Student Name: WENXIN XU

Programme: MSc Geospatial Sciences

Supervisor: Dr James Haworth

Dissertation Title:

Applications of Graph Databases in Urban Rail Transit Analysis

DECLARATION OF OWNERSHIP

• I confirm that I have read and understood the guidelines on plagiarism, that I understand the meaning of plagiarism and that I may be penalised for submitting work that has been plagiarised.

• I declare that all material presented in the accompanying work is entirely my own work except where explicitly and individually indicated and that all sources used in its preparation and all quotations are clearly cited.

• I have submitted an electronic copy of the project report through Moodle/turnitin.

Should this statement prove to be untrue, I recognise the right of the Board of Examiners to recommend what action should be taken in line with UCL's regulations.

Signature: XU WENXIN Date: 2022.09.08

Applications of Graph Databases in Urban Rail Transit Analysis

Wenxin Xu

Supervised by

Dr. James Haworth

2022

Submitted in partial fulfilment of the requirements of the degree of

MSc in Geographical Science

University College London

Dept of Civil, Environmental and Geomatic Engineering

Abstract

Given the rapid increase in demand for public transport, it is questionable whether the existing transport network structure can be adapted to the demand for travel. And as transport networks become more complex, old relational databases do not perform well in the face of complex network structures and large amounts of data. The paper therefore introduces a graph database for transport data storage and analysis that is suitable for the network structure. The article uses a database to store London's rail network and traffic data, and investigates whether the London rail network matches the travel demand by analysing the spatial distribution and coupling of travel demand and network indicators (station density, connectivity and accessibility). It is found that the London rail network is generally compatible with travel demand, but there is still much room for improvement around London. The article also compares the structure, language and query performance of a graph database (Neo4j) and a relational database (PostgreSQL), and finds that Neo4j performs better for transport network structures with a storage size of around 13GB.

Acknowledgement

I would like to thank Dr James Haworth for his guidance and feedback on my dissertation, which helped me to have great content for my essay. I would like to thank my family and friends for their support and help in helping me to get through many low mood periods and to be able to complete my studies successfully.

Page

1 Introduction ... 1

1.1 Rail transit and travel demand ... 1

1.2 Traffic network, graph theory and graph database ... 1

1.3 Research aim ... 2

1.4 Dissertation structure ... 2

2 Literature Review ... 4

2.1 Existing research ... 4

2.1.1 Network Science in Transportation Networks ... 4

2.1.2 Graph databases and transport networks ... 5

2.2 conclusion ... 6

3 Data ... 7

3.1 The rail network ... 7

3.2 Rail traffic flow ... 8

4 Method ... 9

4.1 Quantitative analysis indicators ... 9

4.1.1 Travel demand ... 10

4.1.2 Station density ... 11

4.1.3 Connectivity ... 11

4.1.3 Accessibility ... 12

4.2 Data pre-processing ... 13

4.3 Spatial autocorrelation ... 15

4.4 Coupling ... 17

5 Result and Discuss ... 18

5.1 Analysis of the rail network ... 18

5.1.1 Spatial analysis ... 18

5.1.2 The degree of coupling of all-day travel demands ... 20

5.1.3 Coupling by time period ... 23

5.2 Database representation ... 27

5.2.1 Database structure ... 28

5.2.2 Build of database.. 30
5.2.3 Query performance 30
6 Conclusions and Recommendations 34
6.1 Summary 34
6.1.1 The rail network.. 34
6.1.2 Graph database... 34
6.2 Limitation 35
6.2.1 Analysis of the rail network.. 35
6.2.2 Databases... 35
Reference 36
Appendix A Neo4j database establishment procedures.................. 错误!未定义书签。
Appendix B PostgisSQL database establishment procedures .... 错误!未定义书签。
Appendix C Neo4j database query program.................................. 错误!未定义书签。
Appendix D PostgisSQL database query program .... 错误!未定义书签。

List of Figure

Figure 3-1 London Rail Map 7

Figure 4-1 Methodology Flow 9

Figure 4-2 Station Travel Demand........ 10

Figure 4-3 Station Connectivity 12

Figure 4-4 Accessibility Calculation Process 13

Figure 4-5 Station Accessibility 13

Figure 4-6 Result of KDE 15

Figure4-7 Result of Zonal Statistic........ 15

Figure 4-8 Moran Scatter Plot Partitioning........ 17

Figure 5-1 The Result of LMI........ 19

Figure 5-2 Coupling of Station Density and Travel Demand 21

Figure 5-3 Coupling of Accessibility and Travel Demand........ 21

Figure 5-4 Coupling of Connectivity and Travel Demand........ 22

Figure 5-5 Statistical Scale Diagram of Coupling........ 23

Figure 5-6 Travel demand by time of day 24

Figure 5-7 Coupling of Station Density and Travel Demand at Different Times of Day 25

Figure 5-8 Coupling of Accessibility and Travel Demand at Different Times of Day. 25

Figure 5-9 Coupling of Connectivity and Travel Demand at Different Times of Day. 26

Figure 5-10 Coupling Statistics for Different Time Periods........ 27

Figure 5-11 Graph Database Structure........ 28

Figure 5-12 Relational Database ER Diagram........ 29

Figure 5-13 Relation Database Build Flow 30

Figure 5-14 pgAdmin's traffic network structure query results (1000 rows)........ 31

Figure 5-15 Neo4j's traffic network structure query structure (300 nodes) 31

List of Table

Table 3-1 London Rail Traffic Flow Data 8

Table 4-1 Centrality Indicators.. 11

Table 4-2 GMI Judgement Criteria 16

Table 4-3 Classification of Coupling Degree.. 17

Table 5-1 The Result of GMI 18

Table 5-2 Results for a Given Relationship 32

Table 5-3 Query for Network Analysis... 33

1 Introduction

1.1 Rail transit and travel demand

As an important carrier of urban population mobility, public transport system can be a good substitute for private car trips and alleviate traffic congestion and environmental pollution greatly.[1,2] Whereas urban rail carries a huge volume of tasks in the public transport system, in 2019 London railways (London Underground, London Overground, Docklands Light Railway, Tfl Rail and London Trams) handles around 6 million journeys per day, accounting for 72.4% of trips made by Transport for London[3], and is expected to reach 10.8 million uses per day in 2041[4]. A well-constructed urban rail network will play a very important role in relieving traffic pressure.

It has been 159 years since the first subway was opened in London in 1863, and the network system of London Railway has basically taken shape in the middle and late 20th century, and it can be said that the London rail network has a very old structure.[5,6] However, with the evolution of urban sprawl and the encouragement of government development policies, more and more cities are changing their structure from a monocentric structure to a more complex distribution of activities, which brings about complex trends in population mobility.[7,8] Camille Roth et al. have shown that London has polycentric and very complex mobility patterns.[9] In such a context, it becomes a question whether the London rail network at this stage can meet people's travel needs well.

1.2 Traffic network, graph theory and graph database

Graph theory originated in 1741 with Euler's "Seven Bridges of Königsberg" problem and started to be applied to public transport networks in the 1980s, but in the 1970s there was a tendency to use travel demand models rather than graph theory to study public transport networks.[10,11] With the rise of non-regularized large-size network science in recent years, the identification and processing of complex networks has started to be taken seriously. Some studies have also started to use network science methods to analyze public transportation systems and found that transportation networks with scale-free patterns and small worlds are complex networks that are more suitable for research using network science methods.[10,12,13]

In this context we discovered graph databases, a graph theory-based data model that stores data in a graph data structure. Compared with traditional relational databases that use foreign key-linked relational tables to store structured data, graph databases built based on graph theory have high associativity and scalability and are more suitable for data storage and query of complex networks.[14,15] Public transportation networks have the characteristics of complex networks with large amount of data and frequent changes, which are very suitable for data storage using graph databases according to the theory.[16,17] Therefore, the study proposes to use a graph database to construct the London rail network, store its travel data, and test the performance of using the graph database.

1.3 Research aim

Based on the above research background, this study will investigate the following three problems

1. Constructing the London rail network and storing passenger flow data using graph database and relational database.
2. Analyze the rationality of the London rail network structure through connectivity, accessibility and other network science indicators based on the current stage of passengers' usage of the London rail network.
3. To compare the query performance of graph database and relational database for the required results to determine whether the graph database can better complete the construction and analysis of the transportation network.

1.4 Dissertation structure

The structure of this dissertation is as follows. Chapter 2 presents the background of this study by reviewing existing research and summarising the shortcomings of previous traffic network studies in terms of database and network topology metrics. Chapter 3 summarises the characteristics of the traffic data used in the dissertation and the data sources. Chapter 4 is the methodology section, detailing the topology metrics data acquisition and pre-processing, data spatial distribution and coupling degree analysis methods. Chapter 5 is the results and discussion section, which is divided into two parts, rail traffic network analysis and database operational performance comparison. The analysis of the rail network includes the spatial distribution of data, the coupling of network topology indicators and travel demand throughout the day and at different times of the day. The database runtime performance is divided into a comparison of the database structure and the performance of query language. Chapter 6 summarises the conclusions of the study and lists the limitations of the paper.

2 Literature Review

In the context of traffic networks and graph databases, this chapter summarises previous research in both network science and graph databases in the context of traffic networks. It also discusses the shortcomings of transport network research in the context of network science research in order to argue that it is valuable.

2.1 Existing research

2.1.1 Network Science in Transportation Networks

There has been a lot of literature on the network structure of public transport networks based on graph theory, for example Lam and Schuler[18] propose a connectivity index for transit networks based on travel time. Musso and Vuchic[19] proposed a transport network structure based on graph theory, Derrible and Kennedy[20] proposed three network characteristics for transportation networks.

However, it was only recently, driven by developments in network science, that traffic networks started to quantify the topology of networks through topological metrics.[21,13] Derrible and Kennedy[22] studied the complexity and robustness of 33 metro networks around the world, Derrible[12] studied the network connectivity of 28 metro networks through network centrality, Oded[21] analysed changes in the structure of the Stockholm metro network over time through coverage, connectivity, and directness indicators, and Welcha and Mishra[23] analysed the Washington region through connectivity, accessibility, catchment areas, and the Gini coefficient equity of public transportation.

It could be found that previous literature on network structure analysis has focused on scalarizing the network topology of transportation networks using network science topology metrics, such as coverage, connectivity, accessibility and robustness indices. However, these articles on network structure analysis do not reflect the practical applications of transport networks, i.e. they do not take into account people's travel needs. Instead, articles that integrate transport networks and travel demand mainly focus on the field of network structure design, where the design planning of transport networks is carried out through social and traffic indicators such as population, income and travel distance, such as the studies by Laporte and Pascoal[24] and studies by Petit and Ouyang[25].

However, these studies are mainly based on static considerations which uses current travel demand or predicted future travel growth as design indicators but ignores the fact that factors such as changing urban structure, policy direction and transport networks can all change people's travel patterns and travel demand. However, few articles have examined whether the completed transport network structures, especially those that are dated, are compatible with the current travel needs of people.

