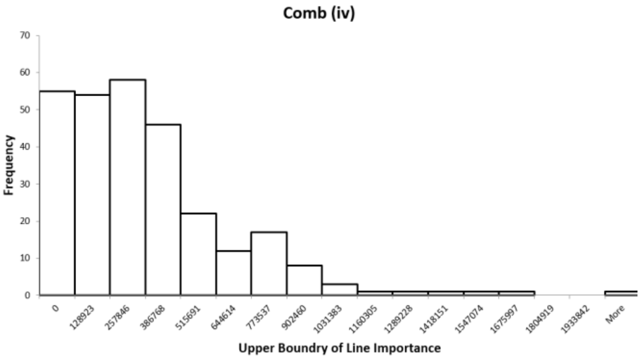


Figure 0-4 Budapest

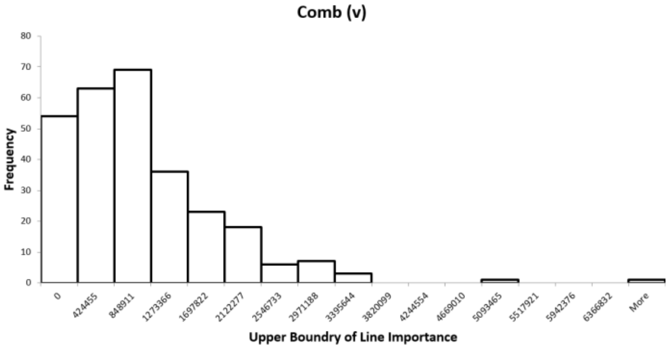




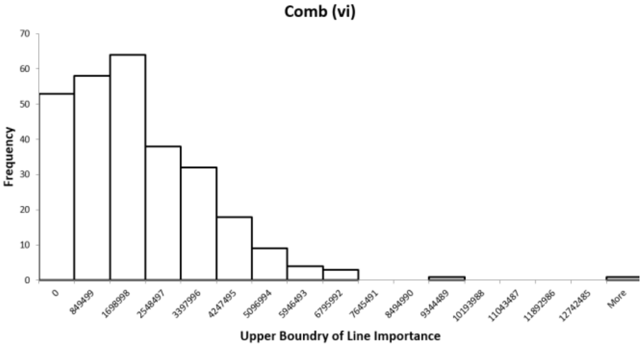
Comb (iii)



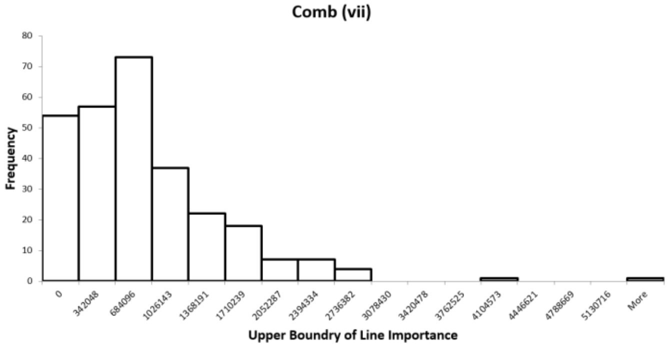
Comb (iv)



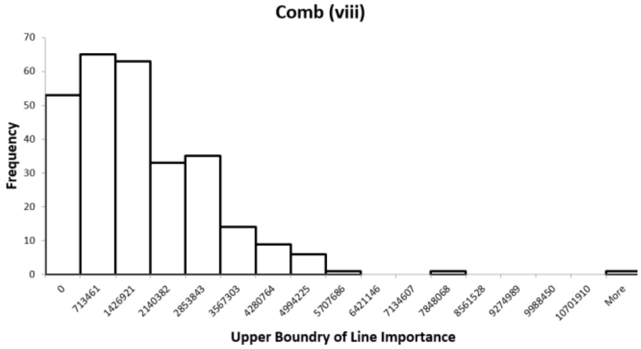
Comb (v)



