Urban Simulation Assessment: investigation of London Underground

April 2023

Contents

1 London's underground resilience

1.1 Topological network

1.1.1 Centrality measures

Degree centrality
Degree centrality is a metric that measures the number of edges attached to a node, and it reflects the level of connectivity of that node. In an adjacent matrix A_{ij}, the degree of a node i is denoted as

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

where j represents the neighbor set of node i, and A_{ij} is the entry value in the adjacent matrix. In the context of London underground, degree centrality is calculated as the number of tube routes that intersect at a given station. A station with a high degree centrality indicates that it is connected to many other tube stations and can serve as a transportation hub for passengers.

Table 1: Top 10 Ranked Nodes by Degree Centrality

Station	**Degree Centrality (Normalised)**
Stratford	0.0225
Bank and Monument	0.0200
Baker Street	0.0175
King's Cross St. Pancras	0.0175
Liverpool Street	0.0150
Earl's Court	0.0150
West Ham	0.0150
Oxford Circus	0.0150
Green Park	0.0150
Canning Town	0.0150

Betweenness centrality
Betweenness centrality is a metric that quantifies the level of intermediate importance of a node in a network, measuring the number of shortest paths that pass through a certain node. It is denoted as

$$x_i = \sum_{st} \frac{n_{st}^i}{g_{st}} \tag{2}$$

where g_{st} is the total number of shortest path between s and t and n_{st}^i=1 if vertex i lies on geodesic path from s to t, 0 otherwise. The normalised betweenness centrality is denoted as

$$x_i = \frac{1}{n^2} \sum_{st} \frac{n_{st}^i}{g_{st}} \tag{3}$$

In essence, nodes with high betweenness centrality lie on many of the shortest paths connecting pairs of other nodes in the network. In the context of the London Underground, a station with high betweenness centrality would act as a an important interchange point between different parts of the tube network. Such a station would be an key node for connecting passengers traveling between different tube lines or transferring between different modes of transportation, such as connections between underground and railway systems. Therefore, stations with high betweenness centrality play a critical role in facilitating the overall efficiency and accessibility of the transportation system.

Table 2: Top 10 Ranked Nodes by Topological Betweenness Centrality

Station	**Topological Betweenness Centrality (Normalised)**
Stratford	0.297846
Bank and Monument	0.290489
Liverpool Street	0.270807
King's Cross St. Pancras	0.255307
Waterloo	0.243921
Green Park	0.215835
Euston	0.208324
Westminster	0.203335
Baker Street	0.191568
Finchley Road	0.165085

Closeness centrality

Closeness centrality is a metric that quantifies the proximity of a node to all other nodes in a network taking into account the geodesic or distance between them. It reflects how quickly a node can access all other nodes in the network, making it a key indicator of the node's accessibility. However, the metric does not work in an unconnected graph, where the shortest path is infinite.

It is denoted as

$$C_i = \frac{1}{l_i} = \frac{n}{\sum_j d_{ij}} \tag{4}$$

where d_{ij} is the shortest path length between i and j and the mean shortest path is $l_i = \frac{1}{n}\sum_j d_{ij}$.

In the context of London Underground, a station with high closeness centrality indicates that it can be quickly accessed from other stations and that it provides passengers a convenient transfer point to travel through.

Table 3: Top 10 Ranked Nodes by Topological Closeness Centrality

Station	**Topological Closeness centrality(normalised)**
Green Park	0.114778
Bank and Monument	0.113572
King's Cross St. Pancras	0.113443
Westminster	0.112549
Waterloo	0.112265
Oxford Circus	0.111204
Bond Street	0.110988
Farringdon	0.110742
Angel	0.110742
Moorgate	0.110314

1.1.2 Impact measures

The global efficiency and clustering coefficient are used to measure the impact of node removal.

