POLITECNICO DI TORINO

Master's Degree in Data Science and Engineering

Master's Degree Thesis

Unraveling fundamental network features of complex urban transportation systems

Supervisors

Dr. Riccardo GALLOTTI

Prof. Enrico BIBBONA

Prof. Moritz MULTENHALER

Co-supervisor

Sebastiano BONTORIN

Candidate

Ulysse MARQUIS

March 2023

Abstract

Transportation networks represent one of the key structural elements behind the functional properties of a large number of natural and man-made systems. In urban systems, they represent some of the arteries that regulate human mobility, especially in commuting scenarios. Moreover, they are by essence linked to the growth and development of urban areas, both in the exploitation of the urban space and its efficiency. Understanding the interplay between networks and urban exploitation represents a step forward in the study of cities and urban areas as complex systems, which in turn can be helpful for decision makers and urban planners. In this thesis we focus on urban transportation networks framing these systems from the mathematical perspective of complex networks. In particular, we study subway systems as they exhibit complex spatial features and we unravel their interplay with urban land use features.

We first provide a general review of the recent literature of cities as complex systems and we frame the problem of metro systems in the context of spatial networks. As a first approach to this analysis, we reproduce some key results presented in literature. To this aim, datasets for real metro systems are essential, although limited datasets for urban networks are available and not longitudinal to several cities. While several sources are available for public transportation information, these are often sparse, decentralized and specific of local providers. As a first task in this work, we describe the process of processing and mapping these data into a unified data structures of spatial weighted networks for a set of cities we have identified.

We then proceeded with the modeling and analysis of these systems. We have analyzed the decomposition of subway systems in a combination of dense core structures with loops and branches in periphery, reproducing some fundamental results obtained in recent works. Then, we show a pitfall in the definition of the center used in the literature. We propose a novel definition of a functional center and show its comparative benefits relative to the previous definition. We introduce also a set of novel metrics describing some of the network's spatial properties and providing novel perspective to the analysis of these systems. This definition carves a more precise decomposition in a core and branch structure and yields clearer structural properties of the network, intimately related to the efficiency of the network structure. Finally, we show that this definition coincides with the spatial maximum of amenities in various urban areas on which we focused.

Acknowledgements

I would like to express my deepest appreciation to my supervisors, Sebastiano Bontorin and Riccardo Gallotti, for their invaluable contribution and helpful advice. To colleagues, friends and family, my most sincere gratitude for their unconditional support.

Table of Contents

List of Tables

List of Figures

Acronyms

POI
: Point Of Interest

GTFS
: General Transit Feed Specification

BC
: Betweenness Centrality

RGG
: Random Geometric Graphs

Chapter 1

Introduction

This first chapter serves as a brief introduction to the study of urban complex systems and spatial networks. We outline the motivations, goals, constraints and main tools that will be used in this work. First, the study of cities and transportation networks as complex systems is introduced. We then provide an introduction on network science and spatial networks, the core mathematical models and tools that will be employed to study these systems.

1.1 Urban complex systems

A complex system can be defined as a set of interacting elements, where each individual element may have well-defined properties and behavior. Its complexity comes from the fact that the knowledge of these individual elements alone is insufficient to fully describe the system. Rather, the system is characterized by collective dynamics, which can be described as a function of the prior knowledge. This results in behaviors that define complex structural and functional properties.

Urban complex systems refer to the interconnected and dynamic systems of cities and urban areas. These systems are defined by a multiplicity of factors such as physical infrastructures and economic activities. They are shaped by a range of factors, including historical development, social and economic forces, political decisions, and environmental conditions. Cities and urban areas are centers of innovation and economic growth, but they also face a range of challenges such as social inequalities or organizational issues.

Understanding urban complex systems is crucial for policymakers, urban planners, and researchers seeking to improve the sustainability, livability, and resilience

of cities and urban areas. By analyzing the dynamics of urban systems and their evolution, it is possible to identify potential solutions to urban challenges and design strategies that can help promote sustainable urban development.

1.1.1 Cities and scaling laws

Cities are open systems, out of their equilibrium state, and can be understood as evolving organisms - they develop through constraints of many kinds, such as spatial, economic, social and environmental issues. To get a better understanding of these systems, researchers try to find common features that they obey to, which helps classifying or clustering them, as well as anticipating their development.

Given a feature, such as energy consumption or road network length, it is crucial to understand its scaling with the city size. One of the fundamental results obtained from the analysis of cities is the discovery that the relationship between several urban features and the population (or population density, or size of the city...) scales with a power law of exponent β. Usually, three regimes are considered :

1. $\beta < 1$ - sub-linear regime : this corresponds to economy of scale - more population requires less and less infrastructure (for instance, energy consumption in function of population density)
2. $\beta = 1$ - linear regime : these features correspond to human needs, such as food consumption
3. $\beta > 1$ - super-linear regime : the increase in population induces an increase in social and economic activity, triggering a gain in productivity (for example, GDP and population)

Classifications of relationship between features, and especially the exact values of these exponents, are not always agreed on, as definitions of cities' morphological structures have not reached a real consensus. However, this type of study is widely explored, for multiple types of factors of interest -for instance socio-economic, structural or functional - for instance in [1] which studies relationship between energy consumption and population density in Great Britain, or [2] which studies the relationship between the development of railway systems in function of economic growth.

1.1.2 Urban transportation networks and human mobility

Transportation networks are an essential feature of urban complex systems, as they rule human and goods mobility, and are shaped by flows, population density,

economic activity and spatial constraints. From the perspective of the last section, understanding the relationship between cities development and their transportation network is fundamental.

Characterizing a city by a small set of features, such as its area, its population and its GDP, and a transportation network by its number of nodes, its total length and its ridership, as done in [2], one can relate socio-economic indicators to its topological properties. We summarize in the following paragraph the model proposed by this work, as it is beneficial to understand the kind of modeling that can be performed to understand these systems. The development of a network can be approached with a cost-benefit analysis, understanding the growth as a succession of equilibrium state satisfying $Z = B - C \approx 0$, where $B = \sum_e B_e$ and $C = \sum_e C_e$, where B_e and C_e are respectively the benefit and the cost of an edge of the network. Considering the ridership of a network R, its ticket price f, its total length L, the cost of maintenance ϵ_L of an unit of length, N_s its number of station and ϵ_s the cost of maintenance of a station, we have $Z = Rf - \epsilon_L L - \epsilon_s N_s$.

Moreover, using hypotheses such as the attraction in a circular area around each station, uniform population density around stations, equilibrium of the system, one can deduce simple relationship between features of the network and the cities. For instance, scaling properties can be found between the total length of the network and its urban area Gross Metropolitan Product. Figure 1.1 shows firstly the generic relationship between a city's wealth and its network's number of stations and secondly the number of lines - expressing the complexity and level of development of a network - in function of the number of stations.

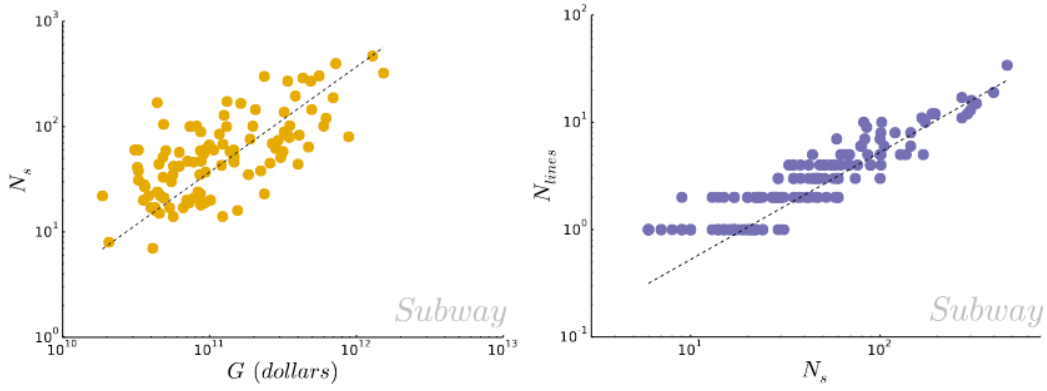

Figure 1.1: Scaling properties in urban structures Relationship between Gross Metropolitan Product of a city and the size of its subway network (left). Relationship between number of nodes and number of lines of a subway system (right). This plot is extracted from [2].

1.2 Elements of Network science

The aforementioned systems are part of what are generally defined complex systems. Complex systems are not just complex because of the number of elements and connections that constitute them, but also because interactions and dynamics take place between these elements. Signals and information are propagated, and quantities (which in the case of urban systems and transportation networks are often represented by people or commodities travelling) are exchanged or moved through nodes in the network. These interactions and connections can be associated with fundamentally different quantities, ranging from opinions, ideas or messages in a social network [3], to data packages in a network of servers, electric signals in a network of neuronal cells [4], and even humans travelling from one node to another in a mobility network (see Figure 1.2).

Despite this complexity, graph theory [3] and network science can provide a simple description of this information. This theoretical framework can be employed to analyze the structure and the dynamics of urban and transportation systems.

Figure 1.2: The World Air-transportation Network represents a fundamental example of complex network. The figure is extracted from [5].

Mathematically, a practical and elegant formalism, and mathematical structure, to describe the often complex and heterogeneous connections between the elements that compose these systems is a graph or network, denoted by $\mathcal{G} = (\mathcal{N}, \mathcal{E})$. In a network, a set of nodes $\mathcal{N}$ identify the elements, and the edges $e = (i, j) \in \mathcal{E}$ define the pair-wise connections and interactions that exist (represented via the set $\mathcal{E}$) between the nodes. Nodes (or vertices) can be graphically depicted or represented as points, and edges can be represented as lines drawn between these points. This information is then embedded into a symmetric square matrix $\mathcal{A}$, the Adjacency Matrix, which can be unweighted or weighted, depending on whether w_{ij} is equal to 1 or a scalar value, respectively.

$$A_{ij} = \begin{cases} w_{ij} & \text{if } (i,j) \in \mathcal{E} \\ 0 & \text{otherwise} \end{cases} \tag{1.1}$$

These weights w_{ij} in an urban network are often used to encode the flux of people moving between points [4] or in a transportation network can encode for example the time necessary to take to travel along that transportation line between points i and j. A network can also be directed or undirected, depending on whether edges are unidirectional or bidirectional.

Connectivity is an important property of a network. A network can be defined as *complete* if all possible connections between pairs exist, while a network is *connected* if a path exists between any pair of nodes, and therefore a single unique giant connected component exists. Sparseness is another relevant property of a network, which encodes the density of connections, and therefore which fraction of all the possible connections that might appear are observed. The degree k_i or connectivity of a node is the number of connections it has.

$$k_i = \sum_{j=1}^{N} A_{ij} \tag{1.2}$$

For an undirected network, this is defined as the sum of the elements in the corresponding row of the adjacency matrix, while for a directed network, we can separate or distinguish between in-degree and out-degree. Nodes with a larger number of connections (with respect to the average connectivity in the network) are known as hubs.

Regarding information diffusion, sequences of edges (each adjacent edge sharing a common vertex) define a path which mediates indirect information exchange. A path is defined as an ordered set of nodes, where edges connect adjacent nodes. If a set of paths between two nodes exists, we define the shortest path length L_{ij} between nodes i and j as the minimum number of edges that need to be traversed

to move from node j to node i, in a general undirected network. Knowledge of this pathways information allows to introduce global metrics that allow the classification and understanding of networks and their spreading properties. Information such as the diameter can be introduced: $d_{\mathcal{G}} = \max_{i,j} L_{ij}$. Moreover, a fundamental metrics such as the average shortest path length can be introduced:

$$\langle L \rangle = \frac{1}{N(N-1)} \sum_{i,j} L_{ij}. \tag{1.3}$$

The diameter and average shortest path length are important measures defining the ability of a network to spread information quickly.

Another important metric on nodes in connected graphs is betweenness centrality which computes $g(v)$, defined as:

$$g(v) = \sum_{s \neq v \neq t} \frac{\sigma_{st}(v)}{\sigma_s t}, \tag{1.4}$$

where σ_{st} is the number of shortest paths from s to t and $\sigma_{st}(v)$ is the number of shortest paths from s to t passing through v. It is relevant in our work because it computes the importance of a node when navigating through shortest paths on a network. It reflects both the influence of a node on the network and the importance of the node for the connectivity of a network. A vertex tends to have high BC values because it connects otherwise disconnected components/clusters, and thus are included in all the shortest paths between those components.

1.2.1 Spatial networks

Spatial networks are a kind of complex network that embeds spatial information into the network structure. Unlike traditional networks, spatial networks are influenced by physical proximity and spatial relationships between nodes which fundamentally change the properties from their traditional counterparts. Specifically, nodes in a spatial network are characterized by specific positions (in some physical or metric space) and connected by edges that reflect the spatial relationships between them. These networks can be found in a wide range of real-world systems, both artificial and natural.

One of the peculiar features of spatial networks is that they often exhibit spatial autocorrelation, which means that elements and nodes that are spatially (or geographically) in proximity to each other are more likely to be connected than nodes that are farther apart. This is in contrast to the paradigm of traditional

complex networks, where edges are typically drawn based on non-spatial factors such as shared attributes or any kind of physical or non-physical interaction.

Fundamental properties of spatial networks

As mentioned, spatial networks are often characterized by certain properties that are induced by their geometrical properties, contrarily to other features of graphs that are only deduced from their topology. Some of their typical features are spatial autocorrelation - as mentioned earlier - clustering, or robustness. This set of features is explained by many factors, such as the fact that spatial networks are often designed to optimize some fort of flows, of person, goods, or energy, and thus present some characteristics proving their efficiency.

A typical spatial graph null model are random geometric graphs (RGGs) [4]. They are a type of graph model used to represent spatial networks where nodes are distributed randomly in space and edges connect nodes that are within a certain distance of each other. This distance is known as the connection radius. Depending on the setup, this parameter can vary, as the number of nodes, or the density. Some works investigate the behavior of some features when the density tends to infinity.

RGGs can exhibit a variety of interesting properties depending on the value of the connection radius, the dimensionality of the space, and other factors. For example, when the connection radius is small, the RGG may consist of isolated clusters or disconnected components, while when the connection radius is large, the RGG may become fully connected [6]. In some cases, RGGs can exhibit a phase transition from a disconnected to a connected state as the connection radius is increased. A slightly different version of the RGG, proposed by Waxman [7], is the soft random geometric graph. For these networks, rather than having all nodes within a distance connected, nodes are connected with probability $\beta e^{-\frac{r_{ij}}{r_0}}$, where r_{ij} is the distance between nodes i and j , and β, r_0 are parameters. In [8], the authors investigate the behavior of radial betweenness centrality distribution over RGGs. In particular, they constrain those networks to particular shapes in space - for instance disks, triangles or square domains containing obstacles. In the limit of infinite density, the authors come up with an analytic expression of the radial betweenness centrality, showing that it decreases quadratically close to the center of the network, and linearly close to its border. Finally, these networks have been used to model a wide range of spatial networks, including wireless sensor networks, social networks, transportation networks, and ecological networks. They can be used to study features such as scaling or disease spreading in a null model for spatial networks, and can help to inform the design and optimization of real-world

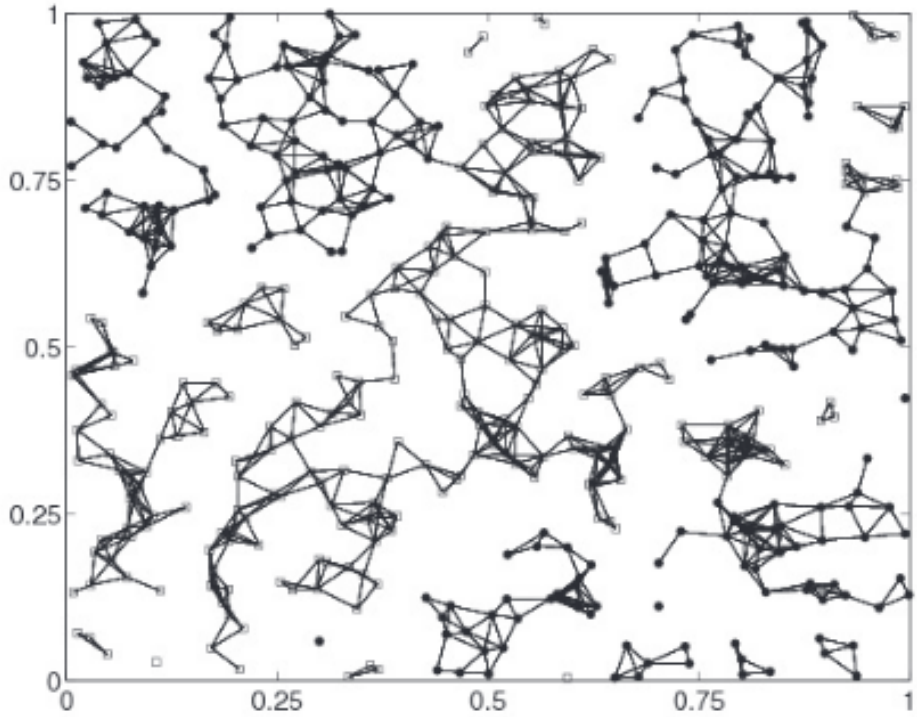

Figure 1.3: 2D random geometric graph with 500 nodes and average nodes' degree $< k >= 5$. The figure is extracted from [4].

networks - for instance ad-hoc networks [9], [4].

Overall, the incorporation of spatial information into network modeling and analysis provides a useful framework for understanding the structure and dynamics of many complex systems that are commonly found in nature. The study of spatial networks properties has important implications for fields ranging from urban planning and transportation engineering to ecology. Out of many, transportation networks are a sub-class of spatial networks which characterize the artery behind the functioning of both man-made and natural (for example the leaves patterns or slime molds), and the study of their properties benefits both the general contexts in which they are found.

Chapter 2

Data gathering

In this chapter, we present the different stages of data collection performed during this work. More specifically, we worked with two different types of data: the representations of subway networks as graphs and the distribution of amenities, as a proxy of Points of Interests (POIs), over various cities.

The first step of our work was to find which dataset permitted an extensive and general analysis of subway networks. In this chapter we describe in the first part why subway networks are relevant in the study of urban complex systems and give some context on their historical development and why they became important tools for large urban areas and give an idea of their typical behavior. We then explain in particular which process we chose to design our dataset, given a number constraints on data availability, network's scale, amount of meta-data and suitedness to a relevant set of features in the scope of the study.

