Assessment for CASA0002 – Urban Simulation on London's Underground Network

Part 1: London's underground resilience

Ⅰ. Topological network

1.1 Centrality Measures

Three centrality measures selected here are degree centrality, betweenness centrality and closeness centrality.

- Degree centrality

Degree centrality $C(i)$ for a node i is given by Equation 1, where n_i measures the number of nodes directly connected to this node and N is the total number of nodes in the network. It's a crude measure of the centres of the network. In the underground network, it indicates regional underground transit centre.

$$C(i) = n_i/N \quad (1)$$

The map shows that the top ten nodes are spatially dispersed, which is a reasonable distribution.

- Betweenness centrality

Betweenness centrality captures how often the node is on a shortest path between any other two nodes. Betweenness centrality of a node v is defined by the Equation 2, where V is the set of nodes, $\sigma(s,t)$ is the number of shortest paths between s and t, and $\sigma(s,t|v)$ is the number of those paths passing through node v (Brandes, 2008).

$$c_B(v) = \sum_{s,t \in V} \frac{\sigma(s,t|v)}{\sigma(s,t)} \quad (2)$$

In the underground network, the higher the betweenness centrality of a node, the more passenger routes will pass through that station and these stations are likely to be

important bridges between different underground lines. Briefly, it indicates the brokers in the network.

- Closeness centrality

Closeness centrality of node u considers the reciprocal of average shortest path distance to u over its reachable nodes, see Equation 3, where $d(v,u)$is the shortest-path distance between v and u, and $n-1$ is the number of nodes reachable from u. It indicates how close a node is to other nodes.

$$C(u) = \frac{n-1}{\sum_{v=1}^{n-1} d(v,u)} \quad (3)$$

In the underground network, the higher the closeness centrality of a station, the better the accessibility of this station to other stations, and the more efficient its transmission. Passengers are more likely to get to other locations fast through this station. It indicates important transport hubs.

The first 10 ranked nodes for three centrality measures are in Table 1.

Table 1. the first 10 Ranked Nodes for the 3 Measures.

Ranking	Degree Centrality	Betweenness Centrality	Closeness Centrality
1	Stratford	Stratford	Green Park
2	Bank and Monument	Bank and Monument	Bank and Monument
3	Baker Street	Liverpool Street	King's Cross St. Pancras
4	King's Cross St. Pancras	King's Cross St. Pancras	Westminster
5	Earl's Court	Waterloo	Waterloo
6	Green Park	Green Park	Oxford Circus
7	Waterloo	Euston	Bond Street
8	Liverpool Street	Westminster	Farringdon
9	Canning Town	Baker Street	Angel
10	Oxford Circus	Finchley Road	Moorgate

1.2 Impact Measure

Considering the resilience and functioning of underground, this study focuses on 1) the interrelationships between nodes and how tight their connection, 2) the efficiency of transporting passengers. Thus, average clustering coefficient and global efficiency are introduced here as the metrics.

Average clustering coefficient, <C>, captures how clustered the nodes in the graph. The coefficient is the probability that two neighbors of a randomly selected node link to each other. A higher clustering coefficient means that the tube network has more triangle loops connecting different stations, and the transport capacity is more robust and flexible. Conversely, a low value means a poor connectivity between stations, making it more prone to performance loss in the event of an attack (Ponton, Wei and Sun, 2013). For other types of networks like social network, it can evaluate the resilience of information spreading among people. For biological network like neural networks, it indicates the presence of functional modules and ability to process information.

Global efficiency, E, is the average efficiency of all pairs of nodes, which is the multiplicative inverse of the shortest path distance between the nodes. It considers how efficiently the passengers are transported between stations, taking travel steps or capacity into account (Latora and Marchiori, 2002). This measure works for biological and social networks either. For instance, high global efficiency for neural network indicates that it interacts well in wide scope and for social network system, it measures information exchanging efficiency (Latora and Marchiori, 2001). Sum up, the two metrics are not specific to the London's underground.

