Door Will Open: A Multidimensional Analysis of Job Accessibility in London

Candidate 21681
MSc Applied Social Data Science
MY498 Capstone Project
7 August 2019
10,000 words

Abstract

London's remarkable vibrancy, productivity, efficiency and innovativeness are shaped by how effectively it can connect workers and jobs to each other. I categorise London's neighbourhoods by job accessibility via public transport, considering the economic sectors of jobs, the social deprivation of potential workers and the novel dimension of travel time reliability alongside an analysis of the overall availability of jobs per potential worker using a two-stage floating catchment area (2SFCA) measure. I find that different areas of London have very different levels of job availability. Central areas rely more heavily on the finance, insurance, real estate, professional and administrative support sectors for jobs. The reliability of travel times and travel modes also varies significantly across London. A k-means clustering of neighbourhoods by these dimensions reveals that low job availability, high social deprivation, high reliance on certain sectors for jobs, low travel time reliability and low modal reliability intersect in different ways across London. The findings support ongoing efforts by Transport for London (TfL) and the Greater London Authority (GLA) to improve transport connectivity and strategically distribute growth in housing and jobs, but also suggest ways to dovetail economic and social strategies to better address location-specific circumstances. Finally, I propose a novel graph-theoretic interpretation of 2SFCA that indicates that the method is more flexible and widely applicable than currently evidenced in the literature.

Contents

Introduction

London's remarkable vibrancy, productivity, efficiency and innovativeness depend in large part on its effectiveness in connecting workers to job opportunities through economic, social, intellectual and physical networks [1]. Job accessibility (that is, the ease with which workers can take advantage of a suitable range of job opportunities) [2] has an intrinsically physical dimension in the form of the spatio-temporal distance between workers and jobs. Recent studies of job accessibility in London have confirmed that the number of jobs that a worker can physically reach within a certain travel time varies widely across space [3], and that the degree of spatial matching between the home and work locations of workers varies by occupation [4]. However, job accessibility also varies across non-spatial dimensions [5, 6, 7, 8, 9], of which two are of special interest to policymakers. Firstly, even if jobs are physically accessible by workers, some workers may face sociodemographic barriers (like a lack of relevant skills) that prevent them from actually occupying those jobs [10]. Secondly, the margin of uncertainty in travel times may force workers to budget more time for travelling in order to be confident of reaching suitable jobs without being late [11, 12].

This study is a multidimensional analysis of job accessibility by public transport in London (Figure 1). I categorise neighbourhoods by their overall availability of jobs, the degree to which they depend on particular economic sectors for jobs, the relative social deprivation of their working-age population and a novel dimension, the reliability of travel times for trips originating in them. In addition to providing insights into how land use and transport interventions can adjust the urban structure of London to improve job-worker spatio-temporal matching, the results can also provide guidance for how to address the non-spatial dimensions, such as through programmes to improve workers' skillsets, encourage job growth in certain sectors and improve the reliability of public transport operations, thus helping to improve job accessibility using more coordinate strategies across a wider range of domains. Finally, the study also proposes a novel graph-theoretic interpretation of the job accessibility metric used, which illuminates significant new domains for its potential application.

Figure 1: London

Black lines are the London rail network (Docklands Light Railway (DLR), National Rail, Overground, TfL Rail, trams, Underground)

Neighbourhoods mentioned in this study are labelled

Literature Review

Hansen [2] defines accessibility as the potential for interaction between people and the opportunities that they seek. Job accessibility thus measures the range of job opportunities that are available for a person to choose from. This differs from connectivity, which measures how easy it is to reach a particular opportunity, and assignment, which is concerned with which opportunity each person should choose given the range available to her and some global optimisation objective.

Many city-scale studies [13, 14, 8, 3, 15, 16, 17, 18, 9, 19] measure job accessibility with "location-based" indicators [20]. These emphasise the physical aspects of job accessibility, namely the spatial distributions of jobs and workers and the cost of travel between different locations. Although most extant studies obtain the required travel costs through network analysis, as far as I can tell the graph-theoretic properties of these measures have been under-emphasised. Specifically, the fact that the measures encode relationships between workers and jobs as bipartite graphs has not been highlighted, leaving substantial avenues for innovative applications of these measures unexplored.

Besides analysing spatial patterns in London's overall job accessibility, I also examine how accessibility varies by different types of workers, jobs and connectivity. To capture spatial variations in sociodemographic characteristics among workers, I follow several studies [14, 16, 18, 19] to use area-level composite scores that summarise multiple dimensions of sociodemographic characteristics, namely the English Indices of Multiple Deprivation (IMD) [21]. I also follow the methodology of Shen [13], Deboosere & El-Geneidy [22] and Shen & Batty [4] to study how job accessibility varies by job type, by disaggregating job accessibility by economic sector.

Studies that examine how accessibility varies by different types of connectivity generally analyse differences in travel times between transport modes [13, 23, 3, 16] or time periods (such as the AM peak and evening off-peak periods) [24, 25, 26]. However, connectivity for the same period (such as the AM peak) also varies across different days, as a trip on any given day may be affected by congestion, vehicles arriving late or breakdowns in the transport network [27]. In reality, travellers have to account for a margin of uncertainty around expected travel times where deciding when and how to travel. Travel time reliability is under-studied in the job accessibility literature, but has received increasing attention in the transport studies literature, driven by the increasing availability of large-scale transport data (such as automated fare collection records, automated vehicle location traces or live trip planning information) that tracks the amount of time needed to complete a trip in real time [11].

In light of the existing literature on job accessibility, this study is positioned to make several contributions. Firstly, in addition to confirming how job accessibility varies across space in London, my analysis illuminates how these accessibility levels vary among workers and jobs with different characteristics. Secondly, I demonstrate how the novel dimension of travel time reliability can be incorporated into an accessibility analysis. Finally, I illuminate the under-explored graph-theoretic properties of the accessibility indicator that I use, which opens up promising new areas of potential research using this method.

Methodology and Data Sources

Choosing a Location-Based Accessibility Indicator

Many types of location-based accessibility indicators exist [20]. "Cumulative potential" indicators are the simplest members of this class, expressing job accessibility in terms of the total number of jobs that can be physically reached from each area within a certain travel cost (also known as a catchment):

$$a_o = \sum_{d=1}^{n} j_d f(c_{o,d})$$

where a_o is the accessibility level of one of m origin areas o, j_d is the number of jobs in one of n destinations d, $c_{o,d}$ is the travel cost (in this study, time) between o and d, and $f(c_{o,d})$ is a distance decay function reflecting the declining attractiveness of the jobs in d to workers in o as $c_{o,d}$ increases.

Cumulative potential indicators treat job accessibility as non-excludable: the fact that any particular worker can reach a job does not affect any other worker's ability to reach the same job. This fails to account for the fact that each job can typically only be occupied by one worker. To address this theoretical flaw, Shen [13] proposed a method to account for the spatial distribution of potential workers, which was then further developed in the public health literature [5, 7] as the two-step floating catchment area (2SFCA) method. This is given by:

$$r_o = \sum_{d=1}^{n} \frac{j_d}{\sum_{o=1}^{m} w_o f(c_{o,d})} f(c_{o,d})$$

where w_o is the number of workers in o, $\sum_{o=1}^{m} w_o f(c_{o,d})$ sums up for each destination d how many potential workers across all m origins can reach it, $\frac{j_d}{\sum_{o=1}^{m} w_o f(c_{o,d})}$ divides the number of jobs in each destination d among the number of potential workers who can reach it, and $\sum_{d=1}^{n} \frac{j_d}{\sum_{o=1}^{m} w_o f(c_{o,d})} f(c_{o,d})$ sums up for each origin o the per-worker job contributions over all n destinations that can be reached.

Like a_o, r_o is a score reflecting the relative availability of job opportunities within o's catchment. However, the 2SFCA method distributes every job exactly once (i.e. there are no leftover undistributed jobs) to potential workers in an excludable way (i.e. if a job is distributed to a particular worker, that job is no longer available to other workers) [13, pp. 363-364]. Thus, r_o reflects both connectivity and competition effects: as the number of jobs within o's catchment increases, r_o will increase, but as the number of potential workers who can reach the same catchment (and thus compete for the same jobs) as o increases, r_o will decrease. Due to these theoretical strengths, I will use the 2SFCA method to assess job accessibility levels in London.

Because the 2SFCA method distributes jobs to workers in an excludable way, the values r_o are meaningful for assessing whether the local supply of jobs is sufficient to satisfy the local demand.

Methods to assess the degree of inequality in distributions such as Lorenz curves can also be straightforwardly applied, by first calculating for each origin o how many jobs are available to it (given by $w_o r_o$), and then sorting the origins in ascending order by r_o values and plotting the cumulative proportion of the potential workers against the cumulative proportion of all jobs that are available to them.

Choosing a Spatial Aggregation Level and a Distance Decay Function

The use of the 2SFCA method requires two choices. The first is a choice of which level of spatial aggregation to use. This is dictated by the limited availability of very finely disaggregated spatial data on one hand, and the need to minimise distortions in the calculation of travel costs due to insufficient spatial resolution[1] on the other. After weighing the options, I set the m origins and n destinations in this study to, respectively, the population-weighted and jobs-weighted centroids of the 983 Middle-layer Super Output Areas (MSOAs) within the Greater London Authority (GLA) boundary. MSOAs are spatial units for statistical reporting created by the UK's Office of National Statistics (ONS) that break up the UK's land mass into areas that contain between 5,000 and 15,000 residents or between 2,000 and 6,000 households.

The second choice involves how to define the distance decay function $f(c_{o,d})$. In many studies, $f(c_{o,d})$ decreases smoothly as $c_{o,d}$ increases [13, 14, 23, 16, 18, 19, 26]. This more accurately reflects how the attractiveness of a destination declines continuously with increasing travel time [28]. However, it is hard to obtain sufficiently detailed and recent origin-destination flow data to accurately calibrate continuous distance decay functions [8, 28], and the resulting raw accessibility values are scores that are only meaningful when interpreted relative to other scores [5, 15].

In contrast, I will follow Luo & Wang [5], El-Geneidy et al [15], Tenkanen et al [24], Deboosere et al [22] and Järv et al [25] to use what Chen & Jia [28] call a binary distance decay, in which:

$$f(c_{o,d}) \begin{cases} 1 \text{ if } c_{o,d} \leq X \\ 0 \text{ otherwise} \end{cases}$$

where X is some travel time cut-off. This undoubtedly distorts reality, as it treats all locations where $c_{o,d} \leq X$ as if they were equally attractive and locations where $c_{o,d} > X$ as if they were totally unattractive. However, the distortion is relatively easy to understand and does not require further assumptions and modelling. At any rate, El-Geneidy et al [15, p. 303] and Chen & Jia [28] have found that accessibility measures calculated using the same value of X are generally highly correlated regardless of the shape of the distance decay function used. Most importantly, a binary measure maintains the raw

[1] Since travel times are calculated between two points, the travel time between two areas is represented by the travel time between their two centroids. The representativeness of centroid-to-centroid travel for area-wide travel conditions decreases as area size increases.

accessibility values as simple ratios that can be interpreted directly [5, 22]. Specifically for this study, the r_o values are interpretable as ratios of the number of jobs available to each potential worker within their catchment. I will term these quantities Job Provision Ratios (JPRs).

Having decided to use a binary distance decay function, the choice of X then becomes significant, as larger values of X will result in greater spatial smoothing of local unevenness in accessibility levels. I follow Chen & Jia's [28] advice to choose X using empirical evidence of people's actual willingness to travel and to reflect policy priorities. In 2017, the ONS Labour Force Survey found that people who worked in London had an average commute time of 46min. TfL [29, p. 224] also uses 45min as a travel time threshold for calculating job accessibility by public transport. As such, I will set X to 45min for this study (see Appendix 1 for an analysis of how sensitive my results are to other values of X).

Obtaining Job, Worker and Travel Time Data

Three components are needed to calculate JPRs. The first two, MSOA-level counts of potential workers (that is, residents aged 16 to 64) and jobs, are easily obtainable from the ONS. I use the most recent data available, which is for 2017.

The third component is a matrix of inter-area travel times. Most extant studies obtain this by compiling a multimodal transport network model and then simulating trips through the model (see Ford et al [3, pp. 129-130] for a step-by-step guide). A significant benefit of this approach is that researchers have full knowledge of the underlying network data used to generate the travel times, and can easily simulate alternative scenarios by adjusting the network and recalculating the times. However, this approach requires an extensive data compilation and cleaning effort to build a completely correct graph representation of the transport network.

A convenient alternative source of travel time data is Transport for London's (TfL's) Journey Planner [30]. This is a data service that provides the public with guidance on how to travel between any two points in London by walking or motorised public transport like buses, trams or trains (but not by car or taxi)[2]. This data source has several compelling advantages. First, it allows me to obtain travel time data without having to build my own transport network model for London, and the results reflect the network and routing assumptions actually used by TfL to assist real travellers in London. Secondly, the guidance comprehensively covers door-to-door travel. Thirdly, the data is highly disaggregated: besides information on the overall fare, distance and time needed for the trip, data on travel time,

[2] The Journey Planner also returns cycling options, which tend to be significantly faster than bus or walking for short trips. However, because only 2% of trips in London were by cycling [29], I omit cycling options from this study to avoid using overly optimistic travel times.

distance and mode are separately reported for each leg[3] of the trip. Finally, the data is freely available in bulk through an Application Processing Interface (API), subject to a limit of 300 trip queries per minute [31].

