I.1. Centrality measures

Degree centrality characterises important nodes as those with the largest number of links to other nodes in the graph. Stations with high degree centrality have more stations that are one stop away.

Degree centrality is defined as follows:

$$C_D(u) = \frac{d(u)}{n-1}$$

where $C_D(u)$ is the degree centrality of node u, $d(u)$ is the degree of node u, n is the number of nodes in the graph and $n-1$ is the maximum number of links for a node.

Rank	Station	Degree centrality
1	Stratford	0.0225
2	Bank and Monument	0.0200
3	Baker Street	0.0175
4	King's Cross St. Pancras	0.0175
5	West Ham	0.0150
6	Canning Town	0.0150
7	Waterloo	0.0150
8	Green Park	0.0150
9	Oxford Circus	0.0150
10	Liverpool Street	0.0150

Table 1: Top 10 ranked stations in terms of degree centrality

Closeness centrality looks at which node has, on average, the shortest distance to all other nodes. The shortest path from node u to v is the one with the fewest edges. This means that commuters at stations with high closeness centrality can, on average, access their destination station in fewer stops compared to an origin station with lower closeness centrality. The actual distance between stations is not considered in the shortest path calculations here as commuters can be argued to be more concerned about their commute time rather than distance, and a path with fewer stops is generally more efficient. Thus, all edges are considered to have an equal distance.

Closeness centrality is defined as follows:

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

where $C_C(u)$ is the closeness centrality of node u, $d(u,v)$ is the shortest path distance between nodes u and v, and n is the number of nodes that can reach u. Closeness centrality is normalized to $\frac{n-1}{|G|-1}$ where |G| is the size of the graph.

Rank	Station	Closeness centrality
1	Green Park	0.115
2	Bank and Monument	0.114
3	King's Cross St. Pancras	0.113
4	Westminster	0.113
5	Waterloo	0.112
6	Oxford Circus	0.111
7	Bond Street	0.111
8	Farringdon	0.111

9	Angel	0.111
10	Moorgate	0.110

Table 2: Top 10 ranked stations in terms of closeness centrality

Betweenness centrality characterises important nodes as those that appear more frequently on the shortest path between all other pairs of nodes. The implicit assumptions are that (1) flows between all pairs of nodes occur with equal probability and (2) all flows occur along the shortest path. In the context of the underground, the first assumption may not hold as flows between pairs of stations vary, but the second assumption is fair as commuters can be expected to minimise travel time and cost. Stations that have high betweenness centrality are important in connecting other stations together, and commuters will frequently pass by these stations as they move from one station to another.

Betweenness centrality is defined as follows:

$$C_B(u) = \sum_{s,t \in V} \frac{\sigma(s,t|u)}{\sigma(s,t)}$$

where $C_B(u)$ is the betweenness centrality of node u, V is the set of nodes, $\sigma(s,t)$ is the number of shortest paths between s and t, and $\sigma(s,t|u)$ is the number of those paths that pass through u. Betweenness centrality is normalised by dividing by $\frac{(n-1)(n-2)}{2}$, which is the total number of pairs of nodes excluding the node itself for an undirected graph.

Rank	Station	Betweenness centrality
1	Stratford	0.298
2	Bank and Monument	0.290
3	Liverpool Street	0.271
4	King’s Cross St. Pancras	0.255
5	Waterloo	0.244
6	Green Park	0.216
7	Euston	0.208
8	Westminster	0.203
9	Baker Street	0.192
10	Finchley Road	0.165

Table 3: Top 10 ranked stations in terms of betweenness centrality

I.2. Impact measures

The size of the largest component provides information on the number of stations that are connected in the main network. This is important when stations are removed and the network starts to fragment into multiple components, where commuters cannot travel between disconnected components. This measure is normalised by dividing by the initial size of the network, 401, to give the fraction of total nodes in the largest component. A larger value indicates less disruption as most stations are still connected, even if some stations are no longer accessible.

The second impact measure is the average shortest path length between any two nodes in the network. A shorter average path length indicates that commuters can travel more quickly from one station to another.

The average shortest path length is defined as follows:

$$a = \sum_{s,t \in V} \frac{d(s,t)}{n(n-1)}$$

where a is the average shortest path length, V is the set of nodes, $d(s,t)$ is the shortest path length from s to t, and n is the number of nodes in the graph. The sum of all shortest path lengths is divided by $n(n-1)$, which is the total number of pairs of nodes.