1.3 Node Removal

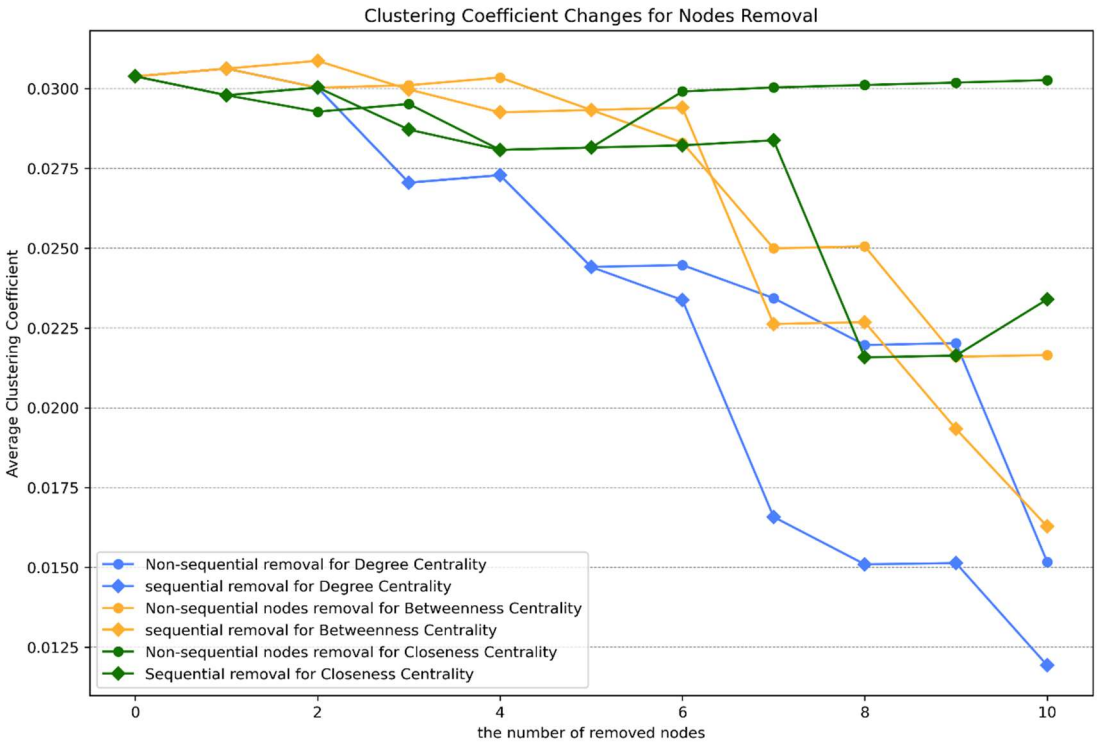

Figure 1. Average Clustering Coefficient Changes

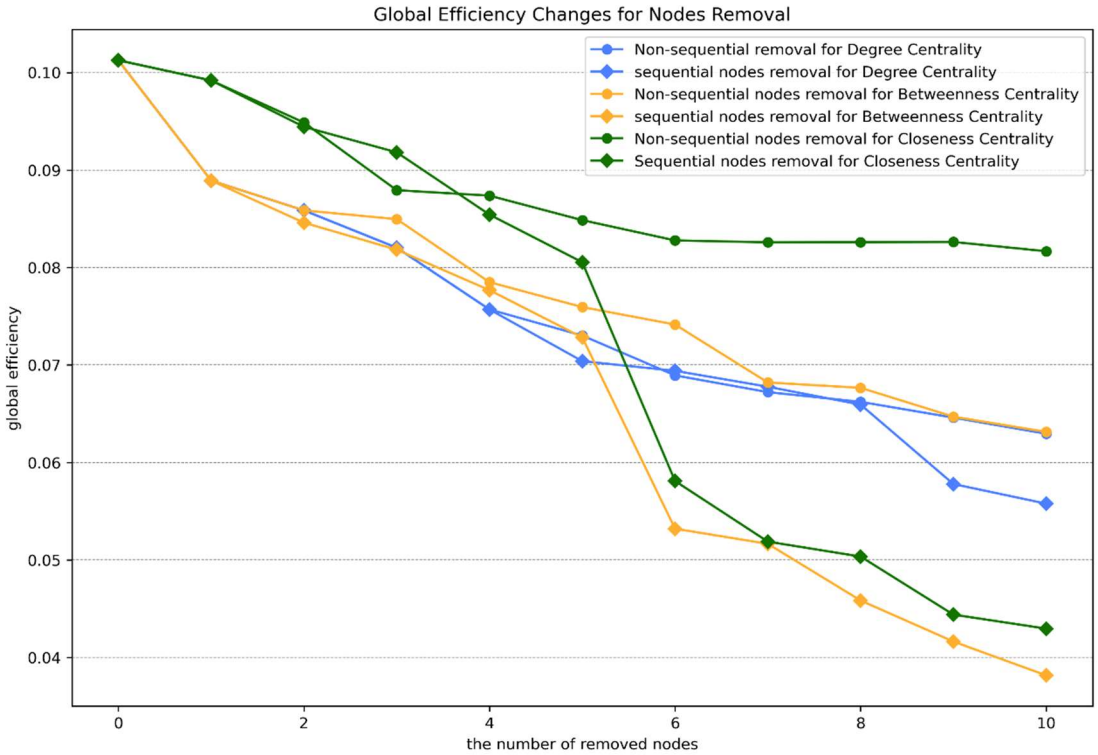

Figure 2. Global Efficiency Changes

Considering the changes of impact measures' metrics (Figure 1-2), degree centrality reflects better the importance of stations for the functioning of underground. When removing nodes, the network performance degrades more and faster for degree centrality, indicating that it captures the impact better. It also indicates the significant impact of some certain sites.