Efficiency
Latora and Marchiori introduced an index to quantify the efficiency of the information exchange within networks, assuming that the communication occurs simultaneously across edges and is most efficient along geodesic paths[3]. The approach estimates communication efficiency to be inversely proportional to the shortest distance between two nodes, using the equation

$$e_{ij} = 1/d_{ij} \tag{5}$$

When there is no path between vertices i and j, resulting in shortest path $d_{ij} = +\infty$, efficiency will be defined as 0 so to avoid non-functional measurements in unconnected graphs. They also defined the average efficiency as

$$E(\mathbf{G}) = \frac{\sum_{i \neq j \in \mathbf{G}} \epsilon_{ij}}{N(N-1)} = \frac{1}{N(N-1)} \sum_{i \neq j \in \mathbf{G}} \frac{1}{d_{ij}} \tag{6}$$

allowing for global analysis of network behaviour even when the network is not fully connected.

Clustering coefficient
The clustering coefficient measures the level of clustering in a graph. It quantifies the probability that two neighbors of a node are connected to each other and is calculated as the ratio between the number of triangles in the graph with node i as one vertex and the number of all possible triangles that i could have formed [1] [6]. The local clustering coefficient is denoted as Ci and is calculated using the formula:

$$C_i = \frac{2L_i}{k_i\,(k_i - 1)} \tag{7}$$

where L_i is the number of links between the k_i neighbors of node i.
The average clustering coefficient represents the degree of clustering of the entire network and is obtained by averaging the clustering coefficient over all nodes.

$$\langle C \rangle = \frac{1}{N} \sum_{i=1}^{N} C_i \tag{8}$$

Analysing the change in the average clustering coefficient can evaluate the impact of node removal on the level of clustering and the overall resilience of the network. A high clustering coefficient suggests a more densely connected graph, where intra-cluster movement is easier and inter-cluster spread is harder due to fewer connections between clusters. In the context of transportation networks, the clustering coefficient can be used to quantify the level of inter-connectivity between stations and their neighbor stations. A transportation with a high clustering coefficient may indicate a more cohesive sub-network, within which passengers can travel with a more efficient option.
It is worth noting that both metric is not exclusive to the London Underground and can be used to examine the spread of disease or information in other networks.

1.1.3 Node removal

According to Table 4, the mean decrease and the standard deviation of decrease in clustering coefficient and global efficiency after node removal resulting from Strategy B were more pronounced than that resulting from Strategy A. Given this, it appears that the impact of node removal using strategy B was more variable across different nodes and therefore strategy B may be more effective in terms of removing nodes that has greater impact on the resilience of the London underground network.

Table 4: Change in Impact Measures after node removal (post-removal-pre-removal)/pre-removal

Centrality Measure	Degree Centrality	Betweenness Centrality	Closeness Centrality
A - Mean Global Efficiency Change	-0.046174	-0.045498	-0.021015
A - Std Global Efficiency Change	0.032148	0.036263	0.023213
B - Mean Global Efficiency Change	-0.057441	-0.090260	-0.078868
B - Std Global Efficiency Change	0.037362	0.070518	0.077744
A - Mean Clustering Coefficient Change	-0.061792	-0.031928	-0.000040
A - Std Clustering Coefficient Change	0.097889	0.053149	0.028015
B - Mean Clustering Coefficient Change	-0.084421	-0.056538	-0.022239
B - Std Clustering Coefficient Change	0.095963	0.087838	0.082979

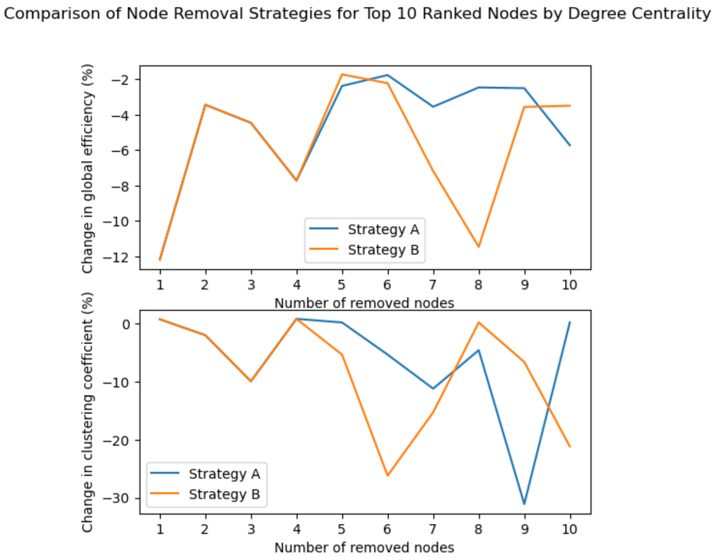

Figure 1: Percentage of change in impact measures after removal of the top 10 ranked nodes by degree centrality

For degree centrality, both strategies A and B resulted in a significant decrease in clustering coefficient and global efficiency. However, the decrease was more pronounced for strategy B, with a 60.71% decrease in clustering coefficient and a 45.05% decrease in global efficiency.