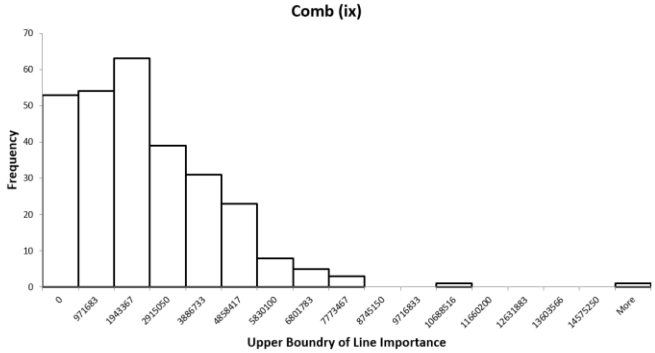
Comb (vi)



Comb (vii)

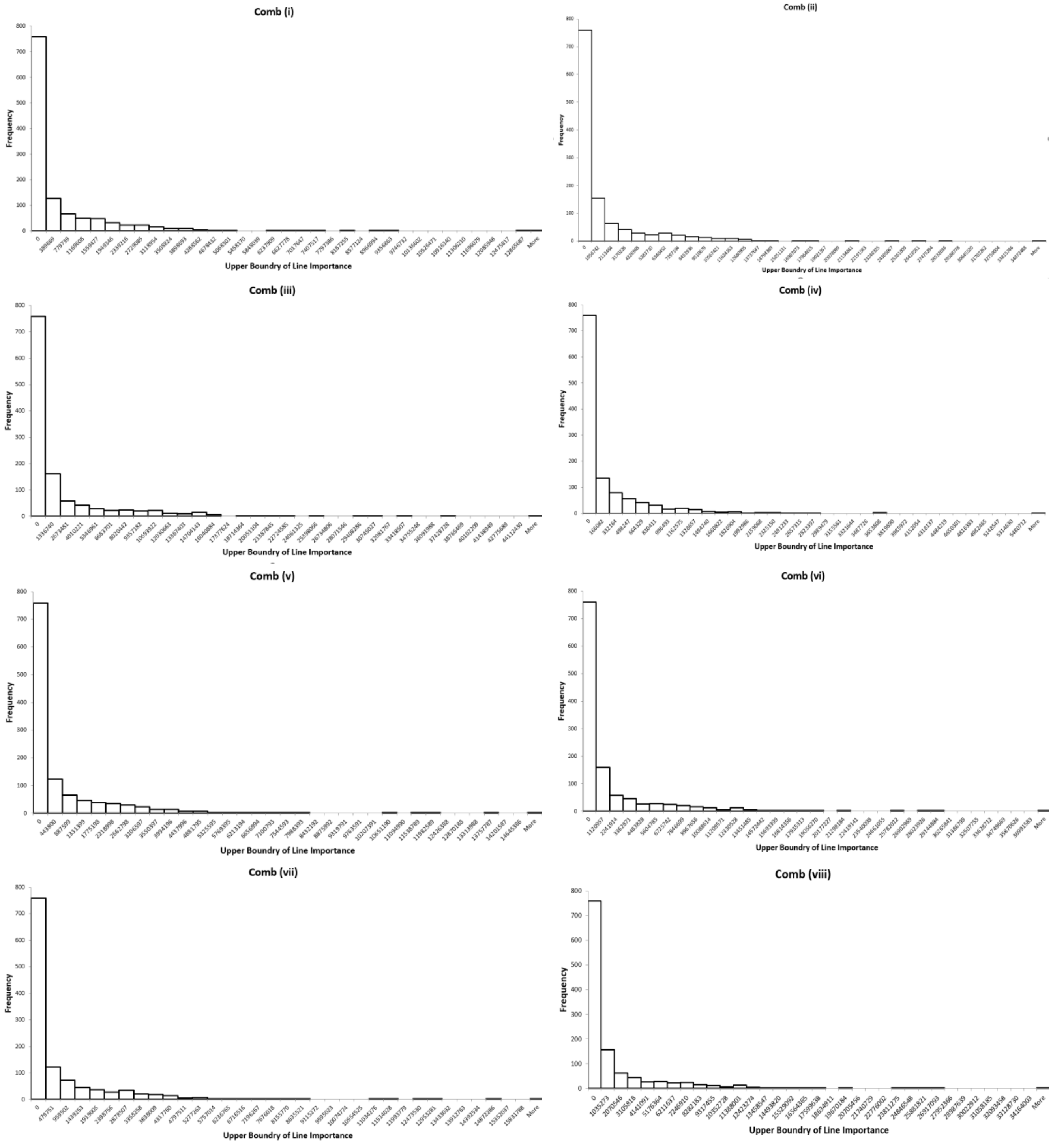


Comb (viii)