Betweenness centrality also performs well and helps distinguish important stations, especially when measuring by global efficiency. Yet in Figure 1, the removal of the first six sites for betweenness centrality did not cause a significant measure change in the network, while the coefficient dropped very rapidly after the removal of the sequential seventh node, Earl's Court, the ninth node, Euston, and the tenth node, Baker Street. This may suggest that betweenness centrality measure will give us a different strategy for thinking about the importance of sites, with a sense of thick and thin. For closeness centrality in global efficiency, its non-sequential removal captures the stations worst and its sequential removal shows similar trend with betweenness centrality but captures worse than betweenness centrality. The variation in average clustering coefficient is so chaotic that it is difficult to articulate a pattern.

In general, taking into account both sequential and non-sequential removal and two metrics, degree centrality is better and more reliable in London's unweighted tube network. Nevertheless, it has an obvious drawback that the formula is crude and only considers the number of connected nodes, resulting in many nodes with the same degree centrality. This may cause discrepancies when removing nodes iteratively.

As can be seen from Figure 1-2, sequential strategy is more effective. Firstly, sequential nodes removal always gives stronger effects on the network's performance. Secondly, when a station is damaged, there might be substitution effects where other stations can take on some of the functions of the failed station, and the degree distribution of the network will change correspondingly. Sequential strategy is therefore more realistic.

Global efficiency is better at assessing the damage after node removal. Firstly, its

definition is more proper for transport networks which often have tree-like networks than clustering coefficient. Average clustering coefficient measures the tendency of stations to form tight clusters, which should be fine for a graph. However, when the whole graph breaks into subgraphs after removing some hubs, value <C> might be higher than original one if each of subgraphs forms a well-connected cluster, even though the transport function of undergrounds is damaged. As the stations being removed further, value <C> will be lower if subgraphs become poorly connected.

As the Figure 1 showing, average clustering coefficient values go up and down for the above reason. Global efficiency values keep decreasing when removing nodes, because shortest path distance increases as the network becomes disconnected. Imagine that travel distances are increasing, and routes are more tortuous, and the operational efficiency of the underground is decreasing steadily, therefore global efficiency is a better reflection of the underground 's transport function.

Average clustering coefficient might be helpful in recognising nodes interrelationship and nodes clustering in the London’s underground network. For example, some of the obscure nodes that may cause a collapse are indicated in the line graph, such as Earl’s Court station, Euston station, West Hampstead.

In addition, when considering the resilience of London’s underground, it is important not to look at nodes and ignore the network as a whole. Damage to the network is gradual and cumulative, and the impact of each node should be measured on top of the damage from the previous stations. Just as the impact of Earl’s Court station would not have been as great if Stratford station and King's Cross St. Pancras had not been removed.

Table 2. Sequential Removal of Top 10 Nodes for the 3 Measures.

Sequence of removed stations	Degree Centrality	Betweenness Centrality	Closeness Centrality
1st	Stratford	Stratford	Green Park
2nd	Bank and Monument	King's Cross St. Pancras	King's Cross St. Pancras
3rd	Baker Street	Waterloo	Waterloo
4th	King's Cross St. Pancras	Bank and Monument	Bank and Monument
5th	Canning Town	Canada Water	West Hampstead
6th	Green Park	West Hampstead	Canada Water
7th	Earl's Court	Earl's Court	Stratford
8th	Waterloo	Shepherd's Bush	Earl's Court
9th	Willesden Junction	Euston	Shepherd's Bush
10th	Turnham Green	Baker Street	Oxford Circus

Ⅱ. Flows: weighted network

2.1 Old vs new measure

The main adjustment is transforming the flows data and adding them as weights in centrality calculations.

The higher the flow between two stations, the more important they are, so the smaller the cost should be. For betweenness centrality and closeness centrality, weights are interpreted as distances to calculate weighted shortest paths. For weighted networks with flows as the edges' weight, the weight should be inversely related to importance of stations and flows.

Degree centrality considers only the number of nodes, so it does not need to be adjusted. Secondly, the numerical range of the flows is quite large. Therefore, the original flows are scaler-normalized first, then using the value of '1 minus normalized flow' as the weights of edges and recomputing the ranking of nodes.

The ranking has changed considerably, with the weighting of flow greatly influencing the centrality of nodes. Westminster and Waterloo are placed at the front due to their extremely high flows.