However, using this data source in this study does impose some constraints. This is because the Journey Planner provides information only for trips that do not involve cars or taxis, and is most reliable for trips within the GLA boundary. As such, my analysis will concentrate only on areas within the GLA boundary and will exclude travel by car or taxi. I recognise that these constraints will distort my results by introducing a systematic tendency to underestimate accessibility levels (particularly in areas with sparser public transport coverage that depend more on car travel) and by omitting the competition effects of the job and labour pools outside the GLA boundary. TfL data [29] can give a sense of how serious these distortions might be: in 2017, 37% of trips within the GLA boundary used cars or taxis, while 20% of trips within the GLA boundary were by people residing outside the GLA boundary.

I used the Journey Planner to build a "baseline" database of inter-MSOA public transport travel times during a typical weekday AM peak, assuming all services run as scheduled. To do this, I submitted requests for information on trips from each MSOA's population-weighted centroid to every MSOA's job-weighted centroid. All requests were for trips ending at 0830 on Friday, 22 March 2019, so that the trips have to occur during the AM peak, but the requests were submitted over 1 month from January to February 2019. As the requests were made far in advance of the intended travel date, the information obtained assumes that all services will run as scheduled on the travel date. After data collection was completed, I assesed for each trip the extent to which the travel time obtained using this method is representative of actual typical travel times during a weekday AM peak (see Appendix 2). No serious anomalies were detected.

Disaggregating Job Accessibility by Economic Sector

The ONS reports not only the number of jobs in each MSOA, but also the breakdown of these jobs by Standardised Industrial Classification Codes (SICCs). This enables me to disaggregate job accessibility into 6 overarching sectors:

1. Primary and Extractive
2. Industrial
3. Retail, Logistics, Food & Beverage and Hospitality (hereafter consumer services)

[3] A leg is a part of the trip that is covered by the same vehicle. For example, a trip that involves a walk from home to bus stop A, a ride on a bus to stop B, another bus ride from stop B to stop C and a walk from stop C to a workplace involves 4 legs.

4. Finance, Insurance, Real Estate, Professional and Administrative Support (hereafter high-end services)
5. Public Sector
6. Others

I then calculate JPRs for each sector s, $r_{s,o}$, which are given by:

$$r_{s,o} = \sum_{d=1}^{n} \frac{j_{s,d}}{\sum_{o=1}^{m} w_o f(c_{o,d})} f(c_{o,d})$$

where $j_{s,d}$ is the number of jobs in sector s in destination d.

Because the 2SFCA method distributes jobs to workers in an excludable way, Lorenz curves of the sector-specific job availability over all potential workers can be constructed and compared to the Lorenz curve for the overall job availability, to see the extent to which the spatial distribution of each sector's jobs contributes to the unevenness in overall job availability. Also, the sum of each MSOA's sector-specific JPRs will be equal to its overall JPR, that is $r_o = \sum_s r_{s,o}$. This allows me to calculate the proportion of o's JPR that is contributed by jobs in each sector s $\frac{r_{s,o}}{r_o}$, which captures the extent to which o is dependent on s for jobs. If an MSOA's potential workers' sociodemographic characteristics do not match the typical profile of employees in sector s, then increasing that MSOA's dependence on s may increase the probability of sociodemographic mismatch between potential workers and the jobs that are available to them [8, 17, 9].

Capturing Variations in Characteristics of the Potential Workers

I also wish to investigate how job accessibility varies across different types of potential workers. Unfortunately, worker counts that are disaggregated by sociodemographic characteristics (like educational qualifications, income or disability status) are only available from Census data, which are 8 years out of date at the time of this study. Instead, I use the IMD to reflect the relative degree of social deprivation experienced by potential workers living in different areas across a wide range of domains, including income, employment, health, education, crime, housing and the living environment. Following the official guidance [21], I aggregate the IMD scores to the MSOA level using weighted averages of the raw scores for smaller spatial units. I then focus on the relative ranking of the areas by IMD index scores, rather than on the raw scores themselves.

Obtaining Inter-MSOA Travel Time Reliability Data

As the Journey Planner provides travel guidance in real time, I also used it to obtain the travel times for selected trips over different days. The overall distribution of travel times per trip constitutes a measurement of its travel time reliability. However, this measurement probably underestimates real trip-level travel time variability.

Journey Planner travel times incorporate access (the time needed to walk to and from stations or between platforms in stations), waiting and in-vehicle times. Access times are based on assumed walking speeds between and within stations, and do not reflect in-station crowding that may reduce walking speeds around and in stations. Waiting and in-vehicle times are calculated solely on service schedules and do not reflect live vehicle locations. In particular, as long as vehicles are running along a route, the Journey Planner assumes all vehicles are running as scheduled, even if they are in fact delayed due to breakdowns or en route congestion[4]. Also, the Journey Planner assumes that travellers will always be able to board the first suitable vehicle that arrives, even though in-vehicle congestion may prevent boarding.

Finally, the Journey Planner only partially accounts for the impact of service disruptions on travel options. If there are service suspensions (i.e. no services running) on sections of rail lines (DLR, National Rail, Overground, TfL Rail, tram and Underground) due to disruptions or planned shutdowns, the Journey Planner will not return travel options that require travel along the suspended sections. However, if services are just delayed but not completely suspended, the Journey Planner will still continue to return travel options via the affected sections as if service schedules were not affected. Also, it assumes that walking and bus routes are never suspended. In particular, the Journey Planner will still recommend using a bus service between two stops, even if that service is actually diverted and service to either or both of the stops in question is suspended[5].

As such, the travel time distributions obtained from the Journey Planner only reflect two sources of variability: the need to reroute travel to avoid service suspensions on rail lines, and fluctuations in wait times during transfers[6] due to a lack of synchronisation between the planned arrival time of the previous vehicle and the planned departure time of the next vehicle. Nonetheless, these two sources of variability are arguably of special interest to transport operations planners, as rail service suspensions have an outsize impact on travel times, and schedule synchronisation is one of the few aspects of transport operations that planners can directly control.

To obtain a distribution of travel times per trip during the AM peak on a typical weekday, I requested for travel information on selected trips beginning between 0600 and 0800 over 50 weekdays between 29 January 2019 and 25 April 2019. I found that I could only process about 7,000

[4] Other TfL open data sources can potentially be used to reflect actual vehicle arrival and departure times – see Appendix 3

[5] If the results from the Journey Planner involve disrupted services, a text alert describing the extent of the disruption accompanies the results, but the user is left to interpret the impact of the disruption on her trip by herself.

[6] That is, the number of non-walking legs minus 1

real-time trip requests in the two-hour window, so I only gathered data on trips starting from all 983 MSOAs' population-weighted centroids that end at the job-weighted centroids of 7 MSOAs that correspond to 7 key transport nodes evenly spaced across the GLA area, totalling 6,881 origin-destination pairs. See Appendix 4 for how the 7 destinations were selected. I varied the start time of each trip every day by randomising the order in which requests for the 6,881 trips were submitted to the API. This allowed the data to capture the trip-specific variation in schedule synchronisation between different services at different points during the weekday AM peak, and the resulting distribution of travel times per trip represents the trip-specific probability distribution of the travel time for any randomly selected start time between 0600 and 0800 on a typical weekday.

Analysing Factors Affecting Inter-MSOA Travel Time Reliability

With the trip-level travel time distributions, I calculate what Uniman et al [11] defined as the Reliability Buffer Time (RBT), which is the difference between the 95th percentile travel time (for an unusually slow trip) and the median travel time (representing a typical trip). The RBT captures the additional travel time that a traveller needs to budget for in order to be 95% confident that a trip on any given day will not end late. Then, I use linear regression to analyse how the characteristics of a trip under baseline conditions affect its RBT. I expect that trips that require more transfers and that cover more distance by certain modes (since some modes, like the bus, are inherently less reliable due to lower operating frequencies and a lack of dedicated rights of way) will have lower travel time reliability.

However, the assumption of independence between observations of the dependent variable does not hold in this application. Because all the MSOAs are connected to each other via the same transport network, the RBTs of different trips are correlated. For instance, two MSOAs that are located along the same rail line will have similar RBTs for trips via that line to the same destination, as both trips will be affected by any disruptions along that line. If the non-independence of the RBTs between trips is not accounted for, the correlations among different RBT observations may distort the coefficient estimates for the covariates included in the model.

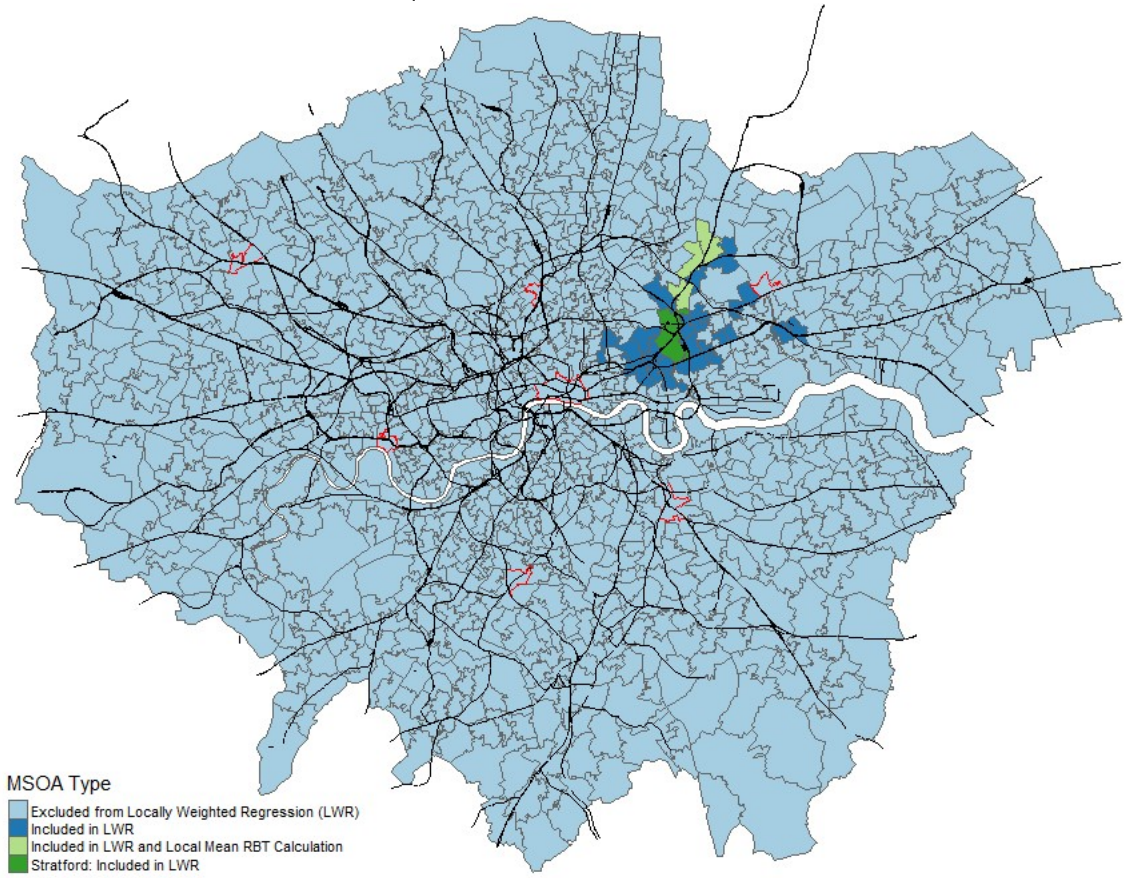
One way to deal with such correlations is to use a spatially autoregressive model, in which the value of the dependent variable in each area partially depends on the mean of the value of the dependent variable in its neighbours [32]. This captures the idea that areas that are near each other tend to be similarly connected to the larger network of areas and thus should experience similar network-related effects. To implement this, I first obtain for each MSOA o a vector of travel times to the 7 destinations from the baseline data, and then calculate the Euclidean distance between the vectors of every pair of MSOAs. Then, for each MSOA o, I identify the 3 other MSOAs (o1, o2 and o3) where this Euclidean distance is minimised; these are the MSOAs whose baseline travel times to the

7 destinations are the most similar to those of MSOA o. Then, for each trip i originating from MSOA o and terminating at each of the 7 destinations d, I calculate the local mean RBT l_i for trips starting from each neighbour o1, o2 and o3 and ending at the same destination. That is:

$$\text{Local Mean RBT for trip i from o to d} = l_i = \frac{RBT_{o1,d} + RBT_{o2,d} + RBT_{o3,d}}{3}$$

For an illustration of where the 3 neighbours are located for the case of one MSOA, Stratford, see Figure 2.

Figure 2:
MSOAs Involved in Local Mean RBT Calculation and Locally Weighted Regression for Stratford
MSOAs outlined in red are the 7 key destinations



The linear regression model that I fit is:

$$Y = \alpha + L\beta + T\gamma + D\delta + \varepsilon$$

where:

- Y is a column vector containing 6,881 values where y_i is the RBT for trip i;
- L is a column vector with 6,881 values where l_i is the local mean RBT for trip i;
- T is a column vector of 6,881 values where t_i is the number of transfers needed for trip i under baseline conditions;

- D is a 6,881 by f matrix where f is the number of travel modes and $d_{g,i}$ is the total distance covered by one of the f modes g for trip i under baseline conditions;
- α is an intercept term, β is a scalar representing the partial effect of the local mean RBT on the RBT, γ is a scalar representing the partial effect of transfers on RBT, δ is a column vector of length f representing the partial effects of distance covered by each mode on RBT, and ε is a column vector of 6,881 idiosyncratic disturbances.