Comb (ix)

Figure 0-5 Berlin



Comb (i)

Comb (ii)

Comb (iii)

Comb (iv)

Comb (v)

Comb (vi)

Comb (vii)

Comb (viii)

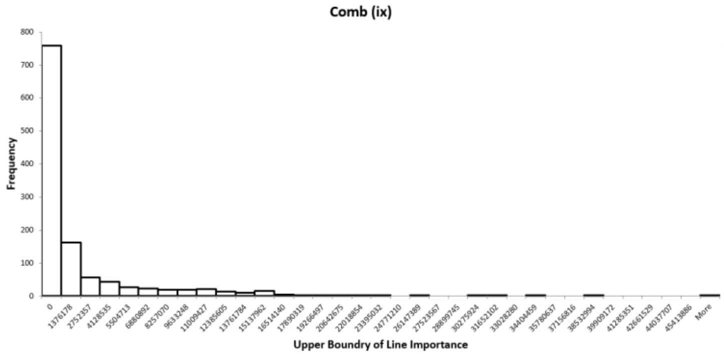
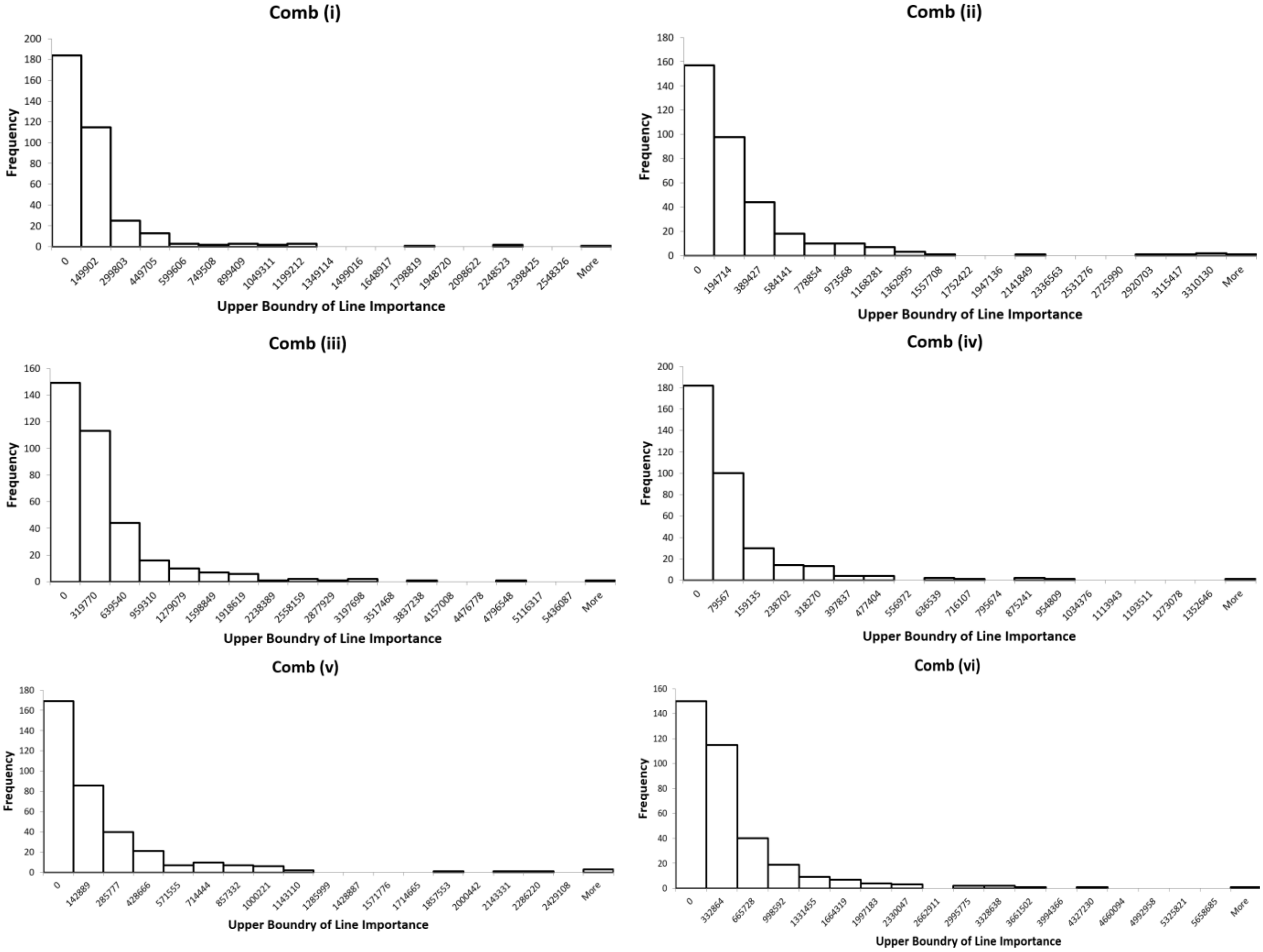
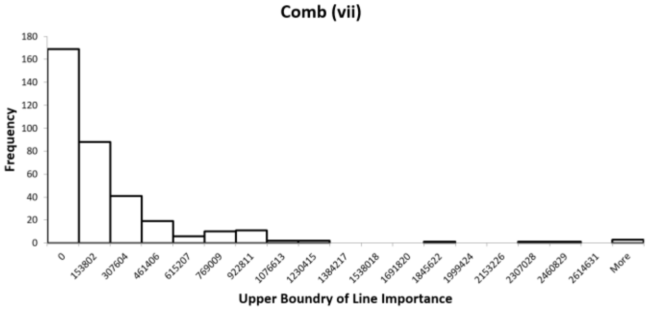
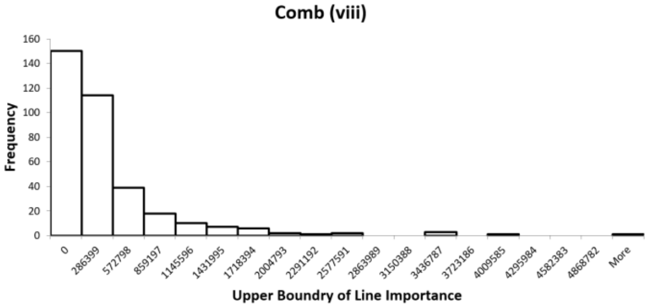