We then go through the explanation of the process we followed to explore which data were available, with a broad point of view, given that the first goal of our work was to design a dataset exploring a certain amount of sources. In particular, before choosing to work with subway networks, we considered the options to work with more general transportation networks, multi-layer or not. After this description, we proceed with an explanation of the format of data we worked with, that is the static General Transit Feed Specification (static GTFS). We explain why we chose to work with this format, especially insisting on its genericity and the amount of information provided. We also describe its structure and main features, to then introduce the next part: the description of the pipeline we built to extract the networks from the raw data we gathered. To finalize the part about the extraction of transportation networks, we give a summary of the obtained result.

To conclude this chapter, we introduce the extraction of amenities instances

through the OpenStreetMap interface, as a proxy for the number of Points of Interest in a urban area. We describe which kind of POIs we chose and the reason behind the decision to add this information to the analysis and the information it provides about the functional structure of a city. Finally we also discuss some of the inherent pitfalls of this type of data to reconstruct the distribution of Points of Interests in a city.

2.1 Subway networks

Subway networks are a relatively recent transportation system. The London Underground, the first built network, is 160 years old. It is shown that cities with a number of inhabitants larger than two millions, are equipped with a subway system over 50% of the times [10]. Urban concentration in the past century enlarged the number of big-size cities and was an incentive to build those networks, given their capability to move around a very large number of passengers over a large area in an efficient manner. On the other hand, only relatively few cities reach the size for which an actual subway network is worth building and developing, which sets a natural limit on the size of the dataset.

Their usual building trajectory is to grow until a certain maximum size and then reach a size "plateau" in time, which means that only a very small number of stations is added after some point. Their growth rate might vary, some of the oldest ones grew over almost one century (example : Paris or London's networks), when some of the most recent ones are built and expanded until their limit shape in just a couple centuries (this is the case in particular for Chinese cities networks). We are interested in fully-grown networks (or almost) and do no study their evolution through time. This topic is explored in a deeper way in literature work and illustrative examples of networks growth trajectories are provided and studied on a larger scale.

2.1.1 Data selection process

Within this section, we describe the data selection process we have performed. Our goal was to construct a comprehensive set of subway systems that met a range of structural criteria. First, we searched for subway networks that we considered to be large enough, meaning containing a sufficient amount of stations and of lines - we wanted to avoid metro networks with a very long metro line, as its topological structure did not enter our framework. Furthermore, we only considered networks that had already achieved a close-to-final shape, as this was essential

for our analysis. To identify such subway networks, we simply sorted the list of subway networks in the world according to their number of nodes ,using Wikipedia [1].

In addition to the size and shape of the subway networks, we also needed access to reliable data about public transportation in the corresponding urban areas. We used a source called Transitland [2], which provides a GTFS description of transportation networks in urban areas. The GTFS format allowed us to extract the subway network information from the source in a generic way, enabling us to obtain the necessary data for our study.

However, it is worth noting that we encountered several challenges during the data selection process. For instance, some of the largest subway networks, such as those in China, were not easily accessible via open access, making them unusable for our study. Moreover, for some networks, such as those in Delhi and Berlin, the data description provided did not allow us to extract the subway network. In addition to these limitations, we had to ensure that the subway networks we selected met specific requirements. For example, we needed to ensure that the growth rate of the network was relatively low - meaning that the network reached a shape close to its limit shape. This allowed us to avoid scenarios such as networks with undeveloped branches.

Despite these constraints, we managed to compile a set of subway networks that met our criteria. However, it is worth noting that we selected Rome's subway network, which had almost no connections between metro lines, as an example of a unique and distinct network. It is also essential to acknowledge that most of the subway networks we used in our study were from European cities, meaning that our findings may contain a bias due to the historical development of these cities.

2.2 Sources

At first, we were interested in knowing which kind of networks we would be able to extract and under which format. At this step, we had not decided whether to work only on subway networks, or other types of networks, like trains, buses in urban area, or even multi-layer transportation, as is done in [11], [12]. Thus, we decided to explore what several data providers had to propose, to have an idea of which kind of networks we could be working with, at which scale, and with which point

[1] `https://en.wikipedia.org/wiki/List_of_metro_systems`

[2] `https://www.transit.land/`

of view (would we be interested in temporal graphs, or on the opposite only spatial graphs, or something in the middle where we take account of frequency or even considering delays in connections).

2.2.1 Data providers

We scraped through a list of possible data providers of transportation networks systems. To briefly list the sources we searched data through the EU data platform, that gathers data published by European states about public transportation systems. a platform named Transitland, gathering open-access GTFS files from transit data, and then National Access Points (NAPs) for France and Italy, and public data transportation for Switzerland. Data lie in different formats (GTFS, GTFS-RT, netex and custom formats), are aggregated at different scales (city, an urban area, province, country) and transportation layers are assembled together in different ways : some providers give the data layer by layer while some others merge each layer together and thus rather gives a multi-layer representation of the transportation network.

data source	link	format	data abondance	granularity	composability
EU data	`https://eu.data.public-transport.earth/`	GTFS / netex	medium	low	low
transitland	`https://www.transit.land/`	GTFS / GTFS-RT	high	high	high
NAP france	`https://transport.data.gouv.fr/`	GTFS / netex	medium	high	medium
NAP italy	`http://dati.mit.gov.it/catalog/dataset/`	many	medium	NA	NA
NAP switzerland	`https://gtfs.geops.ch/#feeds`	GTFS	low	high	very high

Table 2.1: List of a few data providers with an evaluation of the quality of respective properties, as well as the format in which the data is given in. Data abundance relates to the amount of data that could be retrieved, from the point of view of the variety of networks, granularity refers to the property of the data to already being cut in pieces of sub-areas or not, and composability refers to the ease of using this data with data from other sources and other networks.

2.2.2 GTFS data format

The GTFS [3] format is a standardized format for public transportation data and related geographic and temporal information. It was created to help agencies share easily data about the transportation service they managed to third-parties. In particular, its specification allows developers to handle data in a generic way.

[3] `https://developers.google.com/transit/gtfs/reference`

A GTFS description of a given set of transportation system over an urban area consists of a set of text files following a certain specification. Some files are required to follow the specification, while some others contain meta-data, which can be useful for several purposes, such as shapes of paths for a geometric study of a network, the description of the fares over the network, or the description of the connections at transfer points. Inside each file, the same rule apply to fields : some are mandatory, others are not. The level of detail achieved can be high, but this means that the whole description gets heavier and heavier, while the information transmitted through some meta-data is not relevant to all purposes.

We list the main files and fields of interest for our purposes :

- routes
  - route_id
  - route_type
- trips
  - route_id
  - trip_id
- stops
  - stop_id
  - stop_lat
  - stop_lon
  - parent_station
- stop_times
  - route_id
  - trip_id
  - stop_id
  - stop_sequence
  - arrival_time
  - departure_time

Metro lines, abstractly, are represented by routes. Trips represent a specific realization of this route : at what time it starts, in which direction. Parent structure of stops represent a metro station (for instance, a stop can be the platform of a metro station on a line, going in a specific direction). Stop times give information about when which trip reach which station.

2.3 Graph extraction pipeline

Given the goal of having a representation of subway systems in the form of spatial graphs, with nodes representing stations located with coordinates, and edges between adjacent nodes on a metro line. To extract the subway networks from GTFS files, we came up with an algorithm that works generically given a GTFS description of transportation systems, and returns a description of spatial graphs consisting of a set of nodes and edges linking them, associated with additional fields. We first describe the algorithm without giving implementation details, and then summarize the results provided by running the algorithm on the data we could access, i.e through Transitland open-access platform.

2.3.1 Algorithm

We start with a generic description of the transportation over an urban given by GTFS specification. The files and fields we are interested with are described in 2.2.2.

The first step is to identify the metro lines : for this, we select the routes (that is, retrieve route_id attribute) with route_type corresponding to a metro line [4]. Given this information, we can select the trips - retrieve trip_id attributes - of all trips corresponding to the routes we selected. Given this list of trips, through stop_times, we can extract sequences of stop_id, sorted by stop_sequence, and only keep an unique copy of each sequence of stop_id. Given this list of sequences, we iterate through these sequences and each time we encounter a new station, add it to the list of stations, adding its coordinates as attributes, and similarly, we add an edge (if it has not been added yet) joining two consecutive nodes in the sequence. Since we are working with undirected graphs, an edge (n_1, n_2) is equivalent to (n_2, n_1). The post-processing step was to convert longitude-latitude coordinates to Pseudo-Mercator coordinates : we preferred working with projected

[4] `https://developers.google.com/transit/gtfs/reference/extended-route-types`

2D coordinates than longitude-latitude coordinates.

2.3.2 Result

We selected the transportation systems of 18 cities : being Barcelona, Berlin, Buenos Aires, Chicago, Delhi, Hamburg, London, Madrid, Mexico, Milan, New York City, Paris, Philadelphia, Roma, Santiago, Toronto, Vienna, Washington D.C.

We encountered several issues : the layout of the data with Berlin, Delhi, Philadephia and Toronto did not allow us to extract properly a representation of the subway network, due to an issue in the GTFS representation, related to stations and parent stations. For New York City, the GTFS data encountered a problem : two (distinct) stations were located at the same coordinates, but we noticed while working with the wrong representation and corrected it.

We ran the algorithm successfully on the data mentioned above, ad failed on Berlin, Delhi, Philadelphia and Toronto. Among the networks we selected some might look like outlayers given the criterion we mention : Roma is composed of only three lines, Chicago and New York City have a very unusual spatial display, due to the lake and rivers respectively.

2.4 Gathering amenities from OpenStreetMap

Our research in the field of urban complex systems aims to investigate the connectivity between highly populated areas, usually located around the center of cities - areas reached along the branches of the networks - and highly active areas, which are characterized by a high density of amenities. In order to gain a better understanding of the distribution of POIs in the urban areas we were studying, we set out to extract an estimation of their distribution.

To accomplish this, we utilized the OpenStreetMap interface to access a relevant subset of POIs in the urban area. These POIs were provided as a set of points with coordinates within a selected bounding box, which represents the extent of the urban area in question. By analyzing the distribution of these points, we aimed to estimate the spatial density of POIs across the urban area.

However, it's important to note that there are certain biases present in the gathering of POIs over an urban area. First and foremost, these data are collected and transmitted by humans, and the importance placed on the collection of data

Figure 2.1: Display of Milano subway network, as extracted by our pipeline. A random subset of station's identifiers are shown, to give a more understandable view of the network.

varies across different cities. While some cities have organized data collection processes in place, others have not, which can lead to inequalities in the quality and quantity of the data available. As such, we do not expect our estimations to provide an exact representation of reality. An other problem with this data lies in the fact that points tend to be collected more frequently in the center of a city than in the surrounding areas. However, despite these limitations, we believe

that the data we gathered provides a general understanding of the density of POIs across the urban area, and can offer valuable insights into the connectivity and distribution of amenities in urban environments.

The list of POIs that we gathered over urban areas is the following :
'cafe','college','library','school','university','kindergarten','restaurant',
'pub','fast_food','bar','bank','dentist','pharmacy','hospital','clinic','doctor',
'arts_centre','cinema','community_centre','police','post_office','marketplace' .

We show on figure 2.2 an illustration of the density of POIs over Madrid urban areas, restricted to the areas where the subway networks develops.

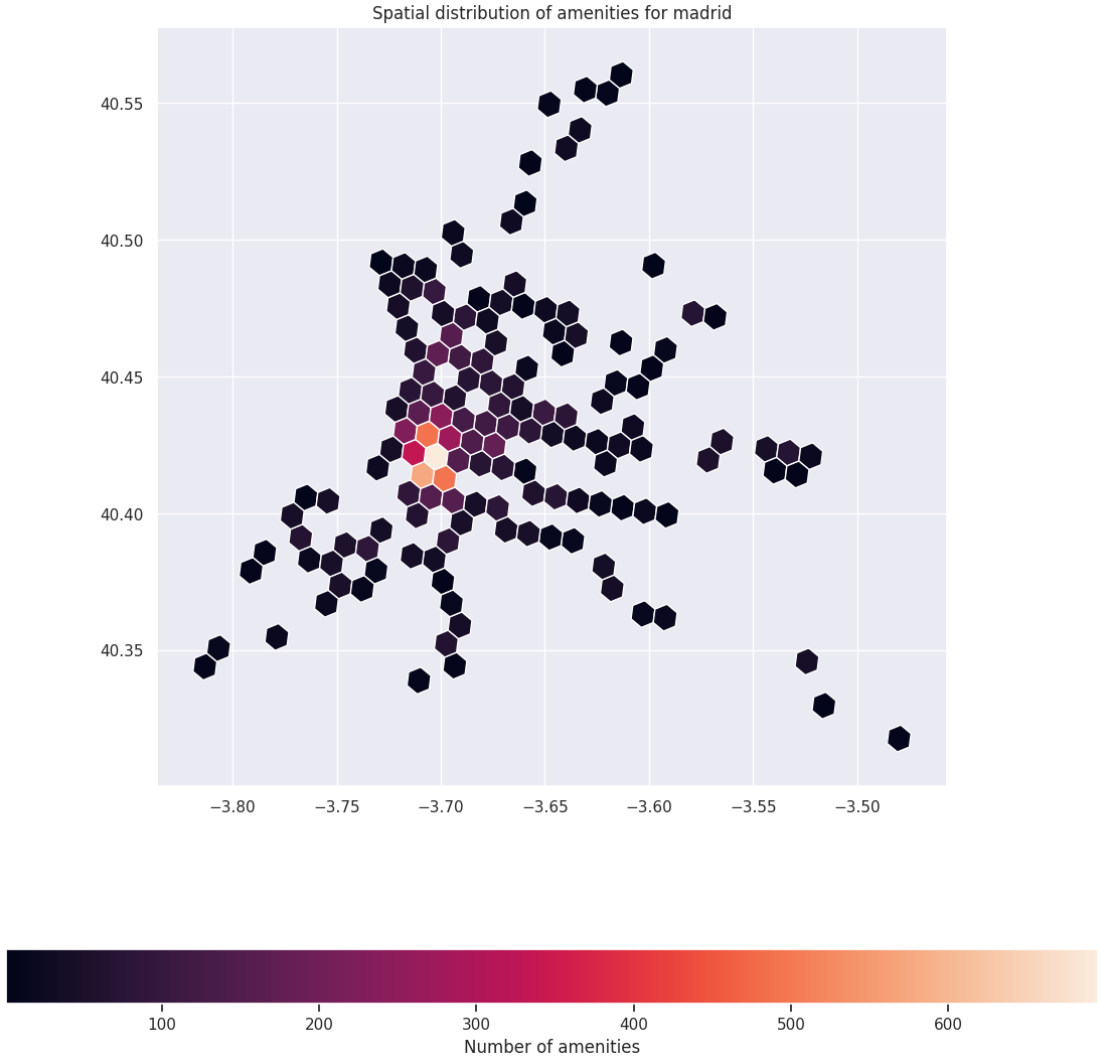

Figure 2.2: Distribution of amenities in Madrid. Density is shown on hexagonal tiling filtered on tiles containing at least a metro station. The distribution peaks very high close to the actual center of the city, and decreases to much lower values as soon as we get further from it.

2.5 Chapter 2: Overview and conclusion

In this chapter we went through the building of the dataset we will be working with , starting with its design, the research of open data, the choice of networks we would be interested with. We then went through the whole construction algorithm, starting from the raw data about transportation systems on an urban area to end up with a spatial network representing stations and their adjacency. In particular,

we pointed out the limit that exists on the size of the dataset, given the fact that the number of very large urban areas is quite limited. Moreover, we explained some of the important features of the networks we are working with, as well as some pitfalls that are inherent with the data, such as not counting for frequency of a line, or more generally, not taking account of the temporal structure over the network. Finally, we gave a description of the extraction of the POIs data through OpenStreetMap interface.

Chapter 3

Analysis of subway network's features

In this chapter we describe the theoretical and computational framework starting with an introduction on the study of subway systems via networks science, supported by a description of some literature works we based our work on. Additionally, we mention a set other works, some even more generally related to urban complex systems and the mathematical analysis of spatial networks. The purpose of this additional part is to give also an introduction to some modeling approaches that we used as foundations, as well as giving us ideas for potential future works and generalizations, with the aim of combining ideas to propose novel analyses of subway networks.

In particular, in this chapter, we give a description about recent literature and foundational works on subway networks. In more detail, we describe the set of hypotheses that are made and a set of results, from the topological and structural properties of these networks to their evolution. We also reproduce these results and then give them a critical look to understand what could be improved or generalized.

Specifically, we report and describe the results reproduced on the dataset we built. We highlight the different concepts we studied, under which kind of hypothesis they can be reproduced, and verify if those results stick with the literature proposed. We first go through a series a very general indicators about the topology and the performances of the network. Then, we go through structural properties such as the core branches structure. We then study the spatial layout of the network around the barycenter through the study of radial distributions of several network metrics. Finally, we proceed with an analysis of some scaling properties, specifically focusing on the property of the radial distribution of the number nodes from the center of the network, which highlights some topological properties of this system.

We then explore how the efficiency metric describes the spatial structure of the network and provides the pivotal metric for giving better understanding of these structures.

3.1 Literature Review

In this section we introduce previous works about subway networks and spatial networks, outlining the fundamental contributions and results obtained that we aim to reproduce.

3.1.1 Subway systems

We first discuss the two main works we got inspired from, specifically references [10] and [13], which approach different aspects of the evolution of subway networks. They focus on temporal evolution of networks, on their static and dynamic properties, propose a decomposition of the subway networks in a core and branches structure and explore its spatial organization, and finally investigate the spatial structure of a metric called efficiency.

About the temporal evolution of subway networks, they discover that, typically, networks size evolve according to a two-step regime: the number of stations first grows until reaching a limiting shape, and then stagnates in time. They conjecture that there exists a temporal stationary limit for those networks and investigate the typical shape of this stationary limit. In particular, they investigate the average velocity of evolution of networks, and the time evolution of the network in a core and branch structure, that we introduce later. The exploration of static properties is an attempt at grasping the topology of the studied networks. They study properties such as the relation between number of lines and number of stations, the average interstation distance, the total length of a network and the proximity between total length and the total length of a "regular" (in the sense, with small degree variation) planar graph with spatially evenly distributed nodes.

The two next concepts we cover are crucial to the work accomplished in the thesis, in the sense that a substantial part of our work was to understand the genericity of these features, trying to improve their generality and we intend to dive more in this topic in our future works.