2.1.2 Graph databases and transport networks

With the use and popularity of public transport technologies such as smart travel cards, passenger counting technologies and vehicle location technologies, the volume of public transport data has increased dramatically and appears to be characterised by high data volumes and real-time recording.[17] And transport data storage relies mainly on traditional relational databases, in addition to interactive networks such as the General Transit Feed Specification (GTFS) and many European data sector open data portals, which are also based on relational database storage and recall.[26-28]

However, a growing body of literature has found that traditional transportation data storage models are difficult to accommodate and analyse the large volumes of real-time data generated by public transport because relational databases are composed of highly structured tables, and relational database queries that obtain results by traversing large numbers of joined tables and populated rows perform poorly when the data set structure is a complex network structure.[26,27,29-31]

On the other hand, graph databases are highly scalable by linking nodes and edges as the basic structure, allowing flexible addition of nodes and relations, and there is no global index between nodes, making query performance independent of data size and ideal for rapid data growth.[16,31,32] This ensures that large complex networks can be run and stored efficiently on graph databases.

Graph databases have been widely used in the building blocks of complex network information repositories. For example, Perçuku et al.[31] applied graph databases to the analysis of data storage in transmission networks and concluded that they outperformed relational databases, Agouti et al.[33] used graph databases to analyse social media influential personalities, Walmart and ebay in the retail industry use graph databases for user behavior analysis[34], Eckman and Brown[35] used graph databases for the management of cell biology. However, there is still a lack of research on the application of graph databases to the field of public transport networks as complex networks.

2.2 conclusion

In general, there have been many papers based on network science approaches to analyse the structure of complex transport networks, often using metrics such as connectivity, accessibility and robustness, and lacking in the integration of analysis with practical applications. There is also a small amount of literature on transport network design and planning that combines transport networks with actual travel demand. However, these analyses are mostly static and do not reflect the suitability of the current travel demand and network structure.

In addition, the literature has found that traditional models for storing public transport data are difficult to meet the explosive growth of traffic data and increasingly complex network structures, while graph databases based on network structures have been widely used to store data in complex network systems with good performance compared to relational databases.

In conclusion, there is a lack of research in the field of storing traffic information through graph databases and analysing whether the transport network is suitable for the current travel demand.

3 Data

As the Covid-19 pandemic outbreak and the corresponding city closure policy severely affected transport travel, especially after the cancellation of the 24h running Night Tube on the London Underground in March 2020, the article argues that the transport data after the outbreak is not a good representation of people's real travel demand. Reviewing the timeline of the covid-19 outbreak, the first patient was detected in China on 31 December 2019 and the first confirmed case was reported in the UK on 31 January 2020, so the article will select data from 2019 for analysis, except where noted.

The data used in the article include two main categories, London rail network structure and rail-related traffic data, which will be presented in turn later in the chapter.

3.1 The rail network

The rail network topology data mainly includes, Tube stations, Tube lines and their connections, as shown in Figure 3-1.

There are 13 London rail lines including London Underground, London Overground, Docklands Light Railway, Tfl Rail and London Trams, and 455 stations. The station geographic coordinates are taken from the Transport For London (Tfl) open data [36] under the General subdivision of Station Location. The Shapefile files for the rail lines are not open to TLF and Digimap, so the lines between stations are used here to represent rail lines.

A special case is the Elisabeth Line, which started operating in part in May 2015, but will not be fully operational until 24 May 2022, so only the operational part in 2019 is considered.

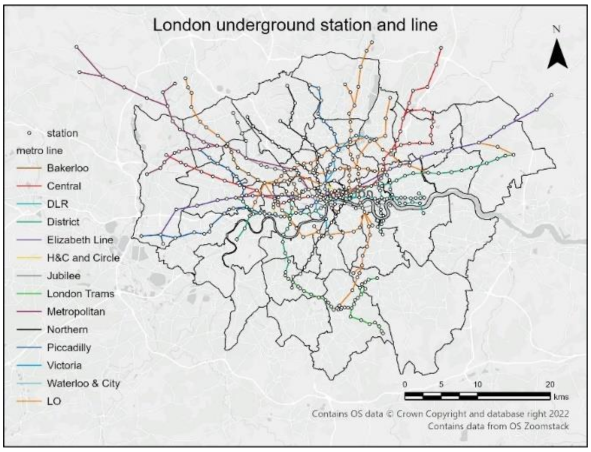


Figure 3-1 London Rail Map

3.2 Rail traffic flow

The stored rail traffic flow data is primarily based on the London Multi-Rail Traffic Demand Comprehensive dataset for a typical Friday in 2019 published by Tfl.[37] The traffic data is broken down by time period which include the flows for the entire traffic day, 6 different traffic periods of the day (Early, AM peak, Midday, PM peak, Evening, Late) and every 15 minute. Specific data includes the number of people on each line (Line Boarder), the number of people on different lines between one station and another (Link Load), and the number of people entering and leaving the station (Station Entry / Exit). Specific statistics are shown in Table 3-1.

Table 3-1 London Rail Traffic Flow Data

sheet name	description	count	feature
Line Boarder	Number of passengers boarding each line	19	Line Name Traffic Flow
Line Load	Number of passengers travelling along a link between one stations and another	1141	Line Name Direction Station Order From Station To Station Traffic Flow
Station Entry	Number of passengers entering a station	455	Station Name Traffic Flow
Station Exit	Number of passengers exiting a station	455	Station Name Traffic Flow

4 Method

The implementation of the article's traffic network analysis is shown in Figure 4-1 which consists of the following parts: acquiring station data, projecting it onto the MSOA scale, analysing the spatial characteristics of the acquired data and comparing the coupling between the data.

In addition, this chapter does not deal with the real-time traffic representation and the creation of a database and the comparison of query effects.

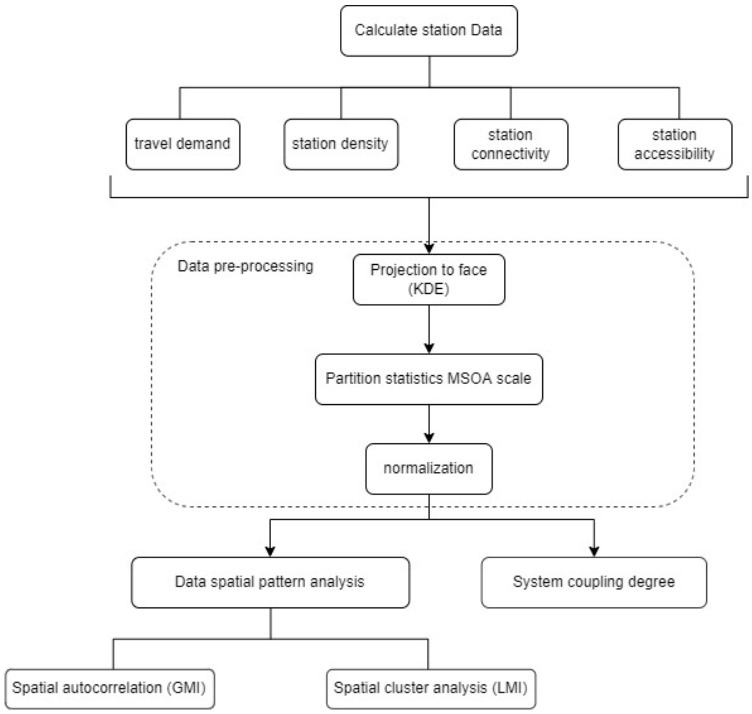


Figure 4-1 Methodology Flow

4.1 Quantitative analysis indicators

The quantitative analysis of transport networks and demand relationships requires a series of specific numerical indicators related to geographical zoning, so in this section we focus on how these indicators are obtained and calculated.

From the literature review in Chapter 2, research found that among the topological analysis metrics of transport networks, connectivity, accessibility and robustness are the most widely used and reflect the topological characteristics of transport networks, which are widely used in network analysis. However, since robustness has little relevance with traffic flow and demand, therefore, connectivity and accessibility are chosen to represent the network characteristics of the traffic network. In addition, the density of stations in the network is also taken into account.

The following section describes the calculation of each of the analysis metrics.

4.1.1 Travel demand

Public transport demand can usually be defined by the number of passengers, [38] but given that station flows are influenced by the structure of the network, the article argues that the flow of passengers at a single station is not representative of demand. Considering that when a single station cannot meet travel demand people tend to choose other stations with high connectivity or accessibility nearby. Therefore, the travel demand of passengers within a certain range is not or little affected by the structure of the transport network and can be represented by regional flows.

On the other hand, the station flow involves inbound and outbound passenger, underground line transit passenger and station transfer passenger. And the latter two do not take the geographical location of the station as the travel target, which are bring by the network structure and can not represent the travel demand of station location. So, the station travel demand using the sum of entry and exit to calculate.

Through the database query calculation, the final station demand is represented on the map results in Figure 4-2.

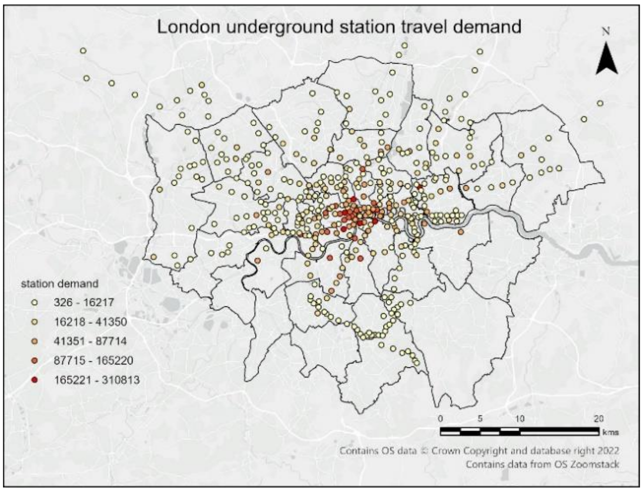


Figure 4-2 Station Travel Demand

4.1.2 Station density

The density of stations on a transport network is also a measure of topological soundness to reflect whether the number of stations in an area is commensurate with travel demand. When demand is high and station density is low, stations can carry more people, leading to overcrowding on nearby underground lines or increasing boarding times during peak periods, and limited stations mean limited rail options, which can limit travel choices and increase interchanges. On the other hand, when demand is low and there are many stations, this usually means that operating costs increase while not meeting the demand for enough trips, resulting in a waste of resources. Station density is calculated using the kernal density method, as described in section 4.2.

4.1.3 Connectivity

Connectivity, a key metric in graph theory, has been widely used to measure the affluence of transmissions within a network after its introduction into transportation networks, usually measured by centrality.[17,39]

Centrality metrics can be subdivided into point degree centrality, eigenvector centrality, intermediary centrality and proximity centrality, which are defined in Table 4-1. Considering the uses of different centrality metrics, the article chooses point degree centrality as a measure of connectivity, which is calculated by equation (4-1) and results in Figure 4-3.

Table 4-1 Centrality Indicators

Indicators	Definition	Characteristics
point degree centrality	Total number of direct connections that a node has in the network	Determine which nodes have the most direct impact in the network
eigenvector centrality	Scoring based on significant nodes connected to other significant nodes	Determine if there are influence clusters in a given network
intermediary centrality	Frequency of nodes lying on the shortest path between each pair of nodes in the network	Identify nodes that are used as connection nodes between nodes in the network
proximity centrality	The inverse of the sum of the shortest path distances from a node to all other nodes in the network	Determine which nodes are most closely associated with other nodes in the network

$$C_d(v_i) = \sum_{i=1,i\neq j}^{n} d_{ij} \qquad (4-1)$$

where n denotes the total number of nodes, v_i denotes the ith node and d_{ij} denotes the connected edge of node v_i.