Figure 0-6 Aachen

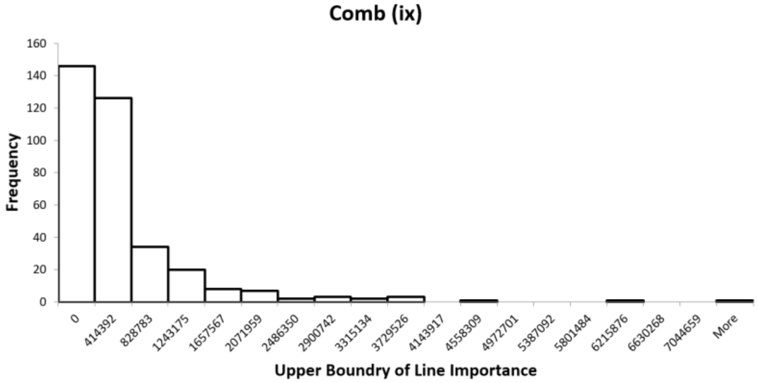




Comb (vii)



Comb (viii)



Comb (ix)

Appendix D: Python scripts developed to compute the line importance function

```
import sys
from tqdm import tqdm
import warnings

warnings.filterwarnings('ignore')

sys.path.insert(0, './library/')
import zipfile
import os
import time
import pymongo as pym
import pandas as pd
import folium
import numpy as np
import requests
import numba
from shapely.geometry import Polygon, LineString, asShape, mapping, Point
import math
import geopy
from shapely.geometry import Polygon, MultiPolygon, Point, mapping
from geopy.distance import geodesic, great_circle
from folium.plugins import FastMarkerCluster
from datetime import datetime
from geopy.distance import geodesic, great_circle

from libAccessibility import arrayTimeCompute, ListFunctionAccessibility
from libHex import area_geojson
from scipy.sparse import coo_matrix
import math
import time
import numpy

inf = 10000000

from numba import jit, int32, int64

city = 'Budapest'  # name of the city
urlMongoDb = "mongodb://localhost:27017/";  # url of the mongodb database

client = pym.MongoClient(urlMongoDb)
gtfsDB = client['PublicTransportAnalysisBudapest']


def setPosField2(gtfsDB, city):
    pos = 0
    for route in gtfsDB['routes'].find({'city': city}).sort([('_id',
pym.ASCENDING)]):
        gtfsDB['routes'].update_one({'_id': route['_id']},
                                    {'$set':
                                        {
                                            'pos': pos
                                        }
                                    })
        pos += 1
        print('{0}'.format(pos), end="\r")
    gtfsDB['routes'].create_index([("pos", pym.ASCENDING)])


setPosField2(gtfsDB, city)

from tqdm import tqdm

for j in tqdm(gtfsDB["routes"].find({'city': city}).sort([('_id',
pym.ASCENDING)])):
    gtfsDB['connections'].update_many({'route_id': j['route_id']},
                                      {'$set':
                                          {
                                              'pos': j['pos'],
                                          }
                                      });
    # print('{0}'.format(j["pos"]), end="\r")
    gtfsDB['connections'].create_index([("pos", pym.ASCENDING)])

stoppss = gtfsDB['stops']

from tqdm import tqdm

toal = gtfsDB['connections'].find({}).count()
for k in tqdm(gtfsDB['connections'].find({'city': city}).sort([('_id',
pym.ASCENDING)]), total=toal):
    timeStart0 = time.time()
    gtfsDB['connections'].update_one({'_id': k['_id']},
                                     {'$set':
                                         {
                                             'distance':
round(geodesic((stoppss.find_one({'pos': k["pStart"]})[

'point']["coordinates"][1],

stoppss.find_one({'pos': k["pStart"]})[

'point']["coordinates"][0]),

(stoppss.find_one({'pos': k["pEnd"]})[

'point']["coordinates"][1],

stoppss.find_one({'pos': k["pEnd"]})[