As this measure requires the distance between all pairs of nodes, it cannot be calculated for a disconnected network. Thus, the average shortest path length will be calculated for the largest component. This creates a problem when the size of the largest component changes as the average shortest path length cannot be compared between graphs of different sizes. This requires the measure to be normalised by dividing by the size of the largest component, so that smaller components are penalised. This allows the measure to be used to compare the effects of node removal on the network.

Both the size of the largest component and the average shortest path length can be used to evaluate the resilience of other networks.

I.3. Node removal

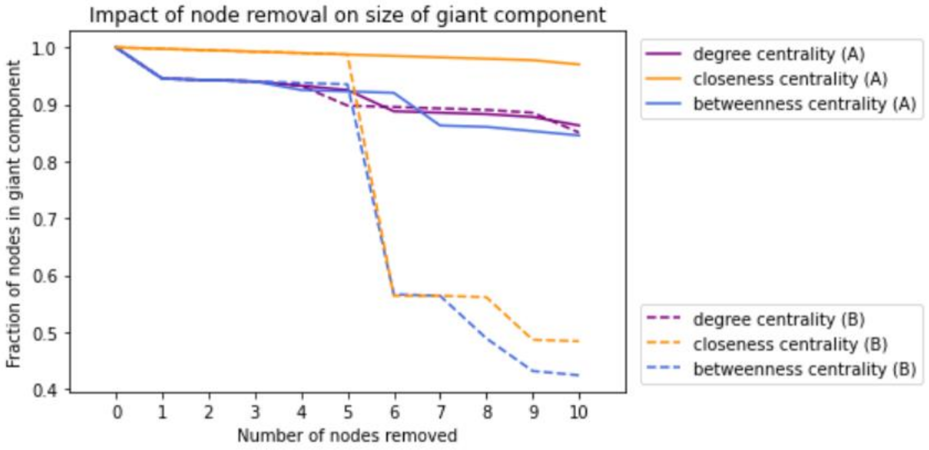

Figure 1: Impact of node removal on size of giant component

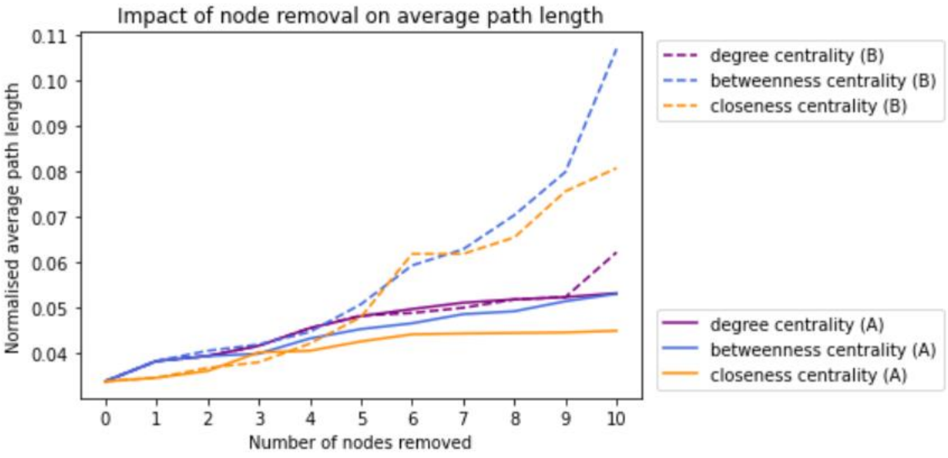

Figure 2: Impact of node removal on average path length

Figures 1 and 2 show that the sequential removal of nodes, strategy B, is more disruptive to the network. In Figure 1, there is a sharp decrease in the fraction of nodes in the largest component when the 6th node is removed, for closeness and betweenness centrality following strategy B. This contrasts with strategy A, where more than 80% of the nodes are still part of the largest component for all three centrality measures, even after 10 nodes were removed. In Figure 2, the normalised average path length for all three centrality measures is higher for strategy B than strategy A, indicating longer average travel time and distance for commuters. Thus, strategy B is better at identifying critical stations to protect as the network structure changes with the removal of nodes and the most important stations must be re-assessed.

Betweenness centrality reflects best the importance of a station to the network, as the removal of stations with the highest betweenness centrality is more disruptive than the removal of stations with the highest degree and closeness centrality. The fraction of nodes in the largest component is the smallest at the end of the node removal process, consisting of less than half the total nodes, while the normalised average path length is the largest among all the centrality measures. This result is not unexpected as the removal of nodes with high betweenness centrality breaks the shortest paths between stations.