While this regression model generates one global estimate over all trips for the partial effects of each covariate on the RBT, there are indications that the strength of each partial effect varies across locations. For example, it is widely reported that National Rail services in the southeast of London are particularly unreliable [33, 34, 35]. To investigate this, I use a form of locally weighted regression to refit the model [36], so that for each MSOA the model is fitted using only the trips originating from the closest neighbours for each MSOA and ending in the 7 destinations. This procedure generates one estimate for each coefficient for each MSOA. I define as neighbours the 30 MSOAs whose travel times to the 7 destinations are most similar to those of the MSOA in question (using the same Euclidean distance calculation as the derivation of the local mean RBT). For Stratford, Figure 2 illustrates which MSOAs are included in the locally weighted regression.

Developing a Typology of MSOAs with K-Means Clustering

Finally, I synthesise the various dimensions of job accessibility into a single result. Some studies attempt to do this by defining a function to summarise values along multiple dimensions of accessibility (such as the spatio-temporal distance between jobs and workers and the sectoral diversity of available jobs) into a single value [8, 22, 9], but this is a lossy process that collapses multiple dimensions into one, thereby sacrificing the nuances in each dimension. The resulting accessibility scores also do not reflect how each underlying dimension contributes to the final score. The results can thus be hard to interpret, especially if the underlying dimensions reflect qualitatively different phenomena.

Instead, I use a k-means clustering analysis to group the MSOAs into k types across 9 dimensions (after standardising the values in each dimension to have mean 0 and variance 1):

- The IMD rank for each MSOA, to capture the sociodemographic characteristics of the potential workers
- Overall JPRs, to capture overall job accessibility
- The percent of each MSOA's JPR contributed by high-end services, to capture sectoral dependence for job allocations

- The RBT for travel from each MSOA to the City, to capture travel time reliability to the most job-rich area in London
- I also identify which of the 6 other key destinations are closest to each MSOA and include the RBT for travel from each MSOA to its respective local destination. In the absence of travel time distributions for every origin-destination pair, this RBT represents to some extent the reliability of travel times from each MSOA to its immediate neighbourhood.
- MSOA-level coefficients for distance travelled by bus, National Rail, Underground and walking from the locally weighted regression, to capture the effects on travel time reliability of the most commonly used travel modes

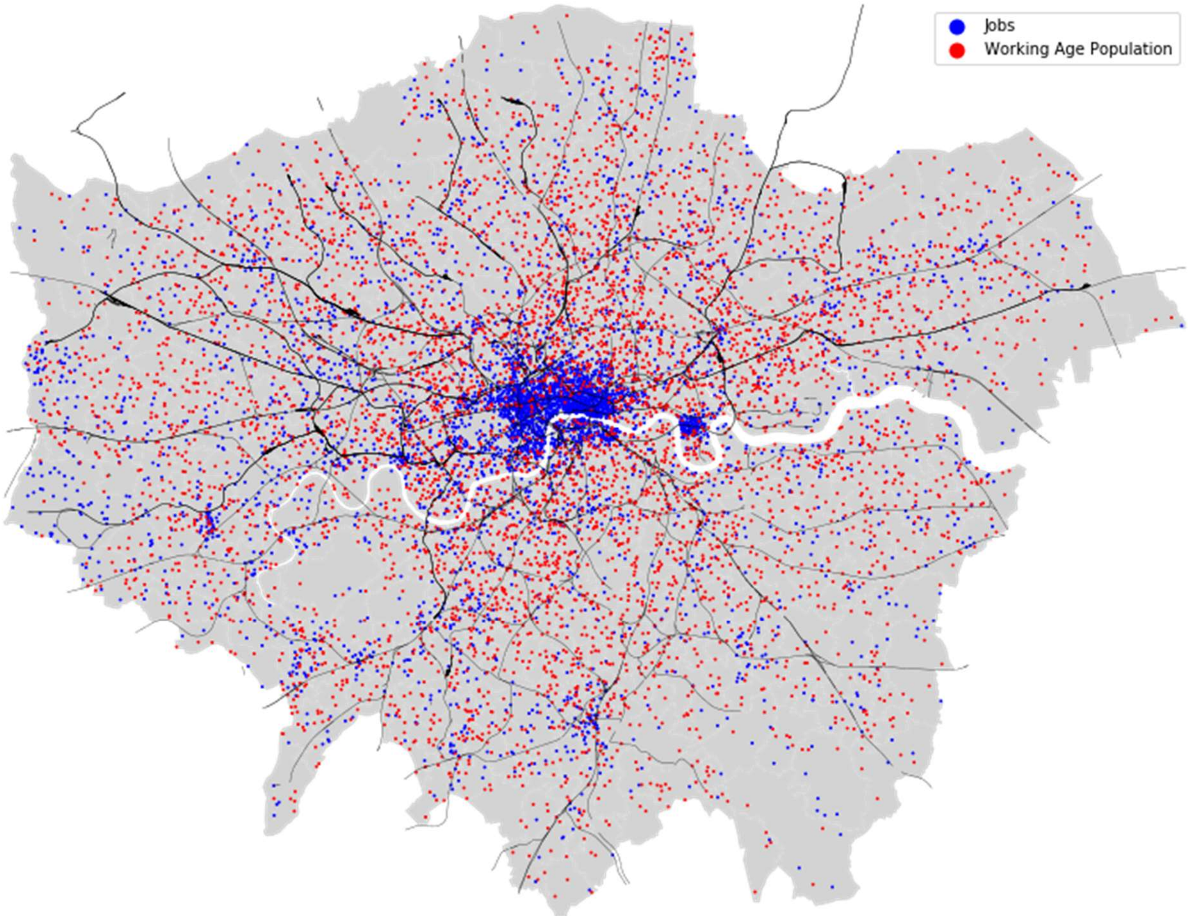
The typology of MSOAs obtained can capture how these dimensions intersect dynamically across London and highlight how each dimension contributes to the accessibility situation faced by each area. These results can then be used to help tailor a suite of interventions for each area to match its circumstances, as well as to prioritise interventions for areas that face more serious problems first.

Results

Overall Job Provision Ratios

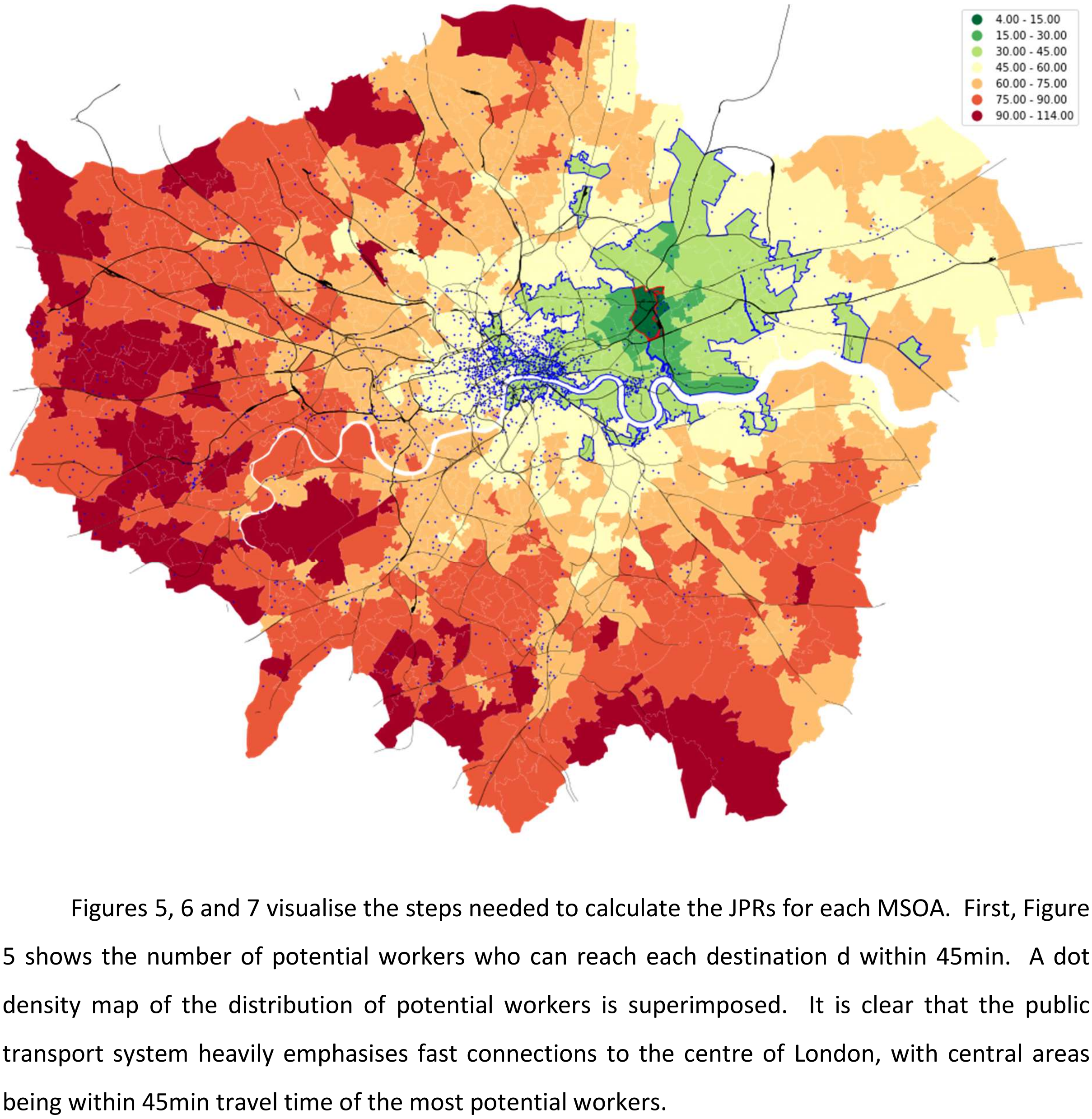
Each area's JPR is a function of 3 components: the spatial distributions of potential workers and jobs, and the connectivity provided by the public transport system. Figure 3 visualises the first two components as a dot density map. The distribution of jobs is much more clustered than the distribution of potential workers, with extreme concentrations in the City of London, Westminster and Southbank. There are also smaller concentrations of jobs around some stations in the GLA's rail network, most notably at Canary Wharf.

Figure 3: Spatial Distribution of Potential Workers and Jobs
1 dot per 1000 jobs or workers



The third component, area-to-area connectivity, is captured in a matrix of inter-MSOA travel times. Each value of $c_{o,d}$ is the minimum time it takes to travel from the population-weighted centroid of MSOA o to the job-weighted centroid of MSOA d to arrive at 0830 on a typical weekday, assuming that all services run as planned. It is difficult to visualise the entire matrix in a meaningful way, but Figure 4 visualises the travel times for all trips originating from the most populous MSOA, Stratford, with all areas within 45min travel time from Stratford (i.e. the areas where $f(c_{Stratford,d}) = 1$) outlined in blue and a dot density map of the distribution of jobs superimposed. In the JPR calculations, each destination d within the blue boundary will contribute some of its jobs to each potential worker in Stratford; exactly how much d contributes depends on how many other potential workers live in areas where they can also reach d within 45min.

Figure 4: Baseline Travel Times (min) from Stratford to each MSOA
Job distribution superimposed: 1 blue dot per 2000 jobs
Red outline indicates the location of Stratford
Blue outline indicates all MSOAs within 45min travel time from Stratford



Figures 5, 6 and 7 visualise the steps needed to calculate the JPRs for each MSOA. First, Figure 5 shows the number of potential workers who can reach each destination d within 45min. A dot density map of the distribution of potential workers is superimposed. It is clear that the public transport system heavily emphasises fast connections to the centre of London, with central areas being within 45min travel time of the most potential workers.

Figure 5: Number of Potential Workers (1000s) who can Reach each MSOA within 45min
Distribution of Potential Workers superimposed: 1 red dot per 2000