Decomposition of core and branches structures

We start describing the decomposition of subway networks in core and branches structures. To give a qualitative description of this decomposition, we imagine a network growing through time. At first, stations and lines are built in the center of the city, creating a denser network in this area. Through time, stations start getting built further and further away from the center, as well as metro lines are created to reach areas in the surroundings. This set of graph extension in a quasi-rectilinear way gives what we call the branches, while the denser part of the graph close to the center is what we call the core. We notice that the boundary between core and branches defines a ring, as shown in 3.1.

We now give a more theoretical definition of those concepts: both works we began with [13], [10] give a slightly different formulation. The former one defines the decomposition as a k-core decomposition [14] with $k = 2$, when instead the latter one adds as well the concept of fork on the branches, that is a node outside of the ring the divides in two sub-branches. To find this decomposition, we apply the following procedure: retrieving the core by iteratively removing nodes of degree one, then the branches are given by the connected components of the graph whole graph, to which we remove the core. Figure 3.1 gives a schematic description of this decomposition and is taken from [10].

These works study the spatial repartition of the core and branches, and we focus in particular on the repartition of the number of nodes at a given distance from the barycenter of the nodes coordinates $N(r)$. [10] shows that $N(r)$ evolves following 3 regimes. Close to the center, it evolves quadratically, then follows a regime with lower growth on the area where new nodes correspond to branches, characterized by average interspacing between stations, and finally ends being constant after passing the furthest node. The study of these properties can be seen as a more advanced version of [15], that explores the development of Paris' subway system at range from the center.
The other proposed scaling model in [13] shows that instead, in the core, $N(r)$ follows a power law with exponent $\gamma_c \approx 1.23 > 1$, while on the branches it follows a power law with exponent $\gamma_b \approx 0.46 < 1$.

Network Spatial Efficiency

We also describe here a metric used in several of the works that inspired us [16], [13], named efficiency. The efficiency E_i of a node i is defined as:

$$E_i = \frac{1}{N-1} \sum_{j \neq i} \frac{1}{d_{ij}}, \tag{3.1}$$

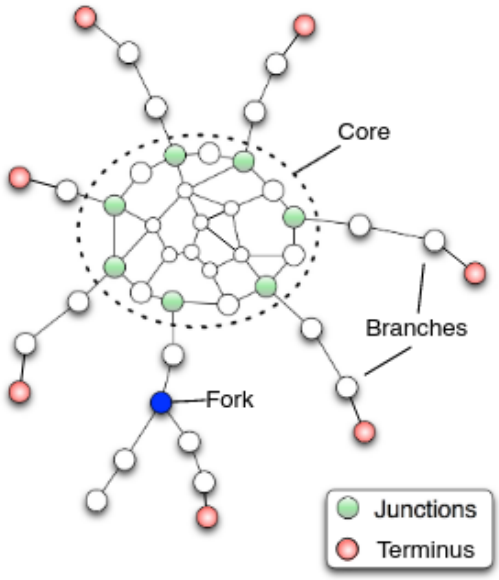

Figure 3.1: Sketch of core paired with branches decomposition. Branches develop around the ring that encapsulates the core. With this definition, branches are composed of a terminus (node of degree 1), a number of junctions (nodes of degree 2) and at most a fork (degree 3). The algorithm we describe allows us to retrieve the decomposition in core and branches of a network. The sketch is taken from [10].

where N is the number of nodes in the graph, the sum is on all the nodes distinct from i and d_{ij} is the shortest path in the graph between i and j. The sum is defined whether the graph is connected or not, taking the convention $d_{ij} = \infty$ if i and j are not connected. E_i are normalized to compare them. For a given network, E_{ideal} is defined as:

$$E_{ideal} = \frac{1}{N(N-1)} \sum_{j \neq i} \frac{1}{l_{ij}}$$

where l_{ij} is the distance between the coordinates of i and j (several distances can be computed, depending on the system of coordinates used to locate the stations). We can understand E_{ideal} as the average efficiency of the complete graph over the nodes of the subway network, with length of edges equal to the distance between the nodes' coordinates (triangle inequality guarantees that for all graph over those nodes, $d_{ij} \geq l_{ij}$). This metric enables us to compute the efficiency of nodes for a

given graph, as the spatial proximity between the node and the rest of the graph, or to evaluate the efficiency of a network relatively to the most efficient possible network over those nodes, as $\frac{\langle E_i \rangle}{E_{ideal}}$. In particular, [16] uses this concept to show how efficient Boston's subway system is, by comparing the number of edges built over the network to the number of edges over a complete graph, and defines this way a notion of cost. [13] provides an illustration of the proximity of the top efficiency nodes to the center of the graph, fact that we will reuse in this work.

3.1.2 Spatial networks

In this section, we give a brief summary of a few references about spatial networks that we have used. We focused in particular on distribution of betweenness centrality (BC) for spatial networks [17], [18]. We mention as well a review about spatial networks that was particularly useful to us, both for introducing many concepts - especially about characterization of the topology of a network and performance indices - and for literature redirection [4]. This work covers a much broader field than what we focused on.

[18] and [17] both discuss BC distribution, from very different perspectives. Let's first recall what betweenness centrality is: it counts the proportion of shortest paths passing by a node. That is, the betweenness centrality of a node is the average over all pairs of distinct nodes of the proportion of shortest paths passing by the given node.

This metric is much more expressive than degree, as it reflects how important a node is relatively to the others, with respect to the navigation over shortest paths over the network. For instance, a cut-vertex node could have a degree as low as 2, which could be comparatively to other nodes, as low as it can get, while it betweenness centrality would have a relatively high value, since all paths connecting nodes belonging to the two connected components have to pass through this node. Still, betweenness centrality does not reflect any spatial property over the network, but only topological structure. While we still use it to evaluate node's importance, we use as well another metric that takes account of node's spatial layout.

Authors in [18] provide a geometrical analysis of spatial behavior of BC , in the framework of extremely dense networks. In particular, with the hypothesis of a circular boundary and dense soft random geometric graphs, they propose a formula on the radial BC distribution when the number of nodes goes to infinity. They show that the BC decreases quadratically near the center and scales linearly close to the boundary. The work in [17] studies the behavior of BC probability on random graphs, relating it to other indicators of the topology of the graph, and concluding

by separating two types of behavior for BC distribution : one for tree-like graph, and one for graph with high density of loops. Since this work does not dive into the general study of spatial networks, our main interest with it was to grasp a better understanding of BC behavior over different kind of graphs. Finally, we used [4] to discover some useful indicators about topology of the graph and performance indicators, such as the transportation performance index. This index is equal to the ratio of the average length of shortest paths over the whole network and the average length of shortest paths over the minimum spanning tree. The smaller, the better the network improves transportation relative to a default structure.

3.2 Reproduction of Literature Results

Given the analytical description of spatial and subway networks introduced above and the data we have at our disposal, we chose to focus solely on studying the final shape of network (i.e viewing the network as having reaching its stationary limit), instead of studying their evolution in time. We mostly investigate spatial, topological and geometric properties of those networks.

We start with a very general description of the network using some useful indicators, to grasp a very general idea about their structure and topological properties. We continue with the analysis of the decomposition of the networks in core and branches structures over different networks, to check if this result is consistent. We see that cities develop spatially in different shapes, and we give a set of hypothesis that are made throughout previous works, for instance, isotropy. We discuss about the barycenter of the networks' nodes we study and try to understand generically how it behaves, with respect to the core and branches structure.

Then, we study the radial distributions of several metrics, like betweenness centrality and number of nodes for example. We also study how these metrics correlate one to another and how relevant they are in the framework we are working with. To continue on the same topic, we dive in detail on the scaling properties of subway systems, understanding how different shaped networks behave : in particular, we see that shapes and length of branches affect the result, as well as anisotropy and non-circular around the barycenter core.

Finally, we describe a spatial network metric called efficiency, and study how this measure yields an accurate description of the network spatial layout.

3.2.1 Network Indicators

In this section, we give a description of a subset of relevant subway networks through a series of very general indicators about those graphs. The networks we discarded were outlayers for our analysis, because they were either not developed enough (that is, their size was not big enough, either in number of lines or number of stations), for instance Rome's subway system, or because their shape was too different of the generical one, that is a structure that does not correspond with a decomposition in core and branches structure - for instance Barcelona. The existence of both those types of networks is explainable, for instance with Rome, difficulties encountered that digging a metro line involves, or with Barcelona, the fact that the urban area is partitioned in spaces reachable with different means of transportation - some parts of the city reachable with metro, while others could only be reached with buses or tram.

We now describe the indicators that we used to characterize the networks. First, the number of nodes and edges. Some of those networks are planar, hence one could compute the number of faces of those graphs with Euler's formula $F + N = E + 2$, where F is the number of faces, N the number of nodes, and E the number of edges. Since we do not use the planarity of the graphs, we chose not to consider this additional metric. Then, we indicate the number of nodes in the core, as computed with the 2-core decomposition algorithm. This is an indicator of the size of the center of the network. Then, we use three indicators α, β and γ, defined for instance [4] and [10]. α, the meshedness, is defined as $\alpha = \frac{E-N+1}{2N-5}$. This indicator measures the number of bounded faces of a planar graph. The closer to 0, the more similar the graph is to a tree, the closer to 1, the more similar it is to a maximal planar graph [1]. γ measures the density of a graph in number of edges, and we use the following definition : $\gamma = \frac{E}{3N-6}$. β defines the proportion of nodes on the branches of the graph, that is $\beta = 1 - \frac{N_c}{N}$.

The next indicator $\langle d \rangle$ is the average degree of the nodes of the graph, while $\langle l \rangle$ is the length of the average shortest path over the graph (considering the graph unweighted). Transport performance and efficiency ratio are defined in section 3.1.2 and section 3.1.1, respectively as $\frac{\langle l \rangle}{<l_{MST}>}$ and $\frac{E_i}{E_{ideal}}$.

We show in table 3.1 the values of these indicators on a set of large subway networks. As expected, the value of α is quite low in general, with a median of $5 \cdot 10^{-2}$, except for the case of Santiago, which is due to a large number of triangles in the graph (that is, two lines going trough the same three nodes, but sometimes

[1] `https://en.wikipedia.org/wiki/Meshedness_coefficient`

by skipping the middle node, which creates a triangle in the graph). This indicates that these networks tend to be similar to trees, which is explainable by the parts of the graph that are tree-like, that are the branches of the graphs. [10] finds that $\beta \approx 45\%$ on the graphs they work with, we find an average value of 53%, explained by the presence of smaller networks like Milan or Washington D.C's, with a large amount of branch nodes. $\gamma \approx 0.38 \pm 0.03$, which indicates that these networks have similar edges densities.

An average nodes degree $\langle d \rangle \approx 2.27 \pm 0.19$, is expected from [10]. Our average value of $\langle d \rangle$ and its variance are increased by the peculiar Santiago network. Efficiency lies $[0.60, 0.80]$, except for New York City's system which has a very low efficiency, partially explainable by its peculiar structure, shaped around both Hudson and East River. Finally, transport performance seems to lie around 0.82 ± 0.08. We notice that Chicago and Washington D.C's subway systems have huge transport performance, explained by the fact that their structure is almost tree-like (E and N being very close), thus the improvement of the length of shortest paths of the network compared to the ones on the minimum spanning tree are marginal. Instead, for the three biggest networks of our study, i.e New York City, Paris and London, there is a noticeable improvement, as transport performances value are lower, around 0.70. This indicator, not taking account of spatial disposition of the network, misses the inefficiency in New York City's system that

	N	E	N_c	α	γ	β	$\langle d \rangle$	efficiency ratio	$\langle l \rangle$	transport performance
Vienna - AT	96	103	35	4.28e-02	3.65e-01	6.35e-01	2.15e+00	7.74e-01	1.04e+01	8.30e-01
Santiago - CL	119	176	102	2.49e-01	5.01e-01	1.43e-01	2.96e+00	8.03e-01	8.76e+00	7.73e-01
Washington D.C - US	97	99	23	1.59e-02	3.47e-01	7.63e-01	2.04e+00	7.51e-01	1.33e+01	9.85e-01
Chicago - US	138	148	37	4.06e-02	3.63e-01	7.32e-01	2.14e+00	7.20e-01	1.37e+01	9.62e-01
New York City - US	472	549	315	8.31e-02	3.89e-01	3.33e-01	2.33e+00	3.03e-01	1.39e+01	6.72e-01
Hamburg - GE	149	162	60	4.78e-02	3.67e-01	5.97e-01	2.17e+00	7.00e-01	1.20e+01	9.17e-01
Madrid - SP	201	222	87	5.54e-02	3.72e-01	5.67e-01	2.21e+00	7.05e-01	1.34e+01	7.43e-01
London - UK	270	316	161	8.79e-02	3.93e-01	4.04e-01	2.34e+00	7.22e-01	1.32e+01	7.08e-01
Paris - FR	308	368	201	9.98e-02	4.01e-01	3.47e-01	2.39e+00	7.66e-01	1.19e+01	7.45e-01
Milan - IT	107	109	21	1.44e-02	3.46e-01	8.04e-01	2.04e+00	7.52e-01	1.24e+01	8.83e-01

Table 3.1: Metrics on a selected subset of subway networks composed of Santiago, New York City, London, Paris, Vienna, Madrid, Hamburg, Chicago, Washington D.C, Milano. We chose metrics and networks which yielded a precise understanding of the prototypical structure of the networks we were going to study. That is, a network close to tree-like structure (translated both by α and transport performance), with a similar edge density, a degree distribution hugely peaked at 2, typically high efficiency value, meaning a certain optimality in space. Number of other insightful metrics have been dismissed, such as η, f_2, number of branches or route factor, described in [10] and [4], given the redundancy of the information they outlined.

the efficiency ratio manages to catch. This feature of efficiency is essential to the rest of our work : it captures spatial pitfalls of a network, and in particular of nodes.

3.2.2 Core and branches structure

In this section, we dive in detail about the core and branches structure over the networks we study, with the wish to approximate their core and branches structures by applying the 2-core decomposition algorithm. We recall the 2-core decomposition algorithm briefly : removing nodes of degree $k = 1$ until there is no more gives the core. The difference of the graph and the core gives the set of branches.

We show in figure 3.2 the result of the algorithm on Paris, London, Madrid and Hamburg's subway networks. Even though we have an idea about the existence of this structure, we notice that the algorithm does not always capture it precisely: in particular, we notice that loops structure create a problem when they lie in the branches, for instance, there are several triangles and loops on London's network that disrupt the core structure (understood qualitatively: a dense structure located in the center of the city). Despite these problems, the visual inspection of the cores computed for Paris and Madrid networks matches our expectation : a dense area of nodes surrounded with radiating branches. For Hamburg, the algorithm fails at capturing nodes included in the central loop, for London, the problem created by the loops causes a huge over estimation of the number of nodes in the actual core, as well as confusion on its dimension.

We will see that these overestimations of number of nodes in the core and size of the core could lead to problems in the following sections.

3.2.3 Barycenter

For studying the network from a geometrical point of view, we need a reference point on the network. Past works use the barycenter of the nodes to define a central point for the network. We show on figure 3.3 a visualization of the four same subway networks as in previous section, with the approximate decomposition in core and branches, and we added the barycenter of the nodes and the circle $Circle(bar, r_c)$, that is the circle with the nodes' barycenter as center, and radius r_c defined as $N(r = r_c) = N_c$, where $N(r)$ is the number of nodes in a disk of radius r and centered on the barycenter, and N_c is the number of nodes in the core as computed with the decomposition algorithm.

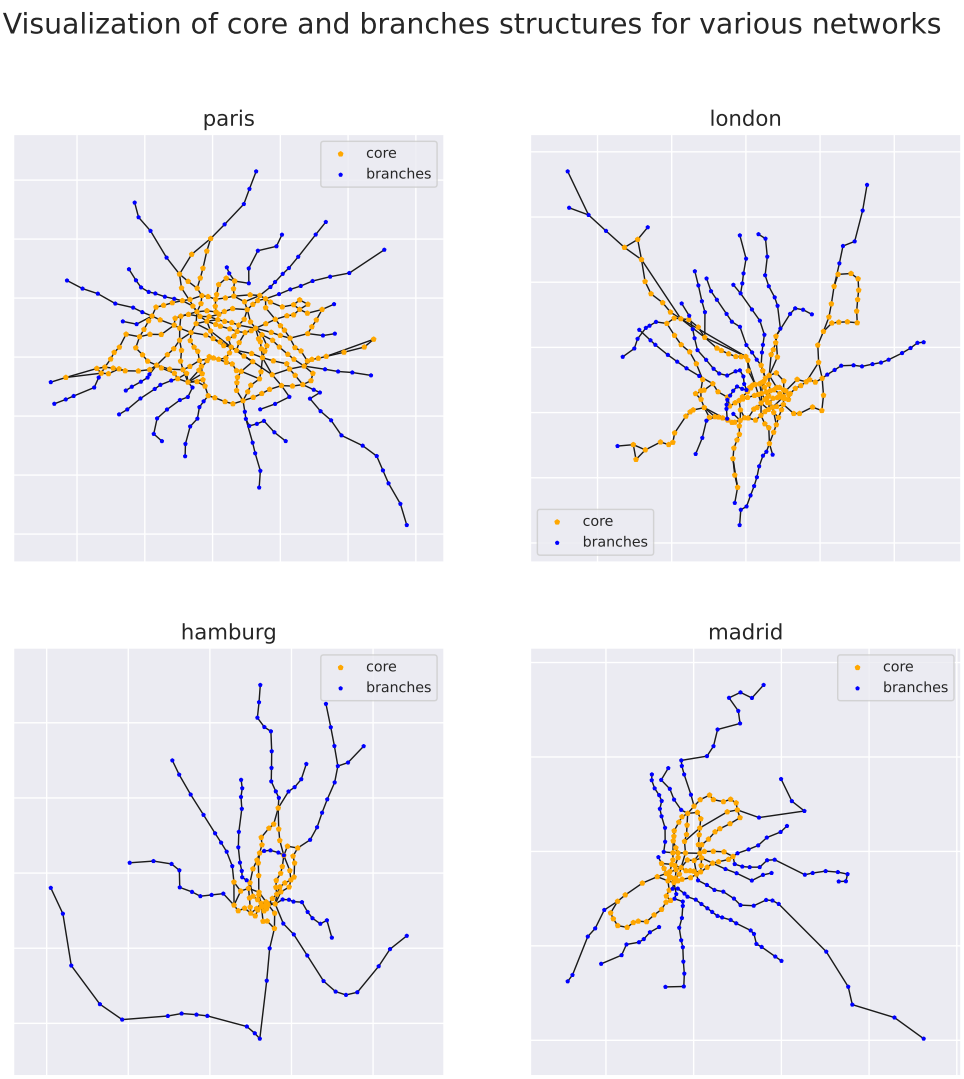

Figure 3.2: Visualization of decomposition of subway networks in core and branches. Decomposition is highlighted for Paris, London, Hamburg and Madrid networks.

