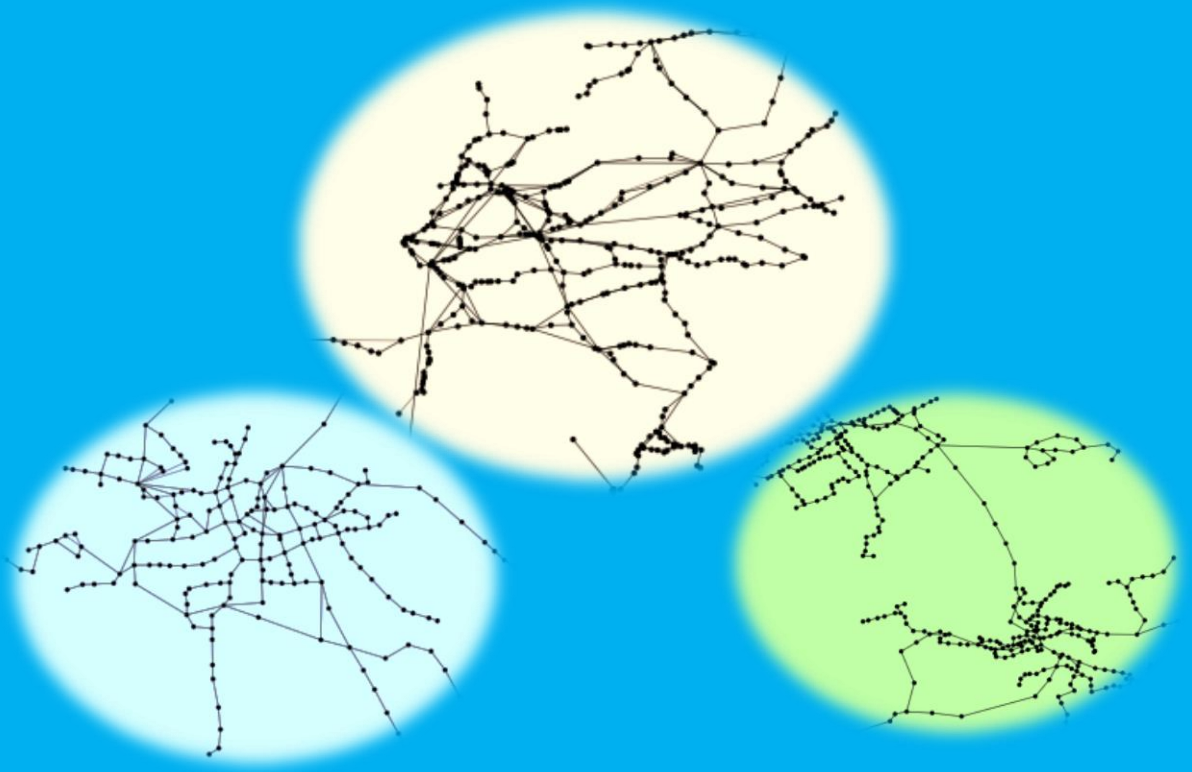
Topological characterizing and clustering of public transport networks

Krissada Tundulyasaree



TU Delft

Lab Smart Public Transport

TOPOLOGICAL CHARACTERIZING AND CLUSTERING OF PUBLIC TRANSPORT NETWORKS

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

Master of Science in Transport, Infrastructure and Logistics

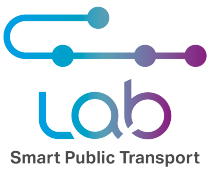
by

Krissada Tundulyasaree

September 2019

Krissada Tundulyasaree: *Topological characterizing and clustering of public transport networks* (2019)

The work in this thesis was made in the:



Smart public transport lab
Department of Transport and Planning
Faculty of Civil engineering& Geosciences
Delft University of Technology

Graduation committee: Dr. Oded Cats (chair)
Dr. Maaike Snelder (daily supervisor)
Dr. Yilin Huang
Ding Luo

PREFACE

Interest in transport and logistics domain drives me to pursue master study in TU Delft two year ago. It was a big challenge to me who was an electrical engineer to get into this field. Under the TIL curriculum, I was prepared and provided with both context and tools to tackle the problems.

My motivation in network studies started with the TIL scientific assignment, a small academic project, supervised by Oded Cats. I was impressed by the power of simplification from complex systems into network to identify pattern or analyze all the relations. This thesis was initiated with the question " Is the Netherlands railway system similar to the metro in metropolitan area"?. Clustering analysis was employed in this thesis to reveal pattern found in the network around the world.

I want to express my gratitude to my exam committee for all their help during this research. I would like to thank Ding Luo for his advice in the technical part of this project regarding the network study and data gathering process. Also, your critical feedback on how I tackle the problem is beneficial. My grateful thanks are also extended to Maaike Snelder for all her support. I appreciated your questions during our meetings and your comments on my report.

I would also like to thank Oded Cats, for the initial idea of the project and guidance during the project definition phase. Your feedback during the committee meetings was constructive and you always provide detailed comments for my thesis report.

Also a word of thanks of gratitude to Yilin Huang as well. Your feedback during the committee meetings was positive and helpful. I have also appreciated your willingness to help and critical feedback especially in my analysis part.

Finally, I wish to thank my family for their support and encouragement throughout my study. Also, I would like to thank friends at afstudeerhok which created a friendly environment to work on this thesis. I also thank other friends who makes my life abroad pleasant.

K. Tundulyasaree

CONTENTS

LIST OF FIGURES

LIST OF TABLES

1 INTRODUCTION

In this chapter, Section 1.1 introduces a problem context in the public transport field and how network studies dealt with. Section 1.2 describes the related works to the structural network studies and highlights the research gap. Section 1.3 states the research goal translated into the main research question and sub-research questions in Section 1.4. An overview of methodology is presented in Section 1.5. The chapter concludes with the outline of this thesis in Section 1.6.

1.1 CONTEXT OF STUDY

Due to the considerable investment cost and societal impacts of public transport (PT) projects, the planning of the system becomes essential. Planners need to understand the current situation of the system before planning any extension or modification. Moreover, they may also want to learn from other system operations in different places by conducting comparative studies. For example, a planner may want to compare his system to those in Japan to identify how they can build on-time train systems. A network study is employed to facilitate those planning purposes. A PT system is viewed as a network, a combination of nodes connected by links. With this simplification, the properties or complex inter-relation in the system can be analysed easily.

There are two lines of research for transportation network studies: structure and dynamics of transportation networks (Ducruet and Lugo, 2013). The former aims to characterize the network structure by network indicators. There are several approaches, including topology, geometry, morphology and traffic flow, to calculate those indicators. Besides, the latter explores changes in the network, identify evolution patterns and the mechanism behind such changes. This thesis aims to contribute to the structural studies of transportation network, specifically public transport networks (PTNs).

1.2 RESEARCH MOTIVATIONS

Comparative study is an approach in structural network studies. They compare different networks to identify similarities or differences. There is much empirical evidence that PTNs in different places share common statistical characteristics (Von Ferber et al., 2009; Derrible and Kennedy, 2010b; XU et al., 2013; Lin and Ban, 2013; Zhang et al., 2013; Haznagy et al., 2015; Wan et al., 2018). For instance, Lin and Ban (2013) found many railway, subway and bus networks in several countries exhibit small-world (Watts and Strogatz, 1998) and scale-free structure (Barabási and Albert, 1999). Since both properties are the significant breakthrough in the complex network studies, more details are in Chapter 2.

Moreover, those mentioned studies adopted a similar approach to characterize the PTNs. They employed graph theory and complex network indicators calculated via a topological approach. Network topology is the abstraction of the real network into only the connection of nodes and links. For PTNs, nodes represent stations while links show the connection between stations. Although this approach leaves out many details, it still leads to progress in understanding collective phenomena in PTNs (e.g. network vulnerability) and identifying important network parts by adapting node centrality concept (de Regt et al., 2017). Moreover, it required a little amount and easily gathered data. The data can be extracted from the network map and timetable usually provided publicly by PT authority or operators. Therefore, analysts can include several networks in their studies.

After the similarities between network were identified, creating classification is a subsequent step. Classification groups similar objects while keeping ones with different properties to other groups. The resulting groups will reveal the pattern of many complex PTNs structure and facilitate further PTNs studies. For example, analysts can lower the number of networks incorporate in the studies because they can still maintain a variety of network structure by selecting a few from different groups. Moreover, it will pave the way for studying each class of PTNs, adding value to the previous comparative studies.

Despite the accumulation of comparative PTNs studies, there were a few attempts to create a classification. Gattuso and Miriello (2005) classified 13 metro networks using combined network indicator scores to rank networks. The scores were calculated from the multi-criteria analysis. However, it is difficult to identify a distinct property of each network as they all combined into a single value. Derrible and Kennedy (2010b) adopted a 2-D graph where each network indicator is on the axis. A few network indicators described each characteristic. They were able to classify 33 metro networks into different groups according to the state of development, interaction with the built environment, and intrinsic structure. Besides, STOILOVA and STOEV (2015) adopted hierarchical clustering to classify 22 European metro networks adopting six network indicators. However, both studies (Derrible and Kennedy, 2010b; STOILOVA and STOEV, 2015) only consider the physical infrastructure side of the PTNs and include a single mode of PT. In their studies, it is assumed that a direct service line always exists connecting a station to any stations; however, that is not always the case. PT operators usually provide several service lines to serve nodes, and the transfer is made at the node where the service line overlapped. Consequently, service lines network is another vital aspect to describe PTNs. Moreover, as only metro networks are considered, the classification result is mode-dependent. In other words, they assume that transport modes influence the network structure.

In summary, few studies classified PTNs and most of them were based on a single PT mode representing on infrastructure networks. Based on such criteria, they assume no similarities between different PT modes and omit the service lines from analysis.

1.3 RESEARCH OBJECTIVES

This thesis aims to classify PTNs employing topological network indicators. PT networks consisting of multiple modes will be considered to explore the similarities across the

modes. In addition to the infrastructure networks, the service aspects will also be taken into account. A quantitative classification will be employed to cluster PTNs because this approach is automatic, scalable and reproducible. Besides, the result of this research can be used by PT network planners to understand their existing PT system and compare to other PT systems. Moreover, the system can further be improved by learning the lessons from the similar successful networks.

1.4 RESEARCH QUESTIONS

The research objectives are translated into the main research question as follows:

"How can public transport networks be quantitatively characterized and clustered from a topological perspective?"

To deal with this question, the main research question is broken down into following sub-research questions:

1. How can the PTNs' structure be quantitatively characterized by network indicators from a topological approach excluding mode-specific properties?

   This question aims to identify a set of network indicators to describe structural characteristics of the public transport network without mode-specific properties.

2. What are the topological clusters of public transport networks and which topological characteristics influence the result?

   This question aims to identify set of clusters of PTNs and which characteristics affect the result.

1.5 APPROACH

In this section, a brief overview of the thesis methodology is described. First, the literature study is used to explore suitable network characteristics to characterize PTNs . Network indicators are then selected to operationalize those characteristics. Subsequently, a selection of PTNs will be characterized by those topological network indicators. The result data set is the compilation of each PTN with their corresponding network indicator values.

The second step is performing classification of the dataset. As there is little prior information of PTNs classification (from structural characteristics), clustering analysis is chosen. From the clustering result, the clustering groups are realized, and the influence of each topological characteristic will be verified.

1.6 OVERVIEW

The remaining of this thesis consists of five chapters. Chapter 2 presents a literature review on the characterization of PTNs and its related network indicators. Moreover, it also

includes the typical approach for topological analysis and related works on PTNs classification. For Chapter 3, the methodology framework of the thesis is presented. It elaborates on the acquisition of PTNs data, network characterization, and network classification. The analysis is conducted and all the results and discussion are presented in Chapter 4. Finally, Chapter 5 concludes the thesis by summarizing and synthesizing main findings and contributions. Additionally, it includes the limitations and recommendations for future research.

2 LITERATURE REVIEW

This chapter reviewed relevant literature two parts in the main research question: PTNs characterization and PTNs classification. First, Section 2.1 identifies relevant structural characteristics of PTNs from complex and spatial network properties. Section 2.2 describes how to operationalize characteristics topologically. Finally, this chapter concludes with the review of previous attempts on PTN classification Section 2.3.

2.1 PUBLIC TRANSPORT NETWORK STRUCTURAL CHARACTERIZATION

The network characterization depends on the network types. From the literature review on the complex topology of transportation networks (Lin and Ban, 2013), several railway, and urban networks are empirically found to exhibit **complex network** properties such as small-world, scale-free, etc. This gives rise to several recent studies (Wei et al., 2019; Wu et al., 2018; Cao et al., 2018; Pagani et al., 2018; Zhang et al., 2018b,a; Bangxang and Jarumaneeroj, 2018; Zanin et al., 2018; Wan et al., 2018; Wang et al., 2017; Cheng et al., 2017) employing complex network properties to analyse PTNs. Moreover, PTNs are also considered as **spatial networks** because their network nodes position and links related to the geographical position governed by Euclidean distance. Also, the concept of planarity is usually applicable to infrastructure networks such as road, rail, and many other types of transportation networks. A planar graph is a graph in which the intersection of link results as a new node (Barthelemy, 2010). For instance, the transfer station connects different service lines. Therefore, PTNs are complex and planar spatial network.

Complex networks usually display non-trivial topological features meaning that they are neither purely regular or purely random network. The two prominent works in complex network field are small-world (Watts and Strogatz, 1998) and scale-free networks (Barabási and Albert, 1999) with a wide range of applications to several fields including technological, biological and social network. The small-world network known as 'six degrees of separation' is highly clustered network with small characteristics path (shortest path) while the vertex connectivity in scale-free networks follows a power-law distribution. Newman (2003) compiled a list of complex network properties as shown in Table 2.1. Derrible and Kennedy (2010c) found that 33 metro networks systems possess both small-worlds and scale-free structure and proposed that highly clustered networks are highly robust. Similarly, in addition to those two properties, bus systems in 330 Chinese cities were found to be degree assortativity, and 30 urban rail transits also possess similar centrality properties (Zhang et al., 2013; XU et al., 2013). Von Ferber et al. (2009) extended the characterization of 14 PTNs in different cities by incorporating property (1)-(3) and (5) -(6) in Table 2.1, but those PTNs show diverse characteristics that classification could not be derived. For the rest of the characteristics, resilience has been explored quite extensively

in many PTNs due to its significance to real operational performance (Zhang et al., 2018c). Many studies (Zhang et al., 2011; von Ferber et al., 2012; Zhang et al., 2018b) found that PTNs are robust against random attack compared to malicious attack.

No.	Network properties	Interpretation	Network indicators
1	Small-world effect	The network is highly clustered and has short characteristic path length.	Average shortest path, Clustering coefficient
2	Transitivity or Clustering	If vertex A is connected to vertex B and vertex B to vertex C, there is a heightened probability that vertex A will also be connected to vertex C.	Clustering coefficient
3	Scale-free networks	The network has a power-law degree distribution.	Degree distribution
4	Resilience	The extent to which network maintains its connectivity, i.e., the existence of a path between pairs of vertices, when subject to removal of network components.	Average shortest path , Network efficiency , Size of the largest component
5	Mixing patterns	The tendency to which vertices pair up with certain vertices.	Assortative mixing coefficient
6	Degree correlations	Mixing patterns according to vertex degree	Pearson correlation coefficient of the degrees

Table 2.1: Complex network properties (Adapted from Newman 2003)

For spatial networks, Barthelemy (2010) compiled network indicators for characterising spatial networks and divided them into two categories: basic and mixing space and topology type. The basic indicators include adjacency matrix, clustering coefficient, assortativity coefficient, average shortest path, discrete Laplacian, and betweenness centrality. For the latter type, strengths, α index, γ index, ringness, route factor, network cost, network efficiency, and modularity were used. Note that since several indicators are the same as those already discussed for the complex network, only non-complex network indicators will be further elaborated.

To see the applications of those indicators to PTNs, related studies will be explored. γ index was used to describe network connectivity in terms of links' density (Garrison, William L and Marble, 1962) while network robustness was quantified by α index (Derrible and Kennedy, 2010c). Since α index counts the number of cycles or alternative routes, it can explain the extent to which a network remains functioning under the disruption when some routes are disconnected. Moreover, Ding et al. (2015); Cats (2017) employed the γ and α index to quantify the connectivity and robustness of the railway network in Kuala Lumpur and Stockholm. Additionally, essential stations can be interpreted as the critical spots in the system because many trips are expected to pass through those

stations. It is crucial to handle them properly. The concept of node centrality, including betweenness, degree, and closeness centrality can identify such stations (Tu, 2013; To, 2015; Chen et al., 2018; Cao et al., 2018). Besides, strength centrality is the extended concept of degree centrality with weighted links. Note that weight can depict the travel distance or travel time between nodes. Cao et al. (2018); Chen et al. (2018) characterized the traffic intensity of Chinese high-speed rail stations by adopting strength centrality. Route factor compared the route provided by a system to the crow flies route between a pair of nodes. It was used to measure the directness of Stockholm metro (Cats, 2017) and the accessibility of stations (Barthelemy, 2010). Similarly, network efficiency is also related to network directness because more direct network tends to provide a shorter and more efficient trip. For instance, (Latora and Marchiori, 2002) employed network efficiency to evaluate Boston network. For the other two indicators, ringness is used to characterize the arterial roads (Xie and Levinson, 2007) while network cost is similar to γ in the sense that it compares the network current to the desired state. In other words, network cost often refers to the total length in the current network compared to the ideal minimum network length, minimum spanning tree network.

In conclusion, PTNs can be characterized by complex and planar spatial network perspectives. Their characteristics can be divided into two groups: complex and spatial properties. Complex properties include **small-world, transitivity, scale-free, resilience, mixing pattern, and degree correlations**. Moreover, **network connectivity, robustness, node centrality, service intensity, network accessibility, and efficiency** are spatial network properties.

2.2 TOPOLOGICAL ANALYSIS OF PUBLIC TRANSPORT NETWORKS

Network topological analysis studies the arrangement and connectivity of network nodes and links (Xie and Levinson, 2007). The basic representation is on a graph consisting of nodes and links. Several graph types can be distinguished depending on link types and network models. For link types, there are two aspects to consider: link direction and links weight. For the link direction, a graph with unidirectional links is a directed graph while the graph with bidirectional links is an undirected type. Moreover, the weighted graph has weighted links. For network models, they refer to what the nodes and links are represented. Although there are several network model available in the literature, only common types for PTNs studies are further investigated.

L-space and P-space are standard network models in PTNs studies (Lin and Ban, 2013). **L-space** graph illustrates stations as nodes while a link connecting nodes exists if there is at least a service line connecting those two consecutive nodes (Von Ferber et al., 2009). In other words, L-space representation depicts the infrastructure side of the PTNs, describing the stations and their interconnection. Sometimes, it is referred to as space-of-infrastructure (Luo et al., 2019) or space-of-stations (Kurant and Thiran, 2006). It is usually applied in work regarding the physical network structure such as the vulnerability of subway network (Zhang et al., 2011; von Ferber et al., 2012), the centrality of stations (Derrible, 2012; Tu, 2013), etc. However, L-space lacks information on the service lines.

To incorporate service line details, **P-space** representation is adopted. A node in this graph still represents a station, but a link exists if there is at least a direct service line linking the pair of nodes (Von Ferber et al., 2009). This implies that a neighborhood of a node is all stations reachable without changing service lines. P-space is often referred to as space-of-service (Luo et al., 2019) or space of transfers (Kurant and Thiran, 2006). For simple visualization, Figure 2.1 illustrate both the L-space and P-space derived from the same network map. Numerous studies employed P-space to investigate different network properties such as accessibility of Chinese high-speed rail network (Chen et al., 2018), hierarchical network property (Wei et al., 2019), structural characteristics of PTNs in 330 Chinese cities (XU et al., 2013). Based on network indicators and their space representation, network characteristics can be operationalized.