Table 3. the New First 10 Ranked Nodes for the 3 Adjusted Measures

New Ranking	Degree Centrality	Betweenness Centrality	Closeness Centrality
1	Stratford	Waterloo	Green Park
2	Bank and Monument	Westminster	Westminster
3	Baker Street	Green Park	Waterloo
4	King's Cross St. Pancras	Bank and Monument	Bank and Monument
5	Earl's Court	Stratford	Victoria
6	Green Park	Liverpool Street	Liverpool Street
7	Waterloo	Victoria	Oxford Circus
8	Liverpool Street	Euston	Bond Street
9	Canning Town	Sloane Square	Stratford
10	Oxford Circus	South Kensington	Warren Street

2.2 Impact measure with flows:

The same is done by adding weights to the two impact measures to adjust. The weight is the same as in 2.1. Global efficiency algorithm in NetworkX ignores edge weight, therefore weighted shortest path distance and its multiplicative inverse are calculated manually, and weighted global efficiency is generated, based on equation in Latora's and Marchiori's research (Latora and Marchiori, 2001).

2.3 Re-experiment with flows

Remove the 10 highest ranked nodes sequentially for degree centrality, and the results of impact measures (Figure 3-4) indicate that Stratford station has the greatest impact on global efficiency, while Earl's Court station has the most significant impact on clustering coefficient in this network. The trends in the two line graphs in Figure 3-4

are generally consistent with that before adjusting in Figure 1-2, though the values differ slightly. This may be due to the data transformation of ‘flows’, not distinguishing well different edges’ weights.

One unreasonable point is that <C> increases after removing Stratford and King’s Cross St. Pancras, meaning that the network is more clustered without them. My interpretation is that there are several high-degree hubs connecting the main underground. Even though the network loses several high-flow connections, the other stations/subgraphs are able to form clusters and operate well, and the underground network becomes instead more assortative.

Another point is that Earl's Court affects the clustering coefficient most among the top ten stations and the same status appears in the betweenness centrality sequential measure in the topological network analysis. I have not been able to find an explanation for it, in terms of either the particularity of the station or methodology.

In a study using Boston’s subway as an example, it was argued that, the subway is not a typical small-world network and has low fault-tolerance, and clustering coefficient C is ill-defined in this case, making it hard to draw conclusion by average C (Latora and Marchiori, 2002). Their alternative method is the quantity, efficiency E.

Our adjusted efficiency measure weighting flows should highlight the stations that have higher capacity from the perspective of passengers. Therefore, Stratford Station’s closure is expected to have the greatest impact on passengers’ mobility, given that global efficiency can measure the damage better.

Additionally, it is just a partial view of real-world transport systems to analyse such a closed underground network in this study, which should be extended by bus, railway system and so on, to complement people’s daily trip. For example, the problem of broken underground station could be solved temporarily by taking bus.