For Paris and Hamburg, given the relative isotropy of the networks, the barycenter is located close to the area of maximum density of nodes, and $Circle(bar, r_c)$, seems to capture the core quite well. Instead, for London, given the error in the estimation of the core, which leads to an error in the estimation in the number

of node it contains, r_c is larger than what it should be and thus large parts of a few branches are captured in $Circle(bar, r_c)$. We notice as well that the barycenter falls quite close to the area of high density of nodes, but not inside. For Madrid, the result looks similar but is actually due to other factors: the core computed by the decomposition algorithm is coherent, but does not follow a circular shape and is instead developed on the south-west/north east diagonal. Thus, $Circle(bar, r_c)$ fails to capture it properly, and contains the beginning of some branches as well. We notice as well that the barycenter is again close to the area of high density of nodes, but not quite inside.

3.2.4 Study of Radial Distributions

Given the geometrical framework we are working with, we study radial distributions of metrics. We start with an illustrative example, in figure 3.4, showing relationship between average interspacing at distance r and distance to barycenter - that we will call as well radius. Average interspacing — $\Delta(r)$ — is the average length of edges adjacent to nodes at radius r. We conclude a correlation between radius and average interspacing, with actually some fairly high R^2 values.

In figure 3.5, we show the radial distribution of BC around the barycenter and its fit with a decreasing exponential, that is

$$\frac{B(r)}{B(r_c)} \approx a e^{-\frac{r}{r_0}} \tag{3.2}$$

The long-range distribution of the BC is very noisy, because only a few branches with a sparser distribution of nodes exist at long distance from the barycenter. The increase in interspacing at long range is shown in figure 3.4, while the decrease of density of nodes at long range is described in section 3.2.5. These two distributions are linked since they are result from the same underlying properties of the network. The characteristic distance given by r_0 does not seem to be correlated with the core radius r_c. Even though radius and betweenness centrality are correlated, the irregularities in the empirical network seem too high to expect that BC gives an accurate representation of the network's structural properties (for instance, there is a number of low BC nodes in the core, while some nodes on the branches have relatively high BC). This is explainable by the fact that BC is computed without taking account of any spatial features, only topology matters.

In the next section, we study the radial cumulative distribution of number of nodes around the barycenter. In the last section of the chapter, we show that efficiency carries spatial properties of the network much better than other metrics.

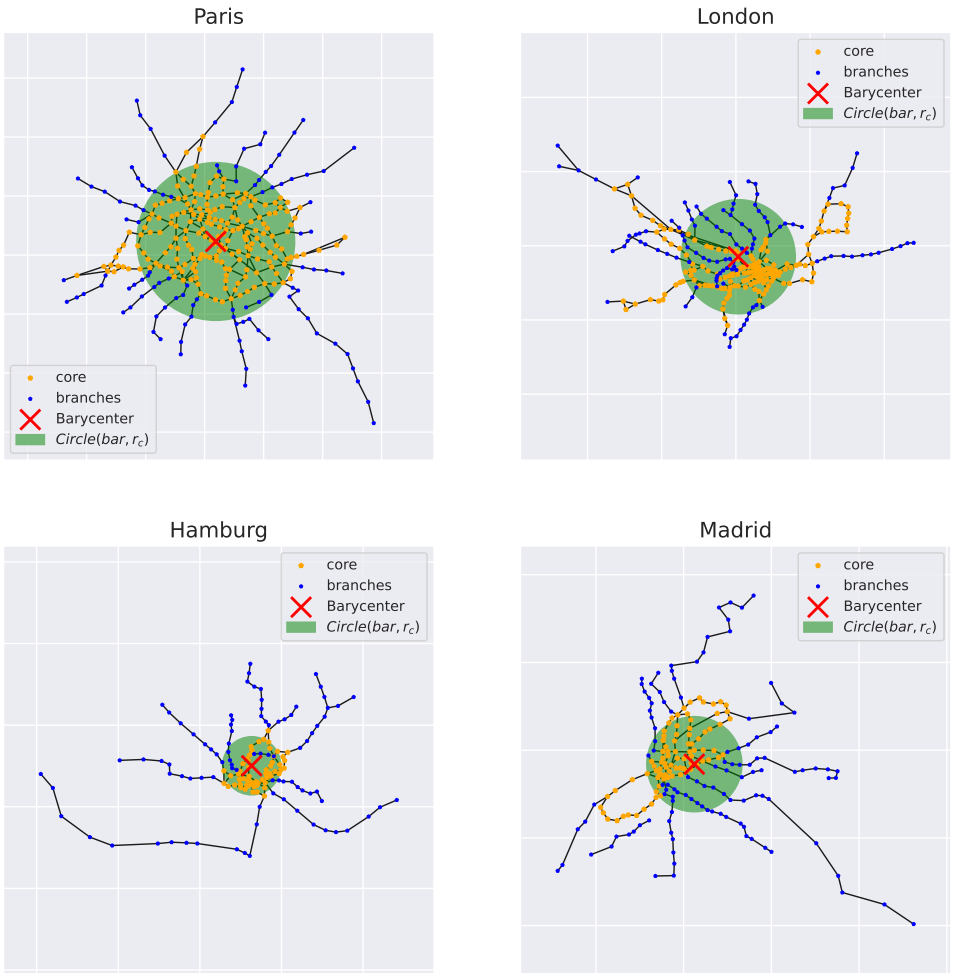

Figure 3.3: Visualization of subway networks of Paris, London, Hamburg and Madrid, partitioned in core and branches structures. We compare the shape of the core to the one of the circle centered on the barycenter and of radius r_c.

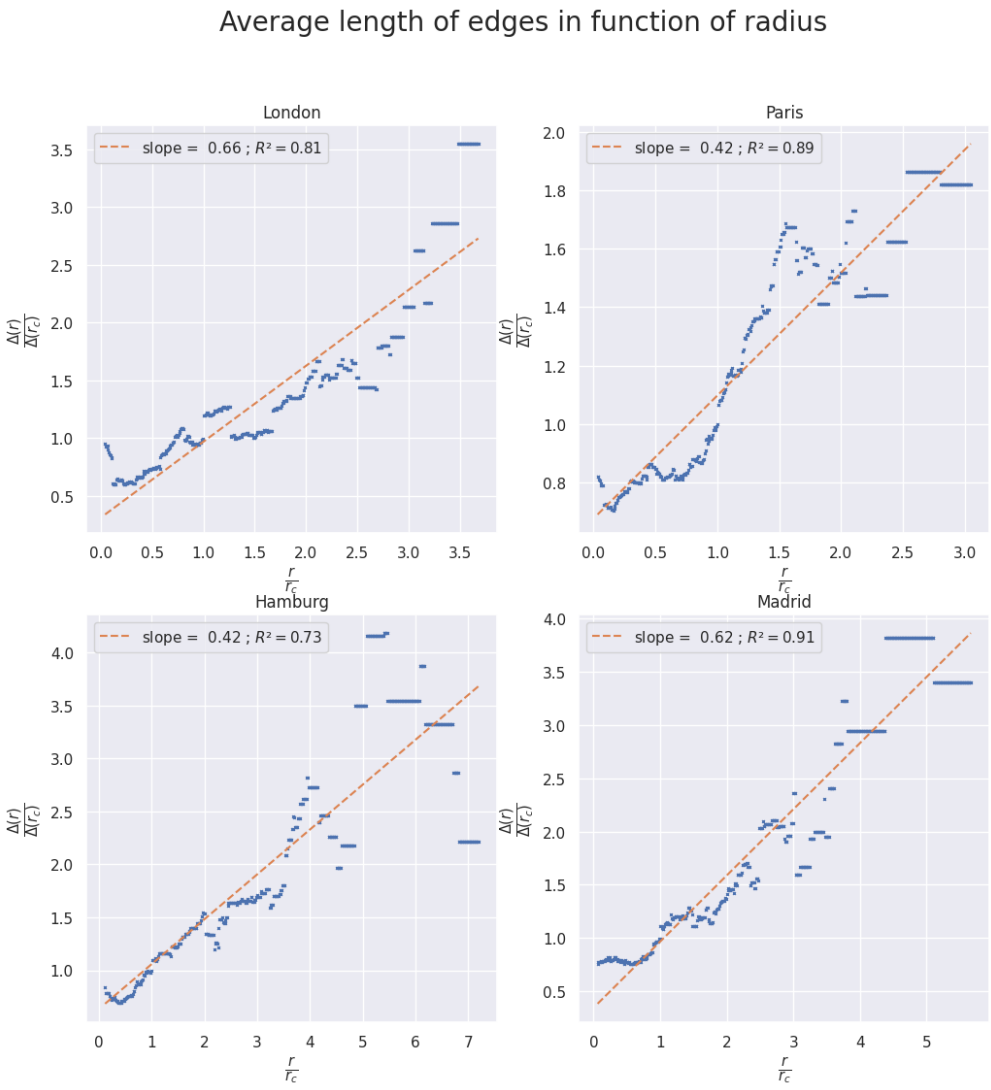

Figure 3.4: Average interspacing in function of distance from barycenter. We average the length of edges laying inside the slice starting at distance r from barycenter and ending at distance $r + dr$. $\Delta(r)$ is used in section 3.2.5 for the estimation the behavior of the number of nodes at distance $r > r_c$. From these graphs, even though noisy at long distance from the barycenter, there is a clear correlation between average interspacing and distance to barycenter.

3.2.5 Scaling properties

In this section, we dig into the radial cumulative distribution of number of nodes around the barycenter. [10] proposes a 3-regime model, using several hypothesis:

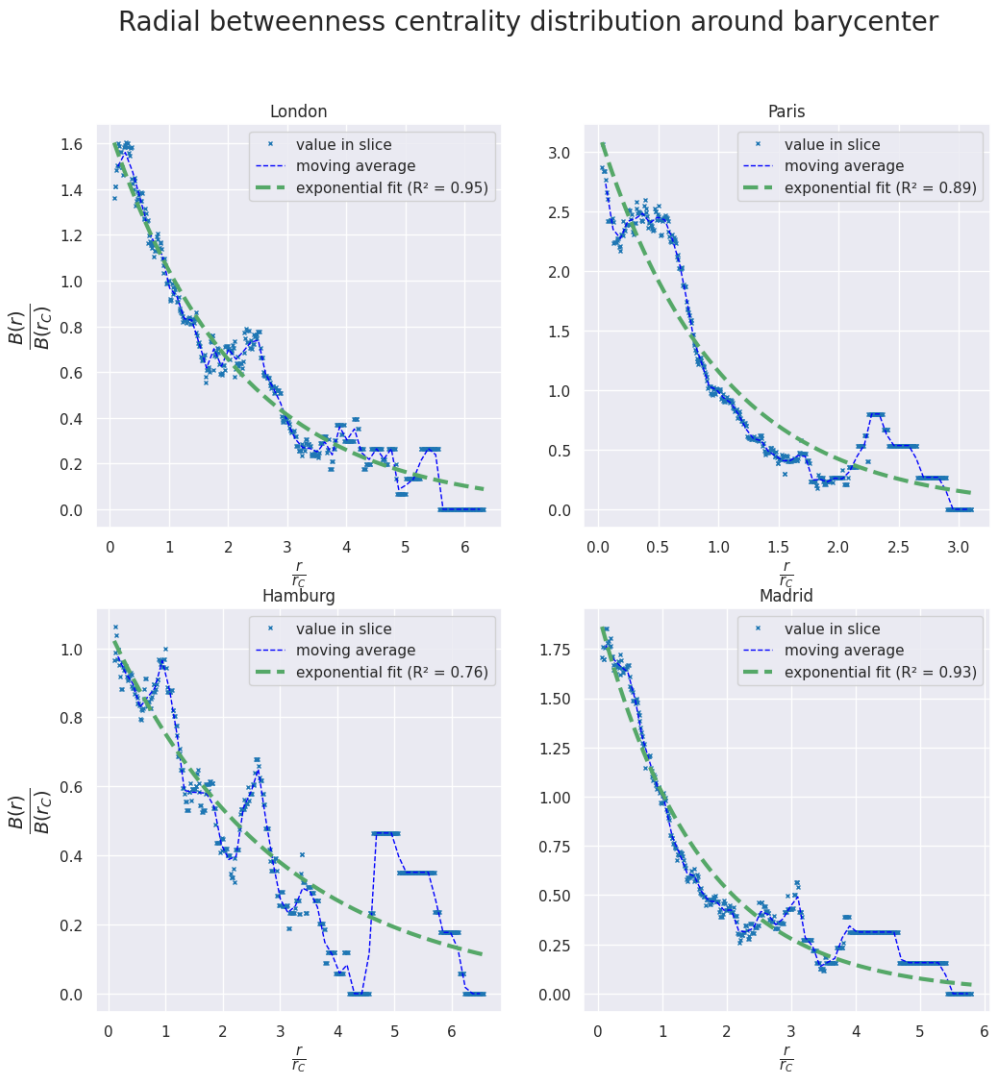

Figure 3.5: Radial distribution of betweenness centrality for London, Paris, Hamburg and Madrid subway networks. Given the expected behavior of BC , at long distance (almost 0 on terminuses, the last node of a branch, which are expected to be the furthest nodes from the barycenter) and in the core, where the top BC nodes should be, and where we expect a large density of nodes, located close to the barycenter. As we saw for London, Madrid and Hamburg, the barycenter are located slightly outside of the densest area of concentration of nodes. This explains the peak close to 0 that we observe for those 3 networks.

the core is circle-shaped and nodes are distributed uniformly in it, while the branch distribution is isotropic.

More specifically, the proposed model is the following:

$$N(r) = \begin{cases} \rho_C \pi r^2 & \text{if } r < r_C \\ \rho_C \pi r^2 + \mathcal{N}_B \int_{r_C}^{r} \frac{dr}{\Delta(r)} & \text{if } r_C < r < r_m \\ N & \text{if } r > r_m \end{cases} \tag{3.3}$$

where ρ_C is the core density, $\mathcal{N}_B$ is the number of branches, $\Delta(r)$ is the average inter-spacing and N the number of stations in the graph.

We show on figure 3.6 the result of the fit of this model on a subset of city of our dataset, chosen with different shapes and structures: Paris has indeed a circular core and isotropic and rectilinear branches, while London has branches of different lengths distributed ununiformly around the center. Hamburg has an isotropic distribution of branches branches but they are not rectilinear. Instead, Madrid has quite rectilinear branches that are not distributed uniformly around the center.

As shown in figure 3.4, $\Delta(r)$ is not constant and quite noisy on the branches. [10] fits the approximation of $N(r_C < r < r_m)$ to a power law. Doing the same thing, we find similar results : $N(r_C < r < r_m)$ fits a power law with exponent around 0.5 with high Pearson correlation coefficient (≥ 0.9).
Instead, we computed the R^2 coefficient between $\rho_C \pi r^2 + \mathcal{N}_B \int_{r_C}^{r} \frac{dr}{\Delta(r)}$ and $N(r)$, between r_C and r_m, which yields high scores. For Hamburg, with extremely long branches, the approximation is of highly precise, as the high R^2 score shows. The fit with a quadratic function on the core is questionable, [13] which is more recent, rather fits the density of the node in the core with a power law with coefficient $1 < \gamma_C < 2$.

3.2.6 Efficiency

In this section, we introduce a metric over spatial network called efficiency. We start by recalling its definition :

$$E_i = \frac{1}{N-1} \sum_{j \neq i} \frac{1}{d_{ij}} \tag{3.4}$$

where d_{ij} is the shortest path distance between i and j. E_{ideal} is the efficiency of the complete graph over the given set of nodes, where edge length is given by the distance between the node it links (for instance, euclidean distance). We will use E_{ideal} for normalization purposes.

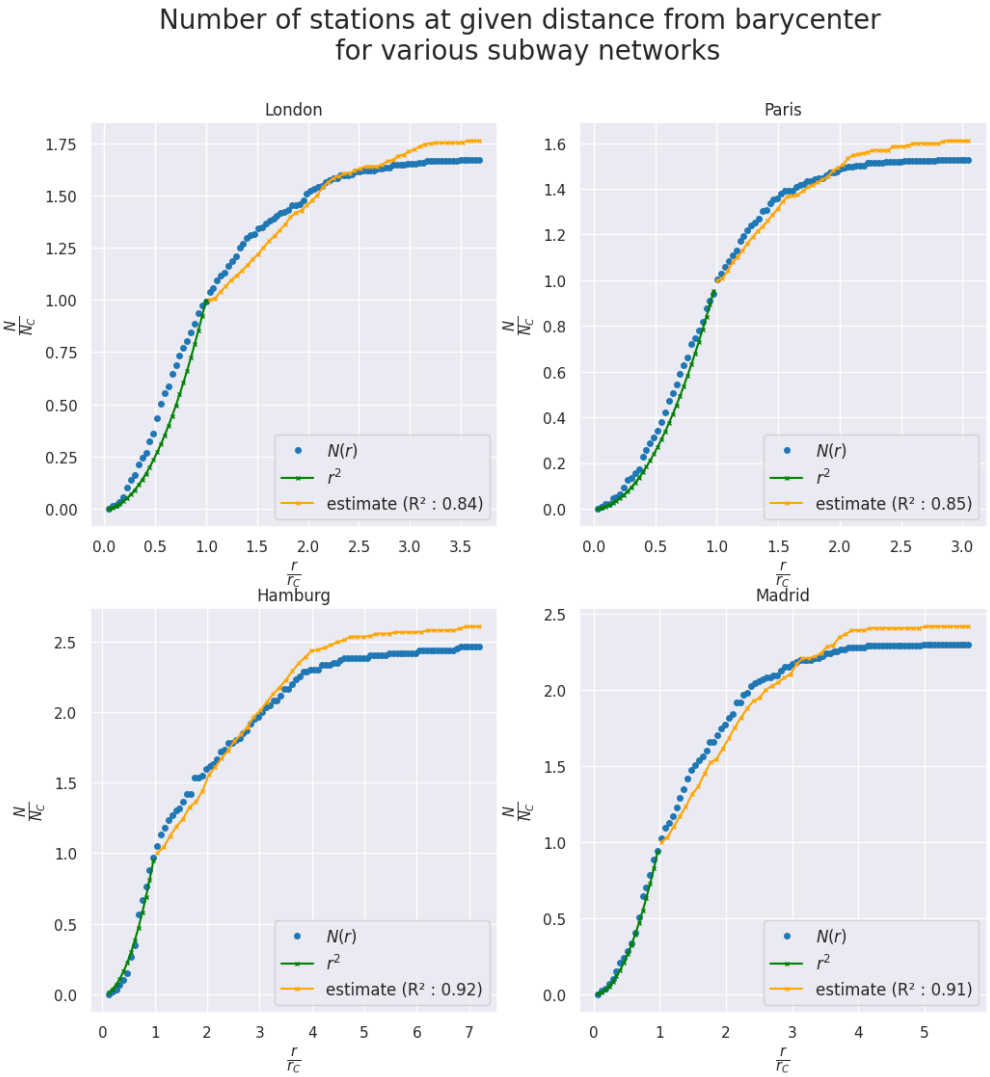

Figure 3.6: Radial cumulative distribution of nodes around the barycenter $N(r)$. Fit with first two regimes is shown: dense core over $[0, r_C]$ and sparse branches over $[r_C, r_m]$. The first regime corresponds to a uniform density of nodes in the core, that is a quadratic regime, the second rather fits with a power law of exponant $\tau < 1$, for instance for Paris, $\tau \approx 0.45$. The quality of the estimation depends of a multiplicity of factors, such as the quality of the estimation of the core radius, the closeness between the core and the barycenter, the isotropy of branches around the core, the proximity of the shape of branches with a line, the quality of estimation of the number of branches and the noise in the estimation of $\Delta(r)$.