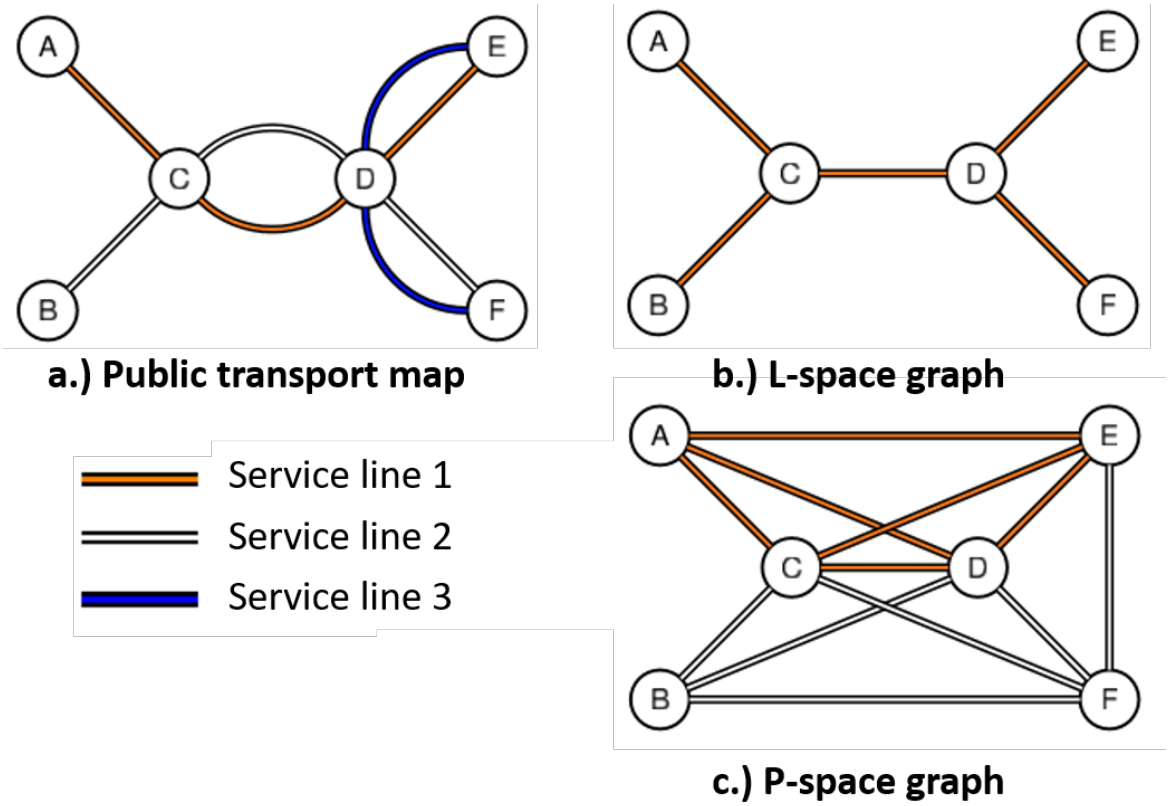


Figure 2.1: Schematic diagram of a typical representation for PTNs (Adapted from Von Ferber et al. 2009)

2.3 PUBLIC TRANSPORT NETWORKS CLASSIFICATION

Early studies classified metro networks according to their geometric form of networks with their quantitative measures. The metro lines were classified as radial, diametrical, tangential, circumferential, trunk, and irregular lines (Musso and Vuchic, 1988; Vuchic and Musso, 1991).Figure 2.2 shows each line type. Based on the line types, the whole

network can be identified in three geometric forms: radial, rectangular, and modified grid. Each type was roughly defined according to its line types. For example, a radial network consisted of radial and diametrical lines intersecting in the city center. As the definition was vague, the geometrical forms were complemented with network indicators value. The proposed networks indicators were grouped into five information categories as follows (Musso and Vuchic, 1988):

1. Measures of network size and forms
2. Indicators of network topology
3. Measures of relationship between network and city
4. Quantity and quality of offered service
5. Measures of service use

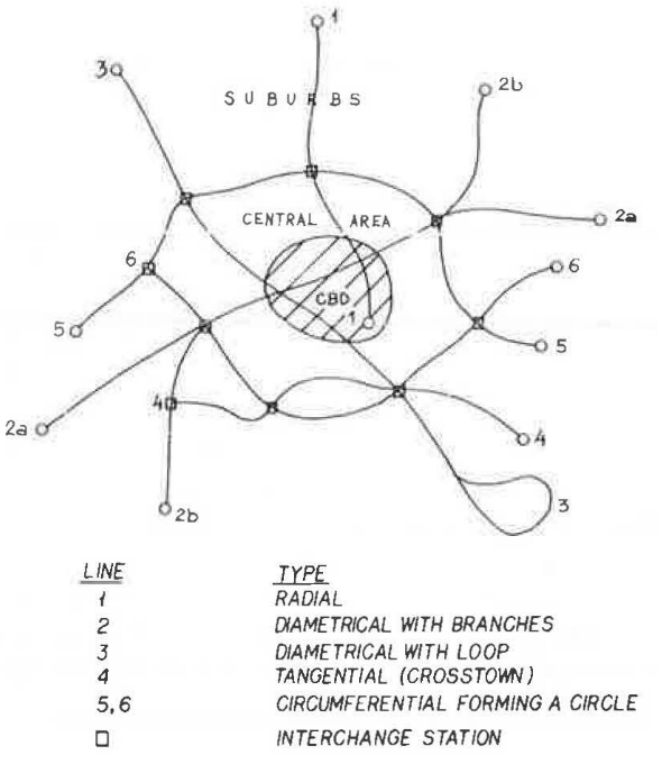


Figure 2.2: Transit line types (Adapted from Musso and Vuchic 1988)

Gattuso and Miriello (2005) distinguished network indicators according to the required input data. There were three types of data: topological, geographical, and operative data. The topological was at the graph level dealing with nodes and links while geographical data includes the information about territory such as length of links, etc. Operative data is the data on operational performance, including commercial speeds and frequencies. Moreover, 13 metro networks were classified or ranked by the multi-criteria score. The score

is the weighted average of all selected network indicators. Although the result describes which network performed best in the chosen criteria, it could not explain or identify similarities or difference among networks.

The network indicators were enriched by adopting space representation.Zhang et al. (2013) employed L-space representation to calculate six global spatial network indicators. They are composed of average degree, average node betweenness, average betweenness of edge, average shortest path, average unit degree betweenness, and clustering coefficient to analyze the urban rail networks in 30 cities around the world. Moreover, they found that any stations in urban rail network directly connect to 2 up to 2.45 stations and between 10 and 16 stations were required when one makes a trip on average. However, this analysis is not useful because their network indicators were not normalized. This implies that the comparison does not consider the difference in network size. For instance, smaller networks tend to require a fewer number of intermediate stations on average. Additionally, Von Ferber et al. (2009) broadens the analysis by incorporating many network indicators in multiple space representations, including L-space, P-space, and C-space. However, they could not identify PTNs division or classification group as there was more diversity in structure than expected.

Derrible and Kennedy (2010a) contribute significantly to PTN classification in their work of classifying 33 metro networks around the world. State, form, and structure are three main characteristics in their study. For each characteristic, a metro network is represented as a coordinate point on a graph where one indicator is on the x-axis, and the other is on the y-axis. Based on the 2-D graph, Derrible and Kennedy (2010a) can classify networks simultaneously considering two network indicators. The final classification result is, as shown in Figure 2.3.

According to Figure 2.3, the order of classification was from the state, structure, and form. State is the current development phase describing by complexity and degree of connectivity indicators. The three-phase state is identified from starting a network (phase 1), gradually expanded (phase 2) to significantly expanded (phase 3). Besides, structure refers to intrinsic properties from connectivity and directness indicators. Similarly, three zones are identified as directness-oriented, connectivity-oriented and integrated. Directness refers to a network providing a fewer number of transfers, while connectivity is the opposite. Last but not least, form shows how the network integrates with surrounding and is quantified by inter-station spacing and average line length. Their analysis result specified three forms: regional accessibility, local coverage, and regional coverage. For regional accessibility, the network focuses more on connecting to the outer layers of the city, while local coverage focuses on servicing the city core. The regional coverage is in between the previous two. After acquiring all sub-groups for each characteristic, they were merged to form the complete classification. However, Derrible and Kennedy (2010a) predefined the groups for each characteristic and classified the metro networks into them. Since their analysis is on the 2-D graph, they arbitrarily drew the boundary lines to define predefined groups. This approach relies heavily on the analyst' insight to identify the number of groups for each characteristic.

Quantitative classification is a systematic approach to alleviate such issue. In this approach, the number of groups is mathematically identified based on the chosen similarity matrix. STOILOVA and STOEV (2015) adapted a hierarchical clustering, one of the quan-

titative classification approaches, to classify 22 European metro networks. They employed ten network indicators which are five newly proposed routing network indicators and the rest from Derrible and Kennedy (2010a). Moreover, they can identify three-cluster groups: complex, simple, and only one metro line. These groups show the stage of development for this sample set of the metro. The network indicators specifying the state and the structure of a network are found to be essential for this grouping.

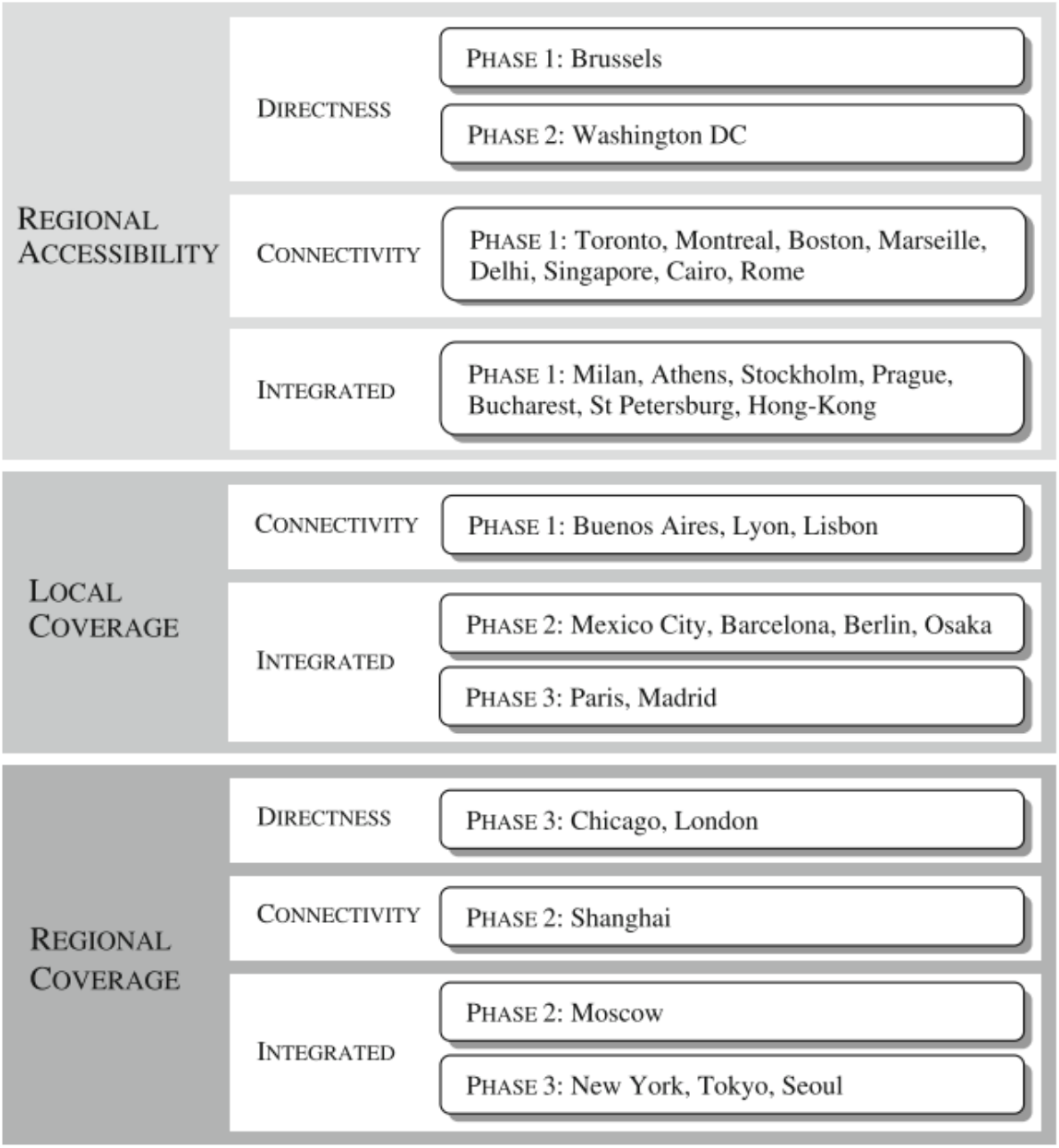


Figure 2.3: An example of metro classification in different cities (Adapted from Derrible and Kennedy 2010a)

2.4 LITERATURE GAPS

In summary, the main research gaps are in the PTNs classification. Although PTNs are characterized by several types of network indicators from graph theory (Musso and Vuchic, 1988; Gattuso and Miriello, 2005; Derrible and Kennedy, 2010b), complex network (Von Ferber et al., 2009; Haznagy et al., 2015; Zhang et al., 2013) and spatial network (Barthelemy, 2010), few went beyond comparison to classify PTNs. Comparison is usually performed per single network indicator even though they were characterized by many indicators. Gattuso and Miriello (2005) employed multi-criteria analysis to combine various network indicators value into a single score, so the networks were ranked by the combined score. However, their result could not identify any similarities or difference among networks. Moreover, Derrible and Kennedy (2010a) adopted a 2-D graph where each indicator is on the axis. They were able to classify metro network into different groups according to the state, form, and structure, but their classification was solely based on the analysts' insight to predefined cluster groups. To alleviate such issue, STOILOVA and STOEV (2015) adopted hierarchical clustering to classify European metro networks and found 3 clusters: simple, complex, and one line network. However, both studies (Derrible and Kennedy, 2010a; STOILOVA and STOEV, 2015) only included metro networks and adapted only space-of-infrastructure. The classification failed to address the variety of PT modes and does not consider the service aspect of the network. Table 2.2 summarize the related research and indicates what this thesis will tackle.

Study	**PTN types**	**Network Level**	**Network model**	**Classification method**
Gattuso & Miriello, 2005	metro	Urban	L-space	Multi-criteria analysis
Von Ferber et al., 2009	bus, ferry, subway tram, urban train	Urban	L-space, P-space, C-space	Single criteria comparison
Derrible & Kennedy, 2010b	metro	Urban	L-space*	2-D graph
STOILOVA & STOEV, 2017	metro	Urban	L-space*	Hierarchical clustering
This thesis	metro, tram train, light-rail	Urban, National	L-space, P-space	k-means, Hierarchical clustering

Table 2.2: Overview of studies on PTNs classification and how this work fits in. Note * refers to the special type of L-space in which only terminal and transfer stations are included in the analysis.

3 METHODOLOGY

This chapter details the approach of this thesis. First, Section 3.1 provides the overview of the methodology and outlines detailed steps in the flowchart. Section 3.2 describes the selected topological approach and details the network characteristics along with their network indicators. Finally, Section 3.3 explains how to classify PTNs by quantitative approach.

3.1 ANALYSIS FRAMEWORK

As a brief overview of the thesis' methodology, Figure 3.1 shows two sub-processes: network characterization and network classification. Network characterization builds a network data set, a collection of PTNs with their corresponding network indicator values. The first step is selecting the desired network characteristics. Each network indicator calculated on an appropriate space representation quantifies each characteristic. After acquiring all network indicators values, the correlation between network indicators is tested. The highly correlated network indicators will be filtered out before storing the rest as the network data set.

The second sub-process is classifying the data set acquired from the network characterization. First, check if the data set is clusterable, otherwise re-select the network indicators. Next, all network indicator values are normalized to ensure they will be weighted equally in the analysis. Then, perform K-means and hierarchical clustering to partitioning the data set into different clusters and identify the matching pattern inside each cluster. After that, evaluate clusters to ensure the quality of the result. Moreover, principal component analysis (PCA) facilitates cluster result visualization. Finally, create radar diagrams and L-space graphs of all networks for further result interpretation. Note that a radar diagram is a circle chart illustrating all network indicator values for each network.

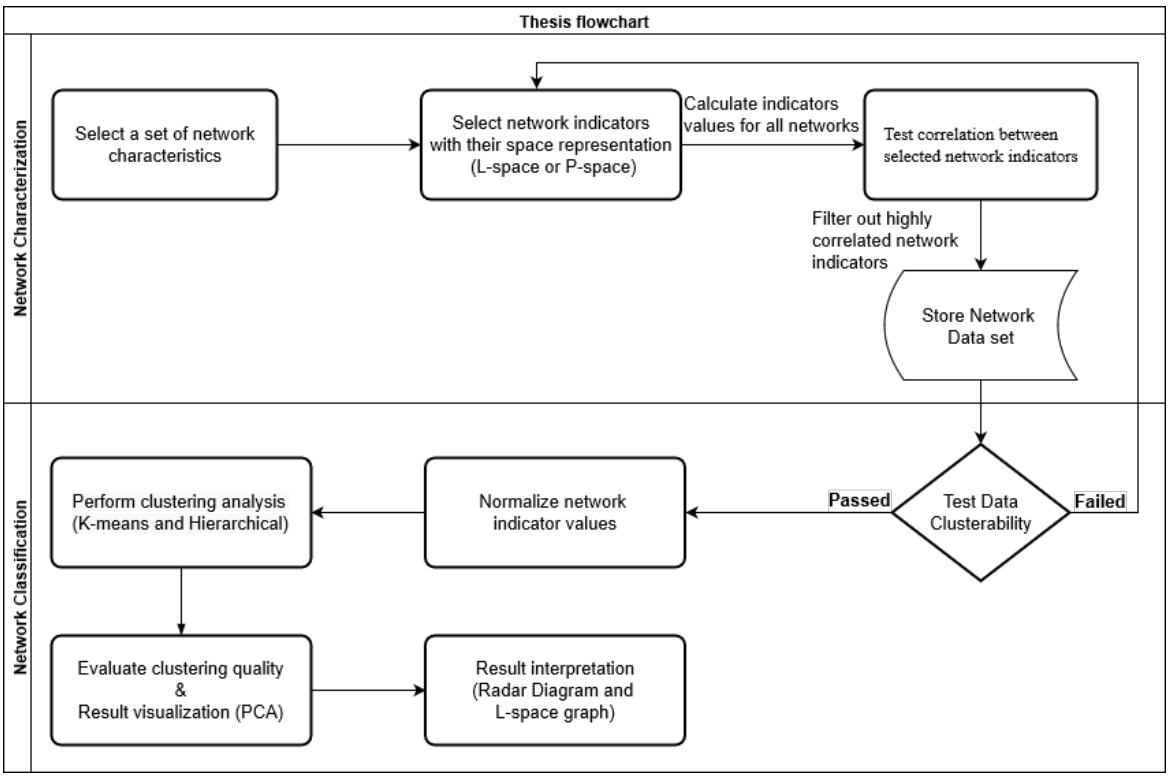


Figure 3.1: Thesis methodology flowchart. *PCA: Principal component analysis

3.2 NETWORKS CHARACTERIZATION

This section elaborates on each step for network characterization as specified in Figure 3.1. The first subsection introduces a brief PTNs representation used in this thesis. Next, network characteristics selection is described and followed by the discussion of each selected network characteristic and their corresponding network indicators. The section concludes with how to determine the correlation between the network indicators.

3.2.1 Network representation

A PTN is represented as a directed graph G representing by $G = (N, E)$ where N is the set of nodes and E is the set of links. A node $n \in N$ represents a station while a link $e \in E$ is defined by an ordered pair of nodes (u, v) in which u and v $(u, v \in N)$ denote the source and sink nodes, respectively. Note that $\mid N \mid$ and $\mid E \mid$ denote the number of stations and links, respectively.

In this thesis, **L-space** and **P-space** graphs are employed to enrich network characterization. The former addresses the infrastructure side of the network while the latter models service side of the network. Moreover, the graph's links have no weight, so all links have identical properties. Since PTNs in this thesis are the combination of different modes such

as metro, tram, and train, the resulting graph does not differentiate the transportation modes.

3.2.2 Network characteristics selection

Since the thesis aims to create PTNs classification considering the combination of PT modes, the selected network characteristics should be relevant to all types of PT. Moreover, they will be assessed when the network is in a normal operation state (without any disruption). The network analysis will take into account all the nodes and links. In a well-functioning network, this is supposed to be a crucial period illustrating the quality of PTNs to facilitate the trips. Besides, this state lasts much longer than the disruption state in most cases. Last but not least, this network characteristic set includes both the infrastructure and service layer properties as the PTNs relied on both layers for functioning.

Based on the above selection criteria, centralization, accessibility, robustness, transitivity, and directness are five selected network characteristics. The first three will be assessed in the infrastructure layer, while the service layer is employed for the last two. The following subsections detail each characteristic and its networks indicator with mathematical formulas. Note that every formula will be in normalized formulation for comparison purpose between different PTNs as this eliminates the dependence of indicators on network size, i.e., a number of stations or links (Zanin et al., 2016).

3.2.3 Centralization

Centrality describes how central a node is in the network. The central notion can be defined up to the application types. For PTNs, an important function of stations (nodes) is to facilitate the trips between any origins and destinations. In other words, they act as intermediate stations in which many trips passed through them. This notion is in line with betweenness centrality from the graph theory.

Node betweenness centrality

Node betweenness centrality is defined by total number of shortest path passing through that node compared to the total number of shortest paths in the network. The node with high betweenness centrality value is more likely to involve in higher number of shortest paths and has a greater control power for any trips in the network. The node betweenness centrality ($C_B^l(v)$) can be calculated by following equation:

$$C_B^l(v) = \sum_{s \in N} \sum_{t \in N} \frac{\sigma_{st}(v)}{\sigma_{st}} \tag{3.1}$$

where σ_{st} denotes the total number of shortest paths between node s and t and $\sigma_{st}(v)$ is number of shortest path between node s and t containing node v. The range of node betweenness centrality is between 0 (no shortest path passed through) and 1 (all shortest paths passed through). Moreover, this indicator will be calculated in L-space representation of the network as specified by the superscript, so the shortest path can be interpreted

as the selected route between any pair of nodes with fewest number of stations in between. In other words, betweenness centrality of nodes in L-space indicates the proportion of the shortest path in a sense of fewest intermediate stations from infrastructure perspective because of the L-space representation.

Betweenness centralization

The node betweenness centrality only concerns at node level. To explore this property at the network level, an extension concept is required. Freeman (1978) proposed the concept of centralization to analyse the distribution of node centrality in a network. The general notion of centralization is that network with high value of centralization is likely to have a single powerful centrality node with most others in the network with low centrality and vice versa. It gives insight to network structure if it resembles the hub-and-spoke (high centralization) or more like regular network (low centralization). Note although this can also be applied to other centrality measures, our focus will be on the betweenness centrality. **Betweenness centralization** (C_B^l) can be realized as follows (Freeman, 1978):

$$C_B^l = \frac{\sum_{v \in N} (C_B^l(v)^* - C_B^l(v))}{\mid N \mid^3 - 4 \mid N \mid^2 + 5 \mid N \mid - 2} \tag{3.2}$$

where $C_B^l(v)^*$ is the maximum betweenness centrality among nodes in the network. The numerator in Equation 3.2 is the sum of the difference between the most central nodes and all the other nodes while the denominator is the sum of the maximum possible difference between the most central nodes and all other nodes. The denominator derived by (Freeman, 1978) as in Equation 3.2 varies according to the number of nodes in the network. As the value in numerator is normalized by the maximum possible value, C_B^l varies between 0 (all nodes with equal value of betweenness centrality) and 1 (few nodes with high betweenness centrality).

3.2.4 Accessibility

After realizing the underlying stations types from centralization, their connection between nodes is assessed. As stations acts as the origin, intermediary and destination in the PTNs, these spots are to be reached or accessed. To evaluate the network options to reach any stations, accessibility property is analysed. In this sense, accessibility referred to "the ease with which any land-use activity can be reached from a location using a particular transport system" (Dalvi and Martin, 1976). For this thesis, the land-use activity only limited to the destination within the PTNs. To quantify the ease to reach a station, node closeness centrality($C_C^l(v)$) is employed and calculated as in Equation 3.3.

$$C_C^l(v) = \frac{\mid N \mid -1}{\sum_{i \in N} d_{vi}} \tag{3.3}$$

where d_{vi} is the shortest topological distance between node v to node i. As $C_C^l(v)$ is calculated in L-space graph, d_{vi} is the number of stations needed for trans-versing from

node v to node s. Closeness centrality of node v is the inverse sum of shortest path length from all other nodes to node v normalized by the number of stations. In other words, it is the inverse of the mean shortest path from node v to all other nodes. The best closeness value is derived from the case where a node is directly connected to all other nodes ($C_C^l(v) = 1$) while the farthest nodes from others possess C_C^l of nearly 0. However, this measure only allows the comparison at node level. Therefore, the global indicators for closeness centrality are derived by finding average value as follows:

$$A^l = \frac{\mid N \mid -1}{\mid N \mid} \sum_{v \in N} \frac{1}{\sum\limits_{i \in N} d_{vi}} \tag{3.4}$$

where A^l is the **Accessibility**. The indicator is calculated in L-space in which the shortest path between nodes represents the number of intermediate stations passed along the route. A^l can vary from 0 to 1. The closer the value of A^l to 1 indicates the better accessibility of all nodes in the network to each other nodes and vice versa.