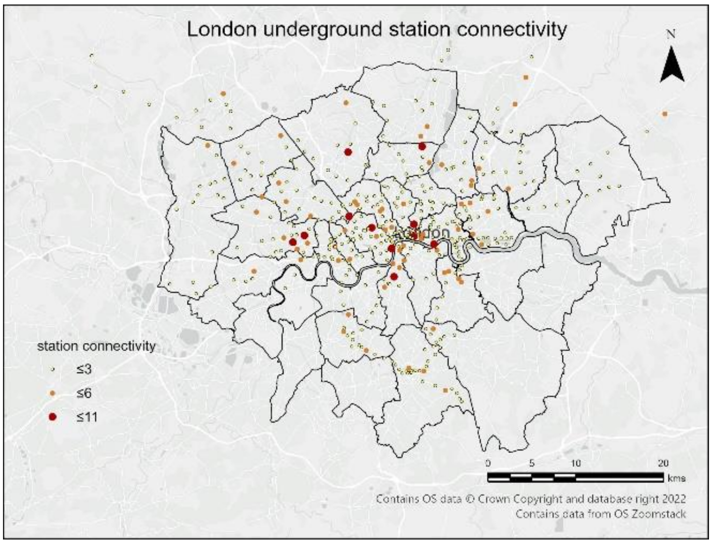


Figure 4-3 Station Connectivity

4.1.3 Accessibility

In graph theory, accessibility refers to the ease of getting from one vertex to another and is usually expressed by the average time or inverse of the distance between nodes travelling to each other.[40] Here the article chooses the inverse of the distance to calculate reachability, with the formula as in equation (4-2).

$$A_i = 1/\frac{\sum_{i=1,i\neq j}^{n} d_{ij}}{n-1} \qquad (4-2)$$

where n denotes the total number of nodes and d_{ij} denotes the distance of the shortest path from the ith node to the jth node.

The article here implements the accessibility calculation through ArcGIS, the specific process is shown in Figure 4-4. First establish the vector layer of the metro line, check the topology rules (no hanging point, no self-intersection and no self-overlap), then establish the network dataset of the relevant transportation network, afterwards create a new OD matrix of the relevant network dataset for the distance calculation between stations, import the starting point, destination for solution, and get the d_{ij} in equation (4-2), then the station accessibility calculation is carried out according to Eq. The result is shown in Figures 4-5.

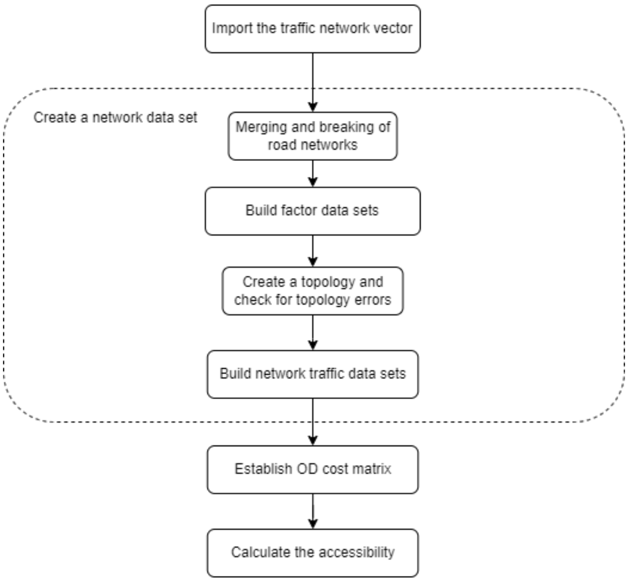


Figure 4-4 Accessibility Calculation Process

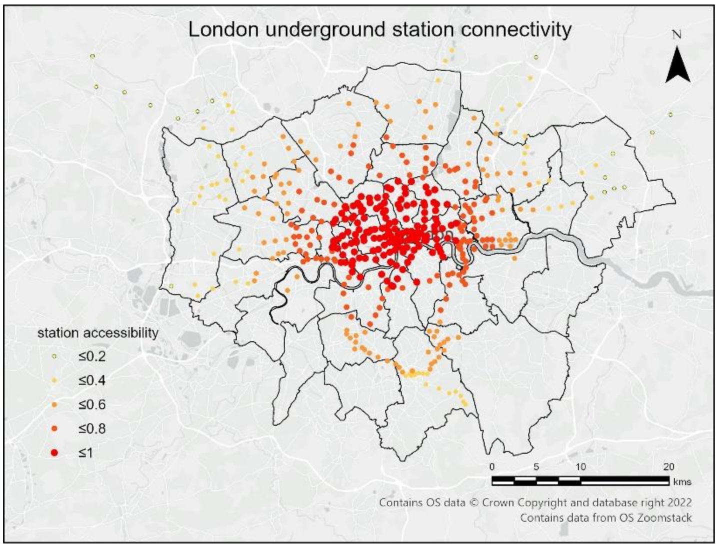


Figure 4-5 Station Accessibility

4.2 Data pre-processing

The above completes the acquisition of the basic station data. As station data is not a good representation of travel demand, as mentioned in 4.1.1, and site density and accessibility are usually represented in regions, research projects these data onto surfaces. Here we choose Middle Super Output Areas (MSOA) as the scale of analysis because the scale is too large and neither Lower Layer Super Output Area (LSOA) nor Output Area (OA) scales are too small, even below walkable accessibility, to reflect regional characteristics well.

The specific projection method is to use the Kernel Density Estimation (KDE) method to obtain the regional density of the points, and then calculate the mean value of the density at the MSOA scale by zonal statistics, and finally normalise the values.

KDE is a method used to estimate an unknown density function and is one of the non-parametric tests, which obtains the shape of the density from the data rather than the parameters. KDE assumes that the density is available anywhere in the study area where the point data are located, rather than only at the location where the event occurs.[41,42] Density estimation is achieved specifically by covering the point with a smooth surface that decreases when the distance to the point increases with the surface. The formula for calculating the kernel density at (x, y) is shown in equation (4-3) and equation (4-4). The density results are shown in Figure 4-6, the partition statistics are shown in Figure 4-7.

$$D(x,y) = \frac{1}{h^2}\sum_{i=1}^{n}\left[\frac{3}{\pi}\cdot v_i\left(1-\frac{(x_i-x)^2+(y_i-y)^2}{h^2}\right)^2\right] \tag{4-3}$$

$$h = 0.9\cdot \min\left(SD, \sqrt{\frac{1}{\ln(2)}}\cdot D_m\right)\cdot sum(v_i)^{-0.2} \tag{4-4}$$

where i=1,...,n is the input point, h is the bandwidth, v_i is the value of the point (x_i, y_i) is the coordinates of the point, D_m is the median distance from the mean centre, and SD is the standard distance.

Normalisation is used to eliminate the effects of different magnitudes between indicators, and the data are standardised to be limited to the range [0, 1] so that the indicators are in the same order of magnitude. Here, the data are linearly transformed using the standardisation of deviations as in equation (4-5).

$$z_i = \frac{x_i - x_{min}}{x_{max} - x_{min}} \tag{4-5}$$

where i=1,...,n, x_i is the ith data, x_{min} and x_{max} are the minimum and maximum values of the data.

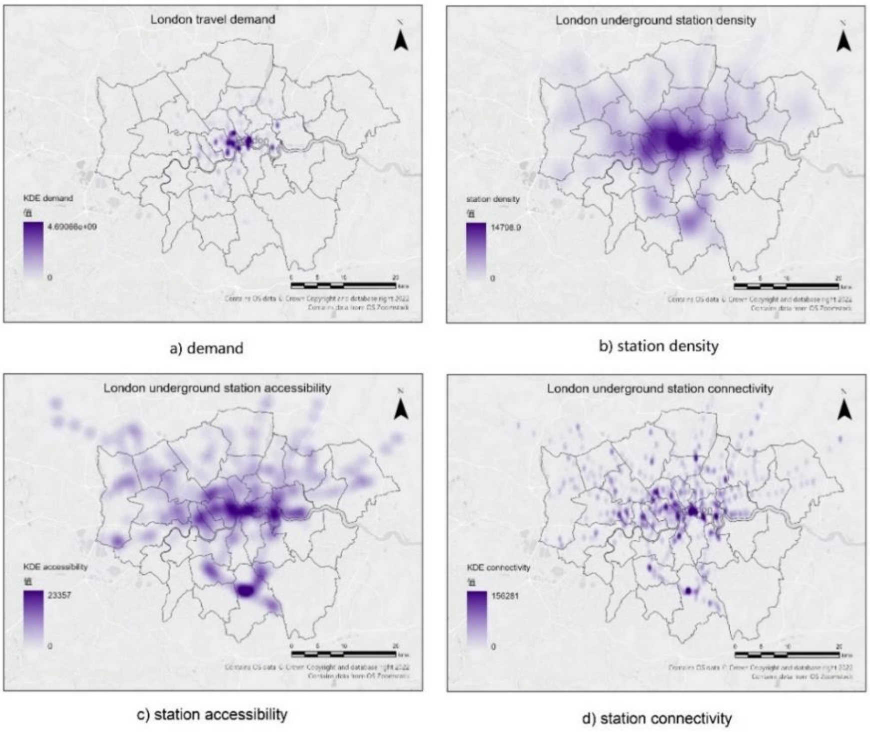


Figure 4-6 Result of KDE

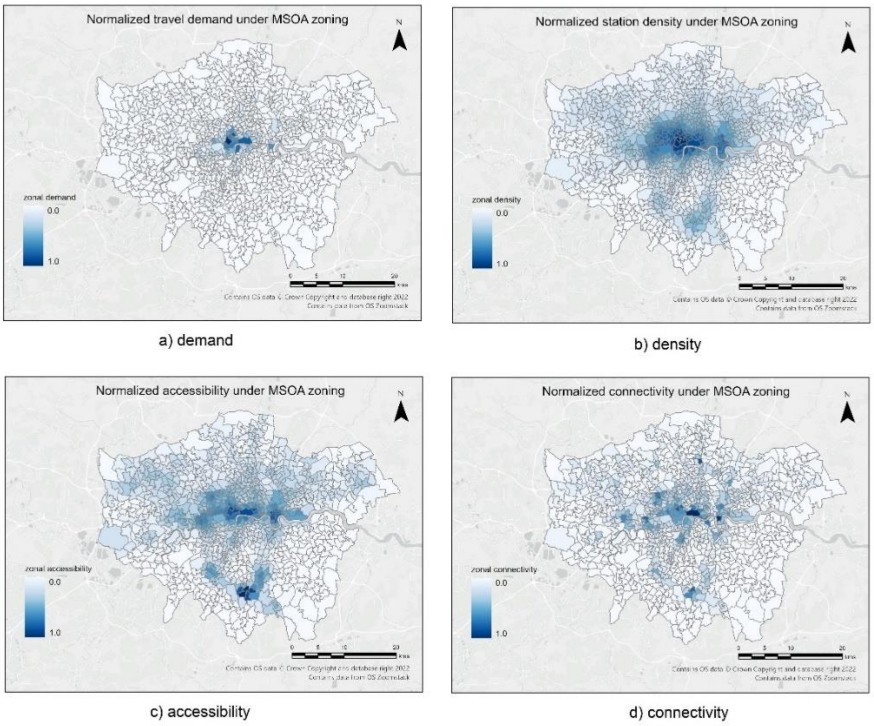


Figure4-7 Result of Zonal Statistic

4.3 Spatial autocorrelation

According to the first law of geography, anything is correlated with something else, but what is near is more correlated than what is far.[43] Geographic data may not be independent of each other and exhibit spatial dependence due to spatial interactions and spatial diffusion.[44] Therefore, for geographical data we need to determine whether the data are spatially dispersed and clustered or random using Global Moran's I (GMI) and to determine the spatial clustering pattern using Local Moran's I (LMI).