The results illustrated in Figure 1 demonstrate that the removal of the initial four nodes following both strategies was identical, resulting in an equivalent reduction in impact measures. However, the ranking of nodes changed beginning from the fifth node. This change may be attributed to the recalculation of degree centrality in the fragmented network as a consequence of node removal, which caused certain nodes to no longer be connected to the same number of tube lines. Furthermore, the analysis revealed that specific nodes with lower degree centrality had a more significant impact on the network. For instance, the removal of the node with the eighth-ranked degree centrality (Willesden Junction) following strategy B led to a substantial decrease in global efficiency, suggesting a less convenient travel experience for passengers.

For betweenness centrality, both strategies also resulted in a decrease in clustering coefficient and global efficiency, but again, the decrease was more pronounced for strategy B. The clustering coefficient decreased by 46.41% and global efficiency decreased by 62.31% with strategy B, compared to 28.73% and 37.64% with strategy A, respectively. In addition, removing an intermediate node from a network can disconnect the network and cause the ranking of the top 10 nodes by betweenness centrality to change immediately due to altered network structure.

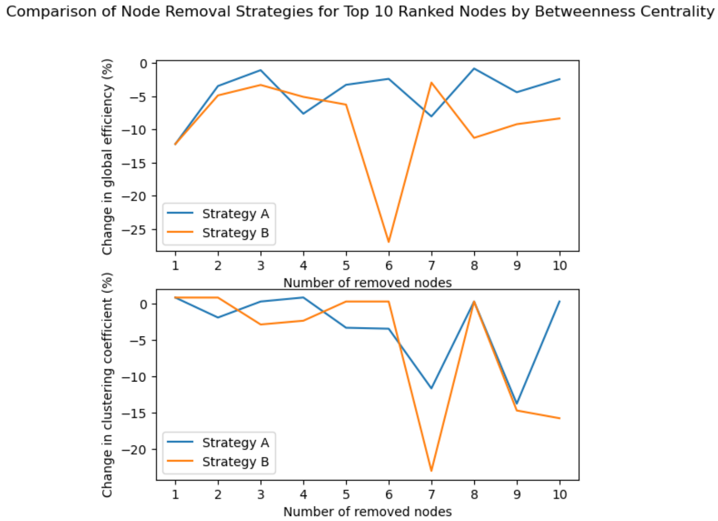

Figure 2: Percentage of change in impact measures after removal of the top 10 ranked nodes by betweeness centrality

On the other hand, for closeness centrality, both strategies resulted in less impact on the clustering coefficient and global efficiency. The change in clustering coefficient was relatively small, with a decrease of 0.39% for strategy A and 22.99% for strategy B. The change in global efficiency was more substantial for strategy B, with a decrease of 57.58% compared to a decrease of 19.34% with strategy A. The results shows that the removal of stations with role as a transfer point will impact the efficiency of traveling within the whole network more than the efficiency of traveling within stations and their neighboring stations.

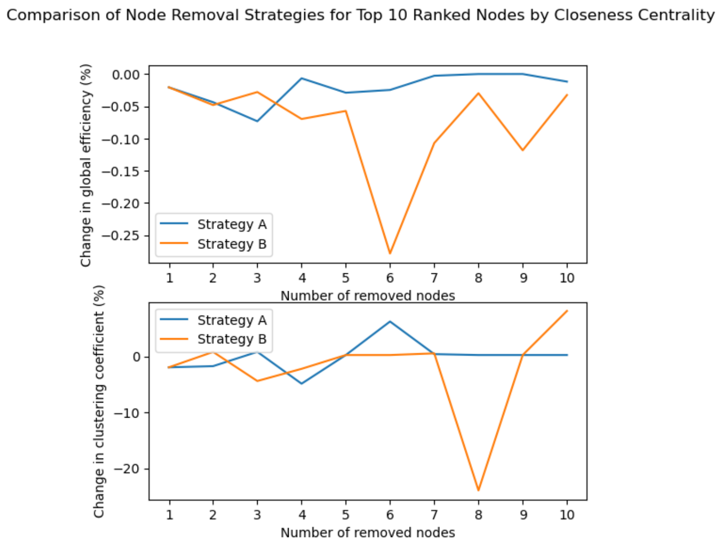

Figure 3: Percentage of change in impact measures after removal of the top 10 ranked nodes by closeness centrality