3.2.5 Robustness

It is unlikely that the disruption to PTNs can be completely prevented. To cope with that, PTNs should be designed to have a robust structure. In other words, the network is able to maintain its functionality to certain extent while the mitigation process is on going. A simple indicator is the number of redundant routes existed in the network because they will become an alternative routes to the disrupted one. This redundancy route concept is related to cycle concept in graph theory. Cycle is the sequence of paths returning to its origin or starting position. In the context of PTNs, a cycle is a series of paths one could take from any stations back to the same station without repeating the same links twice. To quantify number of redundancy routes, **alpha index or network meshedness** (α^l) is used and can be calculated as:

$$\alpha^l = \frac{\mid E \mid - \mid N \mid +1}{2 \mid N \mid -5} \tag{3.5}$$

i.e. the ratio between number of network cycles or loops in a single connected graph and maximum number of network loops in a planar graph with the same number of nodes. Alpha index ranges from 0 for a tree network ($\mid E \mid = \mid N \mid -1$) to 1 for completely connected graph ($\mid E \mid = 3 \mid N \mid -6$).Note that the denominator is derived for the planar network which are for PTNs. The higher value of alpha indicates the more robust the network structure in a sense that network provide greater number of alternative path. In other words, if some links are disrupted, there are more likely to have an alternative paths. Focusing on the infrastructure robustness, α^l is calculated in L-space to reflect the redundant routes from the infrastructure perspective.

3.2.6 Service connectivity

Service connectivity (SC) measures how well the service lines are linked. It compares the current link connection relative to the best connection scenario, which is usually the

complete graph (Derrible and Kennedy, 2010a). Note that the complete graph is a network in which every pair of nodes is connected. The high SC network offers a large number of direct routes between nodes. In PTNs, a direct service route does not require any transfers when transversing between a pair of stations. Moreover, it implies that network nodes cluster together as service lines directly connect them. This concept is in line with a clustering coefficient (c^P) in the space-of-serviceP. Let $Nbh(u)$ is the neighbourhood of a node u ($Nbh(u) = \{v \in N \mid (u,v) \in E\}$) and d_u is the degree of a node u ($d_u = \mid Nbh(u) \mid$). The clustering coefficient can be defined as in Equation 3.6 (Von Ferber et al., 2009):

$$c^P = \frac{\sum_{u \in N} \mid \{\{i,j\} \subseteq Nbh(u) \mid (i,j) \in E\} \mid}{\sum_{u \in N} d_u(d_u - 1)/2} \tag{3.6}$$

c^P measures the whole network service connectivity through the connectivity within the neighbourhood. The denominator in the Equation 3.6 is the maximum number of links in the neighbourhood for node u given the node degree d_u. The range of c^P is between 0 and 1. The closer c^P is to 1, the higher SC level network is.

3.2.7 Directness

For PTNs, it is desirable as a commuter to have a direct route between any pairs of origin and destinations. However, that is not usually possible in most PTNs. The need to transfer is inevitable. To quantify the extent of PTNs directness, number of transfers can be used as a proxy to describe the directness of potential route. PTNs required lower number of transfers are more likely to provide more direct service route. To gain insight for the whole network, an average number of transfers calculated for any pair of nodes is employed. The shortest path length between node i and j in P-space representation (d^p_{ij}) can be interpreted as the lowest number of transfers required between node i and j. Based on that concept, the inverse of d^p_{ij} could be interpreted as the directness of the route between node i and j. Moreover, to analyse the whole network directness, the average value for all pairs of nodes is calculated. The mathematical formulas of the network directness is quantified by network efficiency (E^p) as in Equation 3.7 (Latora and Marchiori, 2002).

$$E^p = \frac{1}{\mid N \mid (\mid N \mid -1)} \sum_{i \in N} \sum_{\substack{j \in N \\ i \neq j}} \frac{1}{d^p_{ij}} \tag{3.7}$$

where d^p_{ij} is the shortest path length from node i to j. The indicator value range from 0 (the most inefficient or direct) to 1 (the most direct network).

3.2.8 Summary of network characterisation

Table 3.1 summarize all the network characteristics, corresponding network indicator, notation and network model.

	Network Characteristics	Network indicators	Notation	Network model
1	Centralization	Betweenness centralization	C_B	L -Space
2	Accessibility	Average Closeness centrality	A^l	L -Space
3	Robustness	Alpha index	α^l	L -Space
4	Service Connectivity	Clustering coefficient	c^p	P-Space
5	Directness	Network efficiency	E^p	P-space

Table 3.1: Summary of network characterization

3.2.9 Network indicators correlation analysis

To verify the relation between network indicators, correlation analysis is employed. The interpretation is made based on the rigorous standard proposed by Jawlik (2016). The criteria are as shown in Table 3.2 and also applied to the negative correlation in the same manner.

Evidence of correlation	Pearson's r (r)
very strong	0.81 - 1.00
strong	0.61 - 0.80
moderate	0.41 - 0.60
weak	0.21 - 0.40
none	0.00 - 0.20

Table 3.2: Threshold for the evidence of correlation with the Pearson's r (adapted from Jawlik 2016)

3.3 NETWORKS CLASSIFICATION

This section describes the network classification procedures as specified in Figure 3.1. First, Section 3.3.1 describes how to test if the data set is clusterable. Next, Section 3.3.2 describes data pre-processing steps to ensure the quality of network clusters. After transforming the data set into a suitable format, Section 3.3.3 elaborates on the two clustering methods,k-means, and hierarchical clustering, employed in this thesis. Section 3.3.4 details on how to evaluate the quality of the clusters and visualize the cluster groups. Finally, **??** describes radar diagram and L-space graph to facilitate the result interpretation.

3.3.1 Testing data clusterability

The clustering tendency indicates whether there is a non-random structure in the data set to test whether such data set contains cluster structure. For example, a data set generated from a uniform distribution does not contain any distinct clustering structure. To check whether a given data set is generated from a uniform data distribution, Hopkins statistics can be used. As a basis for the analysis, the null hypothesis is that the data set, D, is a sample of data generated from uniformly distributed random variable U (i.e. does not contain meaningful clustering structure). The alternative hypothesis is thus D is not generated from the uniform random variable. Let $P = \{p_i \mid i = 1, 2, 3, ..., n\}$ and $Q = \{q_i \mid i = 1, 2, 3, ..., n\}$ are samples of n points drawn from D. The Hopkins statistics, H, can be calculated as follows:

$$H = \frac{\sum_{i=1}^{n} y_i}{\sum_{i=1}^{n} y_i + \sum_{i=1}^{n} x_i} \tag{3.8}$$

where $x_i = \min_{v \in D} d(p_i, v)$ and $y_i = \min_{v \in D, v \neq q_i} d(q_i, v)$. Both x_i and y_i are the distance measured from point v to their corresponding nearest neighbor. The interpretation is that if D is uniformly distributed, both $\sum_{i=1}^{n} x_i$ and $\sum_{i=1}^{n} y_i$ have similar value on average, so H is nearly 0.5. Moreover, in the case that D is highly skewed, H would be close to 0.

However, H value can be sensitive depending on the data sampling process. Alternative approach is to test hypothesis assuming that with H follows the beta distribution with both parameter equal to the number of points selected Adolfsson et al. (2019). The number of points selected are usually around 5-10 % of the raw data set.

3.3.2 Data pre-processing

To classify an unlabelled multi-attribute data set, clustering analysis will be employed. A cluster is a group containing similar data objects which are dissimilar to objects in other clusters. The similarity between objects can be quantified by different measures depending on the type of data. **Euclidean distance** is selected because all object attributes are numeric types, measurable quantity. Let $i = (x_{i1}, x_{i2}, ..., x_{ip})$ and $j = (x_{j1}, x_{j2}, ..., x_{jp})$ be two objects containing p numeric attributes. The Euclidean distance between objects i and j ($d(i, j)$) is calculated as:

$$d(i, j) = \sqrt{(x_{i1} - x_{j1})^2 + (x_{i2} - x_{j2})^2 + ... + (x_{ip} - x_{jp})^2} \tag{3.9}$$

Moreover, each attribute value is normalized, so it is equally treated in the clustering process. Min-max normalization is employed and can be calculated as follows:

$$x'_{1p} = \frac{x_{1p} - \min_p}{\max_p - \min_p}(new\_max_p - new\_min_p) + new\_min_p \tag{3.10}$$

i.e. a value of data object 1 attribute p (x_{1p}) is mapped to a new value x'_{1p} in the new value range from new_min_p to new_max_p. This normalization keeps the relation among

the raw dataset values. After completing data pre-processing, k-Means and agglomerative hierarchical clustering methods are applied to uncover the underlying clusters of the data set. The former method aims to find the inter-relation between clusters while the latter focus on the intra-relation within the cluster itself.

3.3.3 Clustering analysis

k-means clustering

k-means clustering is a partitioning clustering method. For a given data set, D, of n objects, k-Means arranges objects into k mutually exclusive clusters. To assess the partitioning quality, k-means used the following objective function:

$$\min \sum_{i=1}^{k} \sum_{p \in C_i} d(p, c_i)^2 \tag{3.11}$$

i.e. the sum of squared distance between all object p in cluster C_i ($p \in C_i$) and the centroid of the cluster C_i (c_i) is minimized. In other words, the distances between each object in cluster and its cluster centroid are squared and summed. The resulting clusters is supposed to be compact and well-separated. Note that centroid of clusters can be defined in several ways such as the mean or medoid of the objects in the cluster. However, the formulation is NP-hard in general Euclidean space, so the problem will be tackled by k-mean algorithm instead. The step by step algorithm is presented as in Figure 3.2. The implementation is done by Python library named Sci-kit (Pedregosa et al., 2011). Additionally, k-means results will describe the relation of the clusters as a whole. However, they do not contain any information about the relation within the groups, so hierarchical clustering is applied to fill in this gap.

Agglomerative hierarchical clustering

Agglomerative hierarchical clustering provides information within the cluster group as data objects in each cluster are arranged into different levels . Agglomerative refers to a bottom-up strategy when performing clustering. In other words, the method begins with each data point in its own cluster and it is paired up with others in every iteration round until a single cluster is left. Note that there exist divisive approach (top-down strategy) which is the reverse process of agglomerative approach; however, it is not a common approach due to the result accuracy and efficiency (Han et al., 2012). Moreover, this method is flexible because there are a few linkage criteria to choose form yielding different results . In this thesis, three types of linkage measures are applied including ward's method, maximum and average distance. Let $\mid i - j \mid$ be the distance between object i and j, m_i is the center of the data points in cluster, C_i and n_i is the number of data points in cluster i. The criteria can be calculated as in Equation 3.12 - Equation 3.14. Furthermore, the result of hierarchical clustering is plotted on the dendrogram or the tree diagram to show the matching order of data objects. The more similar objects tend to gather first according to their shorter distance. While the x-axis of the diagram will be

Algorithm: *k*-means. The *k*-means algorithm for partitioning, where each cluster's center is represented by the mean value of the objects in the cluster.

Input:

- *k*: the number of clusters,
- *D*: a data set containing *n* objects.

Output: A set of *k* clusters.

Method:

(1) arbitrarily choose *k* objects from *D* as the initial cluster centers;
(2) **repeat**
(3) (re)assign each object to the cluster to which the object is the most similar, based on the mean value of the objects in the cluster;
(4) update the cluster means, that is, calculate the mean value of the objects for each cluster;
(5) **until** no change;

Figure 3.2: k-means algorithm (adapted from (Han et al., 2012))

each network, the y-axis shows the distance between clusters. The diagram will be dissect at the certain distances when it yields the desired number of clusters.

$$d_{max}(C_i, C_j) = \max_{i \in C_i, j \in C_j} \{ \mid i - j \mid \} \tag{3.12}$$

$$d_{avg}(C_i, C_j) = \frac{1}{n_i n_j} \sum_{i \in C_i, j \in C_j} \mid i - j \mid \tag{3.13}$$

$$d_{ward}(C_i, C_j) = \frac{n_i n_j}{n_i + n_j} \mid m_i - m_j \mid^2 \tag{3.14}$$

3.3.4 Result evaluation and visualization

Evaluation of clustering

It is argued that any clustering methods always yield clustering groups however, not all data sets are clusterable. In other words, the clustering result can be random and unreliable with no meaningful interpretation. To evaluate clustering result, both feasibility of the data set and quality of generated result are investigated. In this thesis, three evaluation tasks are performed: **assessing clustering tendency**, **determining the number of clusters**, and **measuring the clustering quality**.

Determining the number of clusters

The number of clusters are important to both compressibility and accuracy of clusters. For example, the maximum compressibility can be achieved by having all data points in a sole cluster while separately differentiate data points into their own clusters will result in the highest accuracy group. However, both cases would not yield reliable or meaningful results. To gain deeper insights about the potential number of clusters a data set possess, two methods are applied: rule of thumb and elbow method. Given that a data set contains n points, the rule of thumb suggests that the number of clusters are $\sqrt{\frac{n}{2}}$ and each cluster has $\sqrt{2n}$. Moreover, the elbow method compares the sum of within-cluster variance (Equation 3.11) and the number of clusters. This is based on the observation that the sum of within-cluster variance is inversely proportional to the number of clusters. The optimal number of clusters are found to be the turning point in the curve (Han et al., 2012).

Measuring the clustering quality

To measure the quality of clustering, there are 2 method: extrinsic and intrinsic method. The intrinsic method is chosen since the ground truth or data labels are not available. Therefore, the intrinsic method mainly assessed the quality of clusters in two aspects: their separation between groups and the compactness of clusters.

The silhouette coefficient is selected to assess the clustering quality. Supposed a data set, D, is partitioned into k clusters $C_1, C_2, ..., C_k$. For each object $v \in D$, silhouette coefficient of v is then defined as:

$$s(v) = \frac{b(v) - a(v)}{\max\{a(v), b(v)\}} \tag{3.15}$$

where $a(v)$ is the average distance between v and all other object in the cluster v belongs to and $b(v)$ is the minimum average distance from v to all clusters v does not belong to (see Equation 3.16).

$$a(v) = \frac{\sum_{v' \in C_i, v \neq v'} d(v, v')}{\mid C_i \mid -1}, \quad b(v) = \min_{C_j : 1 \leq j \leq k, j \neq i'} \{\frac{\sum_{v' \in C_j} d(v, v')}{\mid C_i \mid}\} \tag{3.16}$$

The value of the silhouette coefficient for each object v varies between -1 and 1. While $a(v)$ reflects the compactness of the cluster, $b(v)$ describes the distance between clusters. Therefore, when $s(v)$ is close to 1, the resulting clusters containing object v is compact and far away from other clusters. On the other hand, the negative value of $s(v)$ suggests that object v is closer to objects in other clusters more than objects in its own cluster. The average value of the silhouette coefficient from all objects in the data set can be used to compare different clustering methods.

Result visualization

To facilitate the result visualization, data objects are usually represented on the graph. In general, a graphical representation either in 2-D or 3-D is used in which PTNs will be plotted as coordinate points. Based on such graph, a network comparison can be shown

according to their network characteristics. However, each PTN is described by five indicators, so the data set is a multi-dimension type. For visualization purpose, a **radar diagram** for each PTN can be plotted to show the overall network features; however, this is still hard for the comparison of the whole data set. Moreover,it is not possible to plot all dimensions in an interpretative form such as 2-D or 3-D representation. Therefore, a data reduction technique is needed to do dimensionality reduction while ensuring that most of initial information is retained.

Principal components analysis (PCA) reduces the dimensions of the data set while retaining the variances in the data set as much as possible. To reduce the dimensions, original data set are normalized and projected on a much smaller space. This new space is composed of the principal axes formed by the linear combination of the existing dimension from the initial data set. In addition, the principal axes are the eigenvectors of the covariances matrix of the data set. The axes will be ranked in terms of important for providing high variances comparing the eigenvalue of the axes. Normally, the axes may not specifically represent any specific feature, but correlation analysis helps to indicate the relation between the new axes and the previous data objects features.

4 RESULTS

This chapter presents the analysis result of the thesis. In Section 4.1, provides the network characterization result showing each selected networks and their network characteristics values. Moreover, the correlation analysis is performed on network indicators to identify the relation between selected network indicators. Next, Section 4.2 describes the network classification result from both K-means and hierarchical clustering method. Finally, Section 4.3 interprets the classification results.

4.1 NETWORK CHARACTERIZATION

This section starts with describing the networks data set compiled for the case study. Based on the identified network list, network indicator values are calculated for each network. Before proceeding to the Section 4.2, relation between indicators is investigated via correlation analysis.

4.1.1 Network data set

A network data set is gathered to test the methodology described in Chapter 3. Networks are selected in terms of availability of General transit feed specification (GTFS) data set and the varieties of places. The availability of data depends on the providers such as PT operators, PT authorities. Moreover, several urban PTNs share similar PTNs structure across cities (Lin and Ban, 2013). Additionally, national networks will also be included to explore if they would share similarities with urban networks. All available rail-bound PT modes are selected to represent the PT system of the place because they are mass transit modes, usually carrying a significant number of passengers.

A set of selected PTNs and their corresponding number of stations and edges in directed L-space representation are as shown in Table 4.1. The table provides general information regarding the size and selected structural characteristics of each network. To facilitate the analysis, a set of descriptive statistics is derived as in Table 4.2. Also, a variety of network size ranges from a few stations in Toulouse (37 stations) to many stations in Melbourne (805 stations). Besides, the chosen places are from different continents, including Europe, Australia, North America, and South America, to incorporate more network structure varieties (see Figure 4.1). Note that The Hague and Rotterdam is a metropolitan area which is recently merged for economic cooperation and has a connecting public transport line between the cities (Haag, 2019).

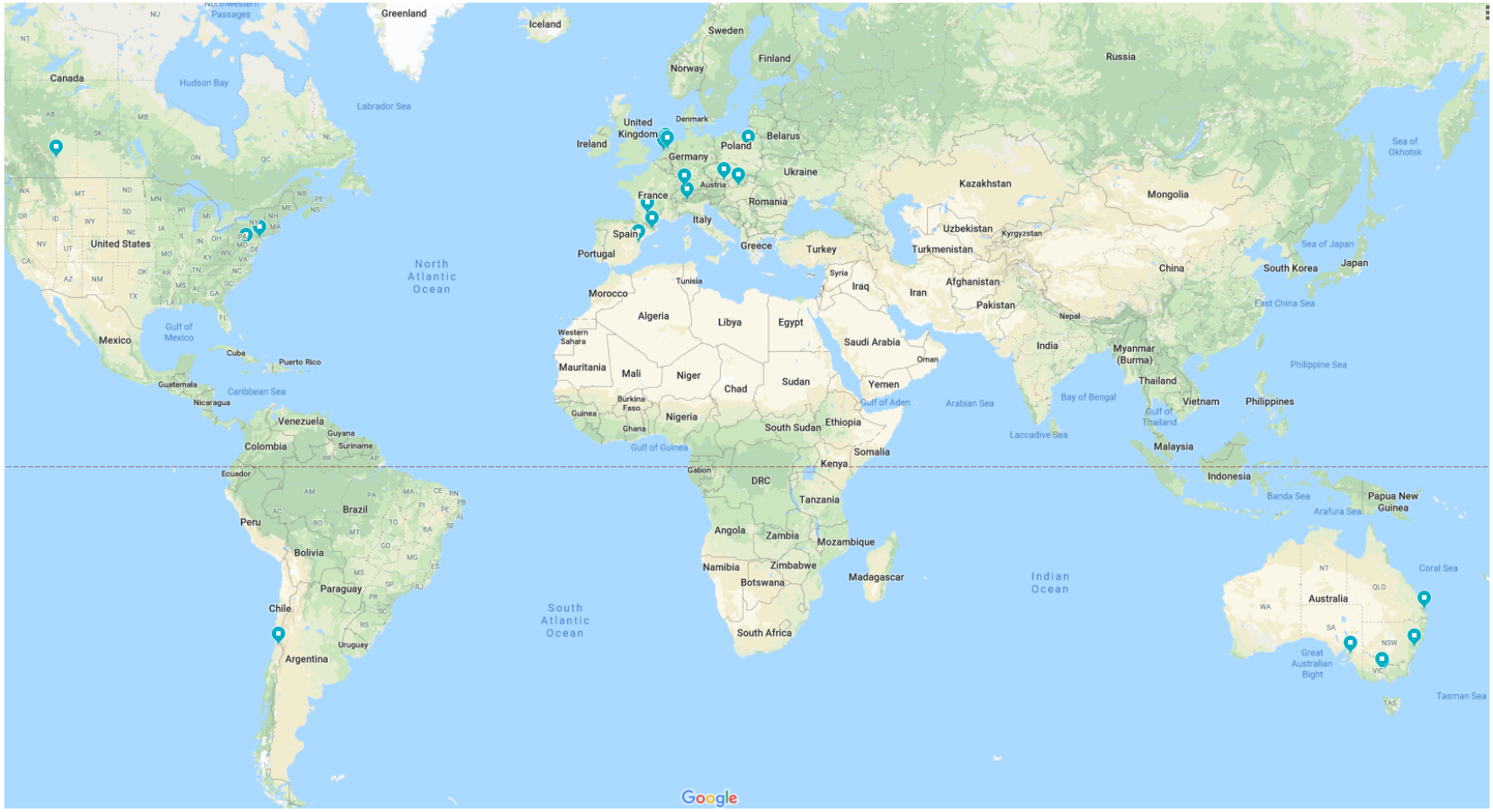


Figure 4.1: Locations of the selected PTNs on the world map

4.1.2 Network indicators result and analysis

All selected PTNs and their corresponding network indicators values are as shown in Table 4.1. Each indicator is analyzed across the networks to gain deeper insights for the selected PTNs and their network characteristics. Moreover, the correlation analysis will identify the relationship between indicators. This relation is essential for network classification to help select important features.