We show in this section that this efficiency metric reflects the spatial structure of the network. We show in figure 3.8 the efficiency spatial distribution distribution. For all these networks, the efficiency decreases smoothly from a central point, located in the core of the network, typically from a maximum value slightly above 1.0, to value of efficiency on terminuses around 0.2 or 0.3.

In figure 3.7, we show the radial distribution of efficiency around the barycenter. For Hamburg, and London it seems there's a peak of radial efficiency between 0 and 1 : we can assume this is caused by the fact that the barycenter is slightly oustide the core, i.e the area of dense nodes, which should have peak efficiency, as shown in figure 3.8. Still, the radial distribution, fitted with a decreasing exponential, similarly to the BC in section 3.2.4. The quality of the fit is higher for the efficiency than it was for BC. This improved match with an exponential decay model can be explained by the fact that BC is more irregular in space than efficiency.

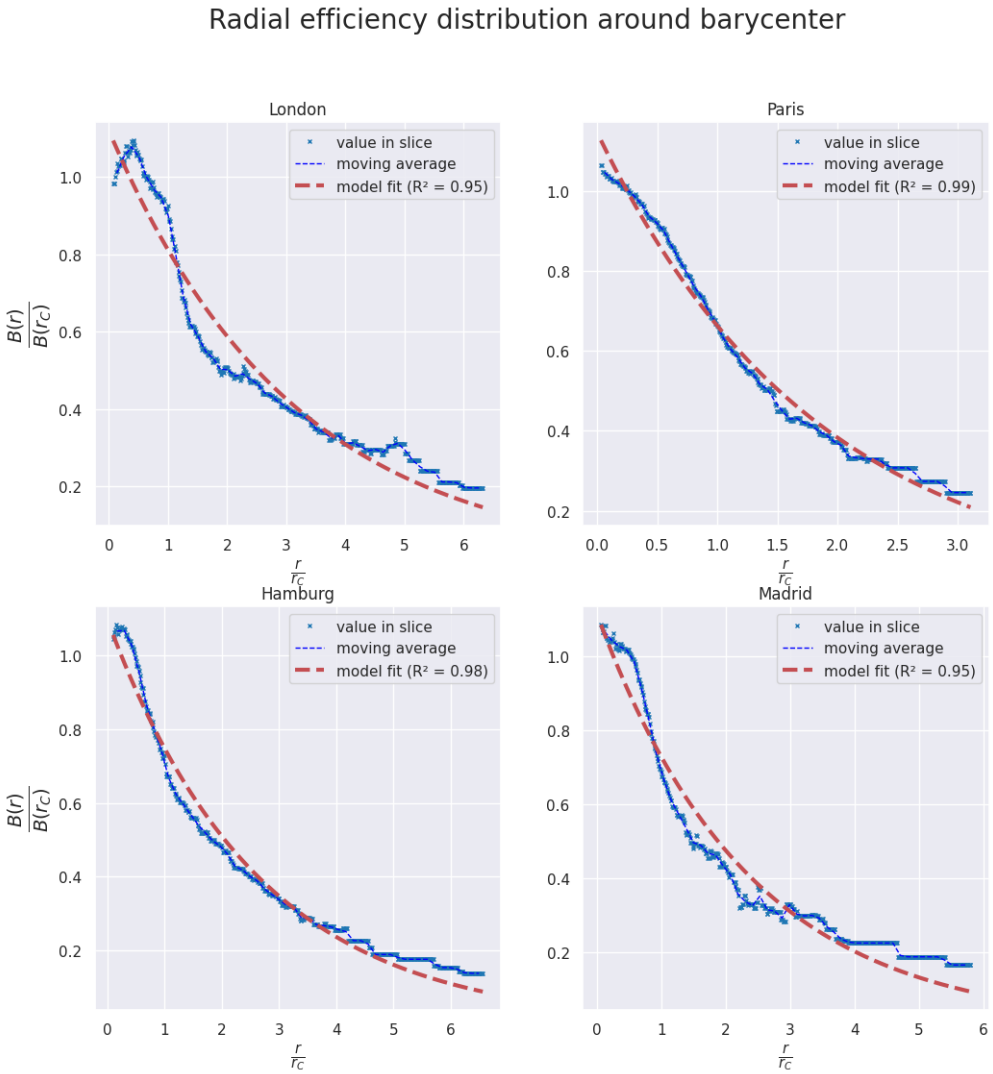

Figure 3.7: Radial distribution of efficiency. We show the moving average of this distribution with the blue line, and a fit with $ae^{-\frac{r}{r_0}}$ with the dashed green line. The fits seem precise , with $R^2 \geq 0.95$. This proves that radius and efficiency are strongly correlated.

3.3 Chapter 3: Overview and conclusion

In this chapter, we presented the theoretical framework we work with, providing an overview of the main results and conjectures that are made about subway networks. In particular, we described the decomposition of the networks in a dense circular core closed by a ring and branches radiating away, beginning along the ring and

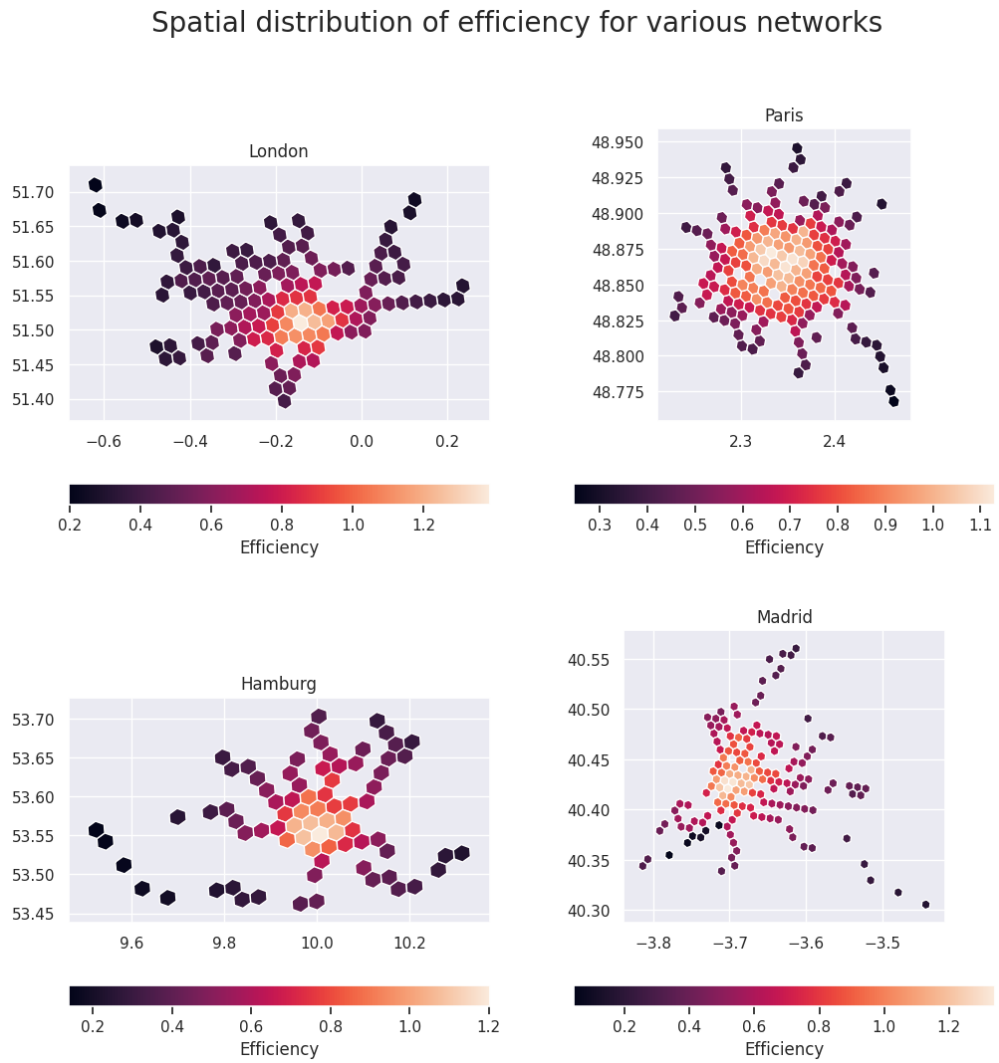

Figure 3.8: Heatmaps of spatial efficiency distribution for subway networks of four urban areas, London, Paris, Hamburg and Madrid. The tiling is restricted to areas where at least one metro station was found. Visually, the high efficiency areas correspond to a circular core around a peak efficiency point.

finishing at terminuses nodes. We briefly gave background on the study of spatial networks, especially within the study of meaningful spatial metrics - in our case, betweenness centrality and efficiency. Then, based on the dataset we extracted (see Chapter 2 for its description), we decided to reproduce these results and to test some of the hypotheses. We then presented the results of the analyses we performed, from the study the spatial efficiency of a network to the study of scaling properties.

In the next chapter we discuss limitations of the current set of hypotheses and introduce some novel analyses.

Chapter 4

A new definition for a functional barycenter

In this chapter we further discuss the analyses and results on subway networks that we performed in the previous chapter, and we point out some pitfalls that we discovered in those approaches and discuss some hypothesis that are made that do not match our results. we then propose a new definition of center of a network, using the spatial efficiency properties we have found. We discuss the possible alternatives that we considered, for instance through an analysis of the isotropy of the network. We then discuss these alternatives on our dataset. We then introduce new metrics to use in our analysis, and that are useful for future models that we develop. Finally, we show the improvement in both qualitative and quantitative ways, using several metrics to show it. In particular, we consider properties on radial, angular and spatial distributions. We examine the results of the scaling properties applied with this new definition of center. Finally, in the framework of the study of distribution of points of interests (POIs) in an urban area, we highlight the proximity of our newly defined center of a subway network and the peak of density of POIs.

4.1 Analysis of pitfalls of previous definitions

In this section, we point out a number of problems we encountered in the previous chapter. We showed in section 3.2.3 that the definition of center of the network did not reflect a point inside the area of dense nodes, which causes problems for analysis like radial distribution. In section 3.2.2, we noticed that the definition of core and branches given by the k-core decomposition algorithm gave us a decomposition that did not always fit correctly with our networks given the difference between actual structures of subway networks and this rather simple model. We noticed as

well that the core structure was not necessarily circular.

We show as well another pitfall of the barycenter : it does not reflect the angular node distribution around the core. Given figure 4.1, the isotropy hypothesis around the barycenter does not seem to hold. Given this fact, two questions are worth asking : how does the point of maximum isotropy behave ? instead of using the isotropy hypothesis, should we consider that the network develops differently in different directions ?

We will try addressing these problems in this chapter. In next section we propose new ways to compute a central point of the network.

4.2 Novel definition of functional barycenter

In this section, we propose several alternatives based on graph properties to the barycenter for finding a point that reflects the functional center of the city. As we showed in section 4.1, the barycenter does not always fall in the core of the network, which causes problems related to gathering spatial information about the network, for instance radial and angular distributions.

The problem with the barycenter is that nodes are not always uniformly distributed around the core, as we saw for Madrid and London's networks, the branches distribution can be anisotropic. Thus, the barycenter will tend to fall between branches, in the preferential direction where they develop. To solve this issue, we propose two mai approaches, that we describe in the next two sections. the first using a minimal anisotropy point as a functional center of the network, while the other using the network efficiency as a measure to identify the functional center. Finally, we compare the behavior of these centers.

4.2.1 The minimal anisotropy area

We first introduce the analytical definition of minimal anysotropy point. We define a minimal anisotropy point as a point x such as :

$$x \in \underset{p}{\operatorname{argmin}} \underset{\theta \sim \mathcal{U}(0,2\pi)}{Var} N_p(\theta)$$

where:

$$N_p(\theta) = |\{x_i : x_i \in \mathrm{Slice(p}, \theta, \mathrm{d}\theta)\}|$$

and:

$$\begin{aligned}&\text{Slice}(p = (x_p, y_p), \theta, d\theta) = \\ &\{(x, y) : \cos(\theta)(y - y_p) - \sin(\theta)(x - x_p) \geq 0 \\ &\text{ and } \cos(\theta + d\theta)(y - y_p) - \sin(\theta + d\theta)(x - x_p) \leq 0\}.\end{aligned}$$

Here $\mathcal{U}(0,2\pi)$ is the uniform distribution on $[0, 2\pi]$. We describe the procedure we used to compute an approximation of this point. We divided the convex hull of the nodes in the network in a grid, and computed the variance of the angular distribution of nodes around the center of the box, for each box in the grid. Then we find a more accurate estimate by repeating the same computation over the square areas found in the previous search, further divided into a grid of smaller square areas, and finally take the location that reaches the minimum variance.

4.2.2 Using the efficiency distribution

As we saw in section 3.2.6, the efficiency describes the spatial structure of the network quite well. In particular, as understood with figure 3.8, it reaches its peak in the core of the network. Figure 3.7 shows in the cases where the barycenter falls out of the core, a peak in radial efficiency distribution that is not exactly in 0 (e.g. London and Hamburg, where in both cases the barycenter is out of the core - is it not so distinct for Madrid where the barycenter is on the boundary) , which is what we would expect if the barycenter was indeed the point of peak efficiency.

Hence, our idea was to use this metric to find a functional center of the network. We came up with several ideas to extract a meaningful central point from the spatial distribution of efficiency over the nodes of the network.
The first idea we had was to smooth the spatial distribution of efficiency over the convex hull defined by the nodes - to avoid selecting the node of highest efficiency - and take a point with maximum value reached over this distribution. This method is actually too sensitive to noise and might select a point on the outer part of the core. To solve this issue, we used a solution balancing the top efficiency points and maximum isotropy in the core. Finally, we decided to go with a third option fitting a gaussian kernel to the efficiency distribution spatially expanded over the core. This fitting gives us a center of the distribution, which is our definition of functional barycenter of the network. This method has naturally also its limitations, in particular we do not expect it to function properly for cities with extreme anisotropy - meaning, network developed only on a part of the plane with respect to the center, such as Milan, New York City or Chicago, because the

spatial distribution of efficiency is skewed towards a direction and thus the gaussian fit is going to anticipate the other. Surprisingly, even given the high anisotropy of Chicago (third-most among our network, with a normalized variance of the angular distribution of nodes around the efficiency center over 1.0), the algorithm is still able to find its core, due to its structure with a very dense core of nodes of high efficiency.

4.2.3 Comparison of centers

We show on figure 4.2 and figure 4.3 the location of these various centers on a subset of graphs. We restricted the networks to subgraphs containing the core. For Paris, for which nodes are distributed quite close to isotropy, the three points are extremely close. For cities with anistropy such as Madrid, London and New York City, the barycenter and the minimal anisotropy point are extremely close and do not achieve in locating the core. In fact, the barycenter and the minimal anisotropy point are always very close. Instead, for London and Madrid, the center of the efficiency gaussian fit finds the center of the core, even if fore Madrid this core is not circular. For New York City, which is extremely anisotropic, the center of the gaussian falls just short of the area of high node density. Since in New York City, it is quite difficult to find a core and there's no really any regular structure, this point is much closer to the area of high density of nodes than the two others possible centers, that fall in the middle of branches. For Paris, the center of the gaussian fit falls in the same spot as the other possible centers, indicating coherence between the three definitions when the network follows the structure of circular core and branches distributed isotropically around the core. Qualitatively, without any hypothesis on the distribution of nodes of the network in space, we could not expect that the barycenter would reflect the properties of "centrality" of the network : the barycenter only has information about the spatial layout of nodes (the point of minimal anisotropy too), none about the topology of the network. Instead, through the fit of the two-dimensional gaussian to the efficiency distribution, we take both account of the spatial layout of the network (through the gaussian fit) and of the topology of the network (through the efficiency distribution).

The study of the proximity of the barycenter and the set of points of minimal anisotropy of points in space could be subject to more study - we did not find any work on this topic.

4.3 Introducing novel metrics

After the definition of the function center, we are interested in extracting information about angular data of subway networks. As the authors of [8] mention, cities tend to develop around fundamental structures such as rivers, hills and ancient streets.

Angular distributions helps grasping a better understanding of how a city behaves, as a complex system, and getting insight about the structure of the subway network around the center of an urban area could help develop a model about the spatial coupling of transportation layers, such as studied in [11]. Understanding an urban area transportation system as a superposition of transportation layers, preferential directions for each layer could be a feature for better comprehension of this complex system.