Overall, the betweenness centrality measure has the highest average change in both impact metrics after node removal,indicating that it better reflects the importance of a station for the functioning of the transportation network. However, among all three centrality measures, nodes with lower rank also had a substantial impact on the network, as evidenced by the significant decrease in impact measures upon their removal. This suggests that centrality measures alone may not be sufficient to accurately predict the importance of a node in a network, and other factors such as network structure and connectivity should also be taken into consideration.

1.2 Weighted Network

1.2.1 Centrality measure

In context of transportation network, the more passenger flows between two stations, the higher level of interaction between two nodes. Therefore, to calculate the betweenness centrality of weighted London Underground network, the weight assigned to each edge is the inverse of the flow between two node. It is important to note that the station with zero flow (Battersea Park) was excluded when exploring the weighted network.

Table 5: Top 10 Ranked Nodes by Weighted Betweenness Centrality (Normalized)

Station	Weighted Betweenness centrality(normalised)
Green Park	0.560110
Bank and Monument	0.527899
Waterloo	0.395173
Westminster	0.347823
Liverpool Street	0.347404
Stratford	0.331705
Euston	0.272078
Victoria	0.265463
Oxford Circus	0.260260
Bond Street	0.254217

Table 5 indicates a change in the ranking of the top 10 important nodes between the weighted and topological networks. This is attributed to the inclusion of weight, as it alters the calculation of the shortest path between nodes and consequently, betweenness centrality.

1.2.2 Impact Measure

The definition of global efficiency in a weighted network involves measuring the shortest path length d_{ij} as the smallest sum of the "distance" through all possible paths in the graph from node i to node j. This definition is valid when the weight is included and will be applied for impact measurement in this section. A Python function has been developed to calculate the global efficiency of a weighted network, which is shown below:

```python
def weighted_global_efficiency(G):
    """
    Calculate the global efficiency of a weighted network.

    Parameters:
    ----------
    G : NetworkX graph
        Weighted network

    Returns:
    -------
    E_glob : float
        Global efficiency of the network
    """
    N = len(G)
    E_glob = 0.0
    for u in G.nodes():
        for v in G.nodes():
            if u != v:
                try:
                    d = 1.0 / nx.shortest_path_length(G, u, v, weight='inv_flows')
                except nx.NetworkXNoPath:
                    d = 0.0
                E_glob += d
    E_glob /= (N * (N - 1))
    return E_glob
```

The clustering coefficient in a weighted network is defined by Saramaki [4] as the geometric average of the subgraph edge weights, which generalizes the clustering coefficient to weighted networks. The formula for clustering coefficient c_u is given by:

$$c_u = \frac{1}{deg(u)(deg(u)-1))} \sum_{vw} (\hat{w}_{uv}\hat{w}_{uw}\hat{w}_{vw})^{1/3} \tag{9}$$

where $deg(u)$ represents the degree of node u, and edge weights $\hat{w}_{uv}$ are normalized by the maximum weight in the network, i.e., $\hat{w}_{uv} = w_{uv}/max(w)$.
In this experiment, the clustering coefficient is calculated using

```python
nx.average_clustering(G, weight='inv_flows')
```

.