No.	Place	Mode	#stations	#edges	C_B^l	A^l	α^l	c^p	E^p
1	Adelaid	Metro & Tram (M+T)	118	247	0.36	0.08	0.56	0.81	0.6
2	Amsterdam	Metro & Tram (M+T)	196	467	0.1	0.11	0.7	0.56	0.58
3	Bacelona	Metro (M)	131	298	0.15	0.1	0.65	0.76	0.57
4	Brisbanne	Rail & Light rail (R+L)	176	367	0.28	0.07	0.55	0.7	0.62
5	Budapest	Tram (T)	237	470	0.2	0.05	0.5	0.72	0.47
6	Calgary	Tram (T)	46	81	0.23	0.1	0.41	0.77	0.82
7	Melbourne	Tram (T)	805	1676	0.14	0.03	0.54	0.62	0.51
8	Milan	Tram (T)	335	693	0.15	0.05	0.54	0.61	0.52
9	New Jersey	Rail & Light rail (R+L)	184	351	0.2	0.04	0.46	0.78	0.33
10	Santiago	Metro (M)	413	1050	0.12	0.11	0.95	0.83	0.55
11	Sydney	Rail & Light rail (R+L)	136	388	0.2	0.05	0.54	0.67	0.49
12	The Hague &Rotterdam	Metro & Tram (M+T)	307	634	0.26	0.05	0.6	0.61	0.46
13	The Netherlands	Rail (R)	424	929	0.2	0.1	0.78	0.51	0.4
14	Toulouse	Metro & Tram (M+T)	37	72	0.27	0.12	0.52	0.95	0.76
15	Valencia	Metro & Tram (M+T)	134	240	0.21	0.06	0.41	0.71	0.62
16	Victoria	Rail (R)	220	449	0.27	0.05	0.53	0.48	0.56
17	Vienna	Tram (T)	385	815	0.13	0.05	0.56	0.58	0.48
18	Warsaw	Tram (T)	271	509	0.14	0.06	0.45	0.48	0.68
19	Washington	Metro (M)	91	186	0.23	0.09	0.54	0.79	0.67
20	Zurich	Tram (T)	190	400	0.18	0.07	0.56	0.6	0.6

Table 4.1: Structural characteristics of the selected network. C_B^l: Centralisation,A^l: Accessibility, α^l: Robustness, c^p: Service connectivity and E^p: Directness

Trend of network indicators

Before investigating each indicator, the overall indicator trends are analysed using descriptive statistics (see Table 4.2) and histogram (Figure 4.2). Next, PTNs are further analyzed per characteristic.

Each indicator is in different value ranges and roughly non-overlapping. From Table 4.2, while both C_B^l and A^l vary in low scale of value (0.03 - 0.36), α^l, c^p and E^p are in the high value range (0.33 -0.95). Also, this is reflected in the different range for all indicators. Moreover, the low-value group (C_B^l and A^l) also has lower standard deviation when compared to the rest of network indicators. Consequently, both centralization and accessibility values are similar among the selected PTNs and vary in low-value range. On the other hand, more variety of robustness, service connectivity, and directness are found across these networks.

	#stations	#edges	C_B^l	A^l	α^l	c^p	E^p
mean	241.80	516.10	0.20	0.07	0.57	0.68	0.56
median	193	424.50	0.20	0.06	0.54	0.68	0.57
SD	174.78	378.06	0.06	0.03	0.13	0.13	0.12
min	37	72	0.10	0.03	0.41	0.48	0.33
max	805	1676	0.36	0.12	0.95	0.95	0.82
range	768	1604	0.26	0.09	0.54	0.47	0.49

Table 4.2: Descriptive statistics of network data set. C_B^l: Centralisation,A^l: Accessibility, α^l: Robustness, c^p: Service connectivity and E^p: Directness

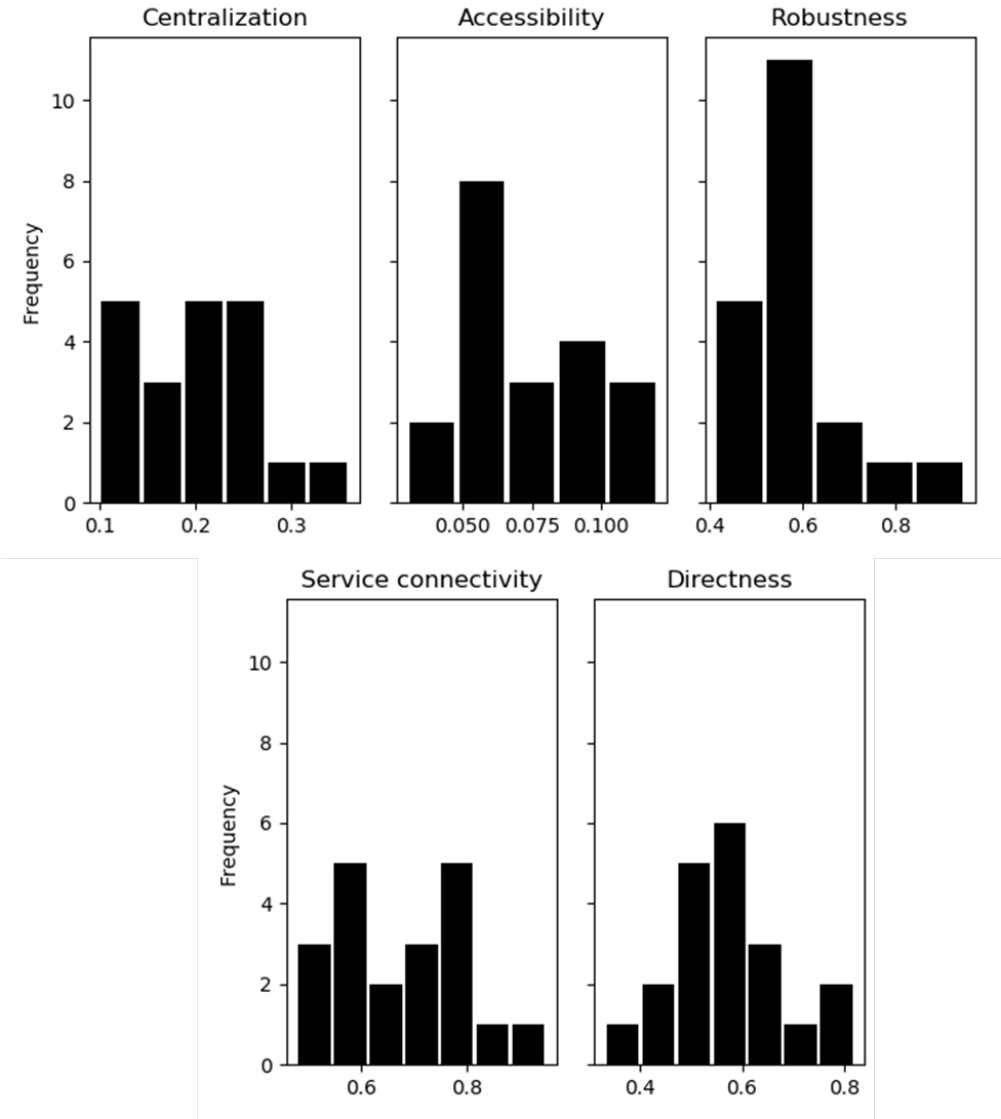


Figure 4.2: Histogram for all selected network characteristics

To further analyse network indicators trend across networks, histograms of all indicators are plotted as in Figure 4.2. There are three distribution patterns for this data set. Both the distribution of accessibility and robustness are skewed right. This implies the network data set contains mostly low-level of accessibility and robustness PTNs. Besides, the centralization and directness resemble the bell-shaped curve, which equally separates the low and high performing network groups. This data set maintains the balance number of networks when considering the centralization and directness. Finally, the service connectivity is found to exhibit bi-modal distribution separating networks into two groups. Surprisingly, this finding is unexpected because the service connectivity (number of direct paths) and directness (average number of transfer) seems to be related so that one would expect similar distribution.

For more specific observations on each network, PTNs are ranked per characteristic as shown in Table 4.3. There are three crucial observations. First, the less centralized network likely results in both highly accessible and robust network as in the case for Amsterdam and Santiago network. In other words, the network accessibility increase when the well-connected nodes are evenly spread in the networks. Besides, the network is also highly robust in the easily accessible network because the connection of nodes in the directed graph generates cycles. Second, the robustness is inversely related to the directness as for the case for The Netherlands and Calgary. This is possible because a network with more cycles will likely require number of transfer (less direct network). Finally, service connectivity and directness are sometimes inversely related. This is surprising since the network with more number of direct routes should result in fewer number of transfers; however, the exception is Warsaw. It is likely possible that although there is fewer direct route, the provided routes do not require that many transfers.

	Centralization	Accessibility	Robustness	Service connectivity	Directness
Top Three	Adelaide (M+T)	Toulouse (M+T)	Amsterdam (M+T)	Adelaide (M+T)	Calgary (T)
	Brisbane (R+L)	Santiago (M)	Santiago (M)	Santiago (M)	Warsaw (T)
	*	Amsterdam (M+T)	The Netherlands (R)	Toulouse (M+T)	Toulouse (M+T)
Bottom Three	Amsterdam (M+T)	Melbourne (T)	Warsaw (T)	Warsaw (T)	New Jersey (R+L)
	Santiago (M)	New Jersey (R+L)	Calgary (T)	Victoria (R)	The Netherlands (R)
	Vienna (T)	*	Valencia (M+T)	The Netherlands (R)	The Hague & Rotterdam (M+T)

Table 4.3: Network ranking by each structural characteristics. * indicates that there are various networks sharing the same rank. M: Metro, T: Tram, R: Rail

Correlation of network indicators

The goals of the correlation analysis are twofold: **correlation between indicators and network size** and **correlation among network indicators themselves**. The former is to evaluate the extent to which network indicators are independent of the network size quantified by either the number of stations and edges. A low level of correlation could facilitate the fair comparison as not only the more extensive network should perform better but also the small well-connected network. Moreover, the latter verifies the correlation among indicators because strongly correlated indicators tend to reveal information. Note that the following discussion only include the statistically significant correlation.

The correlation among the chosen network indicators and both network size indicators, number of edges and stations are as shown in Table 4.4. The blank part of the table (excluding the principal diagonal of one) are elements identical to those in the bottom left part because a correlation matrix is symmetric. Accessibility (A^l), service connectivity (c^P) and directness (E^P) are moderately and negatively correlated to both the number of stations and edges. This is because those indicators' formulas contains the number of stations in the denominator. In other words, all three indicators are smaller as the network grows larger. Since the number of stations and edges are very strongly correlated, the relation also applied to the number of edges. However, it is assumed that these indicators could still provide the acceptable comparisons between PTNs with different sizes as the correlation is moderate.

	C_B^l	A^l	α^l	c^p	E^p	#stations	#edges
C_B^l							
A^l	0.06						
α^l	-0.32	0.47*					
c^p	0.32	0.42	0.002				
E^p	0.23	0.51*	-0.28	0.31			
#stations	-0.34	-0.60**	0.08	-0.54*	-0.53*		
#edges	-0.35	-0.51*	0.20	-0.54*	-0.56*	0.99**	

Table 4.4: Pearson's r correlation result. Note * indicates the statistical significance at the level of 0.05 and ** for the level of 0.01 (two-tailed). C_B^l: Centralisation,A^l: Accessibility, α^l: Robustness, c^p: Service connectivity and E^p: Directness

In addition, the inter-correlation between network indicators are identified. The centralization is positively and moderately correlated to service connectivity and directness. In a high centralized network, the network is comparable to hub and spoke network. Under such network, the service connectivity is high as the network can be viewed as several clusters which are well-connected within the group and weakly link to inter clusters. Also, the directness is also high since most nodes are in cluster even though extra number of transfers are required when tranversing between groups. On the other hand, the robust network (α^l) tends to be less direct (E^p) and low centralized (C_B^l) network as referring to the negative correlation. This is logical since a robust network has many cycles creating longer path between nodes while restricting the occurrence of very centralized nodes. Al-

though this relation reveal insight of how the indicators correlated, they are not statistically significant in this data set.

The only two statistically significant inter-correlation between indicators include accessibility. They are accessibility - robustness and accessibility - directness. The former pair suggests that the nodes in highly accessible network which are closer together (few links between nodes) are likely to create cycles (high robust network) in the context of the directed graph. Moreover, the latter pair suggest that the network will result in lower number of transfers (high directness) if nodes are closer together. These relations confirms the observation found in previous discussion regarding the Trend of network indicators.Note that although pearson's r implies that both pairs are linearly and positive relation, it is not the case here. From Figure 4.3, it can be seen that the relation seems to be similar to an exponential growth curve. In other words, as A^l increases, both α^l and E^P also increase exponentially , so the relations are not linear. Although the extent of correlation is not strong or very strong, the classification process will performed in two cases,including and excluding A^l, to test whether this correlated indicator could make the difference to the realized groups.

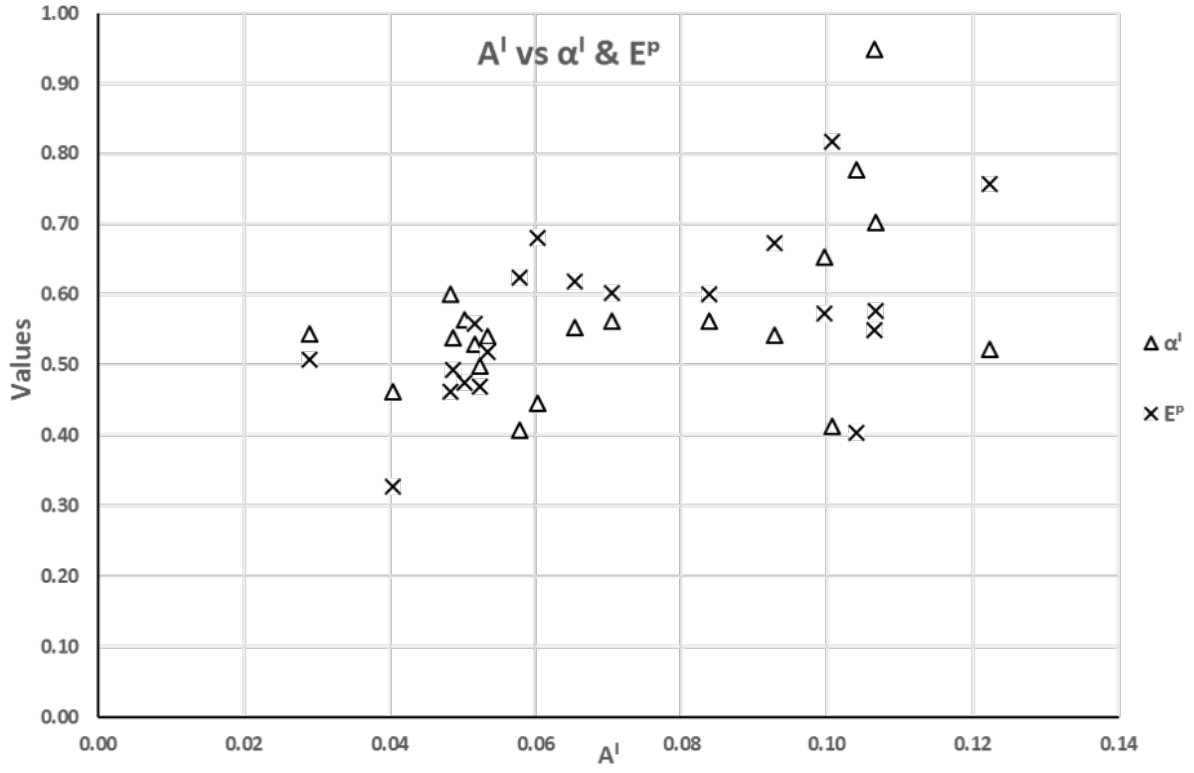


Figure 4.3: Scatter-plot between accessibility (A^l) and robustness (α^l) and directness (E^p)

To summarize, in this section , structural network characteristics are calculated for all network in the data set. To assess the indicators, both the trend testing and correlation analysis are used.

- Although the indicator formulas are normalized for the comparison between networks, their scale and range of values are crucially different. This implies another normalization is required to ensure the clusters are based on all indicators, not influenced by one with greater range of values.
- Accessibility is found to moderately correlate to robustness and directness. Clustering analysis will be separated into two cases: including or excluding accessibility.

4.2 NETWORK CLASSIFICATION

4.2.1 Data pre-processing

Before performing data clustering, three evaluation methods are applied to ensure the quality of clusters. Those three analysis are composed of **assessing clustering tendency**, **determining number of clusters** and **data normalization**.

Assessing clustering tendency

The raw data for clustering analysis is as shown in Table 4.1. Hopkins statistics (H) is measured to analyse if this data set is clusterable. As discussed in Section 3.3, H is compared with the beta quantile value with the number of sampling data as the parameter and the data is clusterable when H value is lower. Moreover, another metric assessing clusterability is the Hopkins value (Adolfsson et al., 2019). This counts the number of times when H value is lower than beta quantile.

The calculation result are shown in Table 4.5. As the data set is quite small, the hopskins value is calculated based on several percentage of sample points to investigate the clustering pattern in the data set. Since the Hopkins value from all simulations is equal to 1, this indicate that the data set is clustable when adapting the statistical significant level of 0.05 for the beta quantile distribution.

Percentage of sample points	Beta quantile	Hopkins value
10	0.865	1.00
15	0.811	1.00
20	0.77	1.00

Table 4.5: Clustering tendency of data set

Determining number of clusters

To ensure the quality of clustering analysis result, a number of clusters in the considered data set is predetermined by two methods: **rule of thumb** and **elbow method**.

From rule of thumb, a set of data containing 20 data points, the number of clusters is $\sqrt{\frac{20}{2}}$ or 3.16 and the number of elements in each clusters is expected to be $\sqrt{2 * 20}$ or 6.32.

Moreover, the graph from the elbow method is shown in Figure 4.4. According to Figure 4.4, the point where the slope of the points change are between three and four. Therefore, the expected number of clusters can be three or four which is in line with the rule of thumb. Therefore, both k-means and hierarchical clustering are performed with the condition of having 3 or 4 clusters.

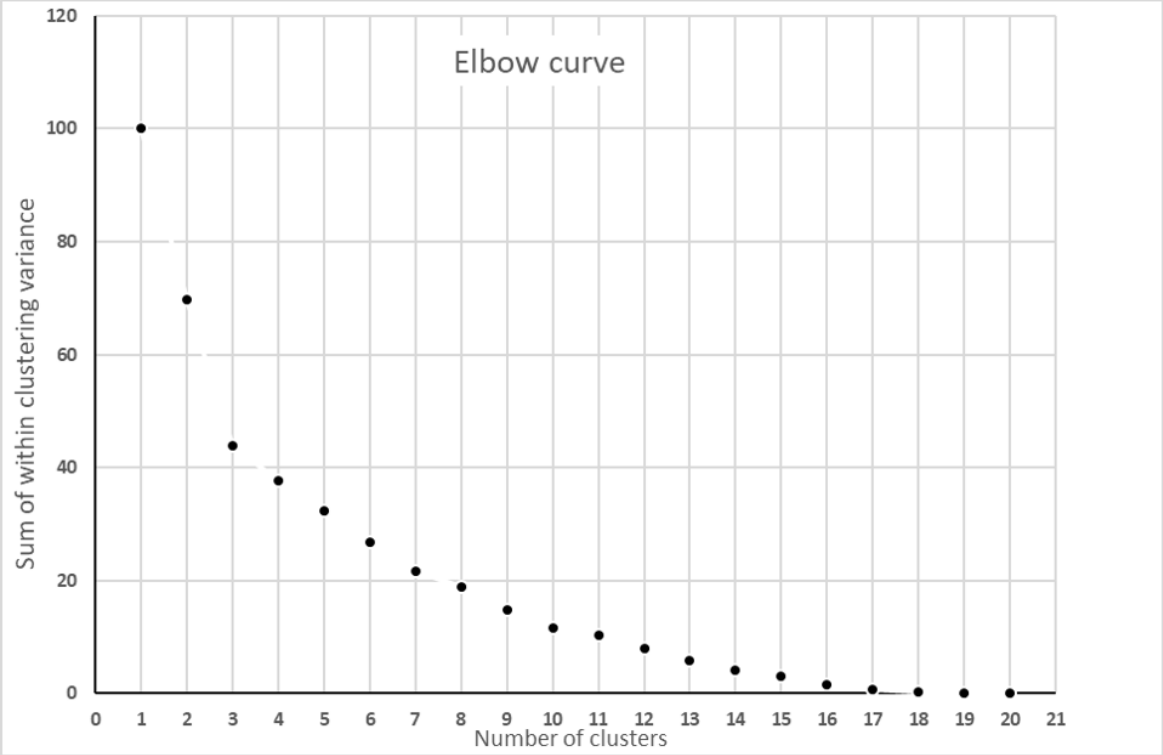


Figure 4.4: Elbow curve

	Rule of thumb method	Elbow method
Number of clusters	3.16	3 - 4

Table 4.6: Summary of predetermined number of clusters

Data normalization

As all selected indicators have different value ranges, this can affect the clustering analysis. The large value-range indicator tends to influence the cluster groups and outweigh other indicators. Min-max normalization (Equation 3.10) is employed to ensure that all

indicators are equally weighted in the clustering analysis. In addition, the new minimum value is 0.05 because this is the smallest value based on the raw data set (see Table 4.1) and 0 would provide misleading interpretation. The processed data set is as shown in Table 4.7. After normalization, all indicators range from 0.05 to 1 while preserving the relationship among the original data values. Both K-means and hierarchical clustering will be performed based on the normalized indicators.

No.	Place	Mode	C_B^l	A^l	α^l	c^p	E^p
1	Adelaide	Metro & Tram (M+T)	1.00	0.61	0.32	0.72	0.58
2	Amsterdam	Metro & Tram (M+T)	0.05	0.84	0.57	0.22	0.54
3	Barcelona	Metro (M)	0.23	0.77	0.48	0.61	0.53
4	Brisbane	Rail & Light rail (R+L)	0.69	0.42	0.31	0.49	0.62
5	Budapest	Tram (T)	0.40	0.29	0.21	0.55	0.33
6	Calgary	Tram (T)	0.51	0.78	0.06	0.63	1.00
7	Melbourne	Tram (T)	0.19	0.05	0.29	0.34	0.40
8	Milan	Tram (T)	0.23	0.30	0.28	0.31	0.42
9	New Jersey	Rail & Light rail (R+L)	0.41	0.17	0.15	0.66	0.05
10	Santiago	Metro (M)	0.11	0.84	1.00	0.76	0.48
11	Sydney	Rail & Light rail (R+L)	0.40	0.25	0.28	0.43	0.37
12	The Hague &Rotterdam	Metro & Tram (M+T)	0.62	0.25	0.39	0.32	0.31
13	The Netherlands	Rail (R)	0.42	0.81	0.70	0.11	0.20
14	Toulouse	Metro & Tram (M+T)	0.67	1.00	0.25	1.00	0.89
15	Valencia	Metro & Tram (M+T)	0.45	0.34	0.05	0.51	0.63
16	Victoria	Rail (R)	0.69	0.28	0.26	0.05	0.50
17	Vienna	Tram (T)	0.13	0.27	0.33	0.26	0.34
18	Warsaw	Tram (T)	0.20	0.37	0.12	0.05	0.74
19	Washington	Metro (M)	0.52	0.70	0.29	0.68	0.72
20	Zurich	Tram (T)	0.33	0.47	0.32	0.29	0.58

Table 4.7: Normalized Structural characteristics of the selected network. C_B^l: Centralisation,A^l: Accessibility, α^l: Robustness, c^p: Service connectivity and E^p: Directness

4.2.2 K-means and principal component analysis (PCA) result

K-means result

The resulting clusters from k-means are as shown in Table 4.8. As A^l moderately correlated with robustness and directness (see Table 4.4), four different cases are derived based on a predetermined number of clusters and the data set with or without A^l. Brisbane and Valencia are assigned to different clusters when comparing the pair of 3-cluster cases while five more networks, including Barcelona, Milan, Victoria, Melbourne and Vienna, are added to the list in the case of 4-cluster pairs. There seem to be distinct differences when excluding accessibility from the data set. Therefore, another measure is needed to compare different clustering cases.