The GMI is a spatial extension of the product-moment correlation coefficient with an exponential result between [-1, 1] and the statistical indicators z-score and p-value to determine the randomness of the data. The calculation formula is shown in equation (4-6) to (4-9).

$$I=\frac{n}{\sum_{i=1}^{n}\sum_{j=1}^{n}w_{ij}}\frac{\sum_{i=1}^{n}\sum_{j=1}^{n}w_{ij}(x_i-\bar{x})\left(x_j-\bar{x}\right)}{\sum_{i=1}^{n}(x_i-\bar{x})^2} \tag{4-5}$$

$$z=\frac{I-E[I]}{\sqrt{V[I]}} \tag{4-6}$$

$$E[I]=-1/(n-1) \tag{4-7}$$

$$V[I]=E[I^2]-E[I]^2 \tag{4-8}$$

where w_{ij} is the spatial weight of i and j, x_i is the value of element i, and $\bar{x}$ is the mean element value.

The resultant judgement is shown in Table 4-2

Table 4-2 GMI Judgement Criteria

spatial pattern	reliability (p-value)	z-score
dispersed	0.01	(-∞, -2.58)
	0.05	[-2.58, -1.96)
	0.1	[-1.96, -1.65)
random		[-1.65, 1.65)
clustered	0.1	[1.65, 1.96)
	0.05	[1.96, 2.58)
	0.01	[2.58, +∞)

LMI is an indicator to identify spatial heterogeneity based on the equation (4-5). The Moran scatter plot is divided into four regions using GMI as the boundary, high value clustering (HH), low value clustering (LL), anomalous high values surrounded by low values (HL) and anomalous low values surrounded by high values (LH), as shown in Figure 4-8.

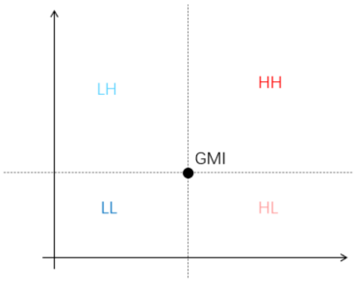


Figure 4-8 Moran Scatter Plot Partitioning

4.4 Coupling

The coupling degree reflects the degree to which multiple systems interact with each other, and the more similar the systems are the higher the coupling degree.[45,46] For two systems, the coupling degree is calculated as in equations (4-9).

$$C = \frac{2\sqrt{U_1 U_2}}{U_1 + U_2} \tag{4-9}$$

where U_1, U_2 represent the values of the two systems respectively.

For the classification of coupling degree, refer to the criteria proposed by song et al.[47] in Table 4-3.

Table 4-3 Classification of Coupling Degree

C value	[0, 0.2]	(0.2, 0.4]	(0.4, 0.5]	(0.5, 0.6]	(0.6, 0.8]	(0.8, 1]
coupling level	Severe disorder	Moderate disorder	Disorder	Coupling	Well coupling	Quality coupling

5 Result and Discuss

Based on the objectives of the study, the results and discussion section are divided into two parts, the results concerning the analysis of the rail network and the results of the graph database.

5.1 Analysis of the rail network

In this section research analyses the spatial patterns and correlations of demand and the three rail network indicators.

5.1.1 Spatial analysis

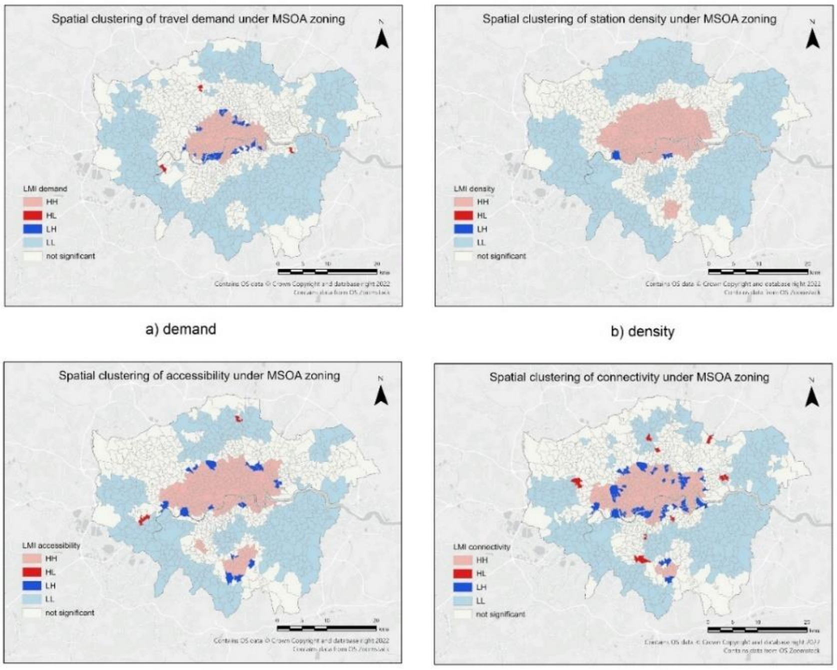
The spatial analysis includes GMI, which explores whether there is spatial dependence in the data, and LMI, which explores spatial clustering patterns.

The GMI analysis allows us to obtain the spatial distribution patterns of the four values after the zonal statistics, as shown in Table 5-1. It can be seen that all of four values show very reliable and obvious spatial clustering, with the probability of a spatial random distribution being approximately 0. It is therefore very important to use the LMI to analyse the data for spatial clustering and outliers.

The LMI results are shown in Figure 5-1, where insignificant refers to regions where the p-value of the Moran index is greater than 0.1, HH denotes high value clustering, HL denotes high values surrounded by low values, LH denotes low values surrounded by hot spots and LL denotes low value clustering.

Table 5-1 The Result of GMI

Indicator	Moran's I	p-value	z-score	Spatial Pattern
travel demand	0.527695	0	43.906825	clustered
station density	0.890356	0	72.67732	clustered
connectivity	0.508547	0	41.632731	clustered
accessibility	0.737778	0	60.249908	clustered



a) demand b) density

c) accessibility d) connectivity

Figure 5-1 The Result of LMI

For the clustering pattern of travel demand, it can be seen that the high values appear to be very clearly clustered, mainly in the central part of London. However, there is an area of MSOA surrounded by low values in the centre of Richmond upon Thames in West London, in the north of Greenwich in East London and in the centre of Barnet in the north. Low values also appear to cluster very clearly, with areas of low demand surrounding the London boundary in areas other than north-west London, and a number of low value areas surrounding the boundary of the central high value cluster.

For Figure 5-1 b) LMI for site density, the pattern is broadly similar to that of the travel demand clustering analysis, with high value clusters in central London and low value clusters in the periphery. However, when comparing demand, the high value clusters for site density are more extensive particularly in the east-west direction and appear to cluster in south London, with the low value clusters being less in the south and more in the north. There are also no anomalous high values, with a few anomalous low values occurring south of the city centre.

As Figure 5-1 c) result of accessibility, the high value clusters are also distributed in the city centre and have a longer range than Figure b) density in the east-west direction, with two areas in the south appearing to have a greater range of high value clusters than site density. The low value clusters are distributed in North, South West and South

East London as in the other figures, but the low value areas are significantly smaller in the South West and South East. In addition, anomalous high values occur in both the west and north, and anomalous cold spots occur in the central and southern high value clusters.

Finally, for connectivity, the high and low value clusters are similar to c) accessibility, but the distribution of outliers is the largest and cluttered, with a very large number of low value areas around the high value clusters and anomalous high values distributed in the inner ring of the cluttered low value clusters.

In conclusion, although the magnitude of the spatial clustering of the four indicators varies, all four indices show high clustering in central London, with low clustering in the north, south-west and much of the east. All three indicators, with the exception of travel demand, show high clustering in the south.

Based on the LMI results, the spatial clustering distribution of the data shows that, with the exception of some anomalies and insignificant areas, the topological indicators of the rail network are generally compatible with travel demand, which means the structure of the rail network can meet the changes in travel demand. The specific distribution is that high travel demand in the city centre corresponds to high topological indicators and a larger and more comprehensive coverage area, probably in view of future development changes. In the case of low peripheral demand, the network structure has a similar geospatial distribution of cold values, with a smaller overall coverage area, which is compatible with travel demand and provides better access for people in low travel demand areas. However, the network structure does not provide good coverage in areas of unusually high values of travel demand and in some insignificant areas.

5.1.2 The degree of coupling of all-day travel demands

In this section research explore the extent to which the transport network and demand are matched by the coupling degree of the three transport network structure indicators and travel demand, as expressed through the coupling degree.

In the calculation of the coupling degree, some regions are calculated to yield a travel demand of 0. These regions are not discussed in this section. The colour of the MSOA region in the coupling degree diagram indicates the coupling degree of the indicator and the overall travel demand, and the point data indicates the normalised result of the three indicators. Refer to Table 4-2 for the specific coupling degree indicator division. coupling degree results are shown in Figure 5-2 to Figure 5-5.

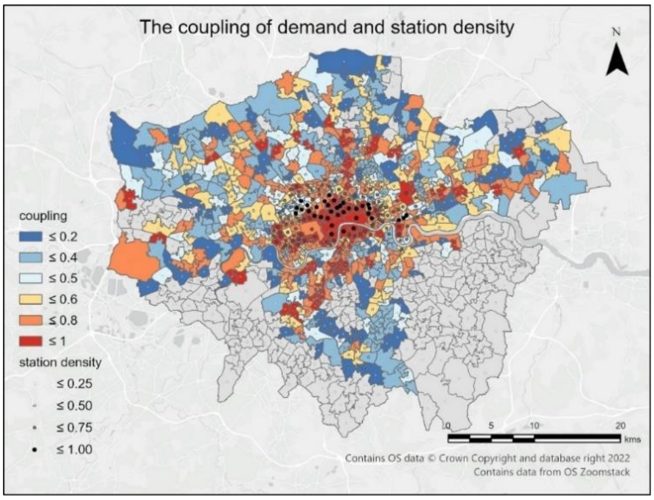


Figure 5-2 Coupling of Station Density and Travel Demand

For the spatial distribution of coupling between site density and travel demand is scattered, but central London is predominantly a high coupling area, with some high coupling interspersed with low coupling in the south. Combined with the station density values, the areas with station densities between 0.75 and 1 are mainly concentrated north of Thames in central London, and the coupling corresponding to these values is largely large at 0.6, i.e. moderate or high coupling, with some exceptions, mainly in the north-west and east of the high coupling area. In contrast, there is no pattern in the coupling corresponding to low station density values.