4.3.1 The angular distribution

Given a point in space p, an angle portion $d\theta$ and a subway network, we aim at computing angular distribution of metrics around p. To do so, we sample the interval $[0, 2\pi]$ uniformly in n points $\{\theta_1, \dots, \theta_n\}$. For each value of θ_i, and for a given metric, we compute the set of nodes located in $Slice(p, \theta, d\theta)$. Depending on the metric, we either aggregate with a sum (for instance, for the number of nodes), or with an average (for instance, for metrics such as betweenness centrality or efficiency). If the set is empty, we have two choices : if we are working with a cumulative metric, such as the number of nodes, we can the value of the angular distribution for this slice to 0. For metrics for which we take the average, it is better to set a NaN value for this slice, since a 0 value would add to the variance of the distribution, and other values (for instance, the mean over non-Nan values) would bias the distribution towards the value, and possibly reduce the variance of the distribution. This allows us to compute the distribution such that its statistics are computed only over areas that contain nodes.

The value of $d\theta$ shapes the distribution, for a large $d\theta$, we compute a sort of average mean over the distribution : it gives an idea of the overall shape of the network around p. The smaller $d\theta$, the closer we are to computing the discrete distribution of nodes around p (i.e, a slice has value 0 if it does not contain any node, else it would tend to the number of nodes that are exactly aligned and aligned with p, in the given slice). This feature is helpful to grasp the distribution of branches around the core : instead of only having the number of branches $\mathcal{N}_B$, this would give an approximation of the proximity of branches, their direction and their length.

In figure 4.4 and figure 4.5, we show the angular node distribution around the efficiency center for four networks of very different shapes, and the angular efficiency distribution around the efficiency center. The former distributions reflect the shape of the network relatively to the center : for Madrid, the north-west is empty, while the east where the branches develop has high number of nodes ; for London, the south-east is empty ; for New-York, the west part of the network is empty and the distribution reflects the branches going in north-east and south-east directions. For Paris, that has a rather isotropic structure, the distribution is close to the average all around. The second plot 4.5, shows that the angular distribution of efficiency around the efficiency peak is rather evenly distributed. For London and Madrid, we see that higher values are reached in the empty directions, which is expected : since these directions do not have branches, that are composed of low-efficiency nodes, the average in this direction only includes nodes from the core, which instead have higher values.

4.4 Improving metrics from the perspective of the novel barycenter

The purpose of this section is to detail the various improvements made possible by the introduction of the novel center definition, specifically in regard of the use of the efficiency distribution. To begin with, we introduce a new definition of the core of the network that is based on efficiency. This novel definition of the network's core takes into account the efficiency of nodes, both spatially and topologically, and provides an a better approximation of the network's core than the approximation given by the 2-core algorithm, which only accounts for topological properties.

Next, we demonstrate how this new definition of the network's center captures more precisely the core of the subway network, which is determined by the network's proximity to efficient nodes. By incorporating this improved definition of the core, we can analyze in a different way the network's scaling feature - that are the radial distribution of nodes around the core, and more specifically the center. To that end, we describe the reproduction of the scaling property experiment around the new center, while taking into account the definition of the core that we propose.

By using the efficiency distribution and the novel center definition, we are able to improve the accuracy of our analysis of the network's properties. We believe that these improvements could help with the analysis of subway networks and more generally with the understanding of the complex systems that transportation

networks over urban areas are. Overall, this section highlights the advantages of our approach and is the foundation of our future works.

4.4.1 Efficiency distribution

We propose a new definition of the core based on the spatial efficiency. Given a network and the efficiency distribution over its nodes E_i, the distribution of efficiency peaks over the core and smoothly decreases over the borders of the core and the branches, as illustrated in figure 3.8. Actually, we show in figure4.6 that the radial distribution of efficiency around the efficiency center is very accurately described by

$$f(r; a, e, r_0) = a\exp(-(r/r0)^e)$$

with $e < 1$.

Thus, we try to provide an other approximation of the core structure of the network based on the efficiency distribution. To do so, we found experimentally that taking the nodes with efficiency above a threshold equal to $\frac{\min(E_i)+\max(E_i)}{2}$ gave a good approximation of the core structure of the network. We show the result of this algorithm on figure 4.7.

4.4.2 Proximity to essential nodes

In this section, we show how the definition of the efficiency center captures in a better way the core structure of the network - that is, achieves to fall closer to most "efficient" nodes. We show on figure 4.8 and figure 4.9 the cumulative distribution of top q% nodes, for BC and efficiency values, in function of radius from barycenter and efficiency center, for values of q in [10,20,40,60,80,100]. The purpose of this is to show the increase of proximity to important nodes gained by the efficiency center comparatively to the barycenter, using both the betweenness centrality that measures the topological importance of a node, while the efficiency measures the capacity in accessing the network from a given node.

4.4.3 Scaling properties

In order to capture better the radial number of nodes distribution around the efficiency center, we tried a new method of computation. Rather than using the barycenter as a reference point, we chose to use the efficiency center. To do this, we employed the same model as described in section 3.2.5, utilizing equation 3.3. However, instead of relying on the core definition provided in existing literature,

we implemented the core definition proposed in section 4.4.1. Our new definition of the core tends to yield a denser structure of nodes, including fewer nodes from the branches. This is illustrated in figure 4.7, which displays the core of the London and Madrid networks compared to the previous method (figure 3.6). Notably, the branches in the new core definition are much longer than in the previous version. Using this updated approach, we obtained a distribution of the radial number of nodes around the efficiency center, which is shown in figure 4.10.

4.5 Correlation with amenities distribution

In our research, we were interested in examining the correlation between a city's network and the distribution of amenities across its urban area. Specifically, we wanted to explore the proximity between the peak of the amenities distribution and the efficiency center of the network. To accomplish this, we used a spatial distribution of POIs, which consists of a collection of points with coordinates. In order to estimate the local density of POIs across the urban area, we divided the area into regular subareas and aggregated the number of POIs within each subarea to count their number. Next, we fitted a gaussian kernel to this estimated density, resulting in an estimation of the center of the distribution. We refer to this center as the peak of amenities.

We illustrate the proximity of the efficiency center and the peak of amenities in figure 4.11, using a subset of networks. We found that in cities with anisotropic distribution of nodes around the core of the network, the peak of amenities tends to lie slightly towards the area with low branch density. This can be explained by the fact that the distribution of POIs contains a larger number of points, spread throughout the urban area, which is not the case for metro networks.

Figure 4.12 displays the density of the distribution of POIs across the urban area of various cities. The distribution has a very high value peak in the center and decreases sharply as soon as a point is further away from it. While the efficiency distribution remains quite high on the ring around the core, the density of POIs may have already reached its lower values, as seen in the case of Madrid. Moreover, it is important to note that the distribution of amenities is defined across the whole urban area, meaning that the peak of the gaussian fit is computed on a different area than the efficiency peak.

4.6 Chapter 4: Overview and conclusion

In Chapter 3, we presented a method to decompose a network into its core and peripheral regions based on the network's topology. While this method provided a relatively simple and easy-to-understand approach for defining the network's decomposition, we identified several issues with it in this chapter. One of the issues we found was that the method failed to capture some essential features of the network, such as the presence of non-central nodes in the approximation of the core. Another issue we identified was the use of barycenter as the center of the network, which may not always be the most representative point of the network.

To address these issues, in this chapter we have proposed a novel definition of the core of a network. The core is defined as the peak of the spatial efficiency distribution. This definition outperforms the barycenter at representing spatial essential nodes of the network, represented by high topology and spatial performances.

Moreover, we demonstrate how this novel definition of the center of the network provides a new understanding of the network when analyzed through it. We show that the core nodes of the network tend to be located in areas with high economic activity, as the peak of POIs and the efficiency center lie close to each other. This finding indicates that the core of the network plays a crucial role in shaping the spatial distribution of amenities and economic activities in urban areas and vice-versa.

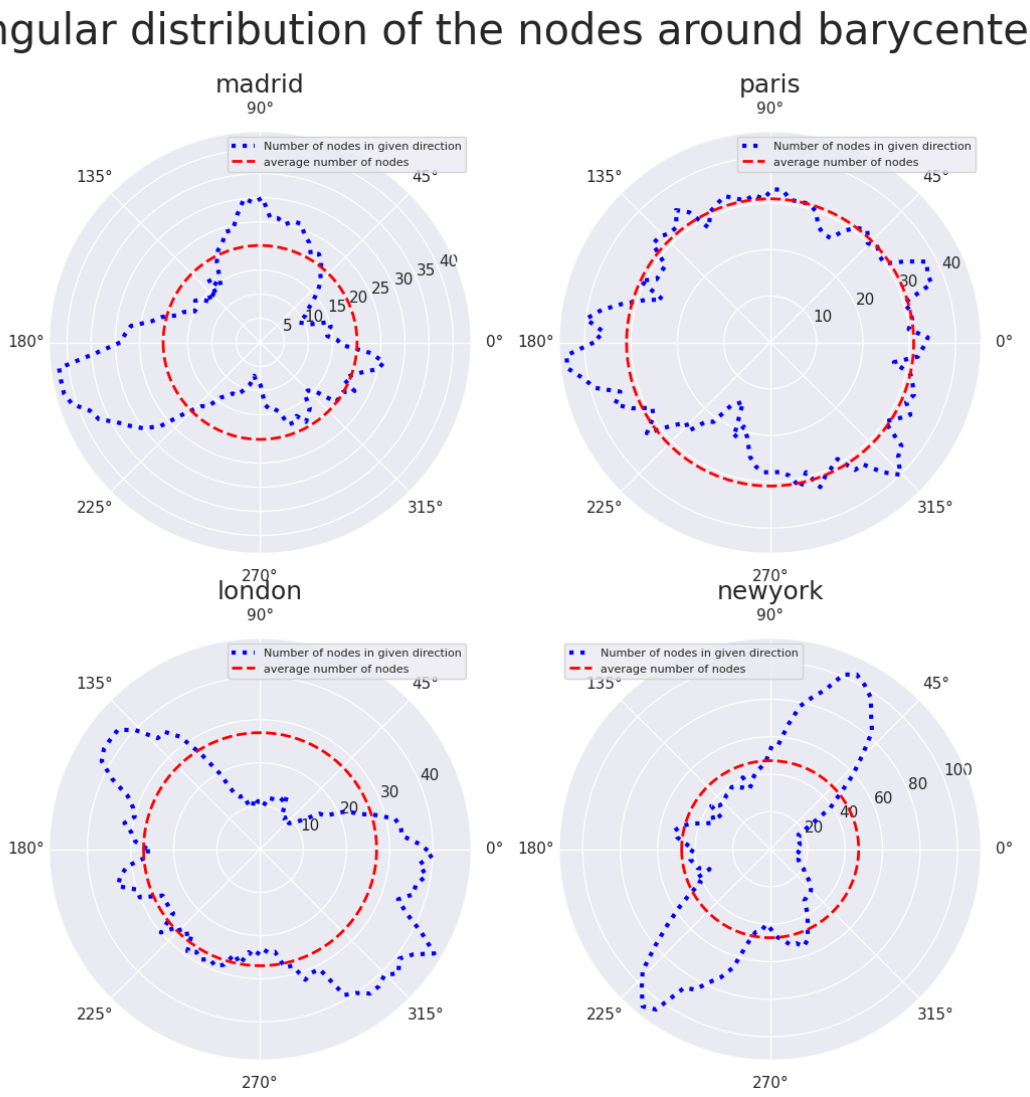

Figure 4.1: Angular distribution of nodes around barycenter for Madrid, Paris, London, New York City. We notice that Paris' network nodes distribution is close to isotropic around its barycenter. Instead, London, Madrid and especially New York City's subway networks display a high anisotropy. This anisotropy can be explained by geographical characteristics in the city, for instance for New York City, the Hudson and the East river lay both sides of center, causing this diagonal development relatively to the barycenter (which is the direction of both rivers). The result for both Madrid and London does not reflect the nodes distribution around the actual center of the city, because of the bias induced by the location of the barycenter outside of the core.

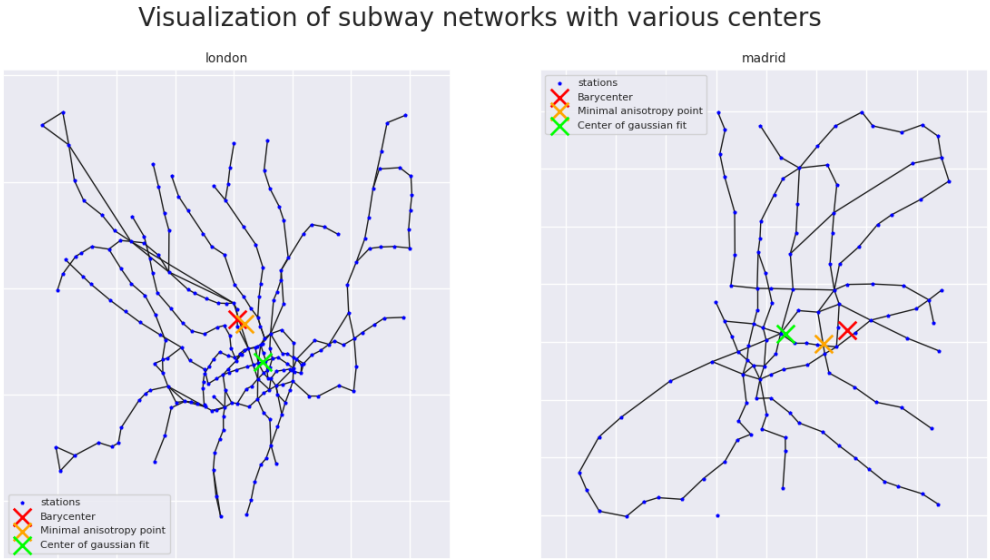

Figure 4.2: Display of center of London and Madrid subway networks, with location of barycenter, minimal anisotropy point and center of gaussian fit. For London, while the barycenter and minimal anisotropy point are located outside of the core, inbetween branches, the center of the gaussian fit lies inside the central loop. The result is similar for Madrid, where the barycenter and minimal anisotropy point lie on the border of the center, where instead the center of gaussian fit is located right in the middle of the central loop.

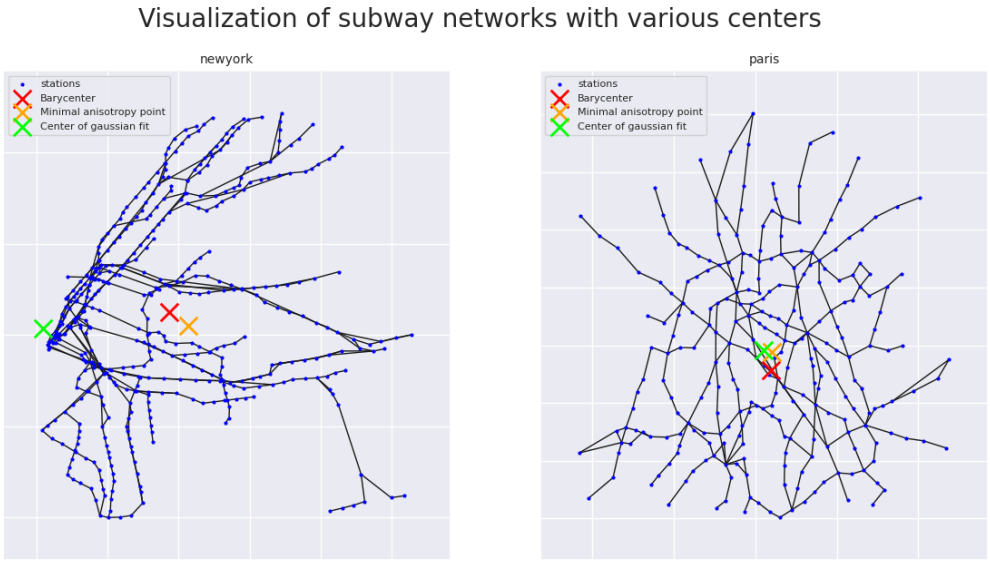

Figure 4.3: Display of center of Paris and New-York subway networks, with location of barycenter, minimal anisotropy point and center of gaussian fit. For Paris, given the isotropy of the network, the three points lie close to each other. For New-York, the barycenter and minimal anisotropy point lie inbetween branches outside of the center. The center of the gaussian fit instead, is located extremely close to the denser part of the network.