The clustering quality for each case can be compared through average silhouette coefficient as shown in Table 4.9. The "Mean" in the last column is the average silhouette coefficient considering all samples for each case. The 3-cluster case, including A^l performs best while 4-cluster group without A^l performs worst based on the "Mean" value. Surprisingly, the cluster containing many same modes of transportation network has high average silhouette value such as the cluster 2 in 3-cluster. This could imply that network mode influences the network characteristics. In the excluding accessibility case, both 3-cluster and 4-cluster result in low silhouette coefficient, because some clusters are not compact as shown by either very close to zero or negative silhouette coefficient. The negative value usually suggests that the assignment of PTNs to that cluster is not suitable because they are more similar to other clusters. Therefore, A^l will be kept for further analysis as it seems to influence the clustering result.

No.	Place	Mode	3-cluster	3-cluster omit A^l	4-cluster	4-cluster omit A^l
1	Adelaide	Metro & Tram (M+T)	1	1	1	1
2	Amsterdam	Metro & Tram (M+T)	3	3	3	3
3	Barcelona	Metro (M)	3	3	3	4
4	Brisbane	Rail & Light rail (R+L)	2	1	4	1
5	Budapest	Tram (T)	2	2	4	4
6	Calgary	Tram (T)	1	1	1	1
7	Melbourne	Tram (T)	2	2	2	4
8	Milan	Tram (T)	2	2	2	4
9	New Jersey	Rail & Light rail (R+L)	2	2	4	4
10	Santiago	Metro (M)	3	3	3	3
11	Sydney	Rail & Light rail (R+L)	2	2	4	4
12	The Hague &Rotterdam	Metro & Tram (M+T)	2	2	4	4
13	The Netherlands	Rail (R)	3	3	3	3
14	Toulouse	Metro & Tram (M+T)	1	1	1	1
15	Valencia	Metro & Tram (M+T)	2	1	4	1
16	Victoria	Rail (R)	2	2	4	2
17	Vienna	Tram (T)	2	2	2	4
18	Warsaw	Tram (T)	2	2	2	2
19	Washington	Metro (M)	1	1	1	1
20	Zurich	Tram (T)	2	2	2	2

Table 4.8: K-means clustering result for 3 and 4 clusters case when include and exclude accessibility. Note the number denotes the cluster number.

	Cluster1	Cluster2	Cluster3	Cluster4	Mean
3-cluster	0.36	0.40	0.25	-	0.34
4-cluster	0.25	0.31	0.09	0.29	0.23
3-cluster omit A^l	0.29	0.31	0.004	-	0.20
4-cluster omit A^l	-0.06	0.28	0.10	0.26	0.15

Table 4.9: Average silhouette coefficient of each cluster for 3 and 4 clusters cases when include and exclude accessibility

In addition to clustering quality evaluation, the physical interpretation for each cluster is another essential aspect as this was not a part of the clustering process. However, the k-means result in tabular form as in Table 4.8 is difficult to interpret, so a graphical representation is adopted. On this graph, each network is plotted according to their properties to analyse their relative position to other networks, yet there are too many dimensions to take into account. Therefore, PCA analysis is adopted to reduce the number of dimensions (network properties) to two or three, which are visually interpretative, and would allow for further interpretation.

Principal component analysis result

Principal component analysis is a dimensional reduction process in which data points are projected on to new axes while maximizing the variance of the projected points. In the following paragraphs, PCA is carried out step-by-step along with the intermediate results. After acquiring principal components (PCs), all 20 networks with k-mean group labels will be plotted on a new axis or PC to illustrate the relationship between PTNs.

PCA is carried out as the following steps. First, compute the eigenvectors and eigenvalues of the covariance matrix (Table 4.10) of all network indicators. The result is shown in Table 4.11. Note these eigenvectors are then called PCs and are ranked according to their eigenvalue (variance as in Table 4.11). The first PC is the one holding the most variance of the data. Second, project the normalized network data (Table 4.7) on PCs and their transformed value are as shown in Table 4.12.

	C_B^l	A^l	α^l	c^p	E^p
C_B^l	0.0548	0.0036	-0.0161	0.0188	0.0116
A^l	0.0036	0.0728	0.0272	0.0285	0.0304
α^l	-0.0161	0.0272	0.0466	0.0001	-0.0135
c^p	0.0188	0.0285	0.0001	0.0621	0.0172
E^p	0.0116	0.0304	-0.0135	0.0172	0.0482

Table 4.10: Co-variance matrix of all network indicators. C_B^l: Centralisation, A^l: Accessibility, α: Robustness,c^p: Service connectivity ,E^p: Directness

	C_B^l	A^l	α^l	c^p	E^p	Variance
PC1	0.24	0.68	0.12	0.53	0.43	0.42
PC2	-0.58	0.38	0.65	-0.20	-0.22	0.29
PC3	-0.35	0.17	-0.34	-0.53	0.67	0.15
PC4	0.69	0.24	0.27	-0.62	-0.06	0.11
PC5	0.08	-0.55	0.61	0.09	0.56	0.03

Table 4.11: Principle component ranking according to their variance or eigenvalue. Note PC denote principal component and the number signifies the importance of the component. C_B^l: Centralisation, A^l: Accessibility, α: Robustness,c^p: Service connectivity ,E^p: Directness

No.	Place	PC 1	PC 2	PC 3	PC 4	PC 5
1	Adelaide	0.389	-0.371	-0.273	0.263	0.036
2	Amsterdam	0.070	0.539	0.244	0.038	-0.084
3	Barcelona	0.259	0.274	-0.016	-0.123	-0.054
4	Brisbane	0.082	-0.237	-0.050	0.135	0.107
5	Budapest	-0.182	-0.127	-0.164	-0.141	-0.058
6	Calgary	0.494	-0.268	0.342	-0.080	-0.030
7	Melbourne	-0.464	-0.020	0.003	-0.195	0.129
8	Milan	-0.294	0.051	0.064	-0.095	-0.001
9	New Jersey	-0.330	-0.181	-0.411	-0.231	-0.172
10	Santiago	0.401	0.687	-0.249	-0.139	0.200
11	Sydney	-0.241	-0.084	-0.103	-0.061	0.022
12	The Hague, Rotterdam	-0.264	-0.106	-0.199	0.192	0.062
13	The Netherlands	-0.047	0.494	-0.104	0.411	-0.158
14	Toulouse	0.852	-0.203	-0.013	-0.085	-0.056
15	Valencia	-0.043	-0.301	0.105	-0.131	-0.018
16	Victoria	-0.303	-0.203	0.094	0.369	0.054
17	Vienna	-0.394	0.147	0.045	-0.121	-0.015
18	Warsaw	-0.275	-0.036	0.490	0.004	0.011
19	Washington	0.377	-0.104	0.033	-0.048	0.002
20	Zurich	-0.087	0.049	0.164	0.036	0.024

Table 4.12: Normalized network indicators values projected on Principal components (PC)

For the graphical purpose, not more than three PCs will usually be selected. Moreover, each PC retains a different amount of data quantified by eigenvalue of the PC. For simple visualization, Figure 4.5 shows the amount of variance kept by their corresponding number of principal components (PC). Note that y-axis shows the cumulative percentage of the variance. Besides, two or three PCs would retain 70.94 and 86.16 percent of amount of information, respectively. Both percentages are acceptable because they retain a significant amount of information.

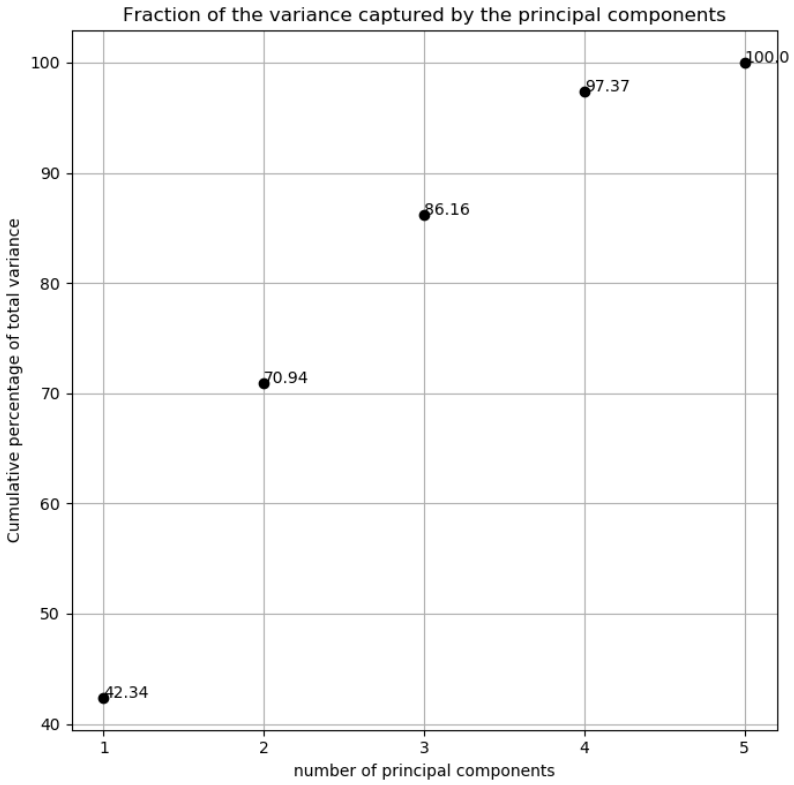


Figure 4.5: Fraction of the variance captured by the principal components

In addition to the amount of information, correlation analysis is applied to identify the relation between PCs and the network indicators. In other words, the analysis aims to investigate the extent to which the first two to three PC can describe or substitute the previous five network indicators. The correlation is indicated by Pearson's statistical significance as shown in Table 4.13. Note that the following observation will only consider the statistically significant correlation. The first PC is strongly correlated to three network indicators: **accessibility, service connectivity and directness**. Moreover, the second PC is very strongly correlated to the **robustness and centralization** of the network. The third PC has a strong correlation with **directness**. For the rest of the PC, they are either moderately or weakly correlated to the network indicators. Since the first two PCs can capture all five network indicators, they are enough to represent this data set for further interpretation.

The projected values on the first and second PC (Table 4.12) for all networks are plotted as in Figure 4.6 and Figure 4.7. Moreover, k-means results from Table 4.8 are integrated into the graph to label each network according to their cluster. Besides, the silhouette coefficient for each network is plotted as a horizontal bar or stripped in its cluster group along with the PCA plotted.

	PC 1	PC 2	PC 3	PC 4	PC 5
C_B^l	0.35	-0.71**	-0.31	0.53*	0.03
A^l	0.87**	0.40	0.13	0.16	-0.18
α^l	0.19	0.86**	-0.33	0.22	0.24
c^p	0.74**	-0.23	-0.44	-0.45*	0.03
E^p	0.68**	-0.29	0.64**	-0.05	0.22

Table 4.13: Correlation analysis between network indicators and principal components.Note * indicates the statistical significance at the level of 0.05 and ** for the level of 0.01 (two-tailed). C_B^l: Centralisation, A^l: Accessibility, α: Robustness,c^p: Service connectivity ,E^p: Directness

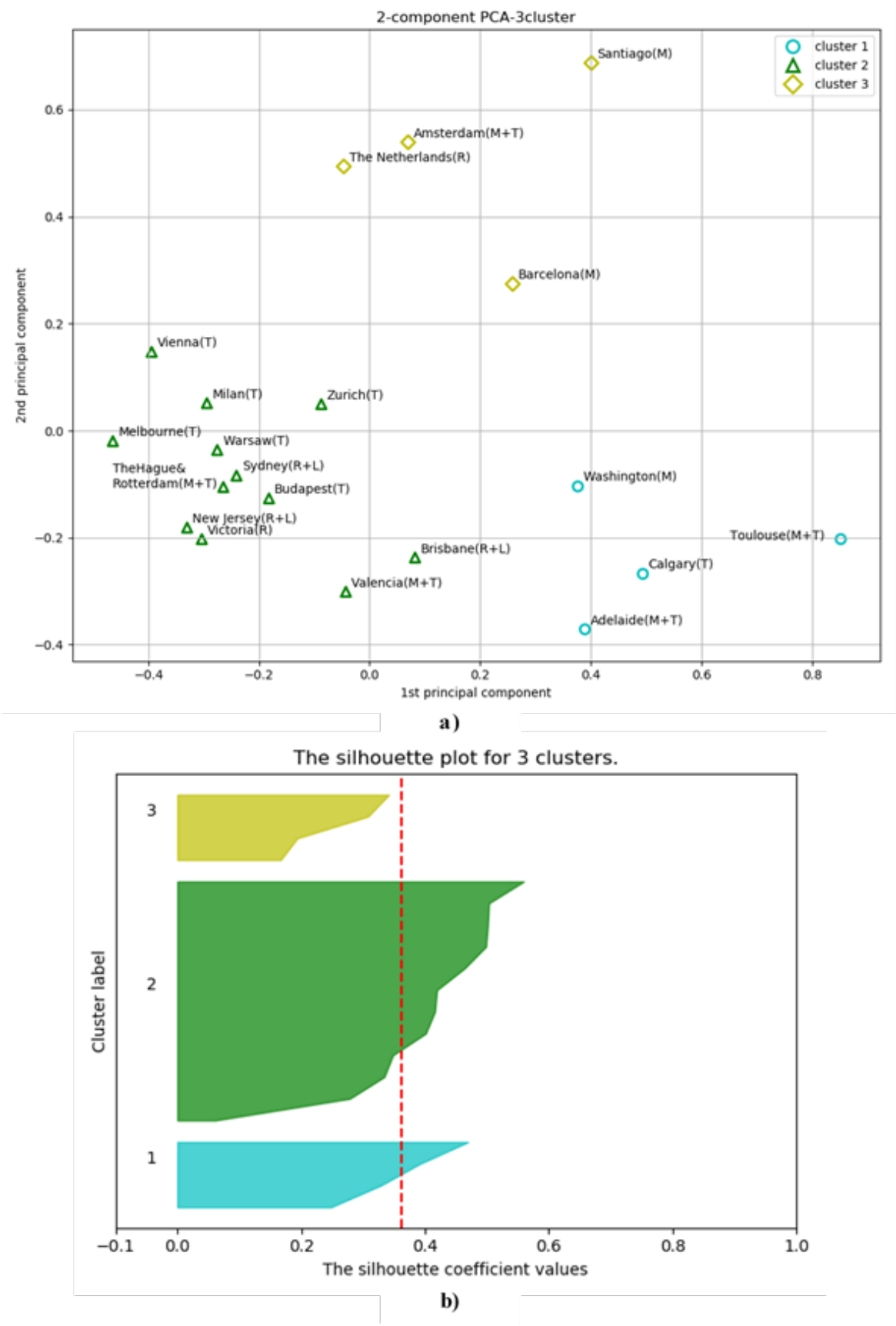


Figure 4.6: 3- cluster group: (a) PTNs plotted on PCA axis with 3-cluster K-mean labels (b) the corresponding silhouette coefficient for this case (The red dotted line show the average value of the silhouette coefficient across all clusters.).Note R: Rail, T:Tram, L: Light-rail, and M: Metro

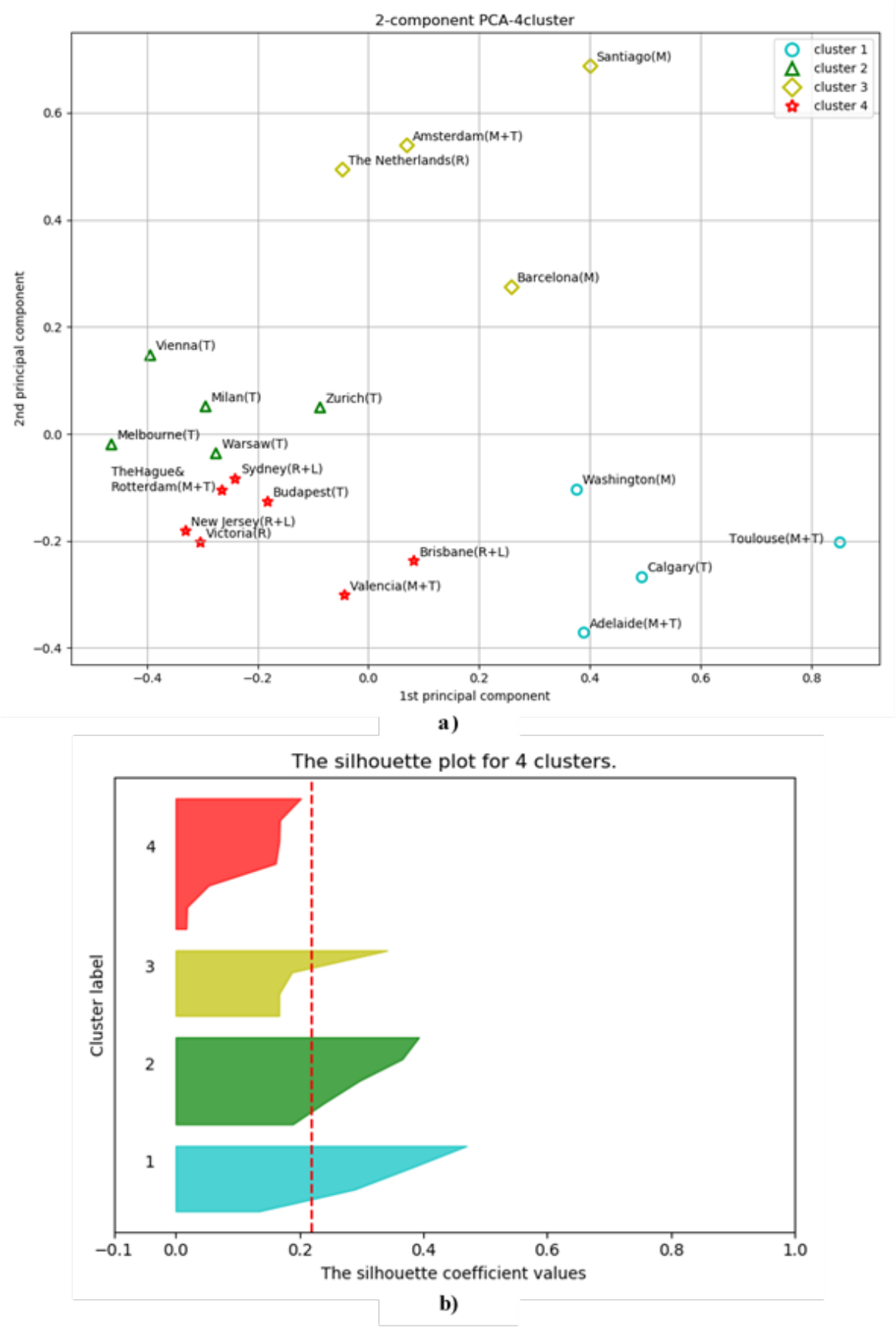


Figure 4.7: 4- cluster group: (a) PTNs plotted on PCA axis with 4-cluster K-mean labels (b) the corresponding silhouette coefficient for this case (The red dotted line show the average value of the silhouette coefficient across all clusters.). Note R: Rail, T:Tram, L: Light-rail, and M: Metro

According to Figure 4.6 (a) and Figure 4.7 (a), the points seems to be organized in either 3 or 4 groups as predetermined in Table 4.6. Cluster 2 in 3-cluster case splits into 2 groups in 4-cluster case, cluster 2 and 4. Interestingly, all network in a new cluster 2 consists of tram networks while the cluster 4 contains the combination of modes, including metro-tram and rail-light rail. Although the average silhouette of 3-cluster group is higher than the 4-cluster case, the number of networks are more evenly spread when compare the silhouette plot for the two-case (see Figure 4.6 and Figure 4.7 (b)). Therefore, the analysis focuses on the 4-cluster case.

For interpretation, the meaning of each axis or PC is required. Each PC is meaningless in the sense that it is the linear combination of all the indicators. However, the correlation analysis between PC and network indicators (Table 4.13) infer the meaning of PC. From the correlation analysis result, the first principal component (PC1) is likely to explain the combination of accessibility, service connectivity and directness while the second principal component (PC2) describes robustness and centralization.

According to Figure 4.7 (a), cluster 1 situates in the lower right quadrant of the graph with positive PC1 value and negative PC2 value. Cluster 2 and 4 stay quite close to each other. While most of cluster 2 networks, excluding Brisbane, are in the lower-left quadrant with both negative PC values, cluster 4 seems to situate in the upper left quadrant. Besides, cluster 1 outperforms cluster 2 and 4 in the combination of three aspects: service connectivity, accessibility and directness (CAD) since the whole group have higher score on PC1 than cluster 2 and 4. Toulouse is the best network while Melbourne is the farthest left when considering CAD aspect. Also, cluster 3 has middle level of CAD which is higher than cluster 2 and 4, but lower than cluster 1.

The data set contains three level of PC2 which describes the robustness and network centralization. Note that the centralization is negatively correlated with PC2, so the network with negative PC2 has high network centralization. As cluster 3 has the highest level among the clusters, their networks have high robustness and low centralization. Cluster 1 and 4 illustrate relative level of PC2 and they are considered to exhibit low robustness and high centralization. Finally, cluster 2 situates in the middle of the high and low group.