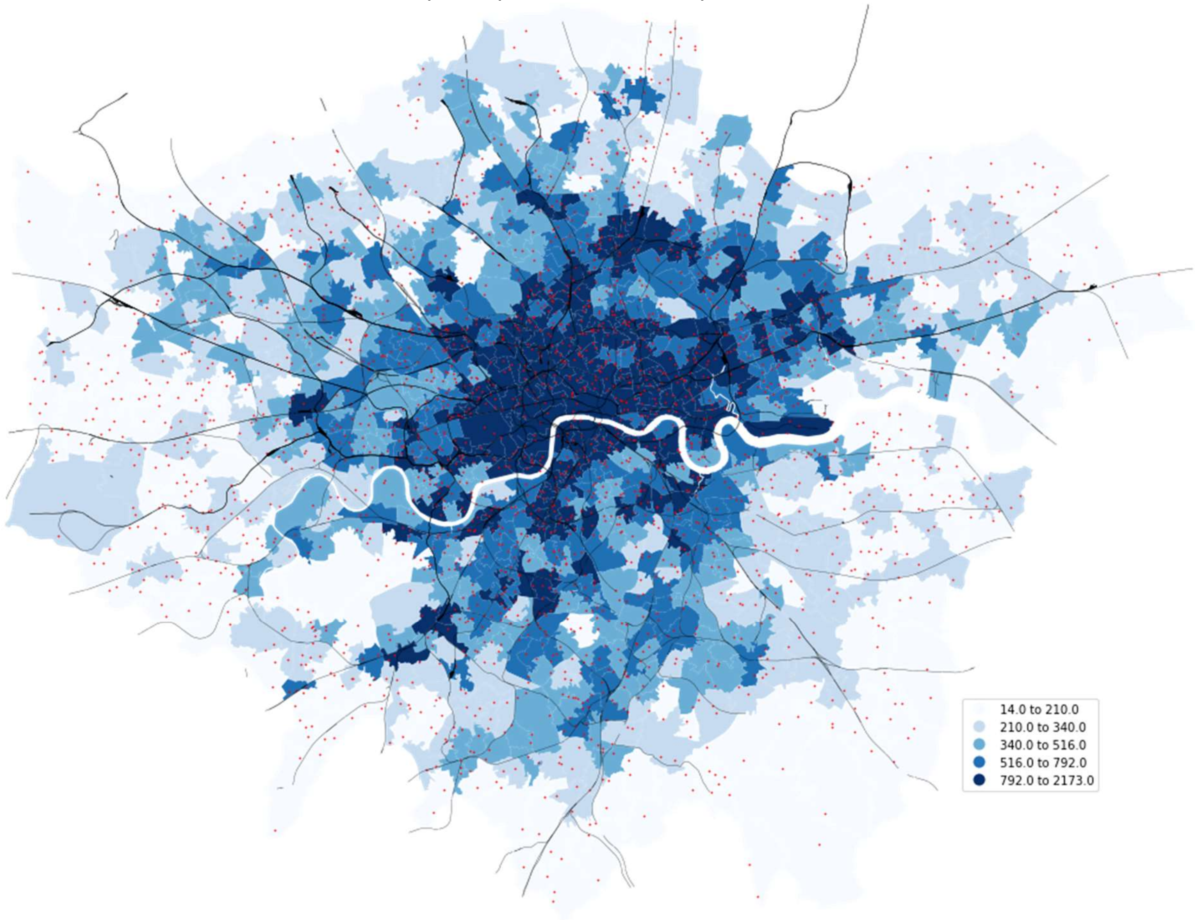
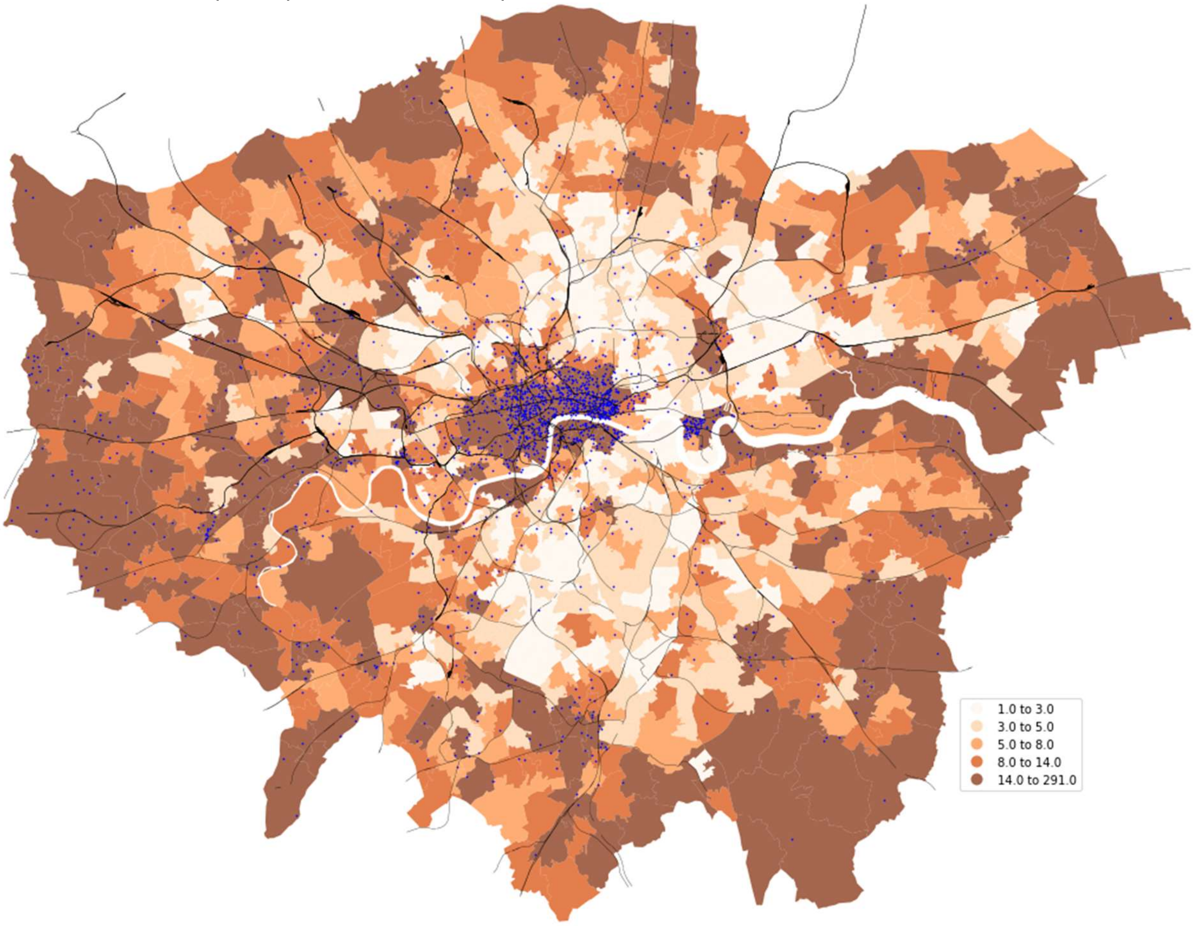


Figure 6 shows the number of jobs in each destination d that is provided to every 1,000 potential worker who can reach it. A dot density map of the distribution of jobs is superimposed. This ratio increases as the number of jobs in d increases but decreases as the number of potential workers who can reach d increases. A clear ring pattern can be seen, as in the most centrally located MSOAs, the sheer volume of jobs available offsets the high number of potential workers who can reach them, while a ring of MSOAs around the centre have few jobs but are easily reachable from many areas around London (which spreads the limited jobs available more thinly across larger catchments), and the MSOAs near the outskirts have few jobs but are also relatively isolated from other areas (so that the few jobs available are shared only among small, sparsely populated local catchments).

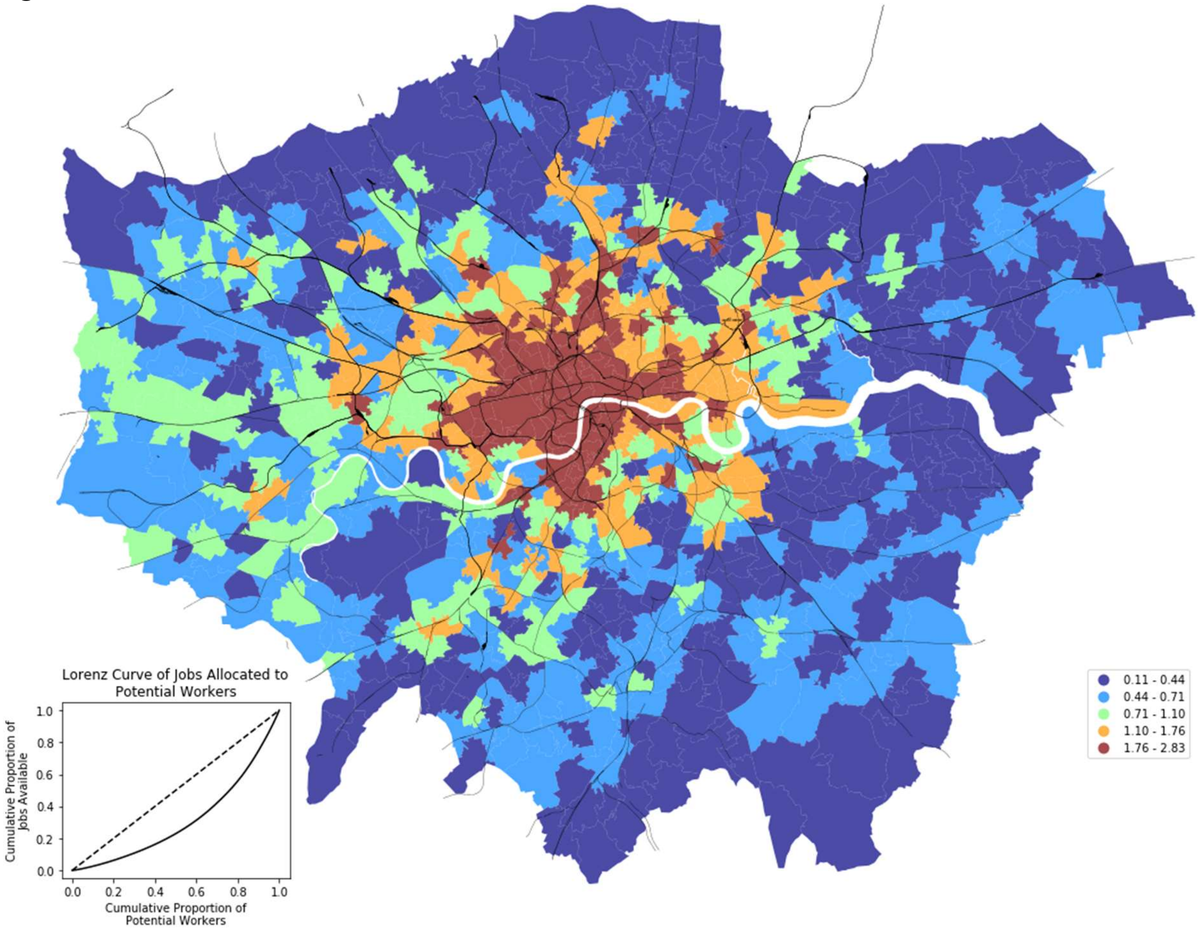
Figure 6: Number of Jobs Provided to every 1000 Potential Workers who can Reach each MSOA
Job distribution superimposed: 1 blue dot per 2000



Finally, Figure 7 displays the JPR for each origin o, which is the sum of the per-worker job contributions over all destinations d that can be reached within 45min. In order to interpret whether an MSOA's JPR is high or low, we can compare it to an ideal case in which jobs, potential workers and connectivity were all evenly distributed throughout the GLA area, which would result in all MSOAs having JPRs equal to R_0, the ratio of the total number of jobs to the total number of potential workers. This value in the GLA area is 0.88. To express the extent to which each MSOA's JPR deviates from R_0, JPR values are binned into 5 classes: 0 to 50% of R_0 (thus there are far too few jobs available to meet the needs of the potential workers), 50% to 80% of R_0 (the number of jobs available is somewhat insufficient for the number of potential workers), 80% to 125% of R_0 (the number of jobs available is proportional to the number of potential workers; this balanced state is visualised with green in the maps), 125% to 200% of R_0 (there are not enough potential workers within reach to fill the available jobs), and more than 200% of R_0 (there are far too few potential workers within reach to fill the available jobs). From Figure 7, we can see that most MSOAs have unbalanced JPRs. Too few potential workers are able to reach the MSOAs near the centre within 45min to fill all the available jobs, while many MSOAs around the outskirts provide too few jobs within reach of each potential worker. The

Lorenz curve in the inset of Figure 7 also captures the overall imbalance, as the bottom 50% of potential workers in the more job-poor areas share only roughly 20% of the jobs, while the top 10% of potential workers in the most job-rich areas have their pick of 20% of the jobs.

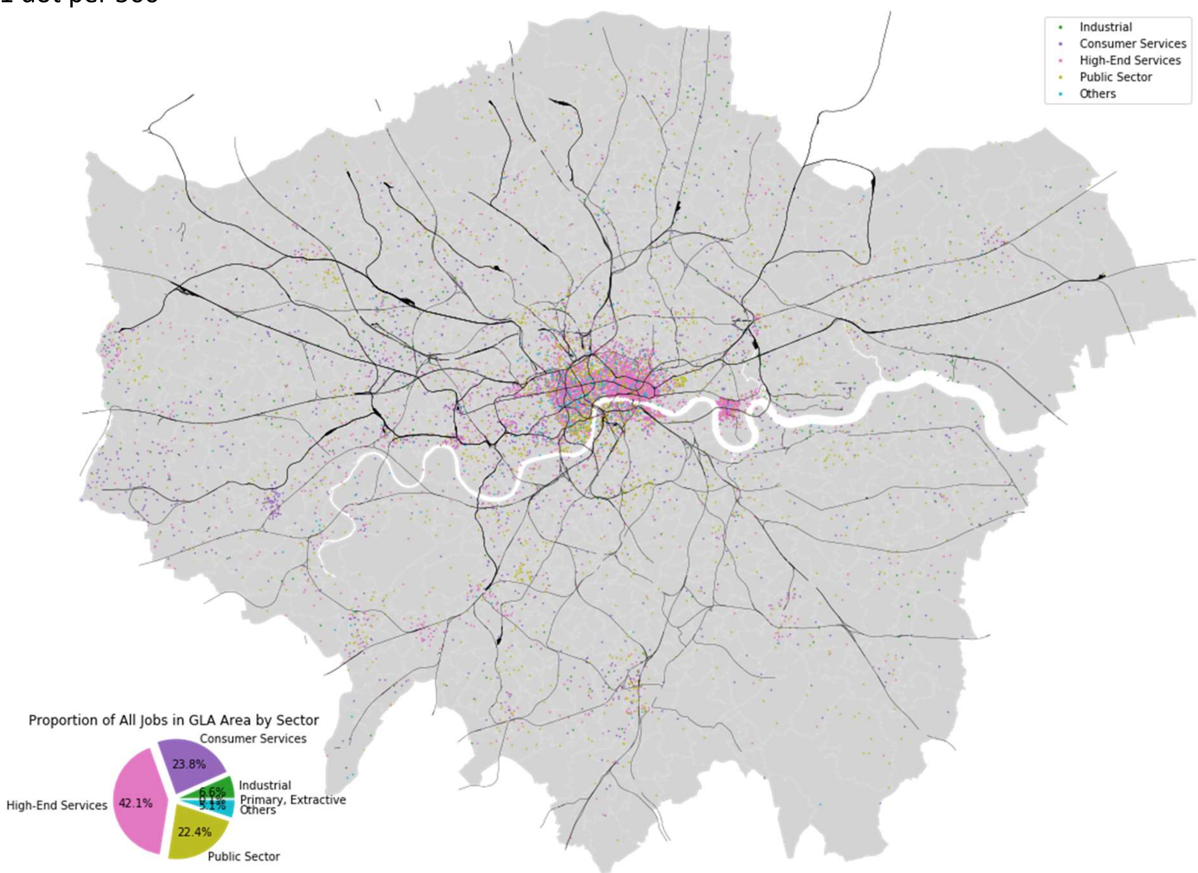
Figure 7: Job Provision Ratios – Number of Jobs Available to each Potential Worker in each MSOA



Job Accessibility by Sector

Next, I analyse how job accessibility varies by economic sector. The pie chart in Figure 8 shows that 42% of all jobs in the GLA area are from the high-end services sector, while consumer services and the public sector make up another 24% and 22% respectively. Figure 8 plots the distribution of jobs in each sector as a dot density map. The spatial distributions of industrial and consumer services jobs are the most decentralised, while public sector jobs are somewhat clustered around some rail interchanges. In contrast, high-end services jobs are mostly concentrated in the centre and at Canary Wharf.