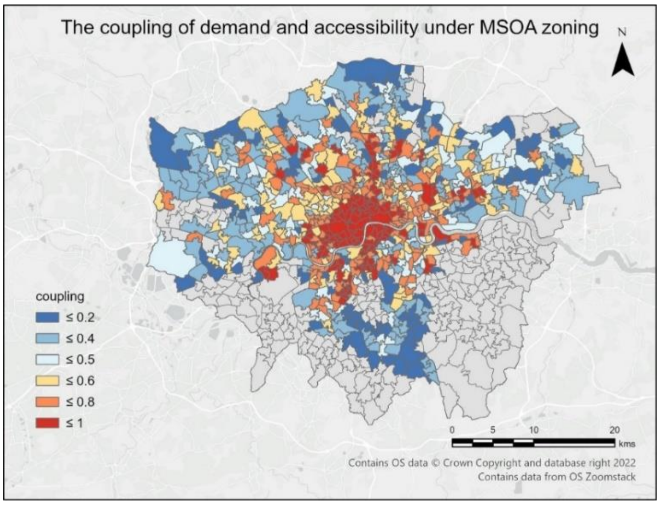


Figure 5-3 Coupling of Accessibility and Travel Demand

For accessibility, high coupling is again mainly concentrated in central London but is more extensive than the sites, low coupling is distributed around the area, the southern area is mainly low coupling distributed, and the northern area is largely decreasing in coupling as the distance to the city centre increases, but interspersed with some areas of high coupling.

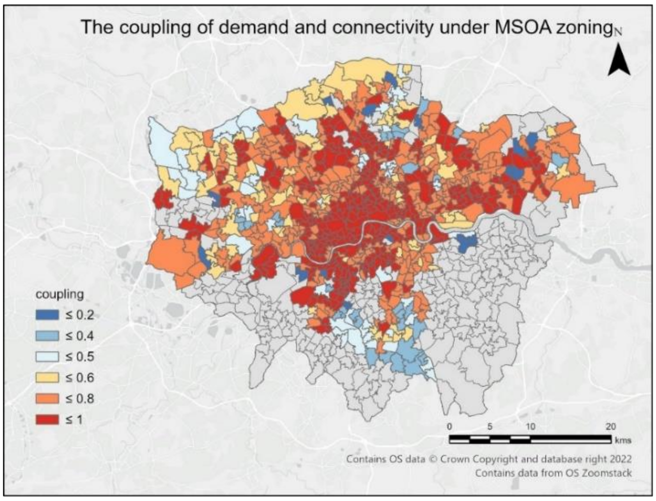
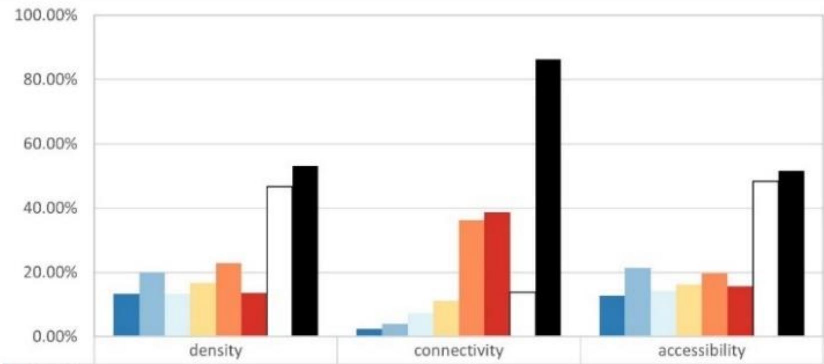


Figure 5-4 Coupling of Connectivity and Travel Demand

As for the coupling of network connectivity and travel demand represented in Figure 5-4, it shows a very good high degree of coupling, with the vast majority of areas showing moderate and high coupling with coupling greater than 0.6, but there are still small areas of low coupling concentration in the south, but these areas do not show severe uncoupling.

The statistical values regarding the coupling degree regions are shown in Figure 5-5, London rail transit has the best coupling degree in terms of connectivity and travel demand, resulting in a coupled MSOA ratio of 86%, station density and accessibility coupling and imbalance situation ratio is basically the same but coupled regions are slightly higher at 53.24% and 51.62% respectively.



	density	connectivity	accessibility
[0, 0.2]	13.38%	2.43%	12.84%
(0.2, 0.4]	20.00%	4.05%	21.35%
(0.4, 0.5]	13.38%	7.30%	14.19%
(0.5, 0.6]	16.76%	11.22%	16.22%
(0.6, 0.8]	22.84%	36.22%	19.73%
(0.8, 1]	13.65%	38.78%	15.68%
[0, 0.5]	46.76%	13.78%	48.38%
[0.5, 1]	53.24%	86.22%	51.62%

Figure 5-5 Statistical Scale Diagram of Coupling

In general, rail design and current travel demand are generally compatible. A comparison of the coupling statistics map and the geographical distribution map shows that, with the exception of south London, connectivity is compatible with people's travel needs, and station density and accessibility are generally compatible with travel needs. And in central London in particular, all three indicators show a high degree of coherence.

And combined with the spatial distribution of the values, it can be found that areas where station density and accessibility are not coupled are largely the result of higher station density and accessibility values than travel demand, reflecting the rail network's confidence in their future growth needs. Based on this characteristic, governments could encourage the development of these high transit design areas with low trip density, i.e. Transit Oriented Development (TOD), which would better integrate transit-driven supply and land use-driven demand while improving pedestrian friendliness.[39,48]

5.1.3 Coupling by time period

In section 5.1.2 research found that the spatial distribution of coupling between transport network indicators and travel demand, with the exception of connectivity, is uneven and there are few areas of high coupling, which may be related to the extreme distribution of travel demand and its high concentration in central London. Considering the fact that the spatial distribution of travel demand may change depending on the travel patterns of people at different times of the day, this may lead to changes in the coupling degree from time to time. This section therefore explores the coupling between the structure of the rail network and demand at different time periods.

Similar to 4.1.1, we obtain the normalised MSOA scale travel demand data for different time periods, as shown in Figure 5-6.

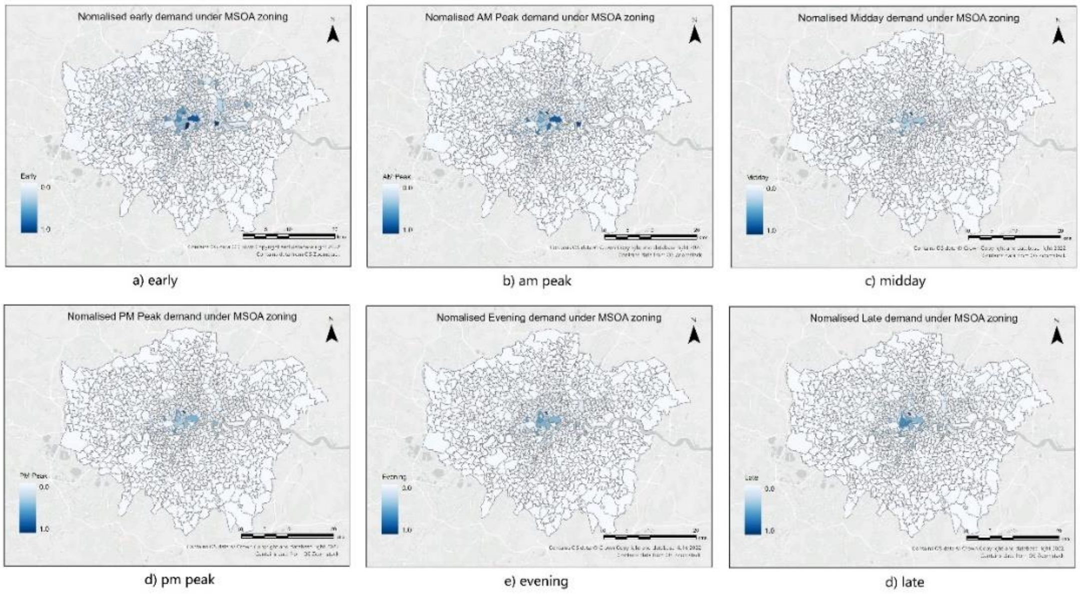


Figure 5-6 Travel demand by time of day

Although the demand at different times of the day remains concentrated in the city centre at different times of the day, the comparison of total trips shows more variation and there is an increase in demand for peripheral trips during the Early and AM Peak phases.

Again following 4.4 the coupling between the different time periods is obtained as shown in Figure 5-7 to Figure 5-9.

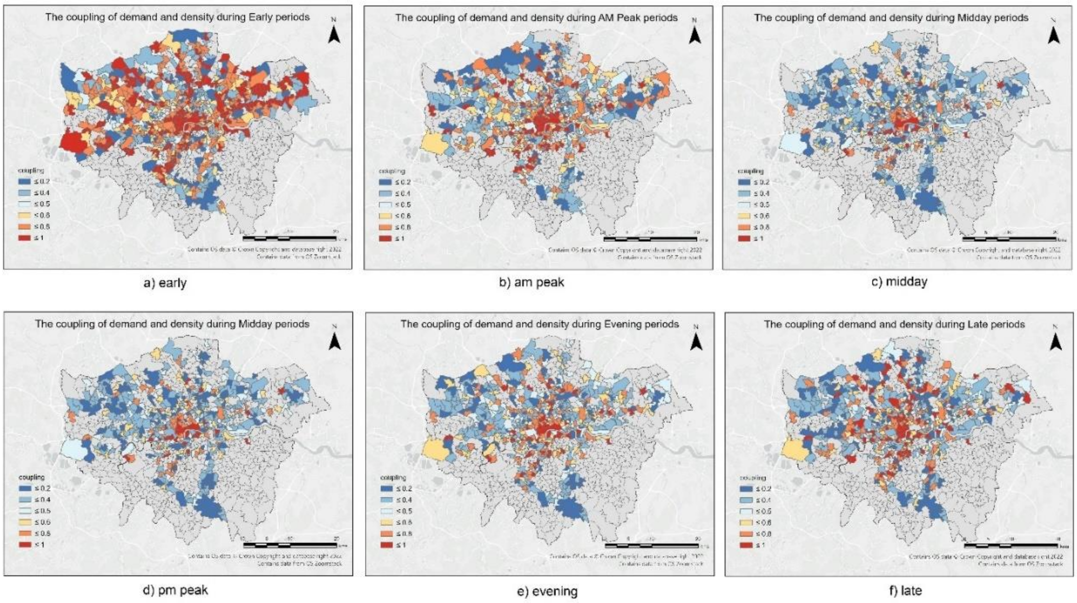


Figure 5-7 Coupling of Station Density and Travel Demand at Different Times of Day

The coupling of station density and travel demand shows temporal variability, with early hours showing a better spatial distribution of coupling than all-day station density (Figure 5-2), with increased areas of high coupling. b) am peak, e) evening and f) late have a similar spatial distribution of coupled areas to the coupling of all-day flows, but late has more areas of high coupling (0.8-1). Finally, c) midday and d) pm peak show poorer results, with far more uncoupled areas than coupled areas and very few highly coupled areas in central London.