Of the two impact measures, the size of the giant component is better at assessing the damage after node removal. This is because the average path length is not comparable for graphs of different sizes, and by normalising it, it is no longer interpretable as the average distance in the network. On the other hand, it is easy to see the fragmentation of the network using the fraction of nodes in the largest component and this is directly comparable to the initial network.

II.1. Weighted betweenness centrality

Stations with more flows between them should be closer to each other, thus, the flows must be inverted before they can be used to calculate the shortest paths. This way, the shortest path between two stations is the path with the most flows.

Stations with the highest weighted betweenness centrality lie on the paths with the most flows and their closure will be disruptive to a larger number of commuters.

Rank	Station	Weighted betweenness centrality
1	Green Park	0.573
2	Bank and Monument	0.505
3	Waterloo	0.416
4	Westminster	0.381
5	Liverpool Street	0.337
6	Stratford	0.331
7	Bond Street	0.292
8	Euston	0.284
9	Oxford Circus	0.271
10	Warren Street	0.254

Table 4: Top 10 ranked stations in terms of weighted betweenness centrality

II.2. Impact measure with flows

The fraction of nodes in the largest component can still be measured for a weighted network as it does not involve links between stations. An alternative measure to account for the flows in the network is the total flows in the largest component. This can be normalised by dividing by the total initial flows.

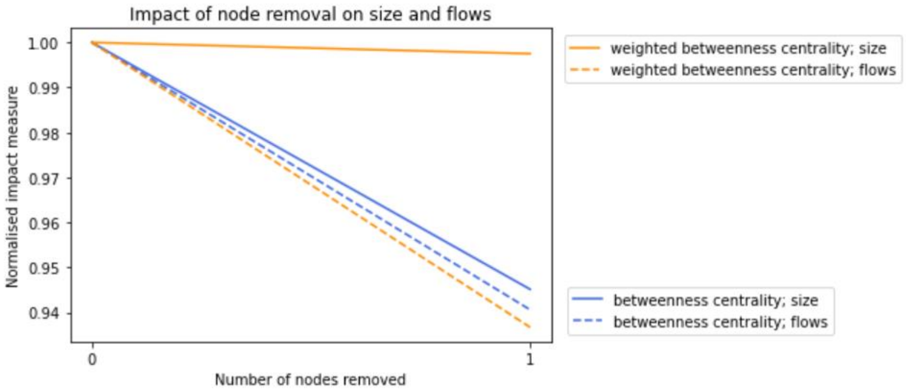

Figure 3: Impact of node removal on size and flows

Figure 3 shows that the removal of the station with the highest betweenness centrality (in blue) results in the fragmentation of the network, with the largest component accounting for about 95% of total stations and flows. On the other hand, the removal of the station with the highest weighted betweenness centrality (in orange) does not fragment the network, yet the reduction in total flows is greater. Thus, weighted betweenness centrality is better at identifying stations which, if removed, will cause the most disruption for commuters as they lie on the paths with the most flows and are the most heavily utilised by commuters.

III.1. The Family of Spatial Interaction Models

The generalised gravity model is defined as follows:

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

where T_{ij} is a measure of the interaction between zones i and j, O_i is a measure of the 'mass term' associated with zone i, D_j is a measure of the 'mass term' associated with zone j, d_{ij} is the generalised cost of travel between zones i and j, and k is a constant of proportionality. k, α, γ, and β are parameters to be estimated.

α and γ determine the effects the origin and destination 'mass terms' on the estimated flows. β is the friction of distance parameter that ensures that as the cost of travel, or distance, increases, flow decreases.

k ensures that total flows estimated by the model add up to the total observed flows, such that

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

where

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

This is known as the **unconstrained model**.

If the total flows, O_i, originating at each zone i are known, this can be added as a constraint in the model, such that

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

k can then be replaced by a set of balancing factors, A_i, to give the **production-constrained model**:

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

where

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

Alternatively, if the total flows, D_j, terminating at each zone j are known, this constraint can be written as

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

and k can instead be replaced by a set of balancing factors, B_j, to give the **attraction-constrained model**:

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

where

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

Lastly, if the constraints on both the origins and destinations hold simultaneously, this gives the **doubly-constrained model**:

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

where

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

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

III.2. Production-constrained model

The production-constrained model is chosen as this model will be used to estimate new flows when the attractiveness of a destination changes. Thus, the number of flows originating at each zone should remain the same, and the only change is where these people travel to.