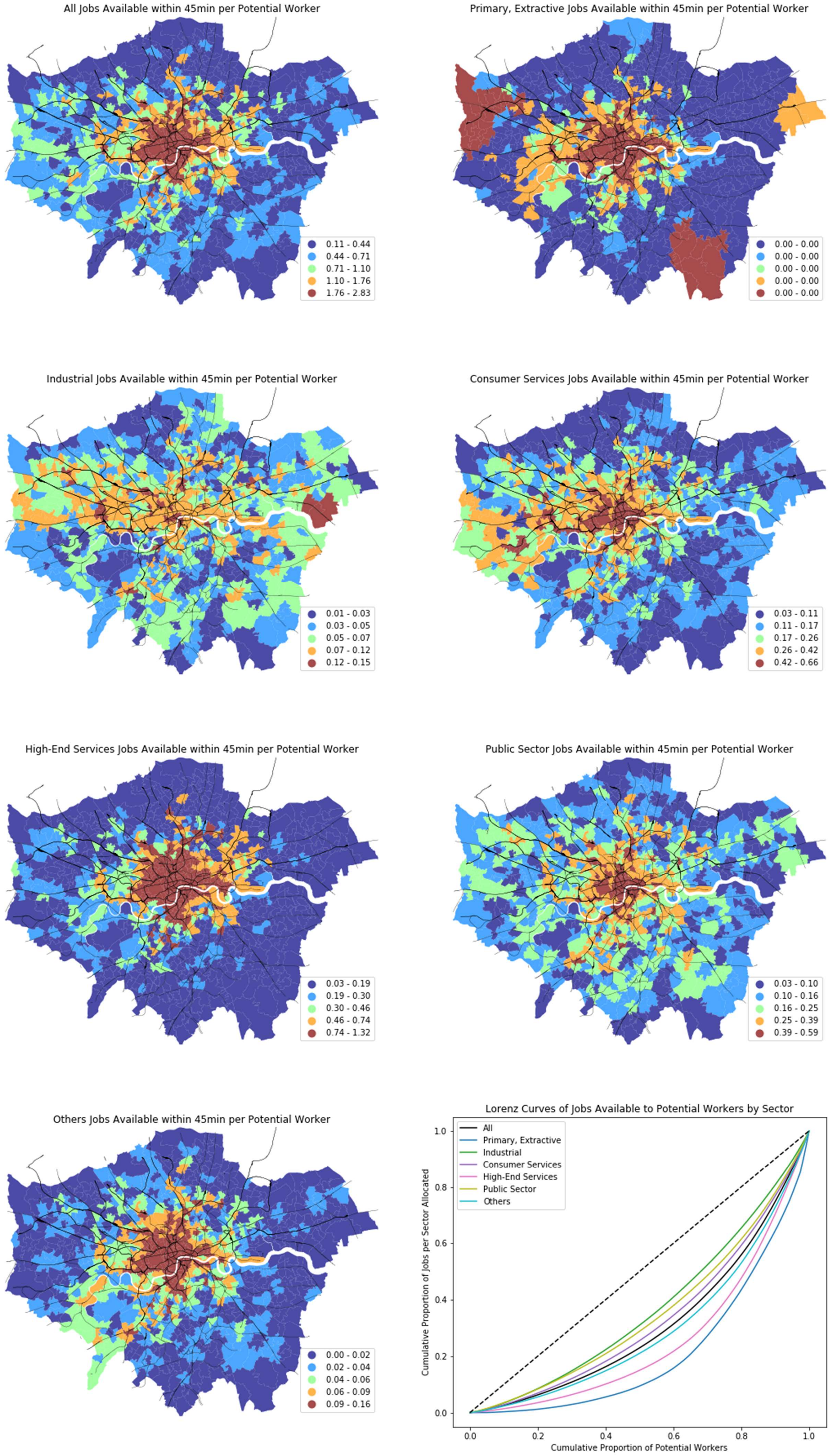
Figure 8: Spatial Distribution of Jobs by Sector
1 dot per 500



The top left panel of Figure 9 reproduces the map of the JPRs for all jobs from Figure 7 for comparison. Each of the other 6 maps in Figure 9 shows the JPRs for each sector and MSOA $r_{s,o}$. Each map uses the same binning scheme as the top left panel, except that the values of R_0 are now sector-specific. From the maps, we can see that the availability of industrial, public sector and consumer services jobs across all potential workers is relatively more balanced. In contrast, the availability of high-end services jobs is very unbalanced, with almost no jobs from this sector available to potential workers living in large swaths of MSOAs in the south, east and north, while the number of jobs from this sector that is available per potential worker living in MSOAs in the centre is between 200% and 360% as high as the GLA average.

When we plot Lorenz curves of the sectoral job distributions to all potential workers (lower right panel of Figure 9), we find that the curves for the industrial, consumer services and public sectors are closer to the diagonal than the curve for all jobs, but the curves for the primary & extractive, high-end services and other sectors are below the curve for all jobs. This indicates that the distributions of jobs in the latter three sectors tend to make London's overall job distribution more unbalanced. Particularly, because the high-end services sector is London's largest, the high concentration of jobs

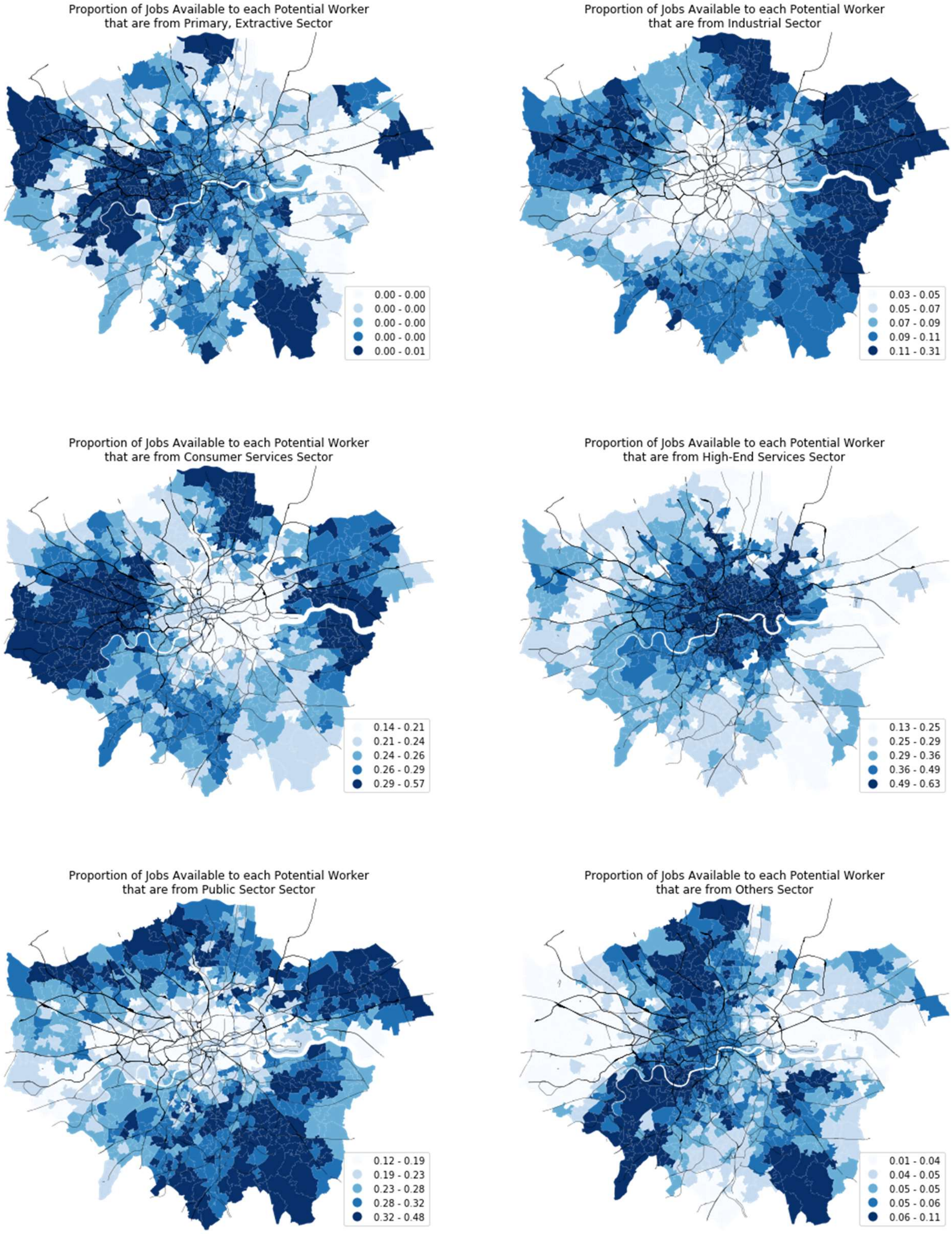
Figure 9: Sectoral Job Provision Ratios



in this sector in the centre of London tends to drive the imbalance in accessibility to jobs in London.

Finally, we can decompose each MSOA's JPR sectorally and analyse the extent to which potential workers in each MSOA are dependent on each sector for job availability. Each map in Figure 10 visualises $\frac{r_{s,o}}{r_o}$ for a particular sector s, with the MSOAs binned into quintiles for visualisation in each map. Areas near the outskirts in the northwest, north and east are especially reliant on industrial jobs, areas near the outskirts in the north, east and west (particularly around Heathrow) are unusually reliant on consumer services jobs, and two swaths along the northern boundary of London and south of the Thames are particularly reliant on public sector jobs. In contrast, the most centrally located areas are disproportionately reliant on the high-end services sector.

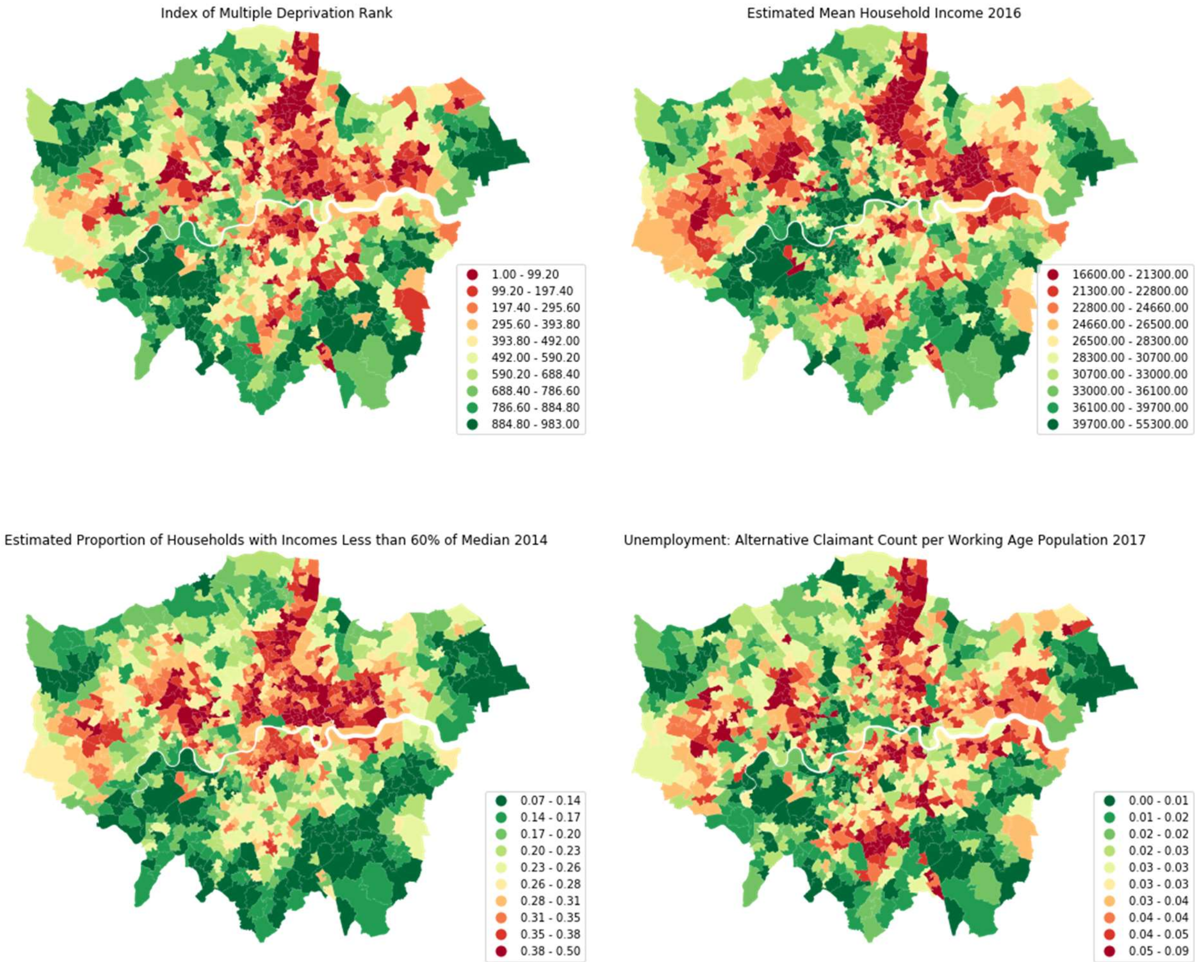
Figure 10: Potential Workers' Sectoral Dependence for Job Availability



Sociodemographic Differences among Potential Workers from Different MSOAs

Potential workers living in different areas of London have different sociodemographic characteristics. MSOA-level variations in these characteristics are reflected in the IMD data. The top left panel of Figure 11 visualises the ranking of each MSOA by IMD scores, with MSOAs binned into deciles so that the 10% of MSOAs that are most (least) deprived are visualised in dark red (green). We can see a swath of more deprived MSOAs running north-south along the Lea Valley from Enfield to Limehouse, another swath running from Dagenham to Clerkenwell along the northern bank of the Thames, and other significant clusters around Shepherd's Bush, Elephant & Castle and Croydon.