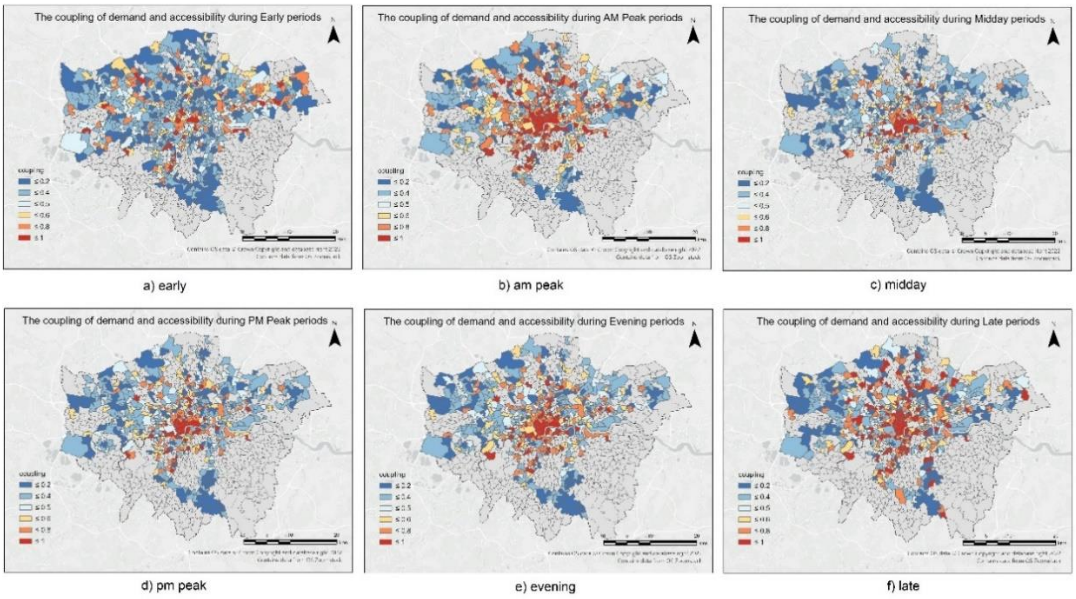


Figure 5-8 Coupling of Accessibility and Travel Demand at Different Times of Day

The accessibility results are generally poor, with the b) AM peak spatial distribution being similar to the all-day travel demand results, but all other time periods show worse coupling, especially the Early time period, which is essentially uncoupled across the whole domain and has very few central coupling areas (0.6-1).

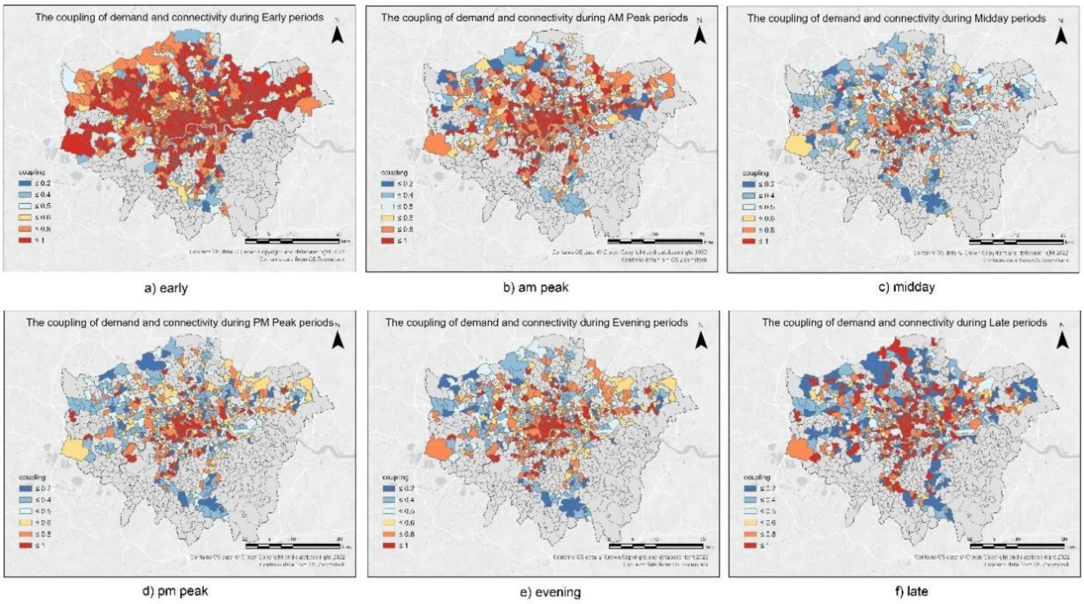


Figure 5-9 Coupling of Connectivity and Travel Demand at Different Times of Day

The coupling between connectivity and travel demand shows a large temporal variation, but the coupling is high in the central London area at all times. The spatial and temporal distribution of coupling in the early hours is similar to that in Figure 5-4, showing a high coupling across the whole area. The spatio-temporal distribution of coupling is similar for AM peak, but slightly less so than for Early. The spatial-temporal distribution of coupling is similar for the three time periods c) Midday, d) PM peak and e) Evening, with a heterogeneous distribution of low coupling areas dominating around the central high coupling. Finally, for the f) Late hours, the uncoupled areas are similar to the results for c), d) and e), but the coupling level increases with a large number of highly coupled (0.8-1) areas.

Combined with the all-day trips we perform a statistical analysis, as shown in Figure 5-10.

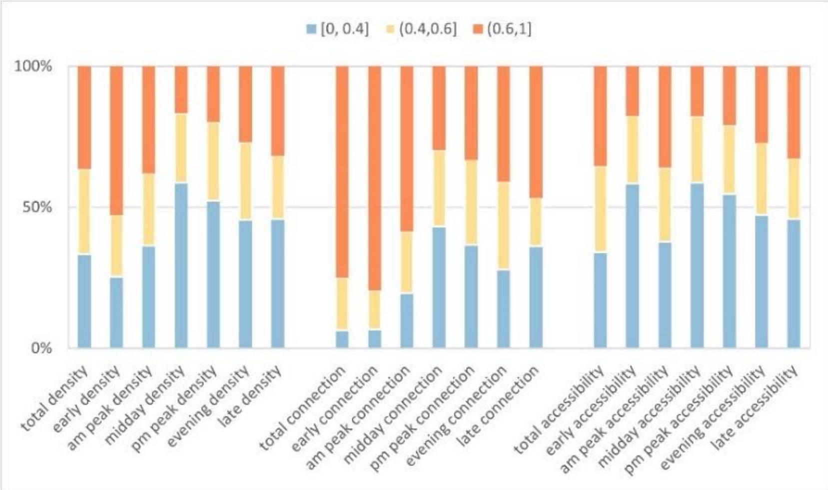


Figure 5-10 Coupling Statistics for Different Time Periods

From a temporal point of view, the different time periods show variability, with the coupling of the different indicators increasing in the order of Midday, PM peak, Evening, Late and AM peak. For the Early time period, the best results are obtained for site density and accessibility, but the coupling for accessibility is the worst among the time periods.

In terms of metrics, connectivity has the best results, but the coupling and non-coupling results are about the same for all time periods except for the Total and Early time periods, which have high coupling rows across the domain. Station density is the second best result, with more uncoupled rows for midday, PM peak, Evening and Late. Lastly, accessibility, coupling with travel demand is poor across all time periods.

In general, connectivity for all hours, station density for Early and AM peak and accessibility for AM peak are generally compatible with travel demand. Furthermore, although the spatial distribution of travel demand is similar across time periods, it still has an impact on the degree of coupling in the transport network and therefore the network design needs to take into account the impact of time periods. In conclusion, the subsequent development of the network structure needs to focus on accessibility and station density and to take into account the demands of the four periods Midday, PM peak, Evening and Late.

5.2 Database representation

The database in this thesis is primarily used to store raw rail traffic data and to perform the resultant queries required for traffic network analysis. The performance of the graph database and the relational database used in relation to the traffic network is compared by the content required for the query analysis.

5.2.1 Database structure

The graph database was built using Neo4j Desktop and its own Neo4j Browser via the edge connection node, and the relational database was built using PostgreSQL and its development platform pgAdmin implementation via the foreign key connection attribute table. The specific structure is shown in Figure 5-11 and Figure 5-12.

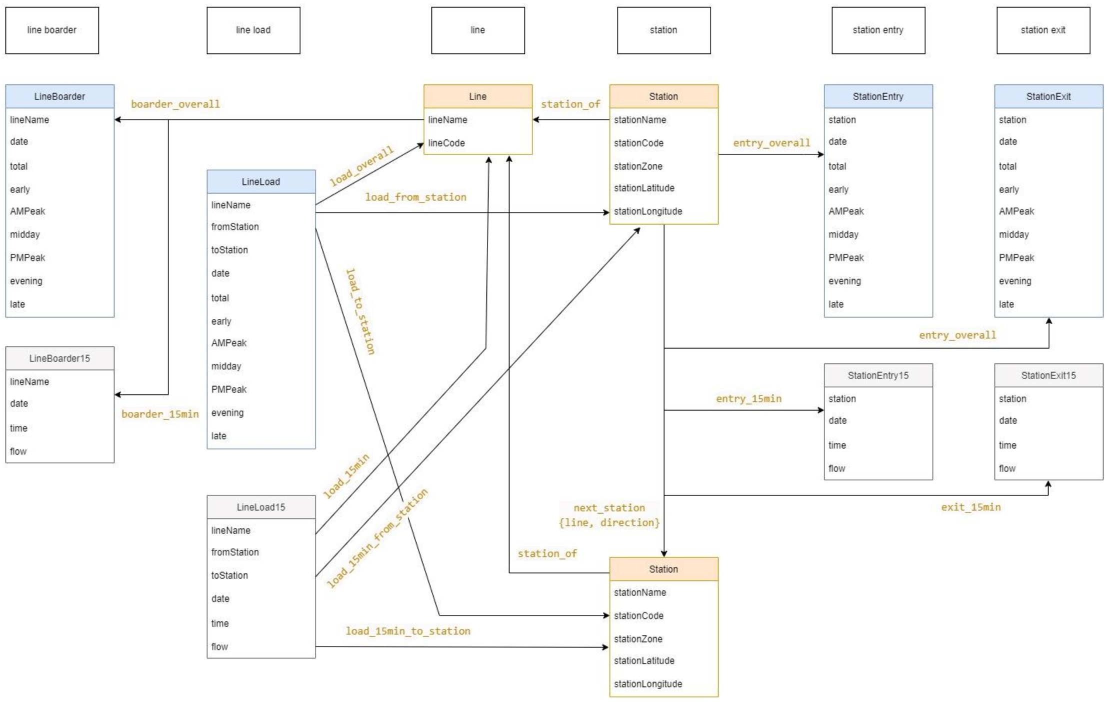


Figure 5-11 Graph Database Structure

According to Figures 5-10, the boxes represent the types of nodes in the graph database and the connecting arrows indicate the edges connecting the nodes in the database. The graph database is constructed in two parts, with the structure of the rail network including stations and lines and the connections between them represented in orange, and the passenger flows on the transport network represented in blue and grey, with a total of 11 types of nodes and 14 types of edges.