Apart from the network characteristics, k-means clusters indirectly differentiate networks according to the size of the network and mode of transportation. From Table 4.14, most large networks belong to cluster 2 while the small ones are in the cluster 1. The rest clusters are similar size and in between those two. In addition, cluster 2 consists of only tram networks (Figure 4.7 (a)). Moreover, all networks except Victoria rail contains either tram or light rail, another form of tram modes. The rest two are the mix of metro and tram modes.

Even though networks are classified into clusters which possess similar network characteristics, there are some varieties inside the cluster. For example, one cluster in the 3-cluster case split into two clusters for the 4-cluster case (see Figure 4.6 and Figure 4.7 (a)). Surprisingly, the new group only consists of tram networks.Therefore, to further investigate the characteristic within each cluster, hierarchical clustering is employed.

	Number of station			Modes
	Min	Max	Average	
Cluster 1	37	118	73	Metro + Tram
Cluster 2	190	805	397.2	Tram
Cluster 3	131	413	219	Metro + Tram
Cluster 4	134	424	240.3	Tram-related

Table 4.14: 4-cluster characteristics summary

To summarize, in this subsection, k-means algorithm is used to find clusters for this PTNs data set. The suitable 'k' (number of clusters) is found to be 3 and 4. From the combination of this fact with accessibility correlation findings, four cases are derived which are 3 and 4-cluster (with accessibility) and 3 and 4-cluster (without accessibility). For visualization purpose, principal component analysis is adopted. The main findings are as follows:

- Although the 3-cluster group has the highest average silhouette coefficient, 4-cluster seems to have a well-spread number of networks among clusters.
- Cluster 1 is a metro-tram group consisting of small networks. All networks perform best on the combination of service connectivity, accessibility and directness (CAD) compared to other clusters. They also have low robustness level (and high centralization).
- Cluster 2 is a tram group and most networks are large in size when compared to other clusters. The group performs the worst on CAD, but has high centralization and low robust level.
- Cluster 3 is a mixed group where most networks are medium in size. They all have high robustness level (and low centralization) while their CAD is in between those of cluster 1 and 2.
- Cluster 4 is a tram-related group which also have CAD similar to cluster 2, but has lower robustness level than cluster 2.

4.2.3 Agglomerative hierarchical clustering

Hierarchical clustering is adopted to identify the similarities or difference within clusters. These are identified when networks are matched during the agglomerative hierarchical process. Besides, there are several methods for pairing up data objects, so three linkage methods composed of ward, complete and average are chosen to compare which would yield the best clustering quality and provide an insightful classification. In addition, the stopping criteria for all methods are when four clusters are formed because the results will be compared with those in Section 4.2.2. In other words, the hierarchical tree is cut when four clusters formed.

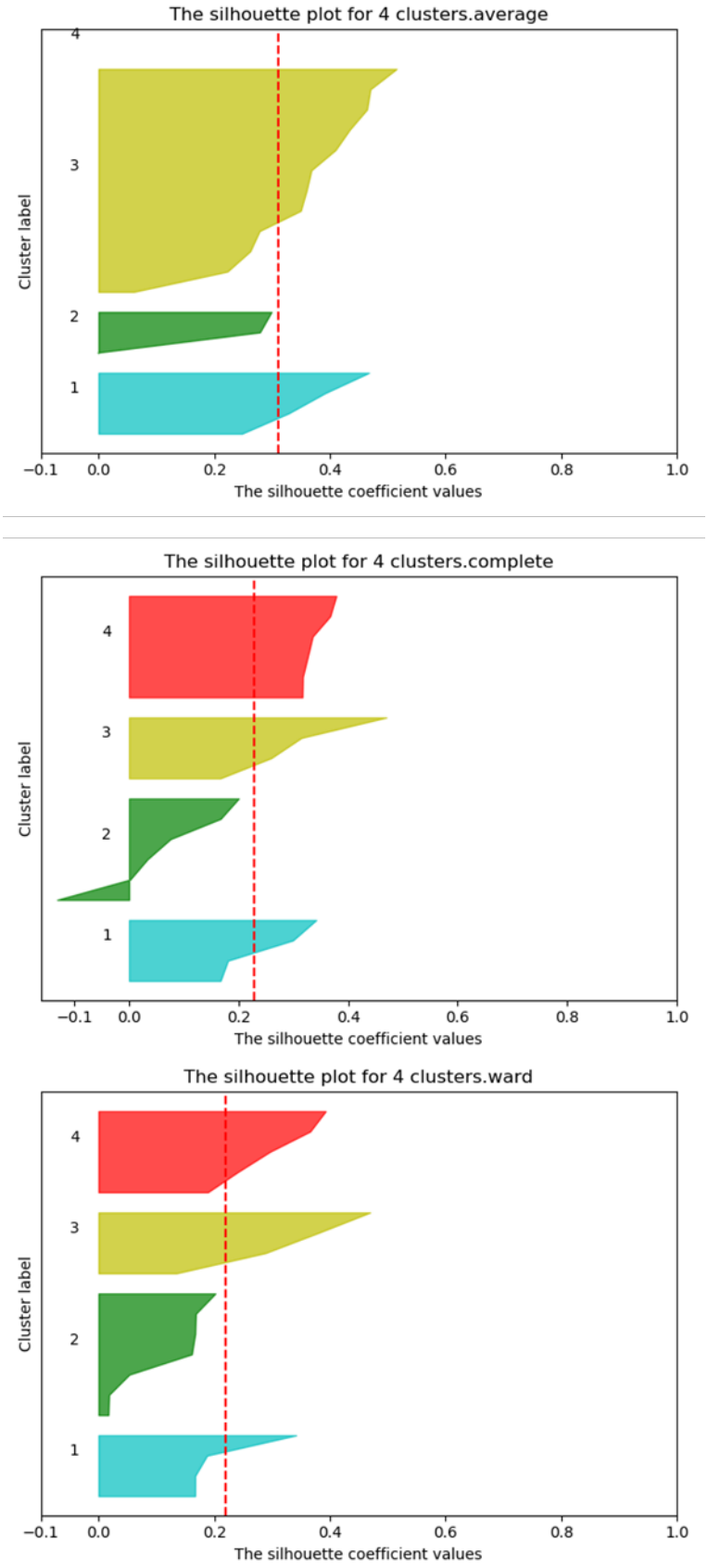


Figure 4.8: The comparison of silhouette plots for all three linkage methods: average, complete and ward (arranged from top to bottom). Note the red dotted line show the average value of the silhouette coefficient across all clusters.

For linkage method comparison, the silhouette plot for each method is illustrated in Figure 4.8. Based on the average silhouette coefficient (as shown by dotted line), the average method has the highest value (0.31) while complete and ward value are 0.23 and 0.22, respectively . However, both average and complete linkage method result in an undesirable result. The networks are unevenly distributed in the average case as cluster 3 contain most networks. Moreover, only three clusters are formed with the average linkage. On the other hand, the complete linkage managed to form four clusters containing equally number of networks, but some of the networks in cluster 2 has the negative silhouette coefficient values suggesting that they were classified in a false cluster. Therefore, although ward method has the lowest average silhouette coefficient value among others, it performs better in equally classifying equally number of networks in each cluster. This linkage method case will be further explored.

The hierarchical results based on ward linkage algorithm are presented in dendrogram shown in Figure 4.9. The tree result is cut to yield four clusters because of the interpretability of the result. Surprisingly, the resulting clusters are identical to the 4-cluster from the K-means result (see Figure 4.7). This allows investigation on each cluster to analyse how each network member is related to each other. For comparison with previous 4-cluster case, each cluster will be referred to as color in the dendrogram with the cluster group number in the parentheses.

According to Figure 4.9, in each cluster network is paired up in an ascending order of their distinct cluster characteristics identified in Section 4.2.2. In other words, network within cluster are situated higher in the tree if it possess higher value of cluster characteristic. For instance, the yellow cluster (cluster 3) networks are all highly robust network and the pairing in the dendrogram shows the rank of the robustness level. Barcelona metro and Amsterdam metro-tram are paired up first since their robustness level are 0.65 and 0.7, respectively. Next, they are matched with the Netherlands train of which robustness level is 0.78. Finally, the three merged with Santiago metro which has the highest robustness level among all networks in this data set. In addition, this pattern is also in other three cluster. The blue cluster (cluster 1) shows the rank of centralization while both the red (cluster 4) and green cluster (cluster 2) exhibit the rank for service connectivity.

To summarize, in this subsection, hierarchical clustering is employed to identify the similarities or differences under each cluster. Three common linkage method, average, complete and ward, are adapted. The main findings are as follows:

- Ward linkage method yields the best clustering quality when considering the formation of four clusters containing similar amount of networks.
- Surprisingly, the result clusters are identical to the result from k-means cluster for the 4-cluster case.
- The resulting tree (dendrogram) illustrates how networks in their clusters are ranked in ascending order according to the cluster distinct characteristic.
- The distinct characteristics can be identified as follows: service connectivity for green (cluster 2) and red (cluster 4) clusters, centralization for blue (cluster 1) cluster and robustness for yellow (cluster 3) cluster.

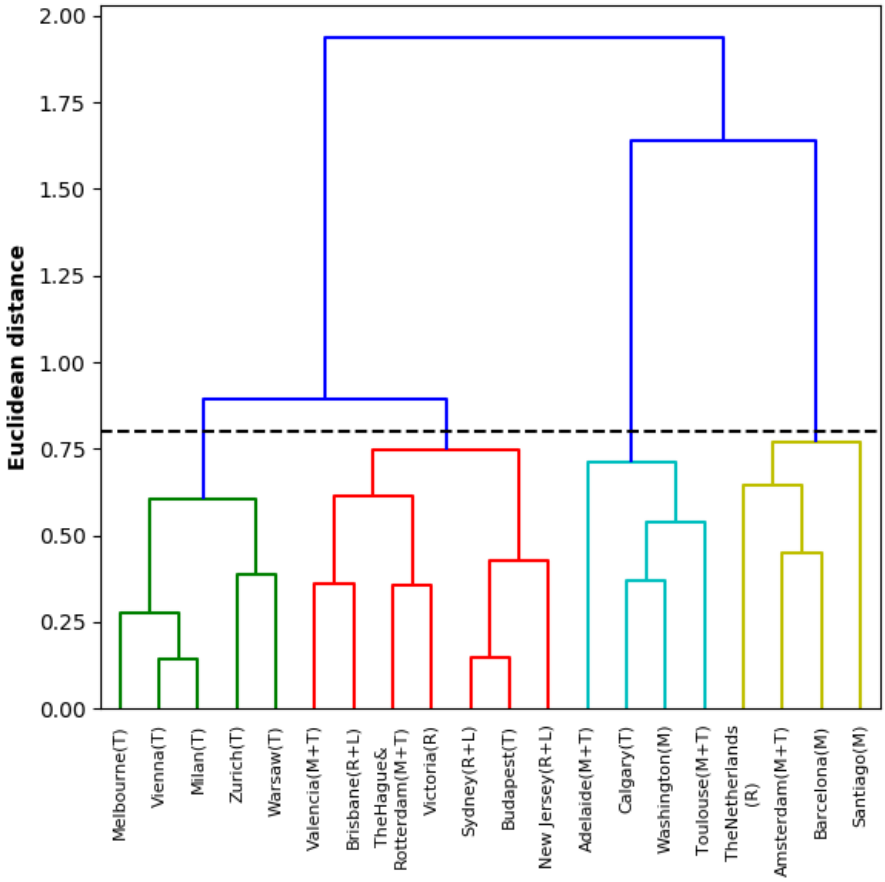


Figure 4.9: Dendrogram of hierarchical clustering adapting ward criteria linkage method. (R: Rail, T:Tram, L: Light-rail, and M: Metro) Note the black horizontal dotted line show the point where the tree is cut to give four clusters.

4.3 RESULT INTERPRETATION

Previously, in Section 4.2, k-means and hierarchical clustering classified the PTNs data set into four clusters. PCA illustrates the former's result on 2-D graphs while dendrogram from the latter method exhibits both the clusters and the ranking of networks within the cluster according to the cluster key properties. However, we do not investigate how the networks are strong in one property, not others. One of the reasons is its network structure. Radar diagram and L-space graph are employed to serve as visual aids describing overall network indicators value and its corresponding infrastructure representation. Note that for comparison purposes, the radar diagram is plotted based on min-max normalization (Table 4.7) to ensure that all indicators span the same whole value range. The radar diagram and its L-space network are as shown in Figure 4.10. Moreover, in the following paragraphs, each cluster will be analysed to identify network features and its characteristics.

For cluster 1 (C1), most members are small-size metro-tram networks. Since the distinct cluster characteristic is centralization as identified in Section 4.2.3, all networks are quite highly centralized because they share similar network structure, radial line structure. They consist of several infrastructure connection lines with very few numbers of overlapping stations or hubs. Consequently, each network possesses very few stations with many links connection, and most of the stations connect to the other two neighbors. Also, the cluster contains Adelaide (M+T), the most centralized network for all networks in this data set.

Network in cluster 2 (C2) are all large-size tram network. They seems to have low value on all the network characteristics. From the L-space graph, every network seems to contain many cycles but it is not enough to compensate for larger number of stations. On the other hand, cluster 4 (C4) seems to score better in overall characteristics (bigger shade region on radar diagram) compared to cluster 2. Cluster 4 (C4) is a tram (or light-rail) related group which refer to the combination of tram with other modes. For those rail and light combination, the networks have higher centralization since their L-space show the radial line structure similar to those in cluster 1.

For cluster 3 (C3), most members are medium-size mixed modes, including metro, tram and rail. They are highly robust and easily accessible compared to other clusters. The former can be realized via the number of cycles in their L-space graph. Moreover, these cycles add extra links or connection to the nodes which describes the high level of accessibility of the networks. Note that although cluster 3 seems to show two distinct characteristics, robustness and accessibility, their level of accessibility is similar to each other. This explain why they were ranked by robustness in Section 4.2.3.

To summarize, in this section, radar diagram describes the overall characteristics of the networks and confirms the findings in the previous section. Moreover, L-space facilitate the visualization of networks and identification of network features contributing to their characteristics. The main findings are as follows:

- Cluster 1 consists of mostly highly centralized networks as their L-space graph looks similar to the radial line structure.
- Cluster 2 and 4 contain similar network size while the latter seems to outscore the former on most characteristics. The combination of rail and light rail results in a high centralized network as in Brisbane and New Jersey.
- Cluster 3 consists of mixed modes with high robustness due to the present of number of cycles in the networks.

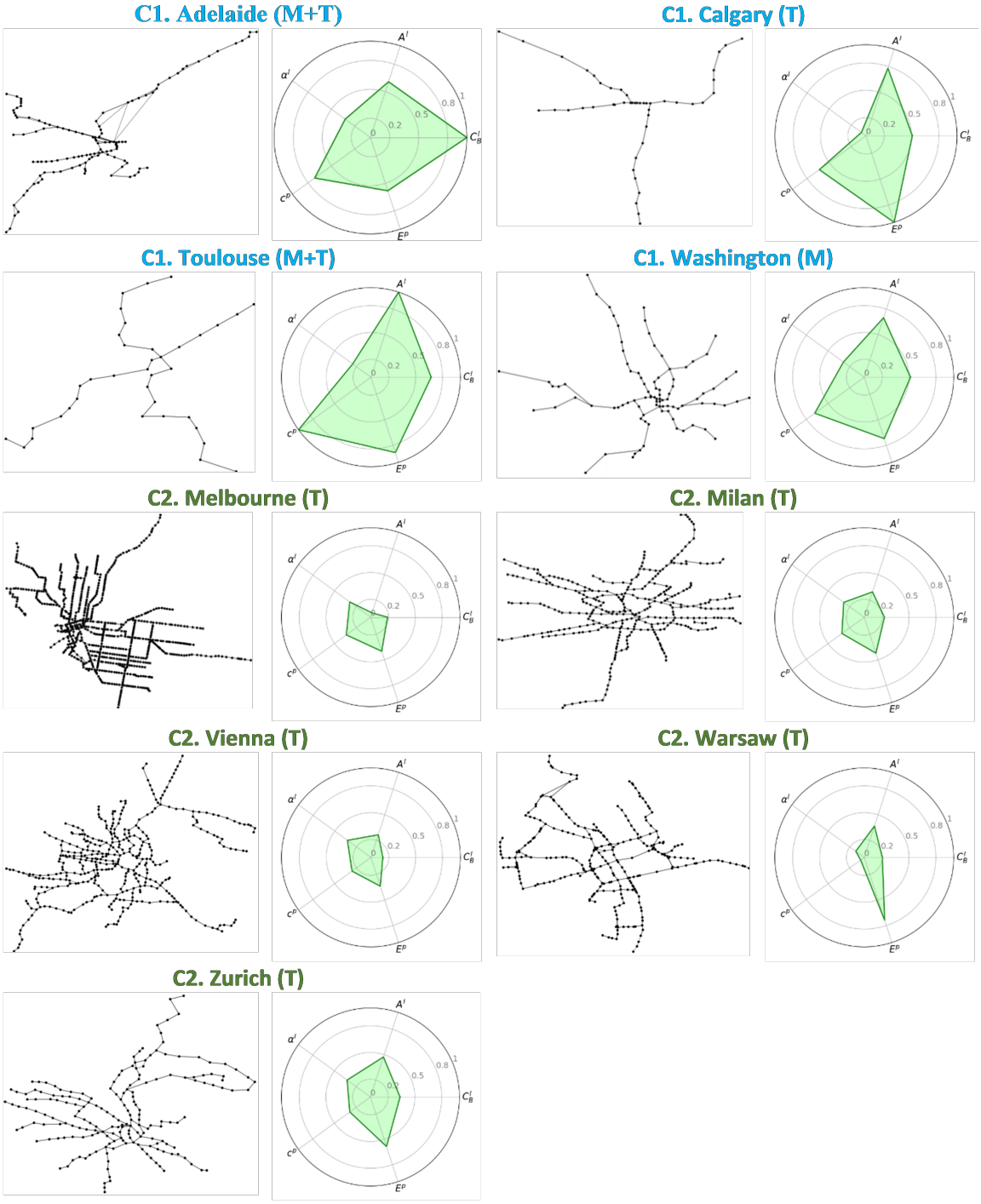


Figure 4.10: Radar diagram and its L-space network for all networks.C_B^l: Centralisation, A^l: Accessibility, α: Robustness,c^p: Service connectivity ,E^p: Directness. Note C1,C2,C3 and C4 indicates the cluster for each network referred to the K-means result in 4-cluster case (see Table 4.8)

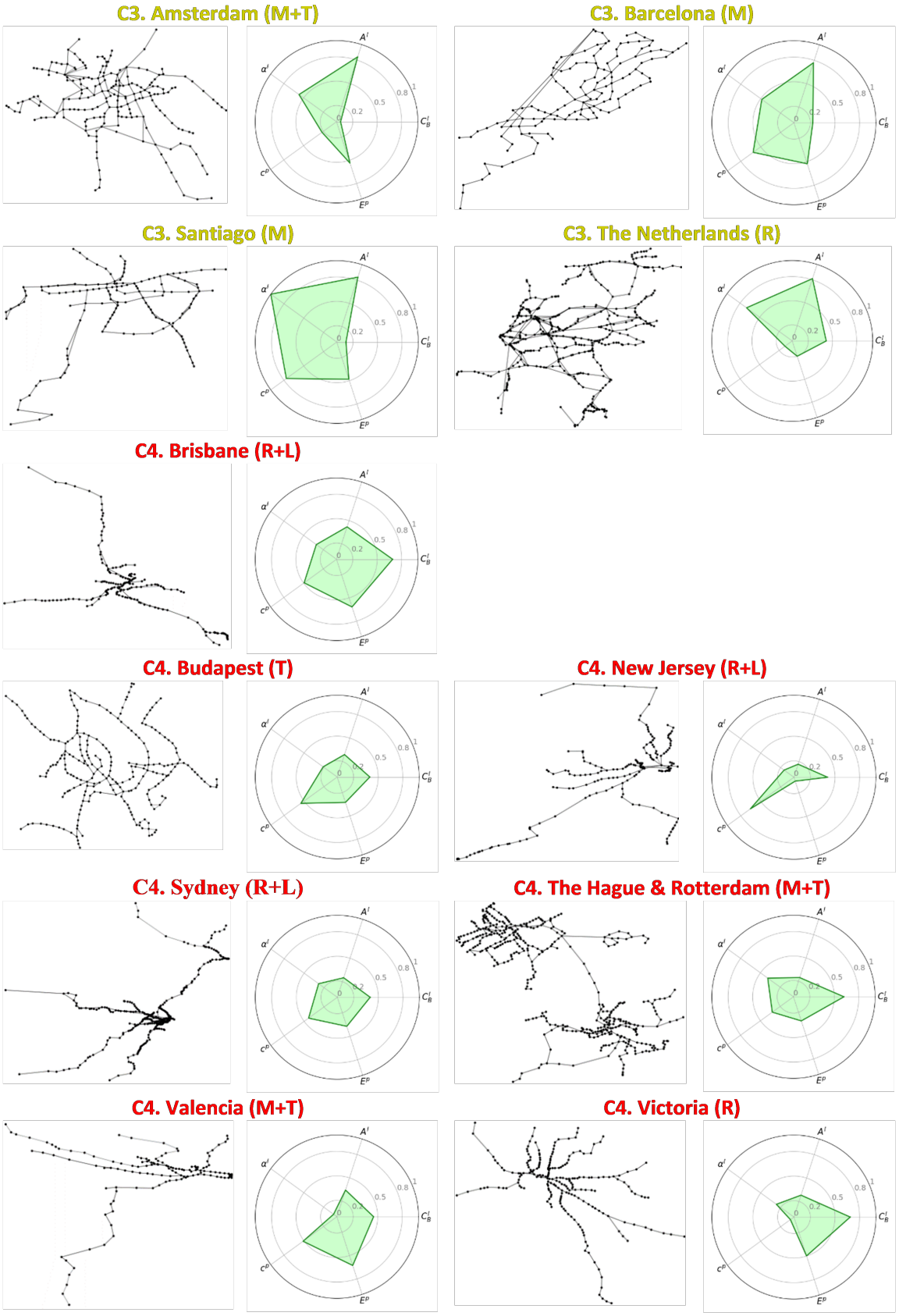


Figure 4.10: (cont.) Radar diagram and its L-space network for all networks.C_B^l: Centralisation, A^l: Accessibility, α: Robustness,c^p: Service connectivity ,E^p: Directness. Note C1,C2,C3 and C4 indicates the cluster for each network referred to the K-means result in 4-cluster case (see Table 4.8)

5 CONCLUSIONS AND RECOMMENDATIONS

5.1 KEY FINDINGS

In this section, main findings of the thesis are discussed. After revisiting the main research question, the sub-research question is introduced with its answer one by one.