Figure 11: Social Deprivation of Potential Workers per MSOA



To verify that the IMD ranks effectively capture variations in social deprivation along multiple dimensions, we can compare the distributions of the ranks and 3 other employment-related indicators of social disadvantage, namely the estimated mean household income in 2016 (post-housing costs and equivalised for household size), the estimated proportion of households in poverty in 2014, and the unemployment rate among working-age residents in 2017. These are visualised in the other panels of Figure 11 using the same binning scheme as the top left panel. In general, the spatial distributions of IMD rankings and the 3 raw indicators of social disadvantage are quite well correlated (absolute Spearman rank correlation coefficients are between 0.8 and 0.9).

Travel Time Reliability

For each trip originating in all 983 MSOAs and ending in one of 7 destinations, I obtained minimum travel times over 50 days. Then, I use the trip-level distribution of travel times to calculate trip-level RBTs, which indicate the degree to which travel times for each trip are reliable. Each panel in Figure 12 plots the RBTs of all trips terminating at one of the 7 destinations (outlined in red in each map). To facilitate comparison of RBTs across all trips, all maps use the same binning scheme to visualise the RBTs, with green shades indicating lower RBTs that are under 5min.

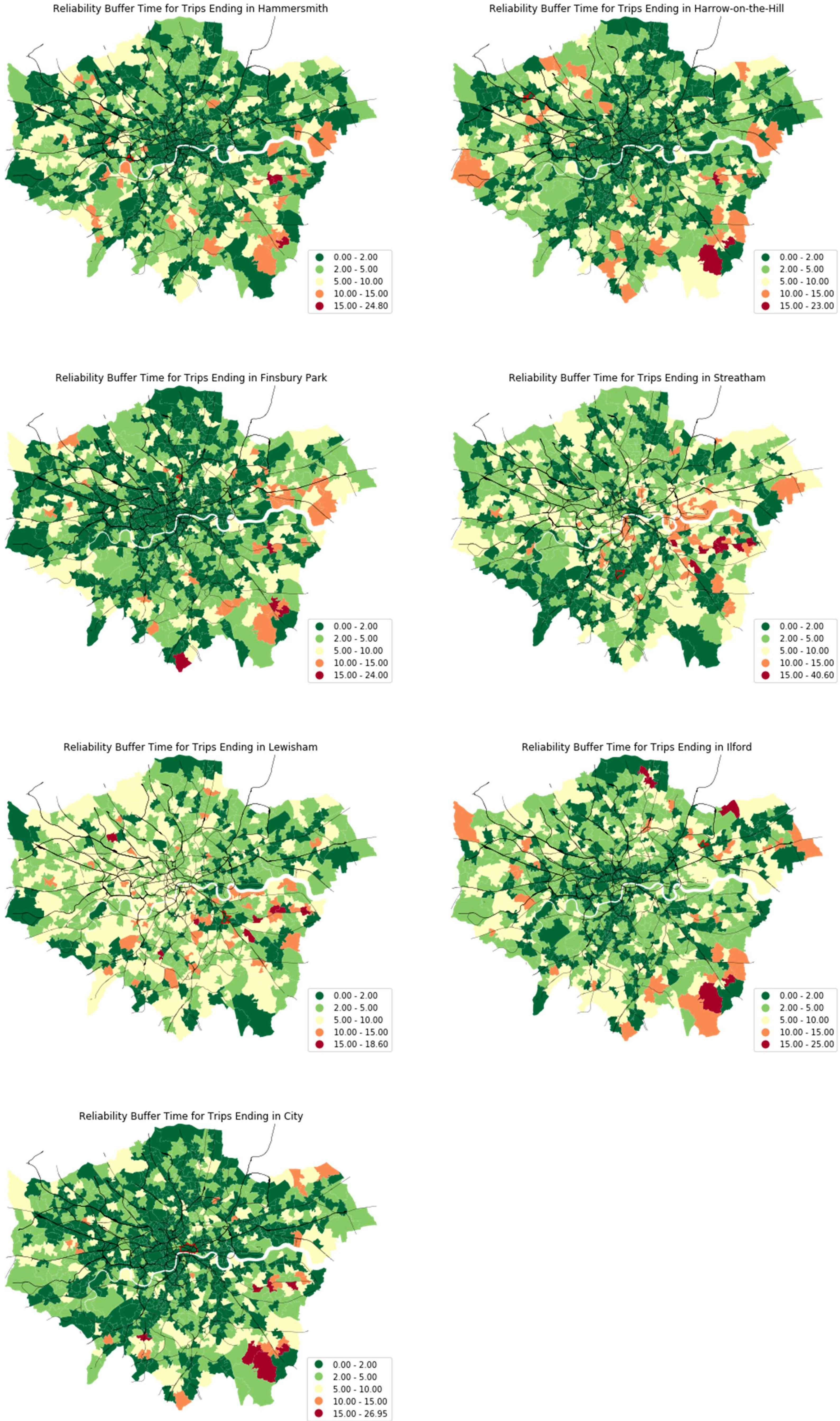
Areas that are directly connected to the destination by rail tend to have lower RBTs, as trips from these areas do not involve transfers, so travel times are only affected by in-vehicle time. Areas that are not directly connected to a destination by rail tend to have higher RBTs, even if they are relatively near the destination. Trips from such areas tend to require more transfers, particularly if bus travel is needed, and this increases the probability of increased waiting times due to schedule desynchronisation between services. Also, trips originating from locations further from a destination tend to have higher RBTs, as these trips tend to involve more transfers too, and long-distance travel tends to rely more on rail-based modes and are thus more likely to be affected by service suspensions.

In addition, trips ending in destinations north of the Thames, particularly in the City but also in Finsbury Park, Harrow-on-the-Hill and Hammersmith, tend to have lower RBTs. This may be due to the more grid-like rail network north of the Thames, which provides commuters with more alternative rail routes if service on a particular section of the rail network is suspended. In contrast, travel times to Lewisham and Streatham tend to be less reliable. This probably reflects two factors: the more tree-like structure of the rail network there (which provides less network resilience against failures along the rail lines leading to these destinations) and reliance on National Rail services to reach these two destinations (these services tend to run less frequently than TfL's rail-based services, thus increasing the likelihood of schedule desynchronisation during transfers). Finally, for all destinations, we can see that trips originating from the southeastern quadrant of the GLA area tend to higher RBTs. This may reflect the effects of the less frequent and less reliable rail services run by Southeastern and Southern Railways.

Factors Shaping Travel Time Reliability

The travel time reliability for a trip seems to depend on the distance covered, the number of transfers, the reliability of different travel modes and location-specific factors. More evidence of these

Figure 12: Reliability Buffer Time for Trips Ending in each of 7 Selected Destinations
Destination of all trips in each map outlined in red



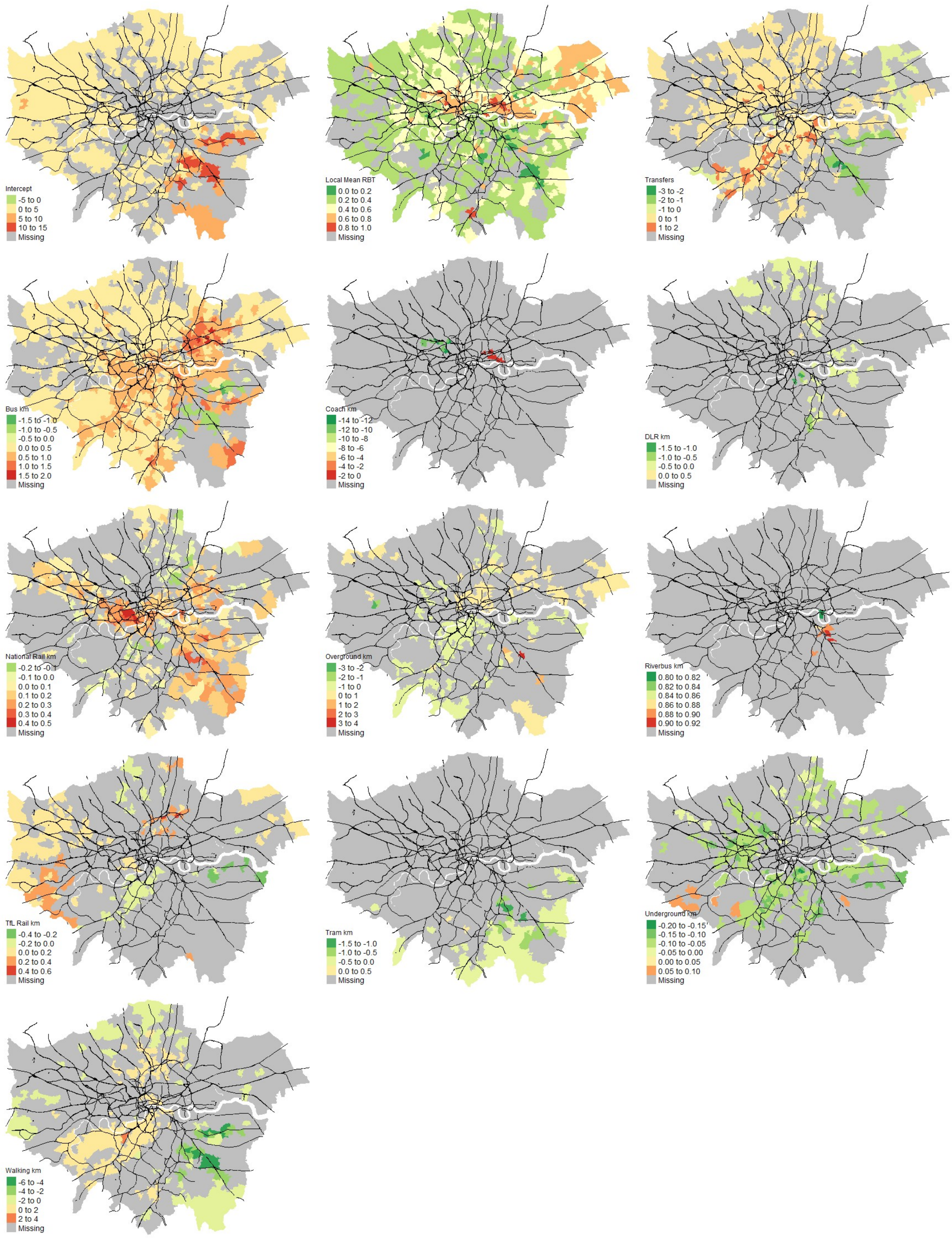
relationships is revealed through a set of linear regression analyses. The results for the global spatially autoregressive model are shown in Column 1 of Table 1. As expected, the local mean RBT captures the correlation between the RBTs of nearby MSOAs that arise due to similarities in how they are connected to the larger transport network. Specifically, as the local mean RBT for a trip increases by one minute, the RBT of the trip in question tends to increase by 0.52min. Also, every additional transfer adds 0.29min to the RBT, reflecting the increased probability of schedule desynchronisation between services during transfers. The mode-specific effects are also as expected, with every additional kilometre travelled by bus or National Rail adding 0.25min and 0.03min to the RBT respectively. This reflects the effects of lower operating frequencies for these two modes, as well as lower speeds for buses and the impacts of service suspensions for National Rail. In contrast, additional travel by Underground, DLR or tram tends to reduce the RBT, indicating that travel times by these modes tend to be more reliable.