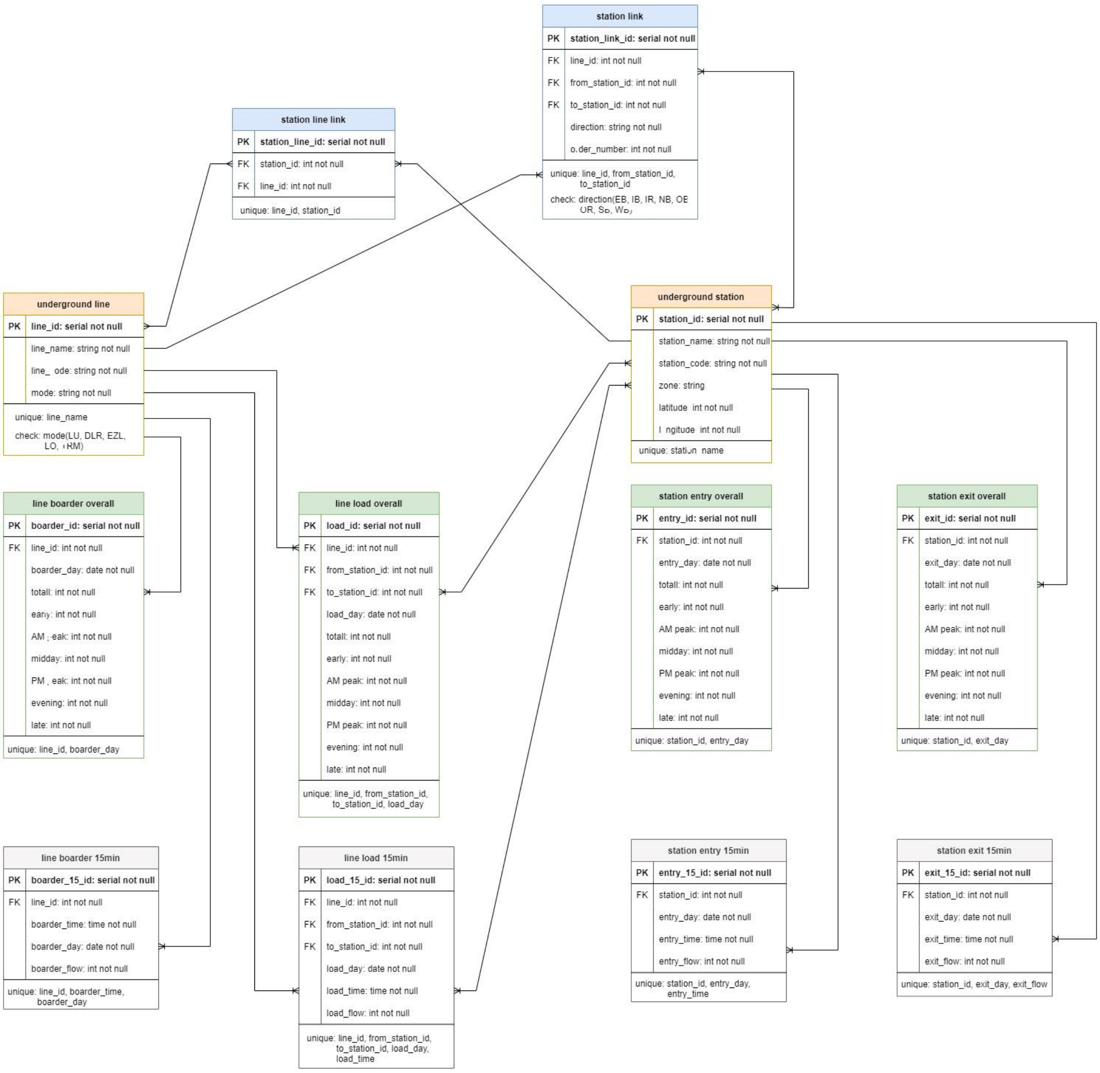


Figure 5-12 Relational Database ER Diagram

The boxes in a relational database list information about the database entities and the connecting lines show the links between the entities. The relational database can be divided into three parts: orange boxes represent traffic network routes and stations, blue boxes represent traffic network connections, and green and grey represent network traffic. The database specifically consists of 12 tables and 17 connections (foreign key).

Both databases contain the same information, which are stations, lines, connections to the transport network and three types of flows at different times (station flow, line load flow and line boarder flow), but the diagram database uses fewer entities and connections due to its highly connected structure.

5.2.2 Build of database

After analysing the structure, this section will compare the implementation of the two database specific builds, the graph database using Cypher statements and the relational database using SQL statements. The graph database in Appendix A and the relational database build statements in Appendix B are compared and analysed.

The graph database has a more streamlined build statement, allowing nodes and edges to be created and data to be inserted directly. However, the relational database does not import data until after a series of tables and constraints have been created, which results in it requiring more statements. The relational database build process is shown in Figure 5-13.

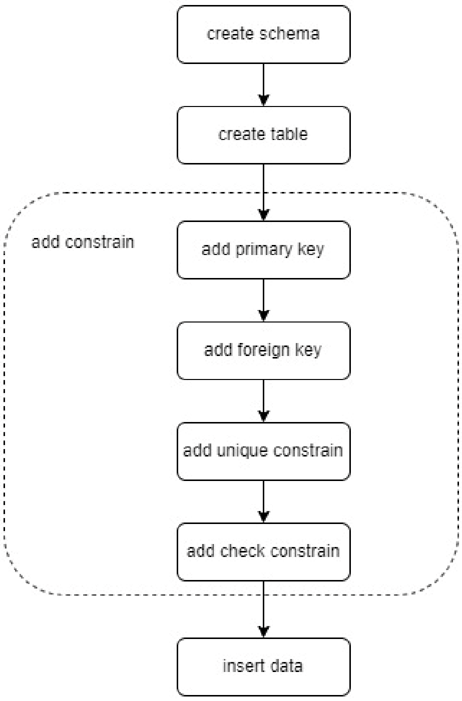


Figure 5-13 Relation Database Build Flow

5.2.3 Query performance

In this section the research compare the performance of the two databases that have been created, by comparing the time of query given different numbers of hops and the query of the data used in the network analysis section of 5.1, using the query statements in Appendix C and Appendix D. Given that a single query will produce errors, the query times are averaged over five queries for each query, with those less than 1ms being calculated as 1ms.

(1) Query language and results

The query language (Appendix C and D) shows that Neo4j is generally more concise and brief, whereas pgAdmin requires a large number of statements to join tables together for queries.

For the presentation of the query structure, pgAdmin presents it in a table, while Neo4j has four representations - Graph, Table, Text and Code - the one that best demonstrates the characteristics of a graph database is the return of results via graph, which is very intuitive, but the graph mode limits the representation of results to 300 nodes. Here the graph mode of pgAdmin and Neo4j return the results of the traffic network structure as shown in Figure 5-14 and Figure 5-15, where it can be seen that the latter is more intuitive.

	station_link_id integer	line_id integer	from_station_id integer	to_station_id integer	direction character varying (20)	order_number integer	line_id integer	line_name character varying (50)	line_cc charac
1	1	1	131	235	NB	1	1	Bakerloo	BAK
2	2	1	235	416	NB	2	1	Bakerloo	BAK
3	3	1	416	135	NB	3	1	Bakerloo	BAK
4	4	1	135	81	NB	4	1	Bakerloo	BAK
5	5	1	81	302	NB	5	1	Bakerloo	BAK
6	6	1	302	293	NB	6	1	Bakerloo	BAK

Total rows: 1000 of 1141 Query complete 00:00:00.065 Ln 14, Col 32

Figure 5-14 pgAdmin's traffic network structure query results (1000 rows)

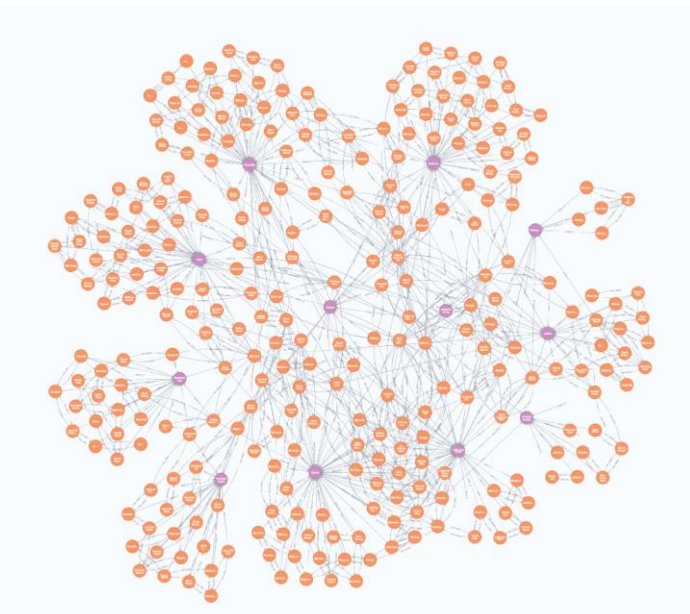


Figure 5-15 Neo4j's traffic network structure query structure (300 nodes)

(2) Queries given relationships

The research will compare the different entities in the database for jumping. Specifically, the number of relational data entities is used as a statistic to compare queries that connect different numbers of entities and involve network structures, the results are shown in Table 5-2.

Table 5-2 Results for a Given Relationship

entities number	entitie and link	result number	Neo4j (ms)	pgAdmin (ms)
1	line	19	1	48.5
	station	455	1	52
2	line-line_boarder	19	6.2	53.6
	station-staion_entry	455	10.9	59.3
3	line-station	575	6.1	55.5
4	line-from_station-to_station	1743	4.6	69.1

As you can see, Neo4j has better query performance compared to pgAdmin, with query times generally within 10ms. In contrast, pgAdmin query times are generally greater than 50ms, even for a joinless entity query against 19 results still averaging around 48ms.

In addition, except for results relating to the number of inbound entries, pgAdmin query times increase with the number of entities and the number of results. Neo4j, however, behaves differently, in queries involving the structure of the transport network (line-staiton and line-from_station-to_station) take less time compared to line-line_boarder, which links two entities with 19 results, despite their more complex structure compared to other queries with more results.

(3) Querying for network analysis

The query structure for the 5.1 network analysis component is shown in Table 5-3. Similar to Table 5-2, Neo4j's queries generally perform better, with query times essentially less than 10ms, particularly for multi-hop queries relating to graph structure (connectivity and accessibility).

Table 5-3 Query for Network Analysis

queries	entities and link	result number	Neo4j (ms)	pgAdmin (ms)
demand	station station entry overall station exit overall	455	11.1	69.1
density	station	455	1	54
connectivity	station station link	455	5.1	56.1
accessibility	station station link	1139	1.1	56.2

In summary, by comparing these queries we found that Neo4j can be written more concisely using the Cypher language, has more options for visualising results and is very intuitive, and that Neo4j performs better than pgAdmin in terms of query speed, especially for multi-hop queries about network structures.