'point']["coordinates"][0])).meters)
                                         }
                                     }
                                     )
    gtfsDB['connections'].create_index([("distance", pym.ASCENDING)])

timeList = [8]  # [7,10,13,16,19,22] # List of starting time for computing the
isochrones
# timeList = [7,10,13,16,19,22] # List of starting time for computing the
isochrones
hStart = timeList[0] * 3600
lenroute = gtfsDB['routes'].find().count()


def makeArrayConnections2(gtfsDB, hStart, city):
    print("start making connections array")
    fields = {'tStart': 1, 'tEnd': 1, 'pStart': 1, 'pEnd': 1, '_id': 0}
    typeMatch = {'city': city, 'tStart': {'$gte': hStart},
                 'tStart': {"$type": "number"}, 'tEnd': {"$type": "number"},
                 'pStart': {"$type": "number"}, 'pEnd': {"$type": "number"}, }
    pipeline = [
        {'$match': {'city': city, 'tStart': {'$gte': hStart}}},
        {'$sort': {'tStart': 1}},
        {'$project': {'_id': "$_id", "c": ['$tStart', '$tEnd', '$pStart',
'$pEnd', '$pos', '$distance']}},
    ]
    allCC = list(gtfsDB['connections'].aggregate(pipeline))
    print("done recover all cc", len(allCC))
    allCC = np.array([x["c"] for x in allCC])
    print("cenverted")
    # arrayCC =
np.full((gtfsDB['connections'].find({"city":city,'tStart':{'$gte':hStart}}).cou
nt(),4),1.,dtype = np.int)
    # countC = 0
    # tot =
gtfsDB['connections'].find({'tStart':{'$gte':hStart},'city':city}).count()

    print('Num of connection', len(allCC))
    return allCC


arrayCC = makeArrayConnections2(gtfsDB, hStart, city)

# ### List of list of the points and stops neighbors

# In[ ]:


from libStopsPoints import listPointsStopsN

arraySP = listPointsStopsN(gtfsDB, city)


@jit((int32[:], int32[:], int32, int64[:, :], int32[:, :], int32[:, :],
int32[:, :], int32[:, :], int32),
     nopython=True)
def coreICSA2(timesValues, timeP, timeStart, arrayCC, S2SPos, S2STime, P2SPos,
P2STime, lenroute):
    # print 'inter'
    # global arrayCC
    # arrayCC = CC
    # global S2SPos
    # global S2STime
    count = 0
    pointRoute = [0.] * lenroute

    routee = []
    timesValuesN = numpy.copy(timesValues)
    for c_i in range(len(arrayCC)):
        c = arrayCC[c_i]
        Pstart_i = c[2]
        if timesValues[Pstart_i] <= c[0] or timesValuesN[Pstart_i] <= c[0]:
            count += 1
            Parr_i = c[3]
            if timesValues[Parr_i] > c[1]:
                timesValues[Parr_i] = c[1]
                if c[1] <= timeStart + 3600:
                    # if distn[Parr_i] == 1:
                    routee.append((Pstart_i, Parr_i, c[4], c[5]))
                for neigh_i in range(len(S2SPos[Parr_i])):
                    if S2SPos[Parr_i][neigh_i] != -2:
                        neigh = S2SPos[Parr_i][neigh_i]
                        neighTime = timesValuesN[neigh]
                        if neighTime > c[1] + S2STime[Parr_i][neigh_i]:
                            timesValuesN[neigh] = c[1] +
S2STime[Parr_i][neigh_i]