A closer look at the data reveals origin stations with 0 population, which is used as the 'mass term' for the origin, and 0 jobs, which is the 'mass term' or the measure of attractiveness of the destination. This occurs when Battersea Park is the origin or destination station, and it is likely due to the way the flow data was processed, since Battersea Park station only recently opened in September 2021. Thus, as these flows are likely to be inaccurate and could cause problems when modelling, they will be removed from the dataset. Furthermore, origin-destination pairs that start and end at the same station are removed.

The cost function can be stated as a power function, $d_{ij}^{-\beta}$, or a negative exponential function, $\exp\,(-\beta d_{ij})$. The model was calibrated with both possibilities, and it was found that the negative exponential function fits the data better, as shown in Table 5. This is consistent with previous research that suggests that power functions fit better in interurban studies, whereas the negative exponential function often fits better in intraurban studies (Wilson, 1971). The negative exponential cost function will be used for the remaining models in this report.

Cost function	R-squared	Root Mean Square Error (RMSE)
Inverse power ($d_{ij}^{-\beta}$)	0.388	102.89
Negative exponential ($\exp\,(-\beta d_{ij})$)	0.468	96.26

Table 5: Model performance using the inverse power and negative exponential cost functions

The calibrated parameter β has a value of 0.000153.

IV.1. Scenario A

New flows were estimated by reducing the number of jobs in Canary Wharf by half. In order to conserve total flows in the system and flows from the origins, the set of balancing factors, A_i, was re-computed using the calibrated β and γ values from part III.2. As a result, flows into Canary Wharf decreased while flows into other stations increased.

IV.2. Scenario B

The doubly-constrained model is used for Scenario B to fix the population at the origins and the jobs at the destinations. This allows us to study the effects of an increase in the cost of transport using 2 different values of β. In the initial calibrated model, $\beta = 0.000154$. Based on this value, 2 additional values of β are chosen, $\beta = 0.0005$ and $\beta = 0.0009$. A larger β value indicates that distance is a larger deterrence.

The new set of balancing factors, A_i and B_j, were iteratively recalculated until convergence and used to obtain the new flows, preserving the total flows in the system and the flows from the origins and into the destinations.

IV.3. Analysis

The correlation coefficient, r, and the root mean square error, RMSE, between the observed flows and the predicted flows for each of the scenarios was calculated. This is used as an indicator of the redistribution of flows. A scenario with a greater impact on the redistribution of flows can be expected to deviate more from the observed flows, resulting in a smaller correlation coefficient and a larger root mean square error.

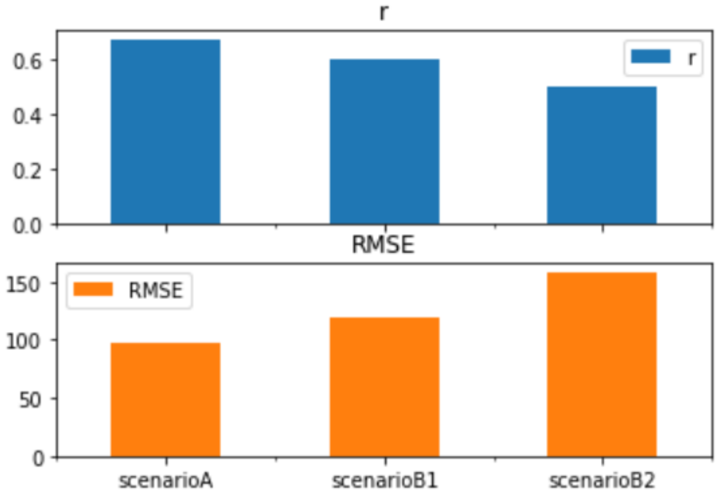

Figure 4: r and RMSE values for each of the three scenarios

Based on the results in Figure 4, Scenario A has the highest r value and the lowest RMSE. This can be interpreted as having the least impact on the redistribution of flows, since the predicted flows are closer to the observed flows, compared to the two other scenarios. This makes sense as only the attractiveness of Canary Wharf was modified, whereas the 'mass terms' for the other origins and destinations was unchanged in Scenario A. On the other hand, an increase in the cost of transport would affect the entire system in Scenario B.

Between Scenarios B1 ($\beta = 0.0005$) and B2 ($\beta = 0.0009$), Scenario B2 had a larger impact on the redistribution of flows, as evidenced by the smaller r value and higher RMSE. This is because the deviation from the calibrated β value of 0.000154 is larger. Thus, Scenario B2 results in the greatest impact on the redistribution of flows of the three scenarios.

Word count: 2500

References

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