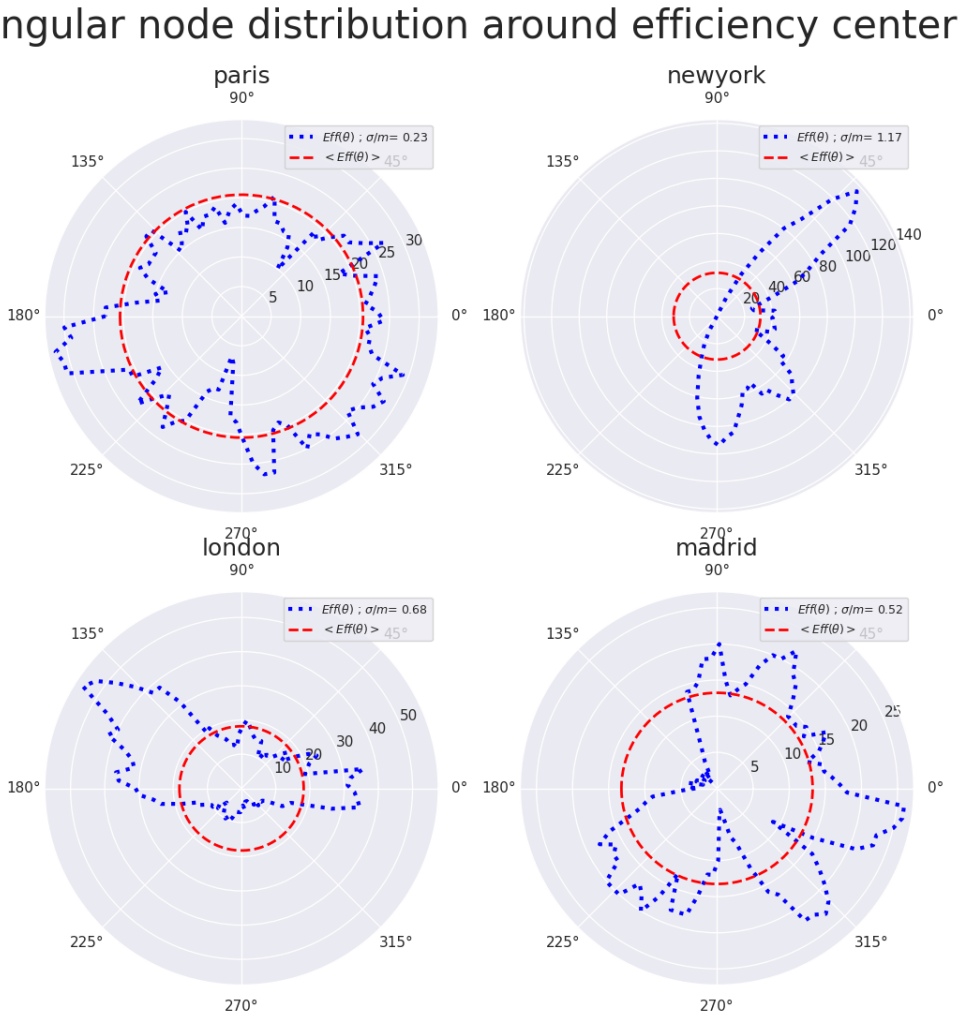

Figure 4.4: Angular node distribution of nodes around efficiency center. The directions where the distribution reach higher values reflect where the branches tend to develop more - for isotropic cities, with branches distributed uniformly all around, there is no particular peak. Instead, for London and Madrid, the preferential directions of the branches reflect through the distribution. For New York City, with extremely high anisotropy

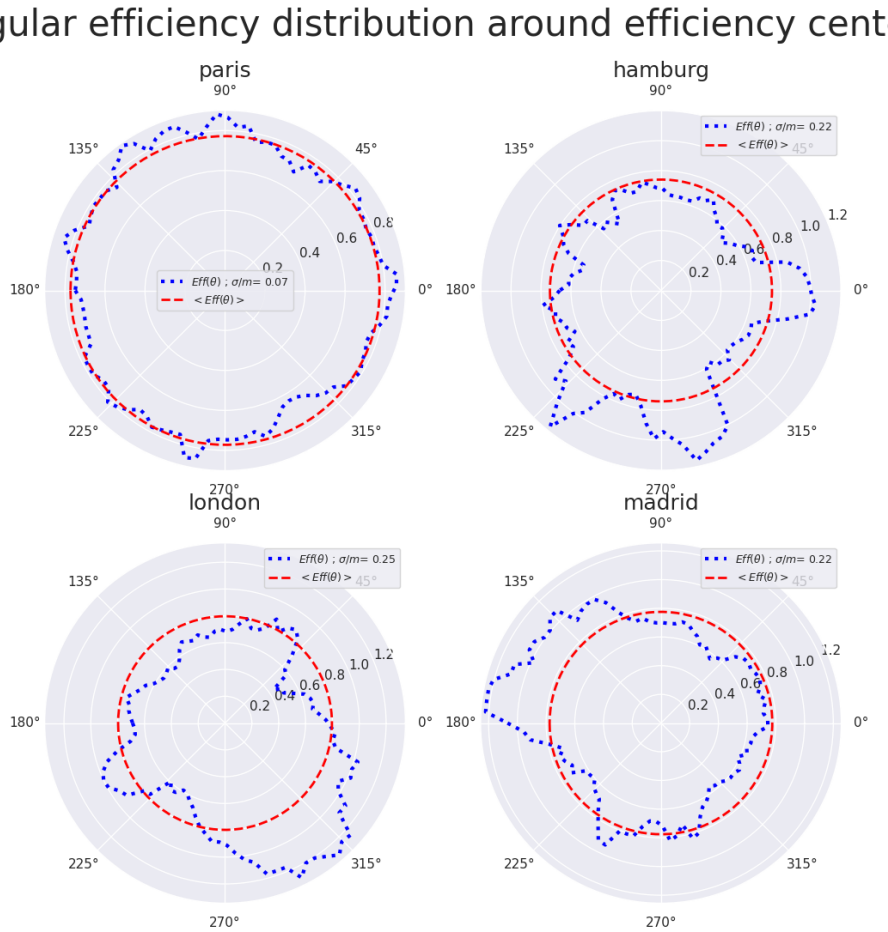

Figure 4.5: Angular efficiency distribution around efficiency center for Paris, Hamburg, London and Madrid's networks. The normalized standard deviation $\frac{\sigma}{\mu}$is lower than the one for the number of nodes angular distribution. For networks with irregular distribution of branches around the core, the area of low density have larger efficiency, since there are no nodes from branches who have lower values in the slices. For an isotropic network, such as Paris, the efficiency angular distribution is extremely close to an isotropic distribution.

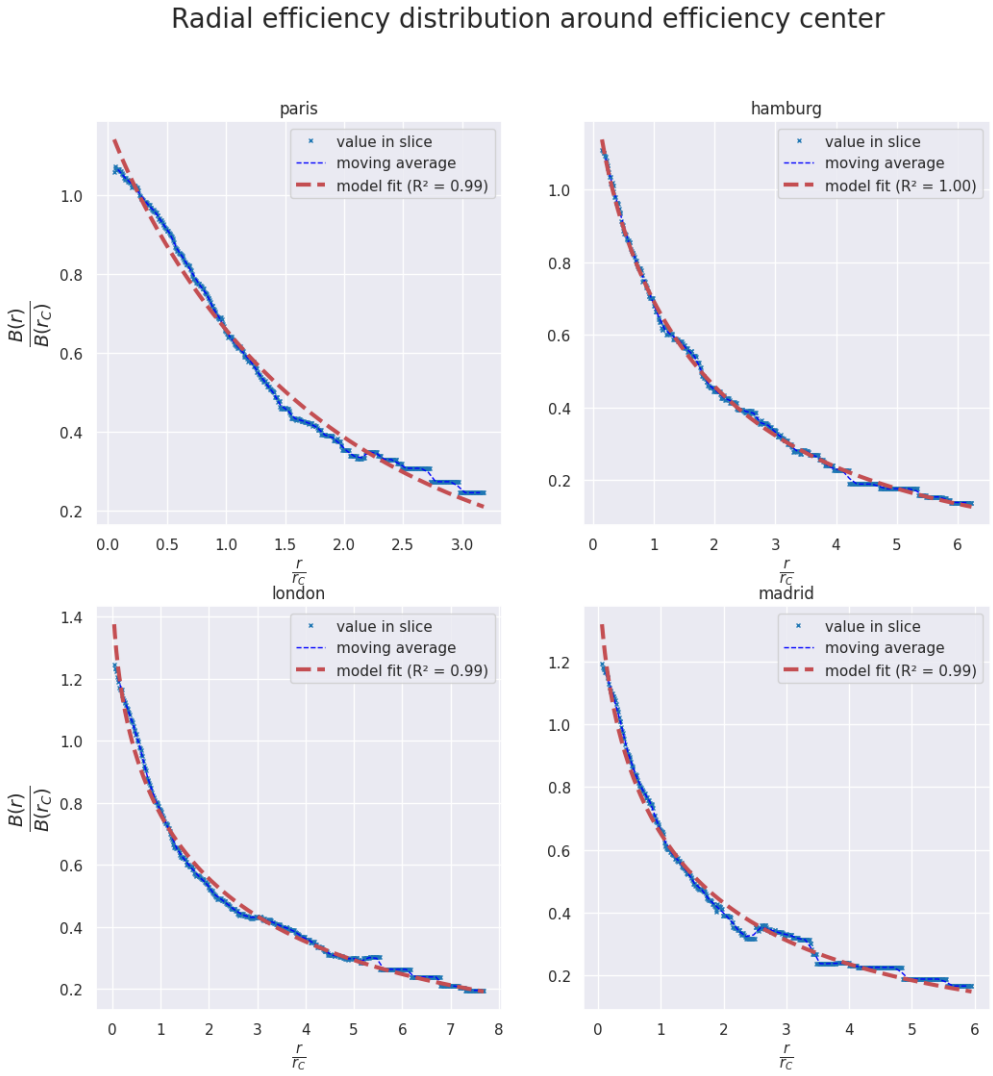

Figure 4.6: Radial efficiency distribution around the efficiency center. This distribution is fit to a law of type $f(r; a, e, r_0) = a\exp(-(r/r0)^e)$ with $e < 1$. The radial efficiency distribution around the efficiency center behaves smoothly and follows well this distribution. In particular, the peak of the distribution is reached in 0, contrarily to the efficiency distribution around the barycenter, such as shown in figure 3.7.

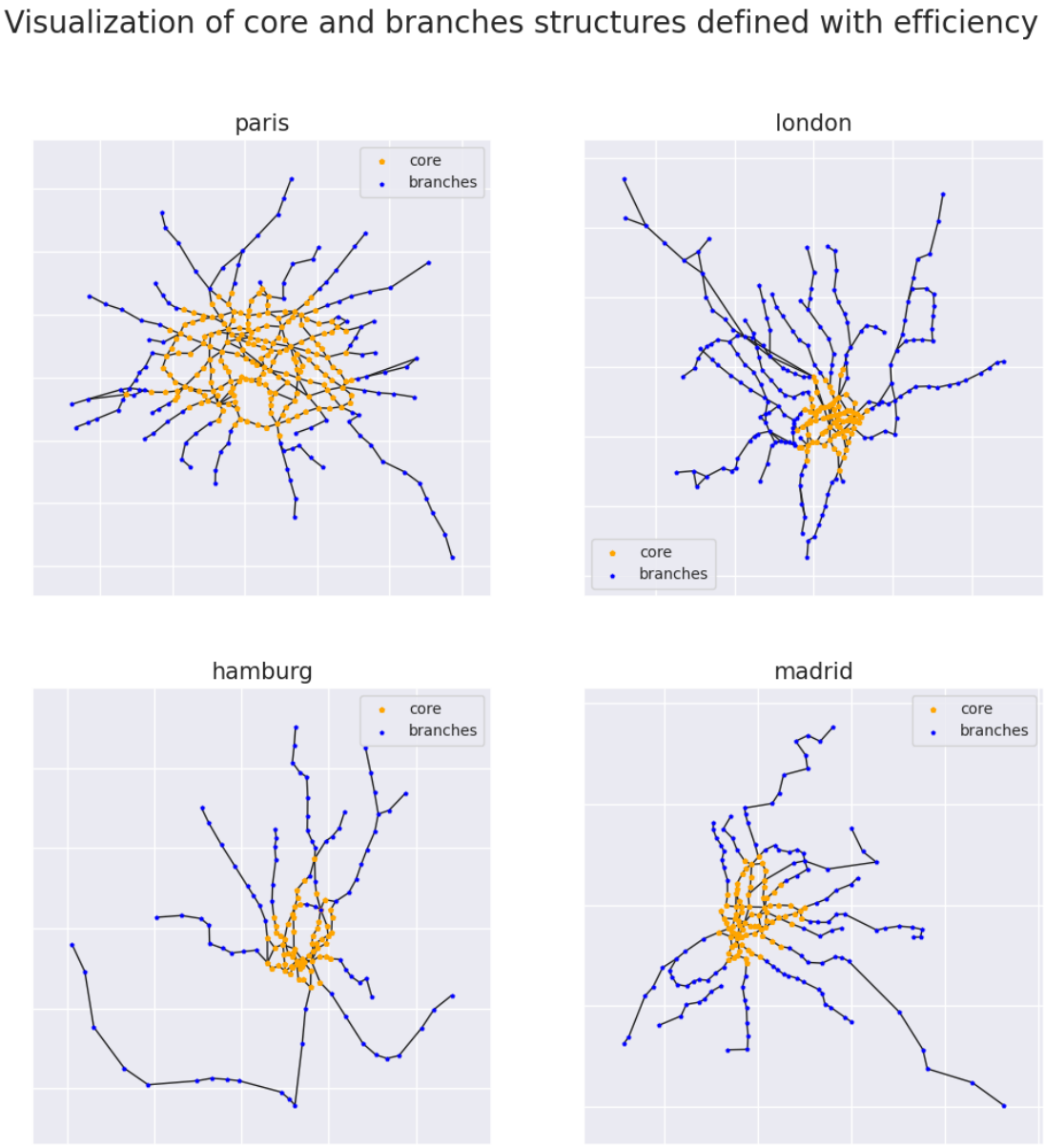

Figure 4.7: Approximation of the core given by the selection of the nodes with efficiency value above $\frac{\mathbf{min}(E_i)+\mathbf{max}(E_i)}{2}$. This procedure gives a smaller core, taking less nodes on the branches, and is able to capture more complex features than the 2-core decomposition of the network. Nevertheless, computing the core with this method gives difficulty in computing other metrics related to the network properties, such as the number of branches. We saw in section 3.2.2 that the core was not necessarily circular.

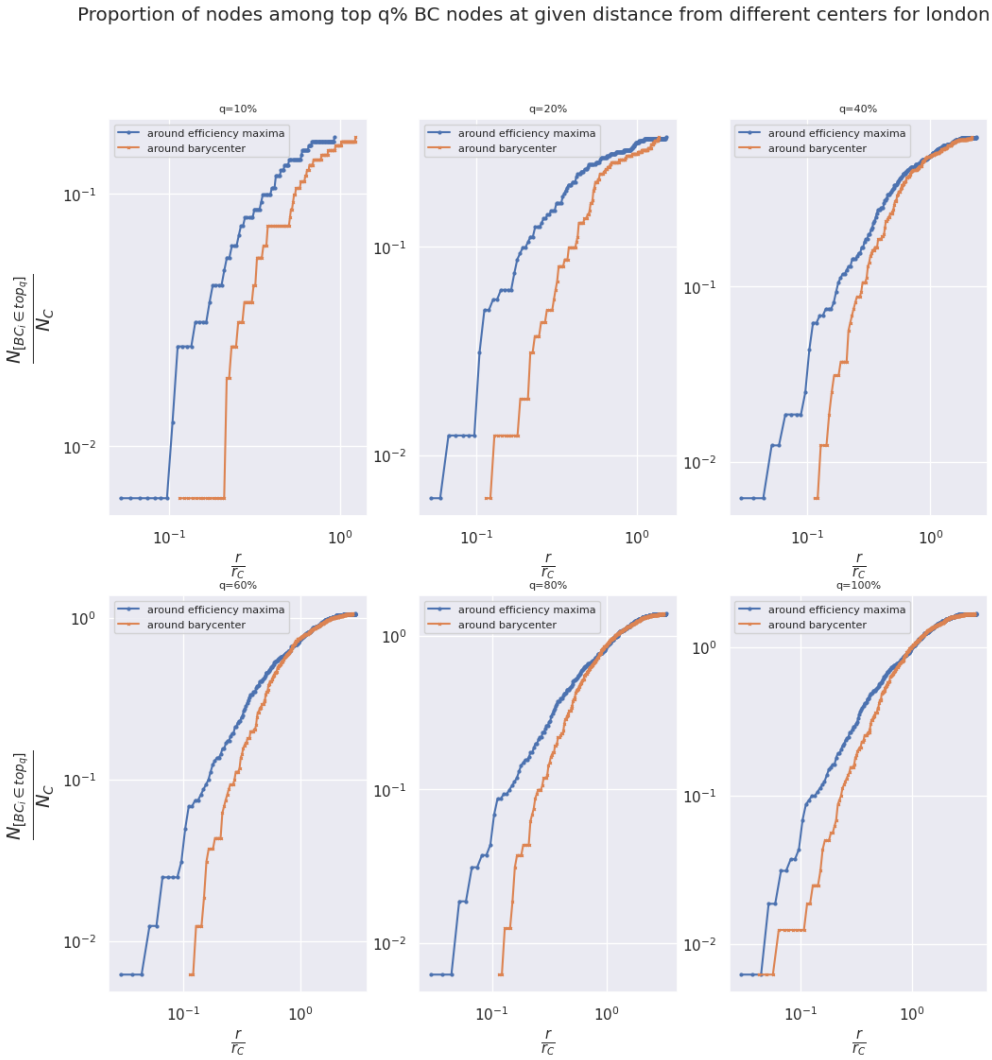

Figure 4.8: Layout of high BC nodes at distance from barycenter and efficiency maxima for London subway network. The y-axis shows the proportion of top-q% BC nodes at a given distance. From left to right and top to bottom, the proportions, of BC nodes taken are 10%, 20%, 40%, 60%, 80% and all nodes. Distance to barycenter is normalized by core radius, while y-axis is normalized with number of nodes in the core. It is noticeable that high efficiency nodes seem to layout closer to the efficiency maxima. The plot with all nodes corresponds to the distribution of number of nodes in increasing size disks.

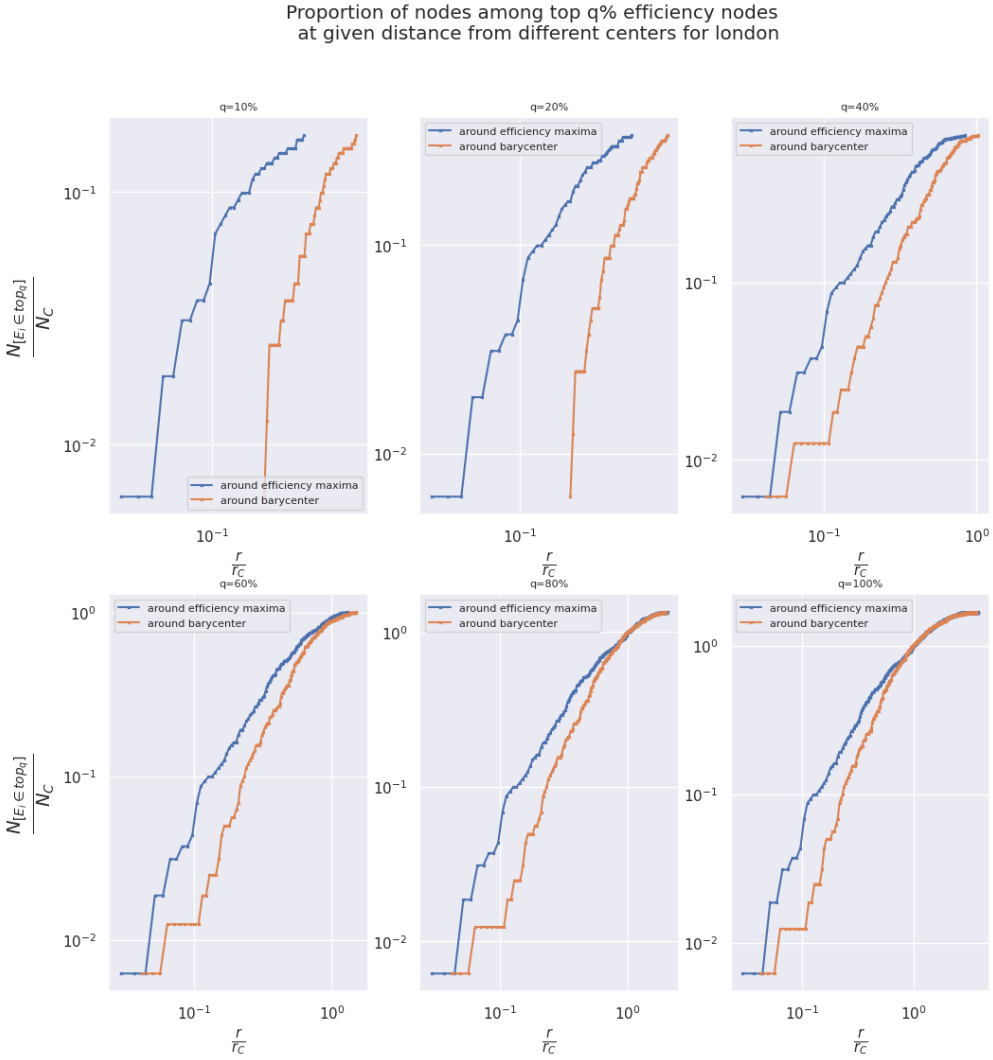

Figure 4.9: Proportion of top efficiency nodes at distance from barycenter and center of spatial efficiency gaussian fit for London subway network. The betweenness centrality and the efficiency do not give nodes the same importance : even if they are correlated, high BC nodes could be low efficiency nodes. The gain given by efficiency with efficiency center compared to barycenter in the top nodes ($q = 10\%, 20\%$) is more important for efficiency than for betweenness centrality. This is explainable by the fact that high efficiency nodes tend to be in the center of the core, while high BC nodes could lay on the ring.