The overarching research question is as follows:

"Which clusters of public transport networks can be identified using the topological characteristics of those networks excluding mode-specific properties? All rail modes (tram, metro, train) are considered."

To dissect this question, two sub-research questions were posed as followed:

1. *How can the PTNs' structure be quantitatively characterized by network indicators from a topological approach excluding mode-specific properties?*

   PTNs characterization from a topological approach requires two components: graph type and structural characteristics.

   Since PTNs were represented on a graph, the first component is to select graph type. It refers to the properties of the links, including direction and weight. While direction specifies whether the links are uni-directional or bi-directional, weight can refer to the importance of links such as distance, frequency, time, etc. In this thesis, a directed unweighted graph was chosen because it will not impose mode-specific properties on the network and also model the link direction in detail. For example, if the distance were weighted on the links, the railway and tram network would behave differently since the two systems usually cover different area scale. For the direction, we want to model PTNs more realistically based on the data gathered without assuming that the links are bi-directional.

   The second component is the structural characteristics. The thesis aims to classify PTNs according to their spatial infrastructure and service connection in the normal network state. Consequently, five characteristics were chosen: centralization, accessibility, robustness, service connectivity and directness. The first three dealt with the infrastructure side while the last two focus on the service network. **Centralization** assessed the distribution of hub nodes throughout the network while **accessibility** measured how far node is from other nodes. **Robustness** measures the network's potential to deal with the disruption. For the service network, **service connectivity** quantifies the number of direct service route and **directness** investigates the average number of transfer in the system. Based on these five characteristics, network indicator and their corresponding network models are selected, as shown in Table 3.1.

   The indicators are standard graph theory indicators which often used to characterize spatial networks while two standard network models were employed in this thesis: L-space and P-space Lin and Ban (2013). The network indicators include betweenness

centralization, closeness centrality, alpha coefficient, global efficiency, and clustering coefficient. For the network model, L-space represents the infrastructure connection, while P-space refers to the service connection. Based on these models, the network indicators are enriched in its interpretation. For instance, global efficiency (E^p) is calculated in P-space to exhibit the network directness. The interpretation is different from the case where E^p is calculated in L-space because that would show the directness provided in the infrastructure network instead of service one.

2. *What are the classification groups of PTNs based on structural characteristics*

   Two clustering techniques, k-means, and hierarchical clustering, were employed to classify the network data set. While the former partition PTNs into non-overlapping groups, the latter also reveal the level in the cluster. Moreover, the dendrogram shows the rank of PTNs according to the cluster's key characteristic in ascending order. This is to say the PTNs with lower scores of such a characteristic situate in the base and vice versa. For both types of clustering, the desired number of clusters is four which were confirmed based on the distribution of silhouette coefficient, elbow method, and rule of thumb. The list of networks for each cluster is as shown in Table 5.1. The k-means results plotted on PCA coordinates are as shown in Figure 4.7 and the dendrogram from hierarchical clustering is in Figure 4.9. Besides, the key observations are as follows:

   - The first cluster is a group of small size combination of metro and tram system. The network size ranges from 37 stations for the smallest network to 118 networks for the largest one. All members exhibit a high level of service connectivity, accessibility, and directness (CAD). On the other hand, these networks are highly centralized but low level of robustness. The dendrogram shows the ranks of network centralization in which networks with larger sizes seem to have high value. Moreover, their L-space graph looks similar to radial lines structure.
   - The second cluster is composed of only tram whose network size is large (average of 397.2 stations). However, this cluster possesses low level of CAD. For the level of robustness, they situate between both the combination of metro and tram cluster (cluster 1 and cluster 3). Centralization ranked the networks in this cluster as shown in the dendrogram.
   - The third cluster is a mixed group of transportation modes including metro, metro-tram, tram, and train. These networks are larger than the first cluster but still smaller than the tram clusters (cluster 2 and 4). All networks are highly robust, but their CAD is in between the first and the rest clusters. From L-space representation, they contain several numbers of cycles in the network. The dendrogram illustrates the cluster ranking for robustness. Moreover, Barcelona metro is more similar to Amsterdam metro and tram than the Dutch national railway, but Santiago metro differs from all others. Also, Santiago is the most robust networks.
   - The fourth cluster is a tram-related group whose network size is similar to second cluster. Tram-related mode refers to the combination of tram modes

with other modes, railway, and metro. They have higher level of centralization compared to the second cluster. Their CAD level is lower than cluster 2 but higher than cluster 1. Centralization ranked the networks in this cluster as shown in the dendrogram.

Cluster 1	Cluster 2	Cluster 3	Cluster 4
Adelaide (M+T)	Melbourne (T)	Amsterdam (M+T)	Brisbane (R+L)
Calgary (T)	Milan (T)	Barcelona (M)	Budapest (T)
Toulouse (M+T)	Vienna (T)	Santiago (M)	New Jersey (R+L)
Washington (M)	Warsaw (T)	The Netherlands (R)	Sydney (R+L)
	Zurich (T)		The Hague & Rotterdam (M+T)
			Valencia (M+T)
			Victoria (R)

Table 5.1: Lists of PTNs in each cluster. M: metro, T: tram, R: Rail and L: Light rail

In summary, apart from each cluster distinct characteristics, they are differentiated by the mode of transportation and the size of the networks. Figure 4.10 shows that network in the same clusters share similar features as illustrating in the L-space graph. Moreover, the similar modes network tends to share similar structure pattern such as the radial lines for the cluster 1 group. For the size of the network, it is implicitly from the network indicators formulas. Although the min-max normalization is used to adjust its range before clustering, the normalization did not include the size of network normalization.

5.2 SCIENTIFIC CONTRIBUTION

This thesis filled the literature gaps in applying quantitative approach to classify PTNs according to various structural network properties. Previous studies, such as that of Von Ferber et al. (2009); Barthelemy (2010); Haznagy et al. (2015) employed several indicators to characterize PTNs to extract more information from the network. However, their analysis and interpretation were only based on single indicator comparison at a time. This means that analysts lack a clear overview of network properties and cannot see how each

characteristic is related. Moreover, this limits the number of indicators in the analysis as it is no use including many.

Few studies try to alleviate this gap. Gattuso and Miriello (2005) adopted a multi-criteria analysis to combine all indicator values into a single score to rank metro networks. However, this method cannot facilitate the comparison or identifying distinct features between networks. Moreover, Derrible and Kennedy (2010b) qualitatively classify metro networks adapting 2-D graph in which network indicator is on each axis, but this analysis is subjected to the analyst's bias. Therefore, this research alleviates the gap by adapting qualitative approach to both network characterization and classification.

For network characterization, five main structural characteristics are selected. Those are centralization, accessibility, transitivity, directness and robustness. Moreover, several transportation modes are included unlike the previous studies (Gattuso and Miriello, 2005; Derrible and Kennedy, 2010b; STOILOVA and STOEV, 2015) which restricted themselves only metro networks. Including only metro networks could not explain the major transportation part of the cities. The considered modes in this work are train, metro, tram and light rail system at both city and national level. Unlike Von Ferber et al. (2009) in which they claim too much diversity in the urban area network and could not derive classification, we found that the network structure can be distinguished according to the mode of transportation and their network size. Our findings suggest that tram, metro and train fundamentally possess distinct structural properties even though the modes label and their related properties such as station spacing were not included. The general properties for each mode are as follows:

- Tram networks perform worse in infrastructural accessibility and both service connectivity and directness. However, they are highly centralized which causes the low level of robustness.
- The tram and train combination modes result in a more centralized network.
- The tram and metro combination can result in two different type of networks. One can be highly robust network or vice versa.

Moreover, this thesis exhibits quantitative classification with an appropriate set of features adding to finding of the previous STOILOVA and STOEV (2015) which only realize the cluster group via size of the networks.

5.3 PRACTICAL IMPLICATION

As public transportation evolves overtime, the set boundary between modes is becoming blurred. For newer version of tram system, the term 'light rail' is adopted instead as can be seen in New Jersey or Sydney. This prompts the question of whether the traditional definition is still relevant in the present context. To refine the current transportation mode concept, the distinct network characteristics can also be tied to the current physical concept. For example, a low level of transitivity, accessibility and directness tends to exist in tram network. On the other hand, metro-tram seems to outperform PTNs of all types

in TAD aspects. These structural network characteristics could serve as supplementary information describing how certain modes usually become. Based on such information, planners have a clearer idea of the mode characteristics. This could be supplementary information to gain more insights on each rail mode.

5.4 LIMITATIONS

The main limitations are in the data set and clustering method. In general, clustering analysis is usually applied to a large data set with more than 50 data objects. To achieve that, GTFS data format seems to be the solution as it set the ground for all PT authority and operators to publish the data set in similar formats. However, the standard itself is quite flexible and changes overtime. This creates difficulty to convert and extract data into usable format. Moreover, the available data set are mostly in Europe and both North and South America, but none are in Asia. It would be interesting to study the extensive networks in China, Japan and Korea.

Clustering results are not so robust and dependent on the sample set. Depending on similarity measures and clustering techniques, data set can be grouped into clusters. In addition, the result will be varying depending on the number of data objects and their corresponding features. However, the finding from this thesis would be some guidance for further analysis when performing clustering results. The findings will be preliminary exploratory result to the bigger data set of PTNs. For example, the pre-assumption from this findings are that same physical transportation modes are confirmed to have similar properties.

5.5 FUTURE RESEARCH

5.5.1 Indicators selection

Since this thesis focuses on PTNs structural characteristics, indicators were picked to reveal network insights. However, PTNs can be evaluated in different stages: normal and under disruption. Since this thesis only selected the former categories of indicators, it is interesting to also the disrupted states when calculated indicators such as resilient or robustness. Moreover, if possible, it is interesting to include as many indicators as possible and adapt the PCA to help picking several significant ones.

5.5.2 Networks selection

In this thesis, the comparison is made based on several rail related modes in different cities and countries. However, it is interesting to include all available PTNs modes to truly assess the city mobility structure. Since city evolution is very complex, these classification may reveal insight suggesting mobility planning in different places. Although, this research indicates that national network is quite different from those urban network, it is interesting

to see how it is comparable to other national networks and how they could be in the same group of urban networks.

BIBLIOGRAPHY

Adolfsson, A., Ackerman, M., and Brownstein, N. C. (2019). To cluster, or not to cluster: An analysis of clusterability methods. *Pattern Recognition*, 88:13–26.

Bangxang, P. N. and Jarumaneeroj, P. (2018). Topological evolution of public transportation network: A case study of Bangkok rail transit network. In *2018 5th International Conference on Industrial Engineering and Applications, ICIEA 2018*, pages 410–414. IEEE.

Barabási, A. L. and Albert, R. (1999). Emergence of scaling in random networks. *Science*, 286(5439):509–512.

Barthelemy, M. (2010). Spatial Networks. *Physics Reports*, 499(1-3):1–101.

Cao, W., Feng, X., and Zhang, H. (2018). The structural and spatial properties of the high-speed railway network in China: A complex network perspective. *Journal of Rail Transport Planning and Management*.

Cats, O. (2017). Topological evolution of a metropolitan rail transport network: The case of Stockholm. *Journal of Transport Geography*, 62:172–183.

Chen, C., D'Alfonso, T., Guo, H., and Jiang, C. (2018). Graph theoretical analysis of the Chinese high-speed rail network over time. *Research in Transportation Economics*, 72:3–14.

Cheng, H. M., Ning, Y. Z., Ma, X., Liu, X., and Zhang, Z. Y. (2017). Effectiveness of rapid rail transit system in Beijing. *PLoS ONE*, 12(7):e0180075.

Dalvi, M. Q. and Martin, K. M. (1976). The measurement of accessibility: Some preliminary results. *Transportation*, 5(1):17–42.

de Regt, R., von Ferber, C., Holovatch, Y., and Lebovka, M. (2017). Public transportation in UK viewed as a complex network. *Transportmetrica A: Transport Science*, pages 1–23.

Derrible, S. (2012). Network Centrality of Metro Systems. *PLoS ONE*, 7(7):e40575.

Derrible, S. and Kennedy, C. (2010a). Characterizing metro networks: state, form, and structure. *Transportation*, 37(2):275–297.

Derrible, S. and Kennedy, C. (2010b). Evaluating, Comparing, and Improving Metro Networks. *Transportation Research Record: Journal of the Transportation Research Board*, 2146(1):43–51.

Derrible, S. and Kennedy, C. (2010c). The complexity and robustness of metro networks. *Physica A: Statistical Mechanics and its Applications*, 389(17):3678–3691.

Ding, R., Ujang, N., bin Hamid, H., and Wu, J. (2015). Complex network theory applied to the growth of Kuala Lumpur's public urban rail transit network. *PLoS ONE*, 10(10):e0139961.

Ducruet, C. and Lugo, I. (2013). Structure and Dynamics of Transportation Networks: Models, Methods and Applications. In *The SAGE Handbook of Transport Studies*, pages 347–364. SAGE Publications, Ltd, 1 Oliver's Yard, 55 City Road London EC1Y 1SP.

Freeman, L. C. (1978). Centrality in social networks conceptual clarification. *Social Networks*, 1(3):215–239.

Garrison, William L and Marble, D. F. (1962). The structure of transportation networks. Technical report, NORTHWESTERN UNIV EVANSTON IL.

Gattuso, D. and Miriello, E. (2005). Compared analysis of metro networks supported by graph theory. *Networks and Spatial Economics*, 5(4):395–414.

Haag, M. R. D. (2019). Gemeenten.

Han, J., Kamber, M., Pei, J., Han, J., Kamber, M., and Pei, J. (2012). Cluster Analysis: Basic Concepts and Methods. In *Data Mining*, pages 443–495. Morgan Kaufmann.

Haznagy, A., Fi, I., London, A., and Nemeth, T. (2015). Complex network analysis of public transportation networks: A comprehensive study. In *2015 International Conference on Models and Technologies for Intelligent Transportation Systems (MT-ITS)*, volume 635, pages 371–378. IEEE.

Jawlik, A. (2016). Correlation – part 1 (of 2). In *Statistics from a to Z : Confusing Concepts Clarified*, volume 1, pages 124–134.

Kurant, M. and Thiran, P. (2006). Extraction and analysis of traffic and topologies of transportation networks. *Physical Review E - Statistical, Nonlinear, and Soft Matter Physics*, 74(3):036114.

Latora, V. and Marchiori, M. (2002). Is the Boston subway a small-world network? *Physica A: Statistical Mechanics and its Applications*, 314(1-4):109–113.

Lin, J. and Ban, Y. (2013). Complex Network Topology of Transportation Systems. *Transport Reviews*, 33(6):658–685.

Luo, D., Cats, O., and van Lint, H. (2019). Can passenger flow distribution be estimated solely based on network properties in public transport systems? *Transportation*, pages 1–20.

Musso, A. and Vuchic, V. R. (1988). Characteristics of metro networks and methodology for their evaluation. *Transportation Research Record*, 1162:22–33.

Newman, M. E. J. (2003). The structure and function of complex networks. *Computer Physics Communications*, 147(1-2):40–45.

Pagani, A., Mosquera, G., Alturki, A., Johnson, S., Jarvis, S., Wilson, A., Guo, W., and Varga, L. (2018). Resilience or Robustness: Identifying Topological Vulnerabilities in Rail Networks. *Royal Society Open Science*, 6(2):181301.

Pedregosa, F., Michel, V., Grisel OLIVIERGRISEL, O., Blondel, M., Prettenhofer, P., Weiss, R., Vanderplas, J., Cournapeau, D., Varoquaux, G., Gramfort, A., Thirion, B., Grisel, O., Dubourg, V., Passos, A., Brucher, M., Perrot andÉdouardand, M., Duchesnay, A., and Duchesnay EDOUARDDUCHESNAY, F. (2011). Scikit-learn: Machine Learning in Python Gaël Varoquaux Bertrand Thirion Vincent Dubourg Alexandre Passos PEDREGOSA, VAROQUAUX, GRAMFORT ET AL. Matthieu Perrot. *Journal of Machine Learning Research*, 12:2825–2830.

STOILOVA, S. and STOEV, V. (2015). An application of the graph theory which examines the metro networks. *Transport Problems*, 10(2):35–48.

To, W. M. (2015). Centrality of an Urban Rail System. *Urban Rail Transit*, 1(4):249–256.

Tu, Y. (2013). Centrality characteristics analysis of urban rail network. In *IEEE ICIRT 2013 - Proceedings: IEEE International Conference on Intelligent Rail Transportation*, pages 285–290. IEEE.

von Ferber, C., Berche, B., Holovatch, T., and Holovatch, Y. (2012). A tale of two cities: Vulnerabilities of the London and Paris transit networks. *Journal of Transportation Security*, 5(3):199–216.

Von Ferber, C., Holovatch, T., Holovatch, Y., and Palchykov, V. (2009). Public transport networks: Empirical analysis and modeling. *European Physical Journal B*, 68(2):261–275.

Vuchic, V. and Musso, A. (1991). Theory and practice of metro network design. *Public Transport International*, 40(3):298.

Wan, D., Huang, Y., Feng, J., Shi, Y., Guo, K., and Zhang, R. (2018). Understanding Topological and Spatial Attributes of Bus Transportation Networks in Cities of Chongqing and Chengdu. *Mathematical Problems in Engineering*, 2018:1–14.

Wang, X., Koç, Y., Derrible, S., Ahmad, S. N., Pino, W. J., and Kooij, R. E. (2017). Multi-criteria robustness analysis of metro networks. *Physica A: Statistical Mechanics and its Applications*, 474:19–31.

Watts, D. J. and Strogatz, S. H. (1998). Collective dynamics of 'small-world' networks. *Nature*, 393(6684):440–442.

Wei, S., Teng, S. N., Li, H. J., Xu, J., Ma, H., Luan, X. L., Yang, X., Shen, D., Liu, M., Huang, Z. Y., and Xu, C. (2019). Hierarchical structure in the world's largest high-speed rail network. *PLoS ONE*, 14(2):e0211052.

Wu, X., Dong, H., Tse, C. K., Ho, I. W., and Lau, F. C. (2018). Analysis of metro network performance from a complex network perspective. *Physica A: Statistical Mechanics and its Applications*, 492:553–563.

Xie, F. and Levinson, D. (2007). Measuring the structure of road networks. *Geographical Analysis*, 39(3):336–356.

XU, Q., ZU, Z., XU, Z., ZHANG, W., and ZHENG, T. (2013). Space P-Based Empirical Research on Public Transport Complex Networks in 330 Cities of China. *Journal of Transportation Systems Engineering and Information Technology*, 13(1):193–198.

Zanin, M., Papo, D., Sousa, P. A., Menasalvas, E., Nicchi, A., Kubik, E., and Boccaletti, S. (2016). Combining complex networks and data mining: Why and how. *Physics Reports*, 635:1–44.

Zanin, M., Sun, X., and Wandelt, S. (2018). Studying the Topology of Transportation Systems through Complex Networks: Handle with Care. *Journal of Advanced Transportation*, 2018:1–17.

Zhang, H., Zhuge, C.-X., Zhao, X., and Song, W.-B. (2018a). Assessing transfer property and reliability of urban bus network based on complex network theory. *International Journal of Modern Physics C*, 29(01):1850004.

Zhang, J., Wang, S., and Wang, X. (2018b). Comparison analysis on vulnerability of metro networks based on complex network. *Physica A: Statistical Mechanics and its Applications*, 496:72–78.

Zhang, J., Xu, X., Hong, L., Wang, S., and Fei, Q. (2011). Networked analysis of the Shanghai subway network, in China. *Physica A: Statistical Mechanics and its Applications*, 390(23-24):4562–4570.

Zhang, J., Zhao, M., Liu, H., and Xu, X. (2013). Networked characteristics of the urban rail transit networks. *Physica A: Statistical Mechanics and its Applications*, 392(6):1538–1546.

Zhang, L., Lu, J., Fu, B.-b., and Li, S.-b. (2018c). A Review and Prospect for the Complexity and Resilience of Urban Public Transit Network Based on Complex Network Theory. *Complexity*, 2018:1–36.

A SCIENTIFIC PAPER

This section contains the paper illustrating the key concept from this thesis.

Topological characterization and clustering of public transport networks

Krissada Tundulyasaree[1] ,Ding Luo[1], Maaike Snelder[1,2], Yilin Huang[3], Oded Cats[1]

***Abstract*— Numerous comparative network studies adopted topological indicators to characterize networks. They were able to identify similar properties between public transport networks (PTNs) such as small-world and scale. However, few studies attempted to create classification. This study proposed methodology to quantitatively characterize and cluster the PTNs. The selected characteristics are centralization, accessibility, robustness, service connectivity and directness. For the clustering part, hierarchical clustering adapting ward linkage criteria is adopted. The proposed method was applied to classify 20 rail-based PTNs worldwide. Apart from the indicator properties, network are distinguished by their mode of transportation and network size.**

I. INTRODUCTION

Comparative study is an approach in structural network studies. They compare different networks to identify similarities or differences. There is much empirical evidence that PTNs in different places share common statistical characteristics [1]–[7]. For instance, Lin and Ban [4] found many railway, subway and bus networks in several countries exhibit small-world [8] and scale-free structure [9].

Despite the accumulation of comparative PTNs studies, there were a few attempts to create a classification. Gattuso and Miriello [10] classified 13 metro networks using combined network indicator scores to rank networks. The scores were calculated from the multi-criteria analysis. However, it is difficult to identify a distinct property of each network as they all combined into a single value. Derrible and Kennedy [2] adopted a 2-D graph where each network indicator is on the axis. A few network indicators described each characteristic. They were able to classify 33 metro networks into different groups according to the state of development, interaction with the built environment, and intrinsic structure. Besides, Stoilova and Stoev [11] adopted hierarchical clustering to classify 22 European metro networks adopting six network indicators. However, both studies [2, 11] only consider the physical infrastructure side of the PTNs and include a single mode of PT. In their studies, it is assumed that a direct service line always exists connecting a station to any stations; however, that is not always the case. PT operators usually provide several service lines to serve nodes, and the transfer is made at the node where the service line overlapped. Consequently, service lines network is another vital aspect to describe PTNs. Moreover, as only metro networks are considered, the classification result is mode-dependent. In other words, they assume that transport modes influence the network structure.

In summary, few studies classified PTNs and most of them were based on a single PT mode representing on infrastructure networks. Based on such criteria, they assume no similarities between different PT modes and omit the service lines from analysis.

This paper aims to classify PTNs employing topological network indicators. The proposed methodology consists of two parts: network characterization and network clustering. For the former part, five network characteristics are chosen and quantified by global network indicators from graph theory and network science. The latter part makes use the hierarchical clustering method. Moreover, this method is applied to 20 networks derived from the general transit feed specification (GTFS) data.