Table 1: Regression Results
Dependent Variable: Trip-Level RBT

Model	**1**			**2**				
				Distribution of Estimates (Percentiles)				
	Estimate	**[95% CI]**		**0**	**25**	**50**	**75**	**100**
(Intercept)	1.330	[1.073,	1.587]	-2.414	0.616	1.474	2.440	14.230
Local Mean RBT	0.521	[0.487,	0.556]	-0.117	0.225	0.336	0.460	0.959
# Transfers	0.285	[0.204,	0.366]	-2.270	0.085	0.296	0.505	1.519
Distance by mode (km)								
Bus	0.253	[0.215,	0.290]	-1.009	0.221	0.348	0.520	1.863
Coach	-0.541	[-2.161,	1.079]	-12.458	-10.794	-1.110	-0.734	1.350
DLR	-0.070	[-0.106,	-0.034]	-1.231	-0.142	-0.065	0	0.810
National Rail	0.025	[0.014,	0.036]	-0.122	-0.016	0.029	0.095	0.425
Overground	0	[-0.031,	0.031]	-2.051	-0.109	-0.017	0.085	3.501
Riverbus	0.276	[-0.458,	1.010]	-0.310	-0.203	0.501	0.729	0.903
TfL Rail	0.012	[-0.008,	0.033]	-0.311	-0.044	0.007	0.061	0.536
Tram	-0.217	[-0.318,	-0.116]	-2.808	-0.353	-0.195	0.092	6.637
Underground	-0.024	[-0.033,	-0.014]	-0.153	-0.051	-0.028	-0.008	0.090
Walking	-0.113	[-0.274,	0.047]	-5.627	-0.347	0.098	0.457	2.130
Adj. R^2	0.234			0.049	0.166	0.234	0.379	0.689

Next, we examine the results for the locally weighted regression. The relevant diagnostics are in Appendix 5. Column 2 in Table 1 displays a summary of the distributions of the coefficient estimates for each covariate across all MSOAs. Figure 13 maps out the coefficient estimates for each covariate and each MSOA. Green (red) shades indicate that the covariate tends to increase (decrease) travel time reliability for trips originating in that MSOA. Grey indicates where the MSOA-specific coefficient

Figure 13: Locally Weighted Regression MSOA-Level Coefficient Estimates
Coefficient estimates are missing if they are not statistically different from zero or if there was no variation in the covariate among MSOAs considered for the LWR



estimates were not statistically significantly different from 0 or were not available (because there was no variation in that covariate among the trips considered for the local regression for that MSOA).

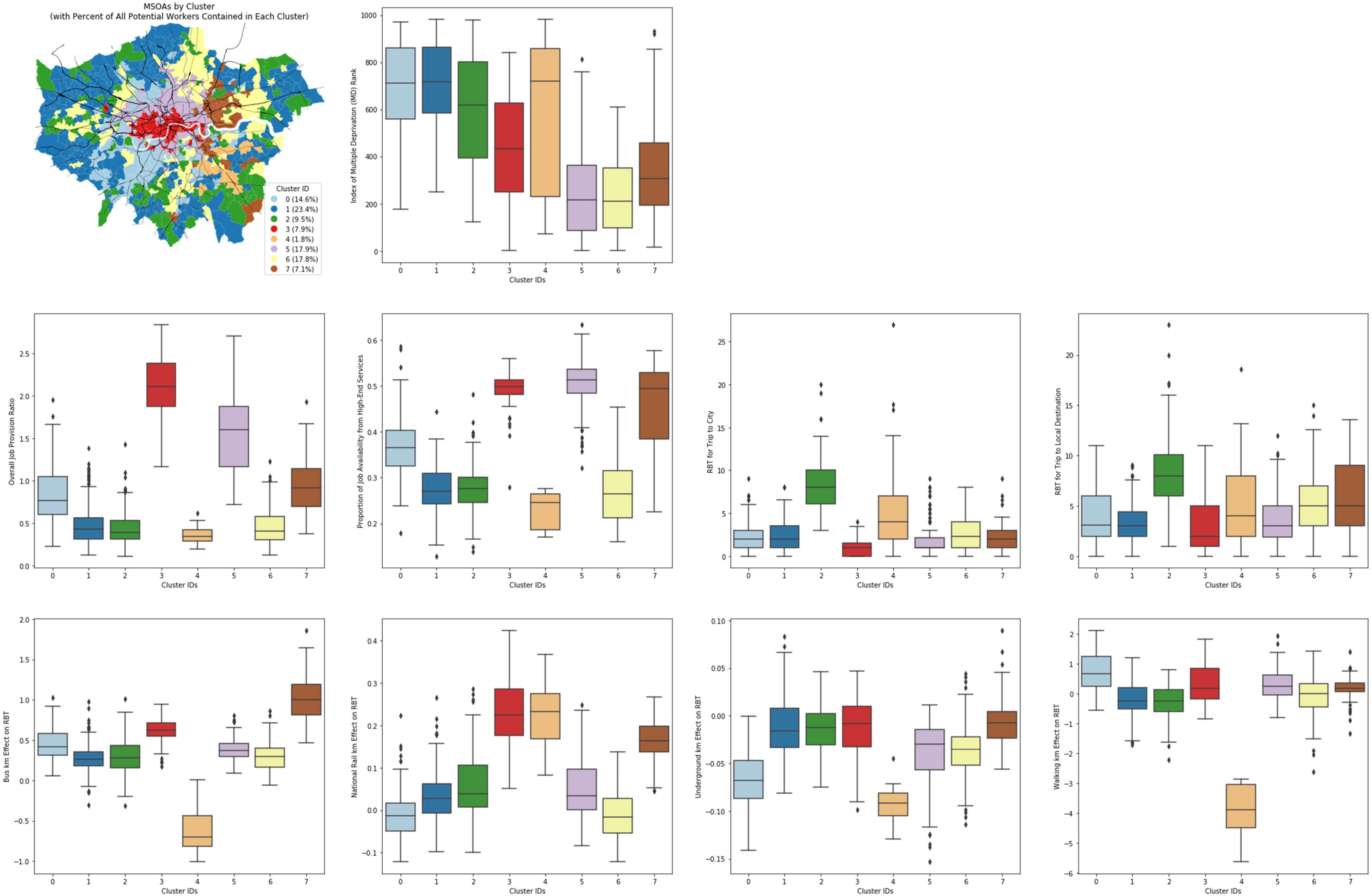
At the local level, increasing travel by bus tends to decrease travel time reliability almost everywhere in London, but particularly in the northeast around Stratford, Barking and Leytonstone, and around the City, Westminster, New Cross, Lewisham, Sidcup, and in a swath from Waterloo to Wimbledon. Travel by National Rail degrades travel time reliability in a large swath in the southeast, indicating further evidence of operational unreliability in Southern and Southeastern services. In fact, National Rail services in these areas seem to be so bad that increasing the number of transfers and increasing travel by bus and walking (and thus reducing reliance on National Rail) tend to improve travel time reliability, contrary to patterns elsewhere. However, travel by National Rail also reduces travel time reliability in a northwestern corridor from Paddington to Ruislip, and around Stratford. This could reflect how most National Rail lines north of the Thames terminate outside the City (in stations like Paddington, Marylebone, Euston, Kings Cross, St. Pancras and Liverpool Street) and do not provide direct orbital connections, so National Rail users need to transfer more to complete their journeys to one of the 7 destinations. On the other hand, the local effects of travel by Underground, tram, Overground and DLR on travel time reliability tend to be positive or not statistically significant.

Typology of London MSOAs by Multidimensional Job Accessibility

Finally, I carry out a k-means clustering analysis to synthesise the findings above into a typology of neighbourhoods in London. The clustering results are stable across different initialisation conditions for the algorithm [37] (see Appendix 6). The results for a clustering with 8 clusters is shown in Figure 14. By social deprivation, three types of clusters are apparent: Clusters 0, 1 and 2 tend to contain MSOAs with lower deprivation, Clusters 5, 6 and 7 tend to contain MSOAs with higher deprivation, and Clusters 3 and 4 are more diverse.

Cluster 3 (with about 8% of the working-age population) contains the most centrally located MSOAs in the City, Westminster and Southbank. As these areas are the most job-rich, MSOAs in Cluster 3 are distinguished by having the highest JPRs, indicating extreme surpluses of jobs available per potential worker. They are also among the most dependent on the high-end services sector for jobs. Their central location gives them access to many public transport routes, especially rail-based routes, resulting in very reliable travel times to the City and to their local destinations. In contrast, the MSOAs in Cluster 4 are distinguished by their combination of job poverty and low travel time reliability, especially to the City. This unreliability is disproportionately due to the unreliability of National Rail services in the southeast. Fortunately, only about 2% of potential workers live in MSOAs in Cluster 4.

Figure 14: K-Means Clustering of MSOAs by JPR, Sectoral Dependence, Travel Time Reliability and Reliability Effects of Travel Modes



Of the three least socially deprived clusters, Cluster 0 (with 15% of the working-age population mostly located in the southwest in areas like Strawberry Hill, Richmond, Wimbledon and Battersea) tends to have the most balanced JPRs. These potential workers tend to rely more on high-end services for their job availability, but because of their low social disadvantage, they may be more likely to have the skills and resources to take advantage of the available jobs in this sector. Travel times to the City and to local destinations tend to be quite reliable, and indeed this area seems to enjoy the most reliable Underground and National Rail services in London. MSOAs in Cluster 1 (with 23% of the working-age population and covering areas near the outskirts) also enjoy relatively reliable travel times to the City and local destinations, but experience job shortages due to their greater distance from the centre and the scarcity of jobs in the outskirts. In contrast, MSOAs in Cluster 2 (with 10% of the working-age population located near the outskirts) experience both insufficient job availability and high travel time unreliability, particularly due to National Rail services.

The situations facing more socially deprived residents are also quite diverse. MSOAs in Cluster 5 (with 18% of the working-age population) are located quite centrally, almost surrounding the MSOAs in Cluster 3. As a result, despite being home to more socially deprived workers, these MSOAs enjoy similar conditions to those in Cluster 3, benefitting from excellent transport connectivity, high travel time reliability and a surplus of jobs available per potential worker. However, these jobs tend to be from the high-end services sector, increasing the risk of sociodemographic mismatch between the potential workers and the available jobs. The extent to which the sheer volume of jobs available can offset the tendency for sociodemographic mismatch with these workers is unclear from this study.

MSOAs in Cluster 6 (containing 18% of the working-age population) are located further towards the outskirts, especially in the Lea Valley in the north, areas from Ealing to Hayes along the TfL Rail line in the west, areas along the Overground line from Sydenham to Croydon in the south, and Becontree in the east. They are characterised by relatively reliable travel times but low job availability. Travel seems to be reliable but too slow to connect these areas effectively to large job clusters. In contrast, MSOAs in Cluster 7 (containing 7% of the working-age population in Stratford, Ilford, Barking, the Royal Docks and Greenwich) have relatively balanced JPRs and reliable travel times to the City, but less reliable travel times to local destinations. The reliability of buses and National Rail in these areas seems especially problematic. These areas are also more dependent on the high-end services sector for jobs. Thus, these areas may also be more vulnerable to a sociodemographic mismatch between potential workers and the available jobs.

In summary, the k-means clustering analysis indicates that accessibility problems in London come in diverse shapes and sizes, as different areas experience distinct combinations of insufficient job availability, dependence on the high-end sector for jobs, heightened social deprivation among potential workers and/or travel time unreliability. The particular combination of challenges in each area suggests that intervention strategies should be tailored to each area's characteristics.

Discussion

Recommendations for Urban Developments in the GLA

The findings from this study can aid in developing multifaceted strategies to improve accessibility to suitable jobs in London, drawing on options from land use development, transport infrastructure and operations and economic and social policy. Firstly, the map of overall JPRs in Figure 7 reveals locations where there are too few or too many jobs available per potential worker. This can directly guide the distribution of future growth to shape the overall urban structure, with new housing supply being directed to locations with job surpluses, and job growth being encouraged in areas with too few jobs available per potential worker. Transport connectivity between nearby areas with very different JPRs can also be enhanced, particularly through bus priority schemes, to bridge local transport gaps that separate areas with many potential workers from areas with many jobs available.

Secondly, even though the measurement of travel time reliability in this study probably underestimates actual travel time variability, the results are still useful for identifying areas where travel times to important destinations like the City are relatively unreliable. The locally weighted regression results also indicate areas where different travel modes seem to be particularly unreliable. These findings can help to target efforts to improve public transport operations.

Thirdly, the typologies of MSOAs derived from the k-means clustering analysis can be used to target social and economic policies to reduce the potential for sociodemographic mismatch between potential workers and the available jobs. For instance, referring to Figure 14, the potential workers of MSOAs in Clusters 5 and 7 may especially benefit from skills training programmes, which can help bridge the gaps between their relatively high social deprivation and dependence on high-end services jobs. Economic incentives or development strategies to encourage job growth in other sectors, especially consumer services and the public sector, can also be targeted to MSOAs in these clusters. The MSOA typology can also be used to prioritise interventions to benefit the most socially deprived areas first, especially since more socially deprived residents may have fewer resources to overcome accessibility barriers themselves and may have fewer options for travel besides public transport.

To a certain extent, these recommendations are in line with what the GLA already plans to do to improve transport connectivity and distribute growth more equitably throughout London. For instance, the upcoming Elizabeth Line is expected to place many more workers in the east and west of London within 45min travel time of the job-rich centre, and this should help to move the JPRs in areas along the route towards a more balanced state. A similar effect can also be expected from the proposed Crossrail 2 project, which runs from the north through the socially deprived and job-poor Lea Valley to the southwest [38]. The Mayor’s Transport Strategy also calls for improvements to bus connectivity and reliability [12, pp. 155-159], improvements to rail reliability south of the Thames (particularly through the conversion of National Rail services to Overground services through "metroisation") [12, pp. 180-181], and the encouragement of housing and job growth in areas with good connections by public transport [12, p. 214]. The findings from this study can strengthen the justifications for these strategies.