                    else:
                        break

    for i, t in enumerate(timesValues):
        if t > timesValuesN[i]:
            timesValues[i] = timesValuesN[i]

    for (org, des, lin, dis) in routee:
        pointRoute[lin] += dis

    return pointRoute


def coumputeTimeOnePoint(point, startTime, timeS, timeP, arrayCC, P2PPos,
P2PTime, P2SPos, P2STime, S2SPos,
                         S2STime, lenroute):
    timeS.fill(inf)  # Inizialize the time of stop
    timeP.fill(inf)
    posPoint = point['pos']  # position of the point in the arrays
    timeP[posPoint] = startTime  # initialize the starting time of the point

    for neigh_i, neigh in enumerate(
            P2PPos[posPoint][P2PPos[posPoint] != -2]):  # loop in the point
near to the selected point
        neigh = neigh
        timeP[neigh] = P2PTime[posPoint][neigh_i] + startTime  # initialize to
startingTime + WalkingTime all near point

    # loop in the stops near to the selected point
    for neigh_i, neigh in enumerate(P2SPos[posPoint][P2SPos[posPoint] != -2]):
        neigh = neigh
        timeS[neigh] = P2STime[posPoint][neigh_i] + startTime  # initialize to
startingTime + WalkingTime all near stops

    # timeSInit = timeS.copy()
    startTime = numpy.int32(startTime)
    arrayCC = arrayCC.astype(numpy.int64)

    routeee = coreICSA2(timeS, timeP, startTime, arrayCC, S2SPos, S2STime,
P2SPos, P2STime, lenroute)

    return routeee


def computeAccessibilities(city, startTime, arrayCC, arraySP, gtfsDB,
lenroute):
    timeS = arraySP['timeS']
    timeP = arraySP['timeP']
    S2SPos = arraySP['S2SPos']
    S2STime = arraySP['S2STime']
    P2PPos = arraySP['P2PPos']
    P2PTime = arraySP['P2PTime']
    P2SPos = arraySP['P2SPos']
    P2STime = arraySP['P2STime']

    maxVel = 0
    totTime = 0.
    avgT = 0
    tot = len(timeP)

    count = 0

    for point in gtfsDB['points'].find({'city': city}, {'pointN': 0, 'stopN':
0}, no_cursor_timeout=True).sort(
            [('pos', 1)]):

        timeStart0 = time.time()

        # Inizialize the time of stop and point
        # print("starting computation")
        routee = coumputeTimeOnePoint(point, startTime, timeS, timeP, arrayCC,
P2PPos, P2PTime,
                                      P2SPos,
                                      P2STime, S2SPos, S2STime, lenroute)

        a_list = list(routee)

        score = {}
        for i, m in enumerate(a_list):
            score[i] = m

        score = {str(k): round(float(v), 2) for k, v in score.items()}

        totTime += time.time() - timeStart0
        avgT = float(totTime) / float(count + 1)
        h = int((tot - count) * avgT / (60 * 60))
        m = (tot - count) * avgT / (60) - h * 60

        gtfsDB['points'].update_one({'_id': point['_id']}, {'$set': {"Score":
score}})
        count += 1
        print(
            'point: {0}, time to finish : {1:.1f}h, {2:.1f} m'.format(
                count, h, m),
            end="\r")


computeAccessibilities(city, hStart, arrayCC, arraySP, gtfsDB, lenroute)
```