1.2.3 Node removal result

After removing the 3 highest ranked nodes for betweenness centrality, the change of global efficiency and clustering coefficient resulted form their removal is listed in the Table 6

Table 6: Change in Impact measures after node removal

Node	Global Efficiency before	Global Efficiency after	Global Efficiency change	Clustering coefficient before	Clustering coefficient after	Clustering coefficient change	No. Connected Components	Diameter of Largest Component
Green Park	0.101248	0.099189	-0.020340	0.00006	0.000060	0.005227	1	36
Bank and Monument	0.099189	0.094963	-0.042606	0.00006	0.000060	0.001689	1	38
King's Cross St. Pancras	0.094963	0.087945	-0.073904	0.00006	0.000061	0.006352	1	42

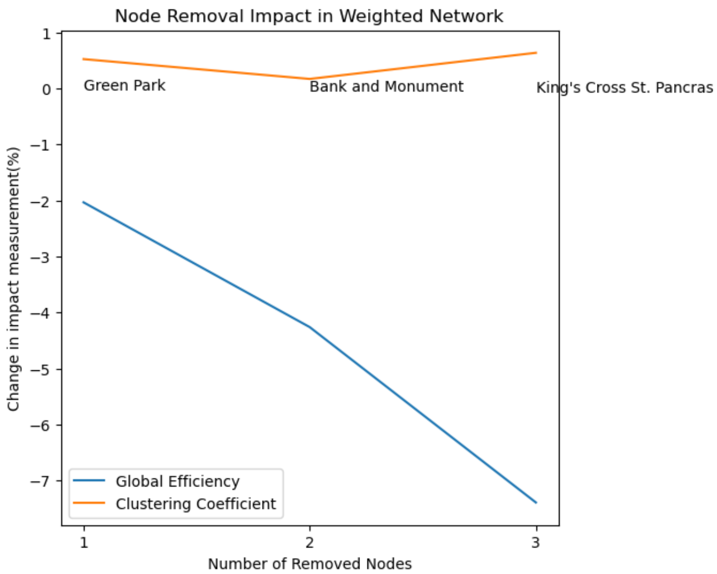

Figure 4: Percentage of change in impact measures after removal of the top 3 ranked nodes by betweenness centrality

Based on the analysis, the closure of three stations is unlikely to impact local travel but would inconvenience travel within the entire network. The results indicate that the closure of King's Cross and St. Pancras would have the greatest impact on the underground network.

2 Spatial interaction models

2.1 Models and Calibration

2.1.1 Gravity model

In 1971, Wilson proposed that gravity models comprise a collection of spatial interaction models instead of a single model[8]. Oshan subsequently developed a Python primer for simulating inter-regional flows based on this concept in 2016[5]. The basic gravity model is used to simulate spatial interactions between an origin and destination, which are proportional to the product of the origin and destination's masses and inversely proportional to the cost of travel between them. This model can be expressed mathematically as follows:

$$T_{ij} = k\frac{O_i^{\alpha} D_j^{\gamma}}{d_{ij}^{\beta}} \tag{10}$$

where

- T_{ij} represents the flow between origin i and destination j
- O_i is a vector of origin attributes related to the emissiveness of origins
- D_j is a vector of destination attributes related to the attractivenss of destinations
- d_{ij} is a matrix of costs associated with flows between i and j (usually distance)
- k is a scaling factor representing the proportionality
- α is a vector of parameters representing the effect of i origin attributes on flows
- γ is a vector of parameters representing the effect of j destination attributes on flows
- β is a parameter representing the effect of cost of travel on flows, where the distance-decay function aligns with the negative power law.

The parameters k, α, γ, and β are to be estimated.

Unconstrained model
In an unconstrained interaction model, the estimated flows are assumed to be equal to the total observed flows used to calibrate the parameters. Thus, the scaling factor k can be computed as the ratio of the sum of observed flows to the sum of all other model elements.

$$k = \frac{T}{\sum_i \sum_j O_i^{\alpha} D_j^{\gamma} d_{ij}^{-\beta}} \tag{11}$$

where:

$$T = \sum_i \sum_j T_{ij} \tag{12}$$

The unconstrained interaction model provides a simple yet powerful way of modeling the flows between origins and destinations.

Production constrained model
When the total of flows originating from all origins is known, the model can be formed as a production constrained model, which can be expressed as:

$$T_{ij} = A_i O_i D_j^{\gamma} d_{ij}^{-\beta} \tag{13}$$

where

$$O_i = \sum_j T_{ij} \tag{14}$$

and

$$A_i = \frac{1}{\sum_j D_j^{\gamma} d_{ij}^{-\beta}} \tag{15}$$

In this case,the balancing factor A_i, which relates to each origin i, will replace k and act as a proportionality factor to ensure that flow estimates from each origin sum to the known total O_i.