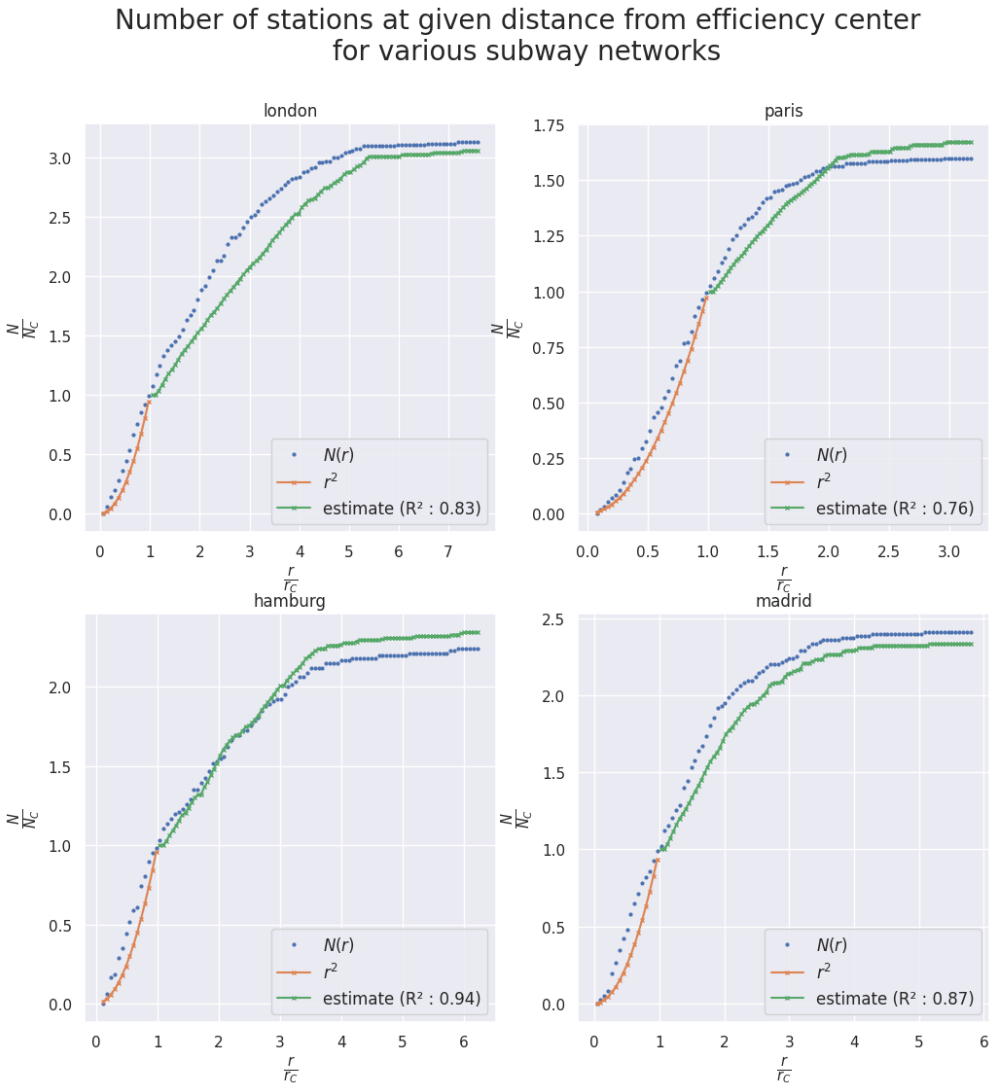

Figure 4.10: Radial number of nodes distribution for London, Paris, Hamburg and Madrid subway networks. We used the novel definition of core, proposed in section 4.4.1. Thus, the branches are longer than the ones shown in figure 3.6, computed with the 2-core decomposition, for Madrid and London. Given the pitfalls due to the simplicity of the fitting model, we do not expect the accuracy of the fitting model to improve.

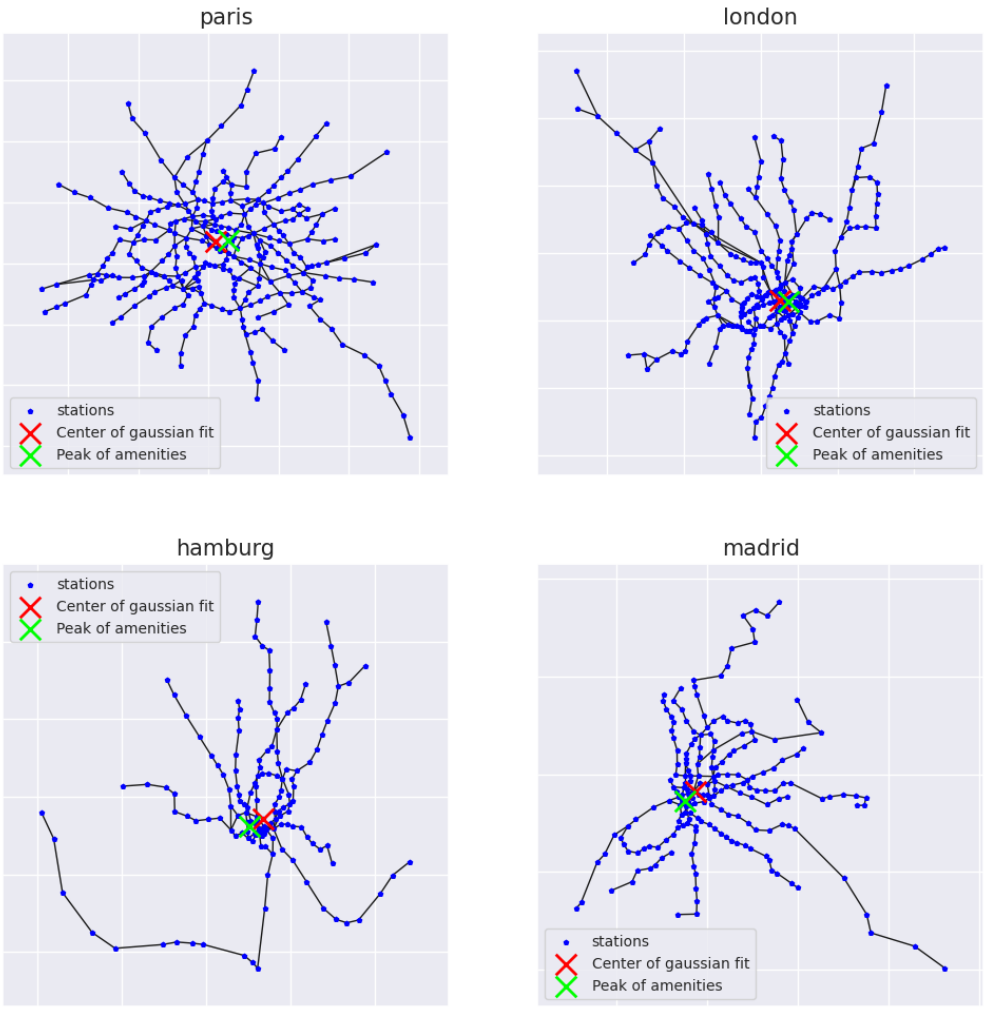

Figure 4.11: Display of the networks with peak amenities point and efficiency center. For the cities displayed, the peak are at proximity. We notice that the POIs peak lies slightly towards area with low number of nodes, compared to the efficiency peak. This is explained by the fact that the distribution of POIs is much more spread over the urban area, and thus the center of its distribution does not have an area of empty values, as it happens with the computation of the efficiency center (in this case, the areas where no node are found do not provide any data for the kernel fit, and thus the gaussian center is skewed in the preferential direction of the network, relatively to the peak of amenities). We observe this for the three cities with anisotropic behavior around the core, that are Madrid, London and Hamburg.

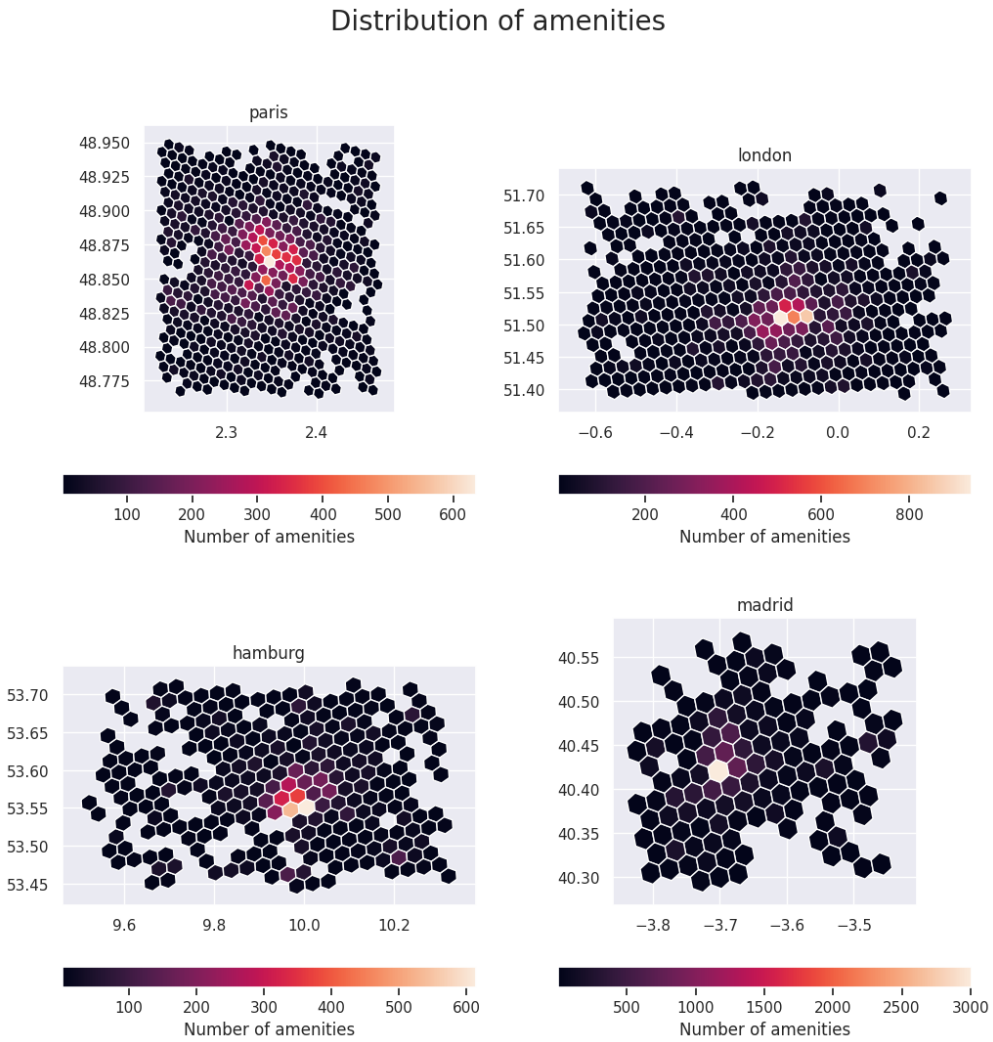

Figure 4.12: Density of amenites on a bounding box over the urban area of Paris, London, Hamburg and Madrid. The amenities distribution is more steeply peaked than the efficiency distribution, which is also due to a natural bias in the data gathering process (the data tends to be gathered more in the center of a city than in the surroundings). The area of high density of amenities correspond to the area of high efficiency : the center of the core of the network.

Chapter 5

Conclusion

In this thesis, we proposed a novel analysis of subway networks as part of urban complex systems. In the perspective of spatial networks, we used structural properties of graphs, combining their geometry and topology to propose a new understanding of subway systems' generic features. We highlight its high efficiency core structure around a functional center and the layout of the network around this point, we expose this property from radial and angular points of view, as well as the coincidence between the core of the network and the most active part of the city. We now give a summary of the different steps of the work, before outlining future works and generalizations.

In chapter 2, we presented the first part of our work consisting of the exploration of open-access data, in the understanding of the different constraints and in the specification of the dataset, and finally in the practical construction of the graph representation of the subway network that we chose to study. We described several sources that we explored and listed their comparative advantages and pitfalls. We explained which features of subway networks we wished to study and which constraints they implied. Finally, we explained in detail the extraction process that we used and mentioned the use and extraction of the amenities data.

Then, in chapter 3, we provided a summary about some results about the study of subway systems from a spatial network point of view. In particular, we described the decomposition in core and branches structures, and gave the definition of several essential metrics. Then, we reproduced literature results on our data, starting with a list of indicators about the topology and the geometry of networks. The network can be understood as a generic spatial graph, decomposed in a dense structure in the center surrounded by radiating branches, or as a geometric structure developing around a point, for instance its barycenter, as we studied in the following chapter. Finally, we outlined how the efficiency gave a precise understanding of the spatial

structure of the network.

The last chapter, 4, starts with an analysis of the results achieved in the previous one. We highlighted several issues in the simplicity of the definition of the decomposition of the network, as well as pitfalls in the use of barycenter as the center of the network. We then provided a novel definition of the core of a network, with a comparison with different alternatives. This definition of the functional center, and the perspective on structural metrics of the network that it introduces, allows a novel understanding of the subway system and the transportation network in the urban space. Finally, we showed the proximity between the center of this network, deduced solely from structural properties , and the peak of amenities over urban areas.

Future developments of this work arise from the extension of this study to a larger set of networks that could be in several directions. For example on a larger set of subway networks, that could be achieved by accessing the numerous Chinese cities' subway systems, or on other types of urban systems networks, such as multi-layer transportation systems, by combining metros and buses transportation layers. Another development would be the use of the information given by the network through the perspective of the efficiency center in the framework of the study of larger complex systems, such as the study of flows in human transportation.

Bibliography

[1] Bruno Osório, Nick McCullen, Ian Walker, and David Coley. «Integrating the energy costs of urban transport and buildings». In: *Sustainable cities and society* 32 (2017), pp. 669–681 (cit. on p. 2).

[2] Rémi Louf, Camille Roth, and Marc Barthelemy. «Scaling in transportation networks». In: *PLoS One* 9.7 (2014), e102007 (cit. on pp. 2, 3).

[3] Mark Newman. *Networks*. Oxford university press, 2018 (cit. on p. 4).

[4] Marc Barthélemy. «Spatial networks». In: *Physics reports* 499.1-3 (2011), pp. 1–101 (cit. on pp. 4, 5, 7, 8, 24–27).

[5] Pascal Klamser et al. «Enhancing global preparedness during an ongoing pandemic from partial and noisy data». In: *medRxiv* (2022), pp. 2022–08 (cit. on p. 4).

[6] Jesper Dall and Michael Christensen. «Random geometric graphs». In: *Physical review E* 66.1 (2002), p. 016121 (cit. on p. 7).

[7] Bernard M Waxman. «Routing of multipoint connections». In: *IEEE journal on selected areas in communications* 6.9 (1988), pp. 1617–1622 (cit. on p. 7).

[8] T Courtat, C Gloaguen, and S Douady. «Mathematics and morphogenesis of the city, a geometrical approach. 12». In: (2010) (cit. on pp. 7, 44).

[9] G Nemeth and G Vattay. «Giant clusters in random ad hoc networks». In: *Physical Review E* 67.3 (2003), p. 036110 (cit. on p. 8).

[10] Camille Roth, Soong Moon Kang, Michael Batty, and Marc Barthelemy. «A long-time limit for world subway networks». In: *Journal of The Royal Society Interface* 9.75 (2012), pp. 2540–2550 (cit. on pp. 10, 21–23, 26, 27, 32, 34).

[11] Richard G Morris and Marc Barthelemy. «Transport on coupled spatial networks». In: *Physical review letters* 109.12 (2012), p. 128703 (cit. on pp. 11, 44).

[12] Riccardo Gallotti and Marc Barthelemy. «Anatomy and efficiency of urban multimodal mobility». In: *Scientific reports* 4.1 (2014), p. 6911 (cit. on p. 11).

[13] Aihui Pei, Feng Xiao, Senbin Yu, and Lili Li. «Efficiency in the evolution of metro networks». In: *Scientific Reports* 12.1 (2022), p. 8326 (cit. on pp. 21, 22, 24, 34).

[14] Stephen B Seidman. «Network structure and minimum degree». In: *Social networks* 5.3 (1983), pp. 269–287 (cit. on p. 22).

[15] Lucien Benguigui and M1 Daoud. «Is the suburban railway system a fractal?» In: *Geographical Analysis* 23.4 (1991), pp. 362–368 (cit. on p. 22).

[16] Vito Latora and Massimo Marchiori. «Is the Boston subway a small-world network?» In: *Physica A: Statistical Mechanics and its Applications* 314.1-4 (2002), pp. 109–113 (cit. on pp. 22, 24).

[17] Marc Barthelemy. «Betweenness centrality in large complex networks». In: *The European physical journal B* 38.2 (2004), pp. 163–168 (cit. on p. 24).

[18] Alexander P Giles, Orestis Georgiou, and Carl P Dettmann. «Betweenness centrality in dense random geometric networks». In: *2015 IEEE International Conference on Communications (ICC)*. IEEE. 2015, pp. 6450–6455 (cit. on p. 24).