The reminder of this paper is organized as follows. Section II details the proposed methodology consisting of two parts. The first part is focused on characterizing five selected PTN features quantified by topological indicators, while the second shows how PTNs are clustered into groups based on the these indicators. Then case study networks, which contain 20 rail-based PTNs worldwide, are introduced in section III-A, followed by the presentation of results and discussion in section III. Section IV concludes the study with main findings, contributions and recommendations for future research.

II. METHODOLOGY

First, we describes the selection of network representation along with the network indicators to characterize PTNs. The second part details how to classify PTNs by hierarchical clustering analysis.

A. Network Characterization

The first subsection introduces a brief PTNs representation used in this thesis. Next, network characteristics and its indicator selection is described.

1) Network representation: A PTN is represented as a directed graph G representing by $G = (N, E)$ where N is the set of nodes and E is the set of links. A node $n \in N$ represents a station while a link $e \in E$ is defined by an ordered pair of nodes (u, v) in which u and v $(u, v \in N)$ denote the source and sink nodes, respectively. Note that $\mid N \mid$ and $\mid E \mid$ denote the number of stations and links, respectively.

[1]Department of Transport and Planning, Faculty of Civil Engineering and Geosciences, Delft University of Technology, Delft, The Netherlands Email: d.luo@tudelft.nl

[2]TNO Netherlands Organisation for Applied Scientific Research, Anna van Buerenplein 1, 2595 DA, Den Haag, The Netherlands Email: m.snelder@tudelft.nl

[3]Section of Systems Engineering and Simulation, Faculty of Technology, Policy and Management, Delft University of Technology, Delft, The Netherlands Email: y.huang@tudelft.nl

In this paper, L-space and P-space graphs are employed to enrich network characterization. **L-space** graph illustrates stations as nodes while a link connecting nodes exists if there is at least a service line connecting those two consecutive nodes [1]. For **P-space**, a node in this graph still represents a station, but a link exists if there is at least a direct service line linking the pair of nodes. Apart from the space representation, the graphs links have no weight, so all links have identical properties. Since PTNs in this thesis are the combination of different modes such as metro, tram, and train, the resulting graph does not differentiate the transportation modes.

2) Network characteristics selection: To characterize PTNs, five network features, including centralization, accessibility, robustness, service connectivity and directness, are chosen. These characteristics were proposed in previous PTN studies and used to characterized different PT modes. The first three features will be assessed in the infrastructure layer, while the service layer is employed for the last two. The following subsection detail each characteristic and its networks indicator with mathematical formulas.

3) Centralization: Centralization measures how nodes with different levels of centrality values in a network are distributed [12]. A network with high value of centralization is likely to have a few powerful nodes with high node centrality, while others with low centrality, vice versa. In this study, the centralization based on nodes' betweenness centrality in the L-space is applied. It is defined as follows.

$$C_B^l(v) = \sum_{s \in N} \sum_{t \in N} \frac{\sigma_{st}(v)}{\sigma_{st}} \tag{1}$$

where σ_{st} denotes the total number of shortest paths from node s and t and $\sigma_{st}(v)$ is number of those paths that pass through v. The global indicator of centralization, denoted as C_B^l, is further defined as follows.

$$C_B^l = \frac{\sum\limits_{v \in N} (C_B^l(v)^* - C_B^l(v))}{\mid N \mid^3 - 4 \mid N \mid^2 + 5 \mid N \mid - 2} \tag{2}$$

where $C_B^l(v)^*$ is the maximum betweenness centrality among nodes in the network. N is the number of nodes.

4) Accessibility: The feature of accessibility measures the ease of reaching all stations on average. Characterized in the L-space, it is determined based on local nodes' *closeness* centrality. The closeness centrality of a node c_i is defined as follows.

$$C_C^l(v) = \frac{\mid N \mid -1}{\sum\limits_{i \in N} d_{vi}} \tag{3}$$

where d_{vi} is the shortest topological distance (i.e., the number of links needed to be traversed) between node v to node i. The nodes are close to others if there exist the direct link connection to it.

According to this definition, node closeness centrality increases for lower travel impedance to other nodes in the network and hence reflects the accessibility of each node to the rest of the network. The average closeness centrality of all the nodes in the network can thus be used as the global indicator of accessibility, denoted as A^l since it reflects the average number of nodes that are passed when travelling between any pair of nodes [13].

$$A^l = \frac{\mid N \mid -1}{\mid N \mid} \sum_{v \in N} \frac{1}{\sum\limits_{i \in N} d_{vi}} \tag{4}$$

5) Robustness: Robustness describes the extent to which networks can cope with the disruption event. A more robust network has higher number of alternative physical paths to maintain the network connectivity while the mitigation process is on-going. To measure this feature, we use the alpha index (or network meshedness) in L-space representation [14]. It quantifies alternative routes existed in the network, which implies that there are more than a single path between node pairs. The global indicator of robustness, denoted as α, is thus defined as follows.

$$\alpha^l = \frac{\mid E \mid - \mid N \mid + 1}{2 \mid N \mid - 5} \tag{5}$$

where $\mid E \mid$ and $\mid N \mid$ represent the number of links and nodes in a graph, respectively. The higher value of α indicates the more robust the network structure in a sense that network provide greater number of alternative paths.

6) Service connectivity: Service connectivity (SC) measures how well the service lines are linked. It compares the current link connection relative to the best connection scenario, which is usually the complete graph [15]. Note that the complete graph is a network in which every pair of nodes is connected. The high SC network offers a large number of direct routes between nodes. In PTNs, a direct service route does not require any transfers when transversing between a pair of stations. Moreover, it implies that network nodes cluster together as service lines directly connect them. This concept is in line with a clustering coefficient (c^P) in the P-space. Let $Nbh(u)$ is the neighbourhood of a node u ($Nbh(u) = \{v \in N \mid (u,v) \in E\}$) and d_u is the degree of a node u ($d_u = \mid Nbh(u) \mid$). The clustering coefficient can be defined as follows: [1]

$$c^P = \frac{\sum_{u \in N} \mid \{\{i,j\} \subseteq Nbh(u) \mid (i,j) \in E\} \mid}{\sum_{u \in N} d_u(d_u - 1)/2} \tag{6}$$

c^P measures the whole network service connectivity through the connectivity within the neighbourhood. The denominator is the maximum number of links in the neighbourhood for node u given the node degree d_u. The range of c^P is between 0 and 1. The closer c^P is to 1, the higher SC level network is.

7) Directness: Directness measures the extent to which network can provide direct service between any node pairs. In the context of PT, we characterize this feature based on the number of transfers required between node pairs. High directness would imply a lower average of required number

of transfers. Thus, we use the average of shortest path lengths in the P-space representation for this feature.

$$E^p = \frac{1}{\mid N \mid (\mid N \mid -1)} \sum_{i \in N} \sum_{\substack{j \in N \\ i \neq j}} \frac{1}{d_{ij}^p} \tag{7}$$

B. Network Clustering

Agglomerative hierarchical clustering groups data objects into different cluster levels. It refers to a bottom-up strategy. Each data point starts in its own cluster and is paired with others in every iteration round until every points in the same cluster. Before any clustering is performed, it is required to determine the proximity matrix containing the distance between each point based on a distance function. Then, the matrix is updated to display the distance between clusters. Let $\mid i-j \mid$ be the distance between object i and j, m_i is the center of the data points in cluster, C_i and n_i is the number of data points in cluster i. In this study, we apply the ward linkage algorithm to compute the distance between clusters as follows:

$$d_{ward}(C_i, C_j) = \frac{n_i n_j}{n_i + n_j} \mid m_i - m_j \mid^2 \tag{8}$$

To measure the quality of clustering, the silhouette coefficient there are 2 method: extrinsic and intrinsic method. The intrinsic method is chosen since the ground truth or data labels are not available. Therefore, the intrinsic method mainly assessed the quality of clusters in two aspects: their separation between groups and the compactness of clusters.

The silhouette coefficient is selected to assess the clustering quality. Supposed a data set, D, is partitioned into k clusters $C_1, C_2, ..., C_k$. For each object $v \in D$, silhouette coefficient of v is then defined as:

$$s(v) = \frac{b(v) - a(v)}{\max\{a(v), b(v)\}} \tag{9}$$

where $a(v)$ is the average distance between v and all other object in the cluster v belongs to and $b(v)$ is the minimum average distance from v to all clusters v does not belong to (see Equation 10,11).

$$a(v) = \frac{\sum_{v' \in C_i, v \neq v'} d(v, v')}{\mid C_i \mid -1}, \tag{10}$$

$$b(v) = \min_{C_j : 1 \leq j \leq k, j \neq i} \{\frac{\sum_{v' \in C_j} d(v, v')}{\mid C_i \mid}\} \tag{11}$$

The value of the silhouette coefficient for each object v varies between -1 and 1. Note that silhouette value for the cluster contain a member is defined as 0.While $a(v)$ reflects the compactness of the cluster, $b(v)$ describes the distance between clusters. Therefore, when $s(v)$ is close to 1, the resulting clusters containing object v is compact and far away from other clusters. On the other hand, the negative value of $s(v)$ suggests that object v is closer to objects in other clusters more than objects in its own cluster. The average value of the silhouette coefficient from all objects in the data set can be used to compare different clustering methods.

III. Result & Discussion

A. Studied Networks

20 rail-bound networks were employed for the case study. These networks were generated using up-to-date GTFS data. Table I shows the location and modes of these networks, along with the number of nodes and links. Various rail-bound modes ranging from rail to tram are included in this analysis.

TABLE I: Overview of the studied rail-bound networks worldwide.

Place	Mode	#nodes	#links
Adelaide	Metro & Tram (M+T)	118	247
Amsterdam	Metro & Tram (M+T)	196	467
Barcelona	Metro (M)	131	298
Brisbane	Rail & Light rail (R+L)	176	367
Budapest	Tram (T)	237	470
Calgary	Tram (T)	46	81
Melbourne	Tram (T)	805	1676
Milan	Tram (T)	335	693
New Jersey	Rail & Light rail (R+L)	184	351
Santiago	Metro (M)	136	388
Sydney	Rail & Light rail (R+L)	307	634
The Hague & Rotterdam	Metro & Tram (M+T)	424	929
The Netherlands	Rail (R)	413	1050
Toulouse	Metro & Tram (M+T)	37	72
Valencia	Metro & Tram (M+T)	134	240
Victoria	Rail (R)	220	449
Vienna	Tram (T)	385	815
Warsaw	Tram (T)	271	509
Washington	Metro (M)	91	186
Zurich	Tram (T)	190	400

B. Network characterization

We computed the five global indicators for all the case study networks, with the histograms of the resulting (original) values first presented in Figure 1. Different distributional patterns can be observed. For instance, the distribution of centralization is skewed with the majority lying on the left side. This means that most studied networks do not exhibit a high degree of centralization. The distribution of efficiency somehow shows values on both left and right sides with a gap in the middle, though there are more networks falling onto the left part. The robustness one displays a skewed distribution again, with the majority lying on the left between 0.5 and 0.6. The service connectivity of all the networks is relatively more evenly distributed compared with the rest, while the directness is most similar to a normal distribution.

Since all selected indicators have different value ranges, this can affect the clustering analysis. The large value-range indicator tends to influence the cluster groups and outweigh other indicators. Min-max normalization is employed to ensure that all indicators are equally weighted in the clustering analysis. In addition, the new minimum value is 0.05 because this is the smallest value based on the raw data set (see Table I) and 0 would provide misleading interpretation. The normalization can be calculated as follows:

$$x'_{1p} = \frac{x_{1p} - \min_p}{\max_p - \min_p}(new\_max_p - new\_min_p) + new\_min_p \tag{12}$$

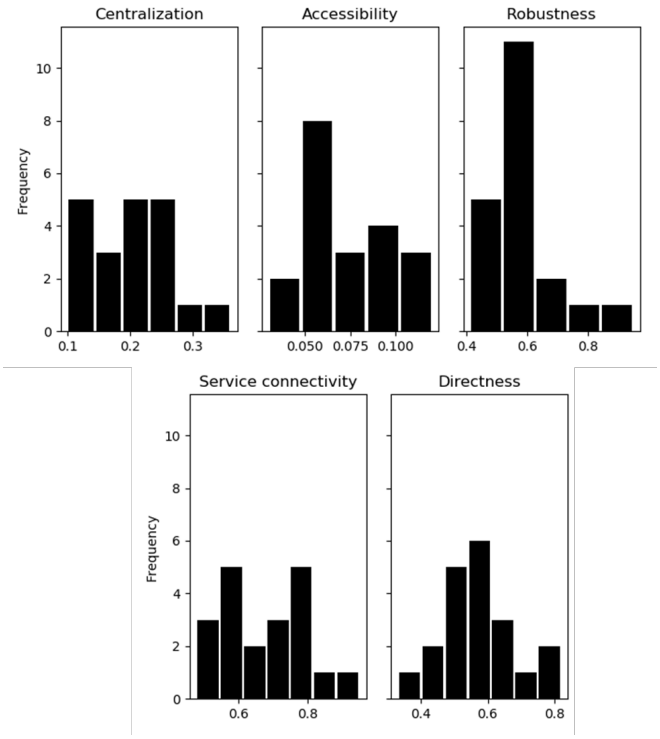


Fig. 1: Distributions of five topological indicators for all the studied networks.

i.e. a value of data object 1 attribute p (x_{1p}) is mapped to a new value x'_{1p} in the new value range from new_min_p to new_max_p. This normalization keeps the relation among the raw dataset values. After normalization, all indicators range from 0.05 to 1 while preserving the relationship among the original data values. Subsequently, hierarchical clustering will be performed based on the normalized indicator values.

C. Network clustering

The hierarchical results based on ward linkage algorithm are presented in dendrogram shown in Figure 3. The number of clusters are determined by the silhouette plot as in Figure 2. These sub graphs illustrate the silhouette coefficient value for every member in the cluster group as specified by the label. The four cluster case is the most appropriate because the number of members in each cluster are equally distributed and no member has negative silhouette value.

Hence, the tree result is cut to yield four clusters. Note that for the following discussion, the cluster group will be referred to as position in the dendrogram (see Figure 3) and the cluster number in parentheses correspond to those in cluster label in Figure 2. The cluster groups are identified as follows:

- The green group (Cluster 2) is a tram group and most networks are large in size when compared to other clusters. Apart from high centralization, the group has low level on all other quality.
- The red group (Cluster 4) is a tram-related group which refers to the combination of trams with other modes xuch railway or metro. This group exhibit higher centralization than green group while the other characteristics remain low. The network size is comparable to cluster 2.
- The blue group (Cluster 1) is a metro-tram group consisting mostly of small networks. All networks perform best on the combination of service connectivity, accessibility and directness (CAD) compared to other clusters. They also have low robustness level (and high centralization).
- The yellow group (Cluster 3) is a mixed group where most networks are medium in size. They all have high robustness level (and low centralization) while their CAD is in between those of cluster 1 and 2.

In addition to the general characteristics for each cluster, Figure 3 shows the ranking of network according to their important properties in an ascending fashion. In other words, network within cluster are situated higher in the tree if it possess higher value of cluster characteristic. For instance, the yellow cluster (cluster 3) networks are all highly robust network and the pairing in the dendrogram shows the rank of the robustness level. Barcelona metro and Amsterdam metro-tram are paired up first since their robustness level are 0.65 and 0.7, respectively. Next, they are matched with the Netherlands train of which robustness level is 0.78. Finally, the three merged with Santiago metro which has the highest robustness level among all networks in this data set. In addition, this pattern is also in other three cluster. The blue cluster (cluster 1) shows the rank of centralization while both the red (cluster 4) and green cluster (cluster 2) exhibit the rank for service connectivity.

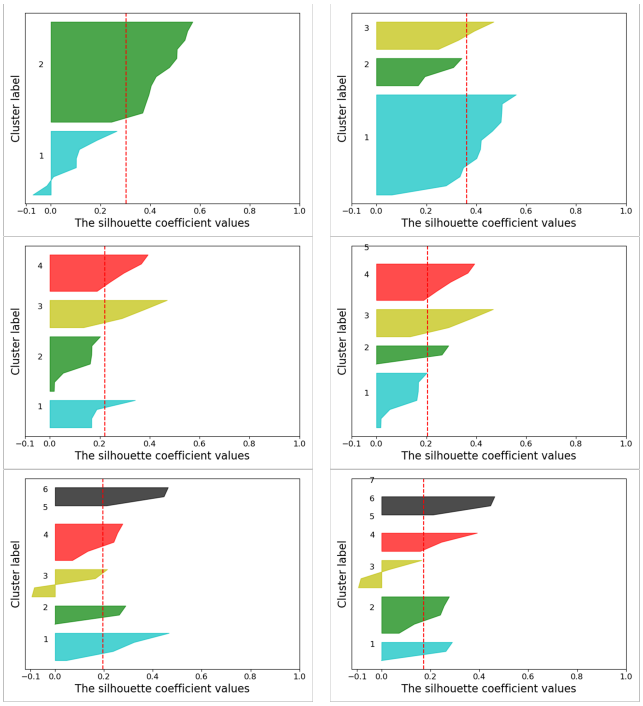


Fig. 2: Comparison of silhouette coefficient value for different number of clusters from two to seven. Note the red dotted line show the average value of silhouette coefficient for all clusters.

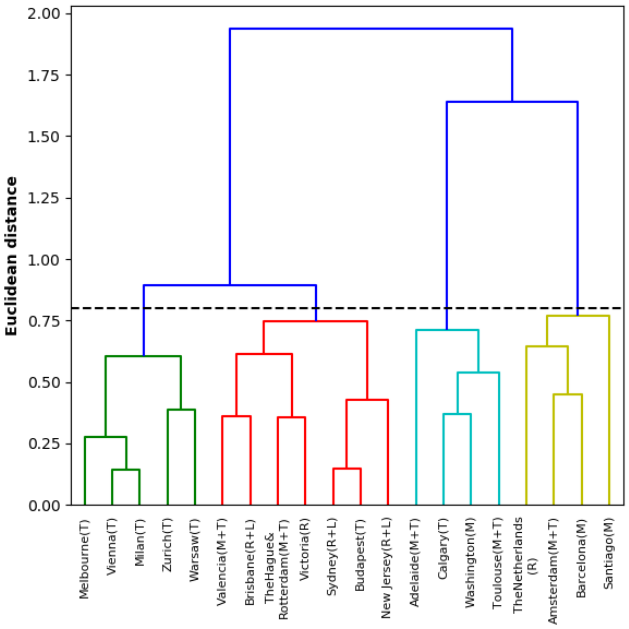


Fig. 3: Dendrogram of hierarchical clustering adapting ward criteria linkage method. (R: Rail, T:Tram, L: Light-rail, and M: Metro) Note the black horizontal dotted line show the point where the tree is cut to give four clusters.

IV. Conclusions

Clustering PTNs based on multiple features has not been well studied in the literature, although such research can bring more insights to the strategic planning of PTNs. In this study, we employed five network features: centralization, accessibility, robustness, service connectivity and directness to cluster rail-bound PTNs. We are able to empirically confirm that highly centralization will result in lower robust network. Moreover, under this set of indicator, network are also distinguished by their mode of transportation and network size. Surprisingly, when metro was combined with tram network, they can result in two distinct group which are highly robust or highly centralized.

References

[1] C. Von Ferber, T. Holovatch, Y. Holovatch, and V. Palchykov, "Public transport networks: Empirical analysis and modeling," *European Physical Journal B*, vol. 68, pp. 261–275, mar 2009.

[2] S. Derrible and C. Kennedy, "Evaluating, Comparing, and Improving Metro Networks," *Transportation Research Record: Journal of the Transportation Research Board*, vol. 2146, pp. 43–51, jan 2010.

[3] Q. XU, Z. ZU, Z. XU, W. ZHANG, and T. ZHENG, "Space P-Based Empirical Research on Public Transport Complex Networks in 330 Cities of China," *Journal of Transportation Systems Engineering and Information Technology*, vol. 13, pp. 193–198, feb 2013.

[4] J. Lin and Y. Ban, "Complex Network Topology of Transportation Systems," *Transport Reviews*, vol. 33, pp. 658–685, nov 2013.

[5] J. Zhang, M. Zhao, H. Liu, and X. Xu, "Networked characteristics of the urban rail transit networks," *Physica A: Statistical Mechanics and its Applications*, vol. 392, pp. 1538–1546, mar 2013.

[6] A. Haznagy, I. Fi, A. London, and T. Nemeth, "Complex network analysis of public transportation networks: A comprehensive study," in *2015 International Conference on Models and Technologies for Intelligent Transportation Systems (MT-ITS)*, vol. 635, pp. 371–378, IEEE, jun 2015.

[7] D. Wan, Y. Huang, J. Feng, Y. Shi, K. Guo, and R. Zhang, "Understanding Topological and Spatial Attributes of Bus Transportation Networks in Cities of Chongqing and Chengdu," *Mathematical Problems in Engineering*, vol. 2018, pp. 1–14, oct 2018.

[8] D. J. Watts and S. H. Strogatz, "Collective dynamics of small-world' networks," *Nature*, vol. 393, pp. 440–442, jun 1998.

[9] A. L. Barabási and R. Albert, "Emergence of scaling in random networks," *Science*, vol. 286, pp. 509–512, oct 1999.

[10] D. Gattuso and E. Miriello, "Compared analysis of metro networks supported by graph theory," *Networks and Spatial Economics*, vol. 5, pp. 395–414, dec 2005.

[11] S. STOILOVA and V. STOEV, "An application of the graph theory which examines the metro networks," *Transport Problems*, vol. 10, pp. 35–48, apr 2015.

[12] L. C. Freeman, "Centrality in social networks conceptual clarification," *Social Networks*, vol. 1, pp. 215–239, jan 1978.

[13] O. Cats, "Topological evolution of a metropolitan rail transport network: The case of Stockholm," *Journal of Transport Geography*, vol. 62, pp. 172–183, jun 2017.

[14] X. Wang, Y. Koç, S. Derrible, S. N. Ahmad, W. J. Pino, and R. E. Kooij, "Multi-criteria robustness analysis of metro networks," *Physica A: Statistical Mechanics and its Applications*, vol. 474, pp. 19–31, may 2017.

[15] S. Derrible and C. Kennedy, "Characterizing metro networks: state, form, and structure," *Transportation*, vol. 37, pp. 275–297, mar 2010.

COLOPHON

This document was typeset using LaTeX. The document layout was generated using the `arsclassica` package by Lorenzo Pantieri, which is an adaption of the original `classicthesis` package from André Miede.