Attraction constrained model
The attraction is another singly constrained model with the total of flows terminating at destination is known. The model can be mathematically expressed as:

$$T_{ij} = D_j B_j O_i^{\alpha} d_{ij}^{-\beta} \tag{16}$$

where flows are constrained as

$$D_j = \sum_i T_{ij} \tag{17}$$

and the balancing factor is

$$B_j = \frac{1}{\sum_i O_i^{\alpha} d_{ij}^{-\beta}} \tag{18}$$

to ensure the convergence of flows.

Doubly Constrained Model If the total of flows is constrained to both origins and destinations, the spatial interaction model can be expressed as

$$T_{ij} = A_i B_j O_i D_j d_{ij}^{-\beta} \tag{19}$$

where the constrained flows are expressed as

$$O_i = \sum_j T_{ij} \tag{20}$$

$$D_j = \sum_i T_{ij} \tag{21}$$

and balancing factors are

$$A_i = \frac{1}{\sum_j B_j D_j d_{ij}^{-\beta}} \tag{22}$$

$$B_j = \frac{1}{\sum_i A_i O_i d_{ij}^{-\beta}} \tag{23}$$

Martyn proposed a solution to the interdependency of balancing factors by initially assigning them a value of 1 and then iteratively computing them until their differences become insignificant.[7].

2.1.2 Modeling and Calibration of parameter

The production-constrained model is a suitable approach for simulating inter-regional flows when the number of jobs in each destination j and the population of workers in each zone i are known. The model assumes that the attractiveness of a zone is represented by the number of jobs available(J_j), and constrains the total flow of people leaving each zone (P_i). The model can be expressed as

$$T_{ij} = A_i P_i J_j^{\gamma} d_{ij}^{-\beta} \tag{24}$$

where T_{ij} would be the flow of people traveling from zone i to work in zone j. By using this model, it is possible to estimate the flow of people travel for working while ensuring that the total flows from each origin are consistent with known data.
The model calibration, based on the re-specification produced by Flowerdew [2], assumes that the flows follow a Poisson distribution with an expected value λ. The model then can be written as

$$\lambda_{ij} = \exp(\alpha_i + \gamma \ln J_j - \beta \ln d_{ij}) \tag{25}$$

where the estimate of T_{ij} is logarithmically linked to the linear combination of other variables, and α_i acts as the balancing factor A_i.
The calibration is performed using the 'glm' function in the Statsmodel.api library,with α modelled as a categorical predictor. The calibrated values of γ and β are 0.769 and 0.878, respectively; The model has a r^2 of 0.388 meaning it accounts for about 38.8% of the variation of flows in London.
However,the doubly constrained model is considered the most appropriate model for simulating transportation systems since it takes into account both the population at origins and the number of jobs at destinations, resulting in more accurate flow simulations. The Poisson Regression model was calibrated with a β value of 0.9097, based on the negative power law. The entropy maximization method produced a prediction with an r^2 value of 0.41. However, the model was not used to simulate the scenarios in the following section because only a single change was made in the destination mass term

2.2 Scenarios

2.2.1 Scenario A

The adjusted vector of numbers of jobs in destinations(J_j) was used to recalculate the balancing factor A_i in order to satisfy the origin constraint after modifying the number of jobs in Canary Wharf.The new balancing factor is expressed as $A_i = \frac{1}{\sum_j J_{jadjusted}^{\gamma} d_{ij}^{-\beta}}$.

2.2.2 Scenario B

Assuming that transport costs are proportional to distance ($cost \propto d_{ij}$), the cost function can be adjusted to reflect the increase in the cost of transport. This can be achieved by modifying the parameter β in the cost function, which aligns with the power law ($f(d_{ij}) = d_{ij}^{-\beta}$).
The adjustment of the parameter βcan be made using the formula

$$(kd_{ij})^{-\beta} = d_{ij}^{-\beta_{adjusted}} \tag{26}$$

, where k is a proportionality constant that reflects the effects of transportation cost on distance. Here, kd_{ij} represents the new travel cost as a whole. After taking logarithmic of both sides, the equation will look like the following

$$\beta(log(k) + log(d_{ij}) = \beta_{adjusted} log(d_{ij}) \tag{27}$$

So the adjust β can be expressed as

$$\beta_{adjusted} = \beta \frac{log(d_{ij}) + log(k)}{log(d_{ij})} \tag{28}$$

To ensure the origin constraint stands in scenario B, the balancing factor was recomputed as $A_i = \frac{1}{\sum_j J_j^{\gamma} d_{ij}^{-\beta_adjusted}}$.