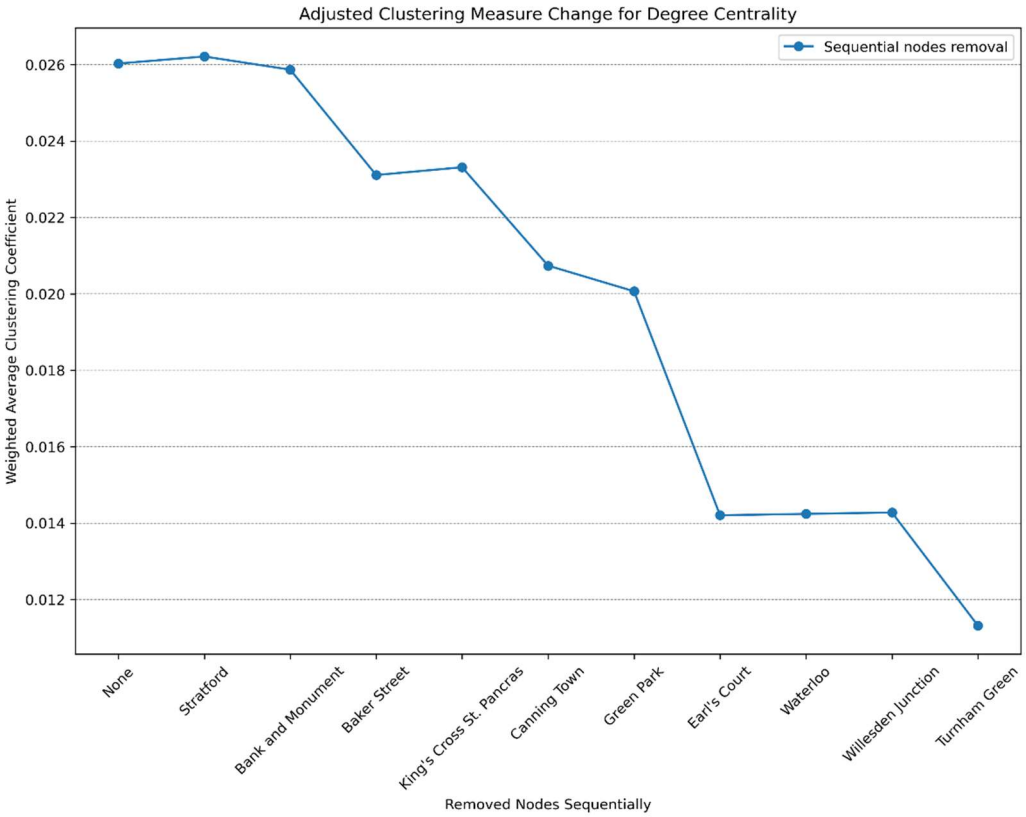

Figure 3. Adjusted Average Clustering Coefficient Changes

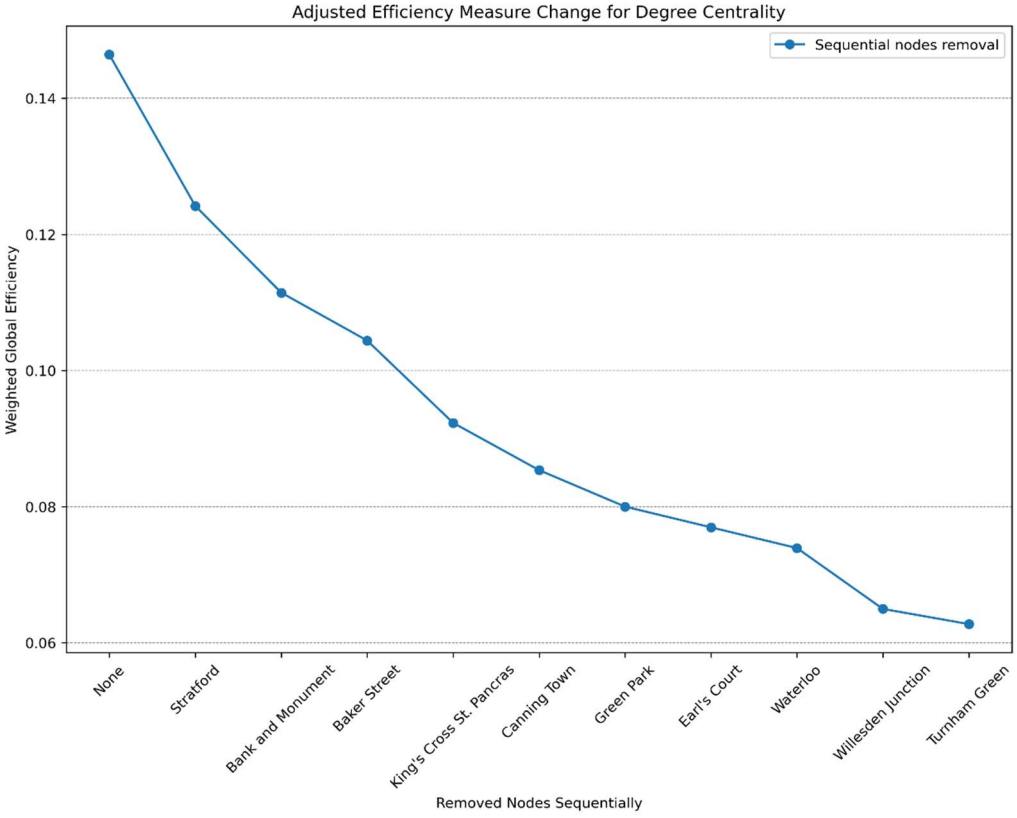

Figure 4. Adjusted Impact Measures Changes for Degree Centrality

Part 2 Spatial interaction models

Ⅲ. Models and calibration

3.1 Spatial Interaction models

The spatial interaction model covered in the lecture is the gravity model, based on Newton's law of gravity. It considers the attraction between two locations is proportional to their masses and the deterrence measure between, which could simulate activity intensity, such as transport trips, flow of money. Thus, the model can be written as Equation 4 and this is extended by several parameters to calibrate the spatial interaction model better.

$$T_{ij} = k\frac{O_i^{\alpha} D_j^{\gamma}}{d_{ij}^{\beta}} \quad (4)$$

T_{ij} is the activity flow between origin i and destination j. O_i is the variable relating to the mass or production measure of origin i, such as the population size. D_j is the variable relating to the mass or attraction of destination j, such as the job amount or salary level. d_{ij}^{β} is a measure of cost between i and j.

Where k, α, γ and β are the calibrated parameters making the model more robust and adaptable to the research context (Wilson, 1971).

- k: a scaling constant to adjust the scale of the model to different context scenarios.
- β: a parameter relating to the deterrence in the cost. It adjusts the fiction of distance, such as travel cost, transport efficiency between locations, to simulate the real situation best.
- α: it measures the strength of production movements in origins, adjusting the original mass O_i.
- γ: it measures the strength of attractiveness in destinations, adjusting the original mass D_j.