6 Conclusions and Recommendations

The previous chapter discussed the degree of matching between the London rail network and existing travel demand and the performance of the graph database in the transport network using a graph database. This section will summarise the results and limitation of the article.

6.1 Summary

6.1.1 The rail network

1. The analysis of the spatial clustering model shows that the three transport network indicators of station density, accessibility and connectivity are able to fit the travel demand in the spatial pattern, showing high values of clustering in central London and low values of clustering in the periphery of London (north, south-west and south-east). The high values of the transport network cover a wider area, reflecting the tolerance of the transport network to the outward expansion of future demand.

2. For the coupling of travel demand and topology indicators throughout the day, the three indicators are generally compatible with travel demand. connectivity has more than 80% regional coupling, station density and accessibility are slightly less coupled but more than half, and central London has a degree of coupling.

3. Analysing the coupling at different times of day, the article finds that there is a large time variation in the results. For all-day connectivity, station density before midday and accessibility during the morning peak match the travel demand. The match between topology indicators and travel demand is generally poor after the end of the morning peak.

Overall, the clustering patterns and indicators suggest that London's rail network is generally compatible with current travel demand. However, there is a need for better integration of transport network planning and travel demand in the surrounding areas, mainly in terms of accessibility and station density, as well as in terms of planning for travel demand after the morning peak.

6.1.2 Graph database

The article compares the use of graph databases and relational databases for traffic networks in terms of structure, language, result presentation and query time. Neo4j is found to have a more streamlined structure than pgAdmin, a more concise language for database creation and querying, more intuitive result presentation, and shorter query times for different entity connection query structures. Neo4j has better performance results for multi-hop queries on network structures.

6.2 Limitation

There are several shortcomings in the article with regard to the comparison of the performance of the rail network and the database respectively, which could be further investigated in the future.

6.2.1 Analysis of the rail network

1. although rail travel accounts for 72.4% of public travel choices, the demand calculated from rail stations is still not fully representative of the true travel demand, and more accurate travel demand can be derived by considering additional transport modes and social research.

2. The article analyses the travel data for Friday 2019 and lacks analysis and discussion of travel patterns for other years and holidays. Changes in travel demand, network structure and their coupling relationship can be analysed by studying data at different times of the year.

6.2.2 Databases

1. The database size created in the article is approximately 13.6GB, and a comparison of different storage size databases is missing.

2. The article graph database was built using neo4j and the relational database was built using PostgreSQL, a comparison of the different databases is missing.

Reference

[1] Du, Q., Zhou, Y., Huang, Y., Wang, Y., & Bai, L. (2022). Spatiotemporal exploration of the non-linear impacts of accessibility on metro ridership. *Journal of Transport Geography*, *102*, 103380.

[2] Chiou, Y. C., Jou, R. C., & Yang, C. H. (2015). Factors affecting public transportation usage rate: Geographically weighted regression. *Transportation Research Part A: Policy and Practice*, *78*, 161-177.

[3] Transport for London. (2019). crowding.data.tfl.gov.uk. From http://crowding.data.tfl.gov.uk/.

[4] Transport for London (TfL). (2018). Underground Services Performance.

[5] London transport museum. (2022). A very short history of the Underground. From A very short history of the Underground | London Transport Museum (ltmuseum.co.uk).

[6] Transport for London. (2022). London underground history. From https://tfl.gov.uk/corporate/about-tfl/culture-and-heritage/londons-transport-a-history/london-underground.

[7] Cats, O., Wang, Q., & Zhao, Y. (2015). Identification and classification of public transport activity centres in Stockholm using passenger flows data. *Journal of Transport Geography*, *48*, 10-22.

[8] Kloosterman, R. C., & Musterd, S. (2001). The polycentric urban region: towards a research agenda. *Urban studies*, *38*(4), 623-633.

[9] Roth, C., Kang, S. M., Batty, M., & Barthélemy, M. (2011). Structure of urban movements: polycentric activity and entangled hierarchical flows. *PloS one*, *6*(1), e15923.

[10] Derrible, S., & Kennedy, C. (2011). Applications of graph theory and network science to transit network design. *Transport reviews*, *31*(4), 495-519.

[11] Derrible, S., & Kennedy, C. (2009). Network analysis of world subway systems using updated graph theory. *Transportation Research Record*, *2112*(1), 17-25.

[12] Derrible, S. (2012). Network centrality of metro systems. *PloS one*, *7*(7), e40575.

[13] Lin, J., & Ban, Y. (2013). Complex network topology of transportation systems. *Transport reviews*, *33*(6), 658-685.

[14] Chaves-Fraga, D., Priyatna, F., Cimmino, A., Toledo, J., Ruckhaus, E., & Corcho, O. (2020). GTFS-Madrid-Bench: A benchmark for virtual knowledge graph access in the transport domain. *Journal of Web Semantics*, *65*, 100596.

[15] Holzschuher, F., & Peinl, R. (2016). Querying a graph database–language selection and performance considerations. *Journal of Computer and System Sciences*, *82*(1), 45-68.

[16] Virgilio, R. D., Maccioni, A., & Torlone, R. (2014, October). Model-driven design of graph databases. In *International Conference on Conceptual Modeling* (pp. 172-185). Springer, Cham.

[17] Maduako, I. D., Wachowicz, M., & Hanson, T. (2019). Transit performance assessment based on graph analytics. *Transportmetrica A: Transport Science*, *15*(2), 1382-1401.

[18] Lam, T. N., & Schuler, H. J. (1982). Connectivity index for systemwide transit route and schedule performance. *Transportation Research Record*, (854).

[19] Musso, A., & Vuchic, V. R. (1988). *Characteristics of metro networks and methodology for their evaluation* (pp. 22-33). Washington, DC, USA: National Research Council, Transportation Research Board.

[20] Derrible, S., & Kennedy, C. (2010). Characterizing metro networks: state, form, and structure. *Transportation*, *37*(2), 275-297.

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

[22] Derrible, S., & Kennedy, C. (2010). The complexity and robustness of metro networks. *Physica A:*

Statistical Mechanics and its Applications, *389*(17), 3678-3691.

[23] Welch, T. F., & Mishra, S. (2013). A measure of equity for public transit connectivity. *Journal of Transport Geography*, *33*, 29-41.

[24] Laporte, G., & Pascoal, M. M. (2015). Path based algorithms for metro network design. *Computers & Operations Research*, *62*, 78-94.

[25] Petit, A., & Ouyang, Y. (2022). Design of heterogeneous flexible-route public transportation networks under low demand. *Transportation Research Part C: Emerging Technologies*, *138*, 103612.

[26] Fayyaz S, S. K., Liu, X. C., & Zhang, G. (2017). An efficient General Transit Feed Specification (GTFS) enabled algorithm for dynamic transit accessibility analysis. *PloS one*, *12*(10), e0185333.

[27] Ma, X., & Wang, Y. (2014). Development of a data-driven platform for transit performance measures using smart card and GPS data. *Journal of Transportation Engineering*, *140*(12), 04014063.

[28] Chaves-Fraga, D., Priyatna, F., Cimmino, A., Toledo, J., Ruckhaus, E., & Corcho, O. (2020). GTFS-Madrid-Bench: A benchmark for virtual knowledge graph access in the transport domain. *Journal of Web Semantics*, *65*, 100596.

[29] Pelletier, M. P., Trépanier, M., & Morency, C. (2011). Smart card data use in public transit: A literature review. *Transportation Research Part C: Emerging Technologies*, *19*(4), 557-568.

[30] Berkow, M., El-Geneidy, A. M., Bertini, R. L., & Crout, D. (2009). Beyond generating transit performance measures: visualizations and statistical analysis with historical data. *Transportation Research Record*, *2111*(1), 158-168.

[31] Perçuku, A., Minkovska, D., & Stoyanova, L. (2017). Modeling and processing big data of power transmission grid substation using neo4j. *Procedia computer science*, *113*, 9-16.

[32] Aggarwal, C. C., & Wang, H. (2010). An Introduction to Graph Data. In *Managing and Mining Graph Data* (pp. 1-11). Springer, Boston, MA.

[33] Agouti, T. (2022). Graph-based modeling using association rule mining to detect influential users in social networks. *Expert Systems with Applications*, 117436.

[34] Webber, J. (2018). Powering real-time recommendations with graph database technology. *Accessed May*, *6*, 2019.

[35] Eckman, B. A., & Brown, P. G. (2006). Graph data management for molecular and cell biology. *IBM journal of research and development*, *50*(6), 545-560.

[36] Transport for London. (2022). Our open data. From Our open data - Transport for London (tfl.gov.uk).

[37] Transport for London. (2019). crowding.data.tfl.gov.uk. From http://crowding.data.tfl.gov.uk/.

[38] Gramsch, B., Guevara, C. A., Munizaga, M., Schwartz, D., & Tirachini, A. (2022). The effect of dynamic lockdowns on public transport demand in times of COVID-19: Evidence from smartcard data. *Transport Policy*, *126*, 136-150.

[39] Tamakloe, R., & Hong, J. (2020). Assessing the efficiency of integrated public transit stations based on the concept of transit-oriented development. *Transportmetrica A Transport Science*, *16*(3), 1459-1489.

[40] Adhvaryu, B., & Kumar, S. (2021). Public transport accessibility mapping and its policy applications: a case study of Lucknow, India. *Case studies on transport policy*, *9*(4), 1503-1517.

[41] Xu, Z., Xu, C., Hu, J., & Meng, Z. (2021). Robust resistance to noise and outliers: Screened Poisson Surface Reconstruction using adaptive kernel density estimation. *Computers & Graphics*, *97*, 19-27.

[42] Shafabakhsh, G. A., Famili, A., & Bahadori, M. S. (2017). GIS-based spatial analysis of urban traffic accidents: Case study in Mashhad, Iran. *Journal of traffic and transportation engineering (English edition)*, *4*(3), 290-299.

[43] Tobler, W. R. (1970). A computer movie simulating urban growth in the Detroit region. *Economic geography*, *46*(sup1), 234-240.

[44] Kumari, M., Sarma, K., & Sharma, R. (2019). Using Moran's I and GIS to study the spatial pattern of land surface temperature in relation to land use/cover around a thermal power plant in Singrauli district, Madhya Pradesh, India. *Remote Sensing Applications: Society and Environment*, *15*, 100239.

[45] Ma, M., & Tang, J. (2022). Interactive coercive relationship and spatio-temporal coupling coordination degree between tourism urbanization and eco-environment: A case study in Western China. *Ecological Indicators*, *142*, 109149.

[46] Xia, D., & Zhang, L. (2022). Coupling coordination degree between coal production reduction and CO2 emission reduction in coal industry. *Energy*, *258*, 124902.

[47] Song, Q., Zhou, N., Liu, T., Siehr, S. A., & Qi, Y. (2018). Investigation of a “coupling model” of coordination between low-carbon development and urbanization in China. *Energy policy*, *121*, 346-354.

[48] Zhang, Y., Marshall, S., & Manley, E. (2019). Network criticality and the node-place-design model: Classifying metro station areas in Greater London. *Journal of transport geography*, *79*, 102485.