Scenario B.1 Assuming that transportation has increased causing the distance to become 1.25 times greater, the new value of beta is calculated as 0.899.

Scenario B.2 Assuming that transportation has increased causing the distance to become 1.5 times greater, the new value of beta is calculated as 0.916.

2.2.3 Analysis

Table 7 compares the observed flows to the flows in scenario A. It reveals that the flow distribution in scenario A has a lower level of variation compared to the observed data. Furthermore, the mean values are similar, but the quantiles at 25%, 50%, and 75% are higher in scenario A. Notably, the maximum flow value has decreased substantially from 15496 to 4480. Overall, these findings suggest a shift towards a more evenly distributed flow pattern in scenario A. The flow between Waterloo and Canary Wharf, as well as London Bridge and Canary Wharf, have decreased substantially, from 8085 to 667.0 and 6165 to 412.0, respectively.

Table 7: Statistics of Flow Distribution

Statistic	**Total**	**Flow Estimate in A**
Count	61413.000000	61413.000000
Mean	25.113298	25.107860
Standard Deviation	131.479668	84.478902
Minimum	0.000000	0.000000
25% Quantile	0.000000	2.000000
50% Quantile	3.000000	7.000000
75% Quantile	13.000000	20.000000
Maximum	15946.000000	4480.000000

Table 8: Statistics of flow distribution in Scenario B

	Total	Flow Estimation in B.1 (distance increased by 1.25)	Flow Estimation in B.2 (distance increased by 1.5)
Count	61413	61413	61413
Mean	25.113298	25.105955	25.105629
Std	131.479668	83.564952	85.431747
Min	0.000000	0.000000	0.000000
25%	0.000000	2.000000	2.000000
50%	3.000000	7.000000	7.000000
75%	13.000000	20.000000	20.000000
Max	15946.000000	4274.000000	4441.000000

Table 8 summarizes the statistics of flows in scenario B. The significant decrease in both standard deviation and maximum flow indicate a more dispersed flow pattern. Noticeably, when the cost of transport has increased even more, the effect on flow is not substantial, suggesting a marginal utility.In scenario B, long-distance flows experienced the largest decrease, while short-distance flows experienced the greatest increase. Overall, the impact of increased transport costs on flow redistribution was greater, as indicated by the mean change in flow of -0.07 compared to -0.05 due to decreased jobs in Canary Wharf.

References

[1] Giorgio Fagiolo. Clustering in complex directed networks. *Physical Review E*, 76, 08 2007.

[2] Robin Flowerdew and Murray Aitkin. A method of fitting the gravity model based on the poisson distribution*. *Journal of Regional Science*, 22:191–202, 05 1982.

[3] Vito Latora and Massimo Marchiori. Efficient behavior of small-world networks. *Physical Review Letters*, 87, 10 2001.

[4] Jukka-Pekka Onnela, Jari Saramäki, János Kertész, and Kimmo Kaski. Intensity and coherence of motifs in weighted complex networks. *Physical Review E*, 71, 06 2005.

[5] Taylor Oshan. A primer for working with the spatial interaction modeling (spint) module in the python spatial analysis library (pysal), 11 2016.

[6] Jari Saramäki, Mikko Kivelä, Jukka-Pekka Onnela, Kimmo Kaski, and János Kertész. Generalizations of the clustering coefficient to weighted complex networks. *Physical Review E*, 75, 02 2007.

[7] Martyn L. Senior. From gravity modelling to entropy maximizing. *Progress in Human Geography*, 3:175–210, 06 1979.

[8] A G Wilson. A family of spatial interaction models, and associated developments. *Environment and Planning A: Economy and Space*, 3:1–32, 03 1971.