3.2 Calibration of model

Production constrained model is selected to estimate the passenger flows in the London's underground network. This study assumes a situation where people living in origins are commuting to work in destination, with population measuring the emission of origin and the number of jobs measuring the attractiveness of destination to people. Where people's place of residence tends to remain stable, while policy changes or industrial changes, such as variations in job availability and transport cost being changed, may lead to changes in commuting. In such scenarios, production constrain model could be applied to explore and predict how underground passenger flow will change under the influence.

In the cost function, negative exponential law for distance is selected. The distance between stations do not affect the interaction seriously in the city, so the distance decay effect should be not as fast in the underground transport situation, compared to inverse power law.

We use Poisson log-linear regression and exponentiate both sides of the log-variable of Equation 4, thus we can fit a model estimate better using a straight line, deal with the non-negative integers data and calibrate the parameters. The production constrained model is shown in Equation 5. The regression equation for calibrating and estimating is Equation 6.

As it is origin constrained, O_i is known. A_i is a balancing factor relating to each origin. The variables are the origin station α_i, the number of jobs in destination station D_j and the distance between origin station and destination station d_{ij}.

$$T_{ij} = A_i O_i D_j^{\gamma} d_{ij}^{-\beta} \quad (5)$$

$$\lambda_{ij} = \exp\left(\alpha_i + \gamma \ln D_j - \beta d_{ij}\right) \quad (6)$$

The results of regression model are shown in Figure 5. The calibrated β parameter is 0.000153 after rounding off. The R^2 0.468 and RMSE 96.263 show the goodness of fit of the model.

Generalized Linear Model Regression Results

Dep. Variable:	flows	No. Observations:	61413
Model:	GLM	Df Residuals:	61013
Model Family:	Poisson	Df Model:	399
Link Function:	Log	Scale:	1.0000
Method:	IRLS	Log-Likelihood:	-9.0994e+05
Date:	Mon, 01 May 2023	Deviance:	1.6477e+06
Time:	20:29:13	Pearson chi2:	2.40e+06
No. Iterations:	8	Pseudo R-squ. (CS):	1.000
Covariance Type:	nonrobust		

	coef	std err	z	P>|z|
origin[Abbey Road]	-2.9143	0.041	-70.509	0.000
origin[Acton Central]	-1.1621	0.029	-39.960	0.000
origin[Acton Town]	-1.6131	0.017	-92.801	0.000
origin[Aldgate]	-2.9430	0.020	-150.138	0.000
origin[Aldgate East]	-2.8548	0.019	-151.960	0.000
origin[All Saints]	-2.8783	0.037	-77.219	0.000
origin[Alperton]	-1.6542	0.026	-64.731	0.000
origin[Amersham]	1.0008	0.030	33.747	0.000
origin[Anerley]	-1.0369	0.040	-26.044	0.000
origin[Angel]	-2.5875	0.017	-156.011	0.000
origin[Archway]	-1.7164	0.015	-117.258	0.000
...				
origin[Woolwich Arsenal]	0.5180	0.013	40.453	0.000
log_jobs	0.7552	0.001	1185.004	0.000
distance	-0.0002	1.88e-07	-814.175	0.000

Figure 5. Regression Results

Ⅳ. Scenarios

4.1 Scenario A

After decreasing Canray Wharf's jobs value to 50% of original, we need to reconstruct the flow distribution while conserving the flows amounts of trips in origins the same. A_i is a balancing factor relating to origin. By adjusting a new A_i for each origin, we ensure that the flows are redistributed globally. New D_j^γ is calculated using new 'jobs' data. New A_i is calculated further based on Equation 7 which is from Equation 5, so it reflects the change of numbers of jobs in destinations. Putting this back to Equation 6, we calculate new flows estimates T_{ij}.

$$A_i = \frac{1}{\sum_j D_j^\gamma \, d_{ij}^{-\beta}} \quad (7)$$

Figure 6A is the estimated result in Scenario A. Checking the right column, it is ensured that there are almost the same number of commuters, around 154k totally, in the system, thus our origin constraints are holding.

odford	Woodgrange Park	Woodside Park	Woolwich Arsenal	All
NaN	NaN	NaN	8.0	597.0
NaN	0.0	NaN	NaN	1226.0
0.0	NaN	1.0	NaN	3750.0
1.0	NaN	1.0	NaN	2886.0
1.0	NaN	1.0	NaN	3167.0
...	...	...	...	...
NaN	NaN	NaN	NaN	4860.0
NaN	NaN	NaN	NaN	532.0
NaN	NaN	NaN	NaN	3102.0
NaN	NaN	NaN	NaN	7889.0
673.0	160.0	427.0	1167.0	1541397.0

Figure 6A. estimates OD flows in Scenario A

odford	Woodgrange Park	Woodside Park	Woolwich Arsenal	All
NaN	NaN	NaN	7.0	595.0
NaN	0.0	NaN	NaN	1226.0
0.0	NaN	1.0	NaN	3744.0
1.0	NaN	1.0	NaN	2885.0
1.0	NaN	1.0	NaN	3160.0
...	...	...	...	...
NaN	NaN	NaN	NaN	4867.0
NaN	NaN	NaN	NaN	532.0
NaN	NaN	NaN	NaN	3100.0
NaN	NaN	NaN	NaN	7893.0
660.0	160.0	422.0	1089.0	1541347.0

Figure 6B. original estimates OD flows matrix

Figure 6. Results of Flow Estimates

4.2 Scenario B

The assumption is that a significant increase in transport costs will affect people's choice of jobs and that the number of jobs at destinations may change, so origin constrained model is still used.

a) Selecting new values for β

The distance decay function considered in this study is Equation 8 based on the negative exponential distance transformation (Rodrigue, 2020). The reason for choosing this has been stated in 3.2.

$$I_{ij} = f(d_{ij}) = \exp(-\beta d_{ij}) \quad (8)$$

Where I_{ij} is some measure of interaction intensity over a distance d_{ij} and $f(d_{ij})$ is a decreasing function of distance (Taylor, 1971). When β gets larger, the decay effect will be more rapid, and the interaction intensity (flows) will decrease more over a distance. We assume that people tend to travel closer in the scenario of significantly

higher cost of transport, and the limitation of distance to the interaction tensity will be stronger. Therefore, values for the parameter β are selected larger than the original calibrated beta, 0.000153.

The two values for β in this study are 0.0002 and 0.0004. The selection is based on Equation 9, where $\Delta\%$ is the percent change expected from one unit (meter) increase in distance (Oshan, 2016). The change percent for different β is in Table 4.

$$\Delta\% = \left(1 - exp(-\beta)\right) * 100 \quad (9)$$

Table 4. The percent change of flows prediction for different β

β (beta)	Δ%
0.000153	0.0152988
0.0002	0.0199980
0.0004	0.0399920

b) Reconstructing the distribution of flows

Setting the new β value, next step is to calculate new d_{ij}^{β} values and A_i balancing vectors just as the calculations in Scenario A. D_j^{γ} values are the same because the jobs of destination stations do not change. Finally, we plug these values back in to Equation 5, and generate new scenario flow estimates for our new β.

4.3 Analysis

a) Flow into Canary Wharf

In Scenario A, the estimated flow into Canary Wharf stations is 29489, where the original estimated flow is 47681. It means that, in the origin constrained model, the flow destined for Canary Wharf does not drop to as much as 50% of the original estimate. Does this indicate the higher attractiveness of Canary wharf?

b) Global measure of the change in OD matrix

To show to what extent have OD flow matrix changed, this study calculates two values as measures, the overall absolute volume of flow change and the average absolute volume of flow change. The two values are calculating based on the absolute change of flows for each pair of nodes in the matrix. The results are in Table 5.

Table 5. Global Measure of Flows Changes in Different Scenarios

		Average flow change	Total flow change	Change rate to total flows
Scenario A	Scenario A	0.46	73712	0.05
Scenario B1	$\beta = 0.0002$	1.70	271164	0.18
Scenario B2	$\beta = 0.0004$	12.12	1929652	1.25

From the perspective of two metrics in Table 5, Scenario B2 has the greatest impact, followed by Scenario B1, with the least impact being Scenario A. Can I speculate that, conditional on larger than calibrated β, the effect on the distribution of flows will be greater as β gets larger? Although the total flow changes in Scenario B2 are too large and even exceed total commuting flows, they do not correspond to reality.

Looking at the results for Scenario B1, when β equals to 0.0002, slightly greater than the calibrated β 0.000153, the absolute change in flow distribution reaches more than 27k, representing 18% of the total flow, which is a considerable change.

Furthermore, the grid maps of flow changes (Figure 8, Figure 9) indicated that Scenario B has a much more global impact than Scenario A, while the impact of Scenario A is concentrated mainly on several pairs of stations.

In summary, I would argue that Scenario B has more impact in the redistribution of flows. And the impact would be reflected more globally, compared to the scenario of one location's attraction change. The real-world context of scenario B might be the impact on public transportation travel caused by changes in transport charging policy, ticket prices and other policy factors.

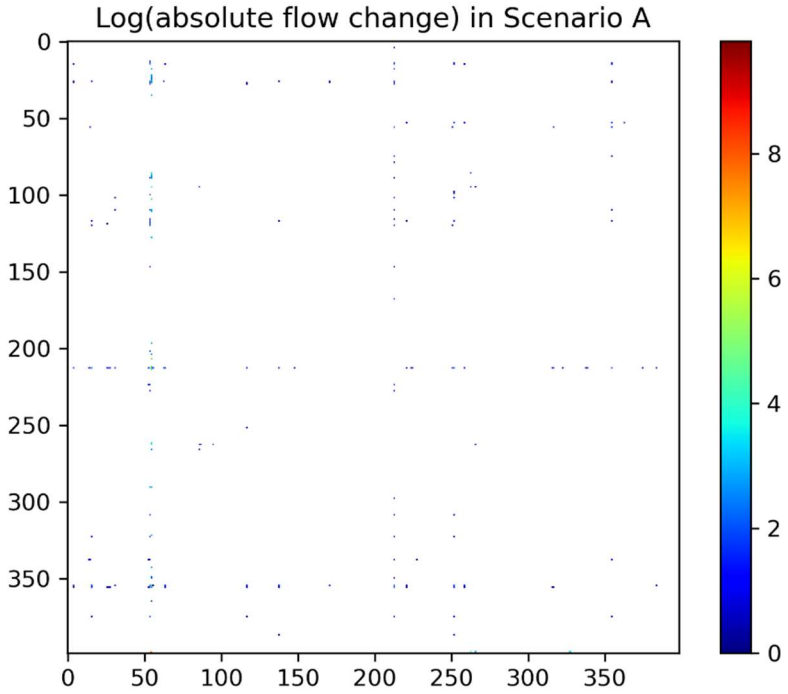

Figure 8. Grid Map of flow change in Scenario A

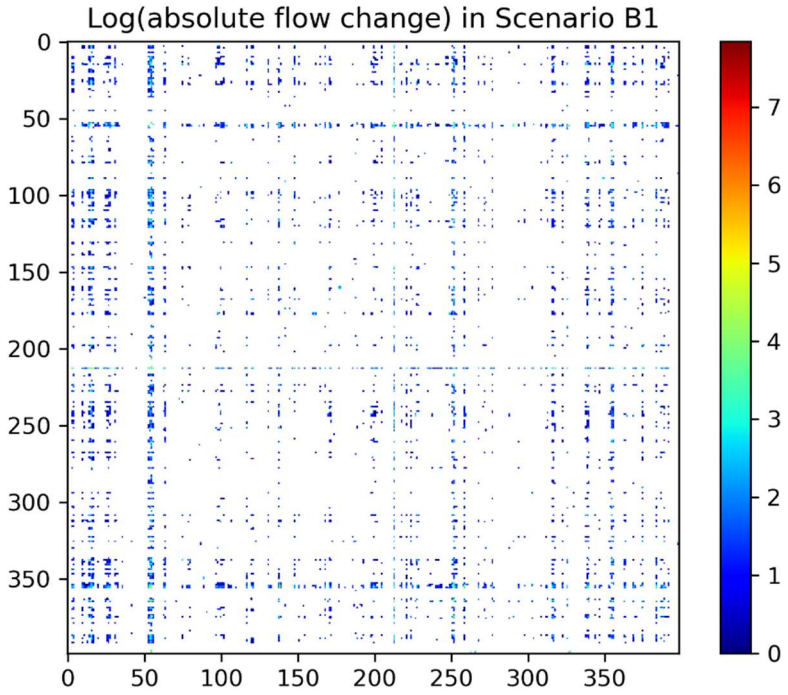

Figure 9. Grid Map of flow change in Scenario B1

References

Brandes, U. (2008) 'On variants of shortest-path betweenness centrality and their generic computation', *Social Networks*, 30(2), pp. 136–145. doi:10.1016/j.socnet.2007.11.001.

Latora, V. and Marchiori, M. (2001) 'Efficient Behavior of Small-World Networks', *Physical Review Letters*, 87(19), p. 198701. doi:10.1103/PhysRevLett.87.198701.

Latora, V. and Marchiori, M. (2002) 'Is the Boston subway a small-world network?', *Physica A: Statistical Mechanics and its Applications*, 314(1–4), pp. 109–113. doi:10.1016/S0378-4371(02)01089-0.

Oshan, T.M. (2016) 'A primer for working with the Spatial Interaction modeling (SpInt) module in the python spatial analysis library (PySAL)', *REGION*, 3(2), p. 11. doi:10.18335/region.v3i2.175.

Ponton, J., Wei, P. and Sun, D. (2013) 'Weighted clustering coefficient maximization for air transportation networks', in *2013 European Control Conference (ECC)*. *2013 European Control Conference (ECC)*, Zurich: IEEE, pp. 866–871. doi:10.23919/ECC.2013.6669250.

Taylor, P.J. (1971) 'Distance Transformation and Distance Decay Functions', *Geographical Analysis*, 3(3), pp. 221–238. doi:10.1111/j.1538-4632.1971.tb00364.x.

Wilson, A.G. (1971) 'A Family of Spatial Interaction Models, and Associated Developments', *Environment and Planning A: Economy and Space*, 3(1), pp. 1–32. doi:10.1068/a030001.

Appendix

Code: https://github.com/JinJiang22/LondonTubeNetwork_simulation.git