The findings can also guide the prioritisation of some of these strategies. For instance, the leg of Crossrail 2 that runs north of the Thames benefits more deprived areas than the southern leg; this should justify building the northern leg first. Also, the Transport Strategy [12] recommends encouraging job growth in four strategic transport interchanges, Lewisham, Stratford, Clapham Junction and Willesden Junction, to put more jobs within easy reach of workers. This is a sensible strategy, as these four locations are surrounded by areas with low JPRs (from Figure 7). However, the number of potential workers that can reach Lewisham within 45min travel (shown in Figure 5) is relatively low, and the RBTs for travel to Lewisham (shown in Figure 12) are quite high; this suggests that travel speeds and reliability to Lewisham should be improved in tandem with the delivery of more jobs there. This strategy of dovetailing transport improvements and development is probably applicable to the other 3 locations as well.

Overall, the results paint a picture of job accessibility in London that is imperfect but hopeful. Problem areas do exist where high social deprivation, low job accessibility, high sectoral dependence and low travel time reliability coincide to varying degrees, but it is important to recognise that there are also areas where more disadvantaged residents enjoy job accessibility conditions that are similar to their more privileged neighbours. In addition, TfL, the GLA and the UK government have recognised the importance of job accessibility to enhancing residents' quality of life, and have demonstrated a commitment to improving the situation especially for more socially deprived areas, through measures that have already been implemented or plans that they are already pursuing. This study affirms the efforts made so far, and provides some guidance for how to tailor ongoing efforts to more effectively address circumstances that are particular to different locations.

A Graph-Theoretic Interpretation of the 2SFCA Method

Since Luo & Wang's [5] seminal description of the 2SFCA method, the indicator has been invariably expressed as two summations, first iterating over all origins o with respect to each destination d (to find out which origins have catchments that include d), then iterating over all d with respect to each o (to sum up the per-worker job allocations from all d within o's catchment). However, the method can be equivalently expressed using matrix operations involving 3 components:

1. J, a column vector of length n storing the spatial distribution of jobs:

$$J = \begin{pmatrix} j_1 \\ j_2 \\ \vdots \\ j_d \\ \vdots \\ j_n \end{pmatrix}$$

2. W, a column vector of length m storing the spatial distribution of potential workers:

$$W = \begin{pmatrix} w_1 \\ w_2 \\ \vdots \\ w_o \\ \vdots \\ w_m \end{pmatrix}$$

3. C, a m by n matrix reflecting the ease of travel between any origin o and destination d:

$$C = \begin{pmatrix} f(c_{1,1}) & f(c_{1,2}) & \dots & f(c_{1,n}) \\ f(c_{2,1}) & \ddots & & \vdots \\ \vdots & & f(c_{o,d}) & \\ f(c_{m,1}) & \dots & & f(c_{m,n}) \end{pmatrix}$$

Then, the accessibility value r_o for each origin o is stored in a column vector R of length m, which is obtained by:

$$R = \begin{pmatrix} r_1 \\ r_2 \\ \vdots \\ r_o \\ \vdots \\ r_m \end{pmatrix} = C \cdot [J \oslash (C' \cdot W)]$$

where $C' \cdot W$ is equivalent to carrying out $\sum_{o=1}^{m} w_o f(c_{o,d})$ across all d, $J \oslash (C' \cdot W)$ is element-wise Hadamard division that yields $\frac{j_d}{\sum_{o=1}^{m} w_o f(c_{o,d})}$ across all d, and $C \cdot [J \oslash (C' \cdot W)]$ yields $\sum_{d=1}^{n} \frac{j_d}{\sum_{o=1}^{m} w_o f(c_{o,d})} f(c_{o,d})$ across all o.

A matrix implementation of the 2SFCA method has computational efficiency advantages. But more importantly, it brings out an interesting but understudied property of the method: it encodes the travel costs as an incidence matrix of a bipartite graph. Specifically, in the matrix C, we can interpret the origins o (in the rows) and the destinations d (in the columns) as two types of nodes, and each value of $f(c_{o,d})$ as an edge weight encoding the strength of the link between any o-d pair. Clearly, there are no edges between nodes of the same type (o-o or d-d).

Three generalisations of 2SFCA can be immediately derived from this observation. First, the nodes in the network need not be spatial units per se but can be objects of any sort which may or may not have spatial properties. In the context of this study, even though 2SFCA is "location-based", we should think of o and d not as locations with workers and jobs respectively, but as workers and jobs with locational coordinates. It then becomes straightforward to expand J, W and C to disaggregate the jobs and worker in each location by type. Secondly, the distances encoded in $c_{o,d}$ need not be spatial, and can incorporate non-spatial distances like dissimilarity indices. Finally, the weight on any edge can be set independently from all other edge weights. This means that each origin-destination pair can potentially use its own distance decay function, opening up vast flexibility for researchers to craft the relationships between origins and destinations.

All this implies that the 2SFCA method is considerably more flexible and widely applicable than currently demonstrated in the literature. One straightforward extension is to incorporate non-spatial distances directly into the incidence matrix C. For instance, $c_{o,d}$ can be the product of the scaled travel time between o and d and the scaled absolute difference between the mean income of the workers living in o and the mean wages of the jobs in d. In this way, the values in C and ultimately the values in R can be affected by both spatial proximity and sociodemographic matching, potentially capturing the idea that the availability of jobs per worker depends on both how distant the worker is from the jobs and how closely the worker's characteristics match those needed by the jobs [8, 9]. More generally, 2SFCA may be applicable to measure resource availability to satisfy demand in partially or even entirely non-spatial settings, such as funding availability from venture capital firms to start-ups (where the distances could measure the degree of congruence between the start-ups' economic sector and those historically funded by each firm), or the availability of positions for PhD candidates in different academic departments (where the distances can reflect the alignment between departments' and candidates' research interests).

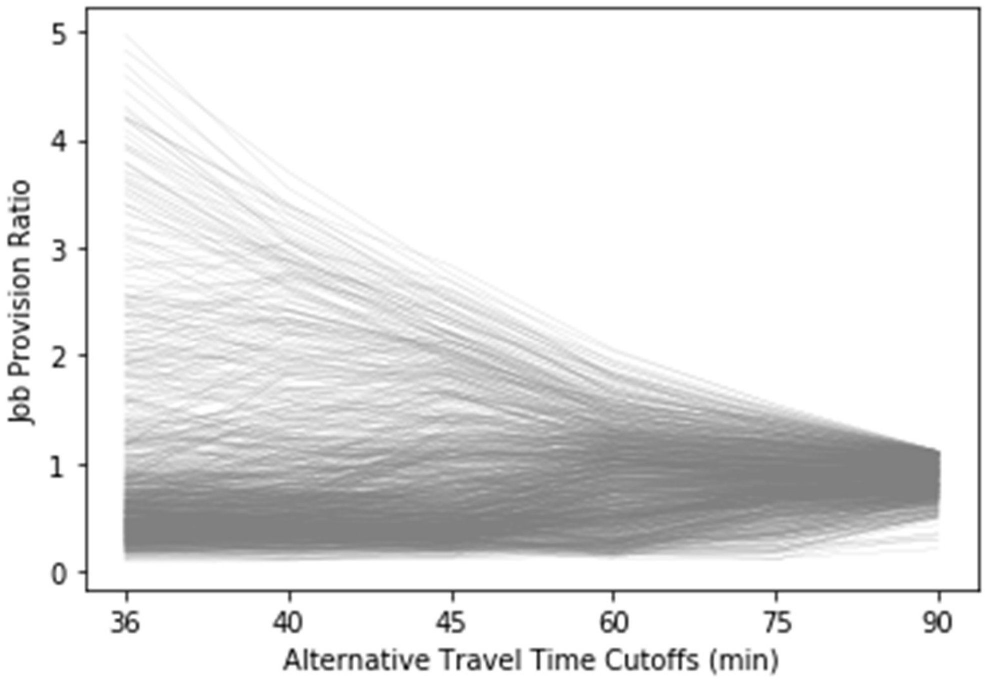
Further, this suggests that methods developed for bipartite graphs are also applicable to accessibility studies. A one-mode projection of the bipartite graph could be used to reveal which workers are supplied with jobs from the same destinations (and are thus competitors) or which destinations supply jobs to workers in the same origins (and thus are members of the same local job market). Algorithms that can detect communities in weighted bipartite networks could be used to identify groups of workers that are similarly reliant on jobs from particular destinations or sectors. We could also go beyond measuring resource availability to assessing resource assignment. For instance, given the set of workers and jobs and the hybrid spatial-sociodemographic distance between each worker-job pair from above, bipartite graph matching algorithms exist [39] that could find a scheme that matches each worker to exactly one job so as to minimise the sum of the job-worker distances over all the pairs thus formed. Planning and policy-making in the urban, social and economic fields can only influence job availability and cannot dictate job assignment, but it may nonetheless be interesting to analyse how the edge weights of the bipartite graph affect the total sum of distances in the optimal matching scheme, which could give some insights into which pairwise distances should be reduced (such as by reducing travel times or increasing sociodemographic matching) in order to make a better global optimum possible.

However, this is only a brief foray into the possibilities for further developments of 2SFCA from a graph-theoretic perspective. Clearly, there is broad potential for further research to improve and apply the 2SFCA method in domains beyond urban and transport studies, and there are also many opportunities for methods developed for bipartite graphs to yield further insights into the structure of accessibility in urban areas. Researchers in these two fields who pay close attention to one another's work will probably find fertile ground for innovation.

Appendix 1: Sensitivity of Job Provision Ratio Results to Alternative Travel Time Cut-offs

In order to test the extent to which JPR results for each MSOA could be sensitive to the choice of travel time cut-off X, I recalculate JPRs for each MSOA using alternative values of X: 36min (which is the maximum of the shortest inter-MSOA travel times, to guarantee that all potential workers in all origins will be able to reach at least one destination), 40min, 45min, 60min, 75min and 90min. Figure 15 plots one line for each of the 983 MSOAs, tracing how each MSOA's JPR value evolves as X changes over this range. As we can see, as X increases, all MSOAs' JPRs converge towards the global average R_0. This reflects the spatial smoothing effect observed by Chen & Jia [28]: as X increases, local unevenness in accessibility tends to get smoothened out more.

Figure 15: Job Provision Ratio per MSOA with Alternative Travel Time Cut-offs
Each line plots the evolution of JPR values for each MSOA over the range of alternative cut-offs



We can see that some MSOAs' JPRs change dramatically as X changes, especially if an increase in X suddenly puts a very job-rich area within reach. However, most MSOAs' JPR values change slowly and in parallel to the general spatial smoothing trend, and the JPR distributions for different values of X remain quite well correlated (Pearson correlation coefficients are generally above 0.6, see Table 2). This is similar to Chen & Jia's [28] finding that the correlation between two sets of accessibility values calculated with 2SFCA using the same distance decay function but two different values of X decreases as the difference between the two X-values increases. However, the degree of decrease in correlation that I found is not as dramatic as what Chen & Jia found, possibly because the job distribution in London is unusually strongly clustered, and because the inter-area travel times in my study vary across a narrower range than the inter-area travel distances that they use in their study.

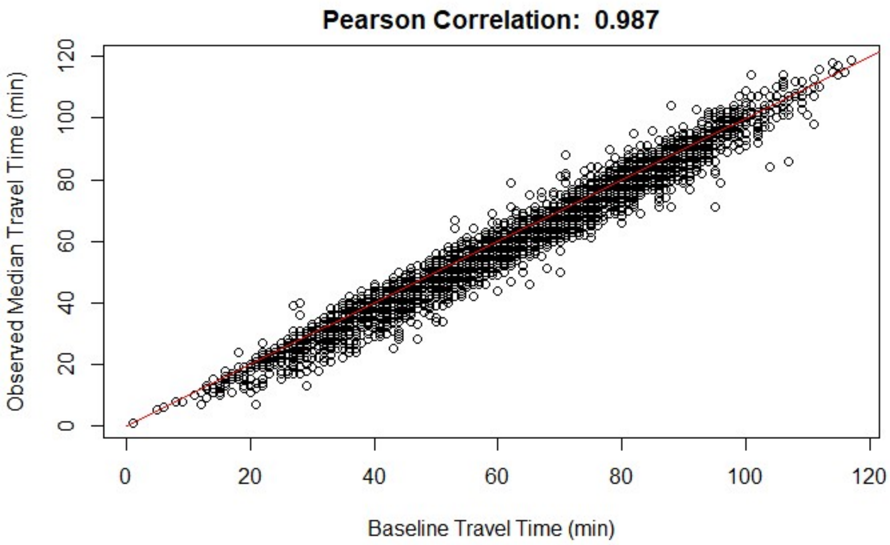
Table 2: Pearson Correlation Between JPR Distributions using Alternative Travel Time Cut-Offs

Cut-Offs (min)	36	40	45	60	75	90
36	1.00	0.95	0.86	0.70	0.67	0.57
40	0.95	1.00	0.94	0.78	0.75	0.66
45	0.86	0.94	1.00	0.87	0.83	0.76
60	0.70	0.78	0.87	1.00	0.95	0.92
75	0.67	0.75	0.83	0.95	1.00	0.96
90	0.57	0.66	0.76	0.92	0.96	1.00

Appendix 2: Representativeness of Actual Travel Times by Baseline Travel Times

For 6,881 trips, I obtained both the baseline travel time (for a trip ending on 0830 on 22 March 2019) and a distribution of actual travel times (for a trip starting at a random time between 0600 and 0800) over 50 days. Thus, I can compare the baseline and actual median travel times for these trips, to assess the extent to which the baseline times are representative of the actual distribution of travel times. From Figure 16, it is clear that the two quantities are highly correlated, although there is a slight systematic tendency for baseline times to be more than the actual median times, and there are isolated instances where the baseline time may be significantly more or less than the actual median travel time.

Figure 16: Correlation between Baseline and Actual Median Travel Times for 6,881 Trips
1:1 line indicated in red



Appendix 3: Potential Improvements to Accuracy of Travel Time Data from TfL's Journey Planner

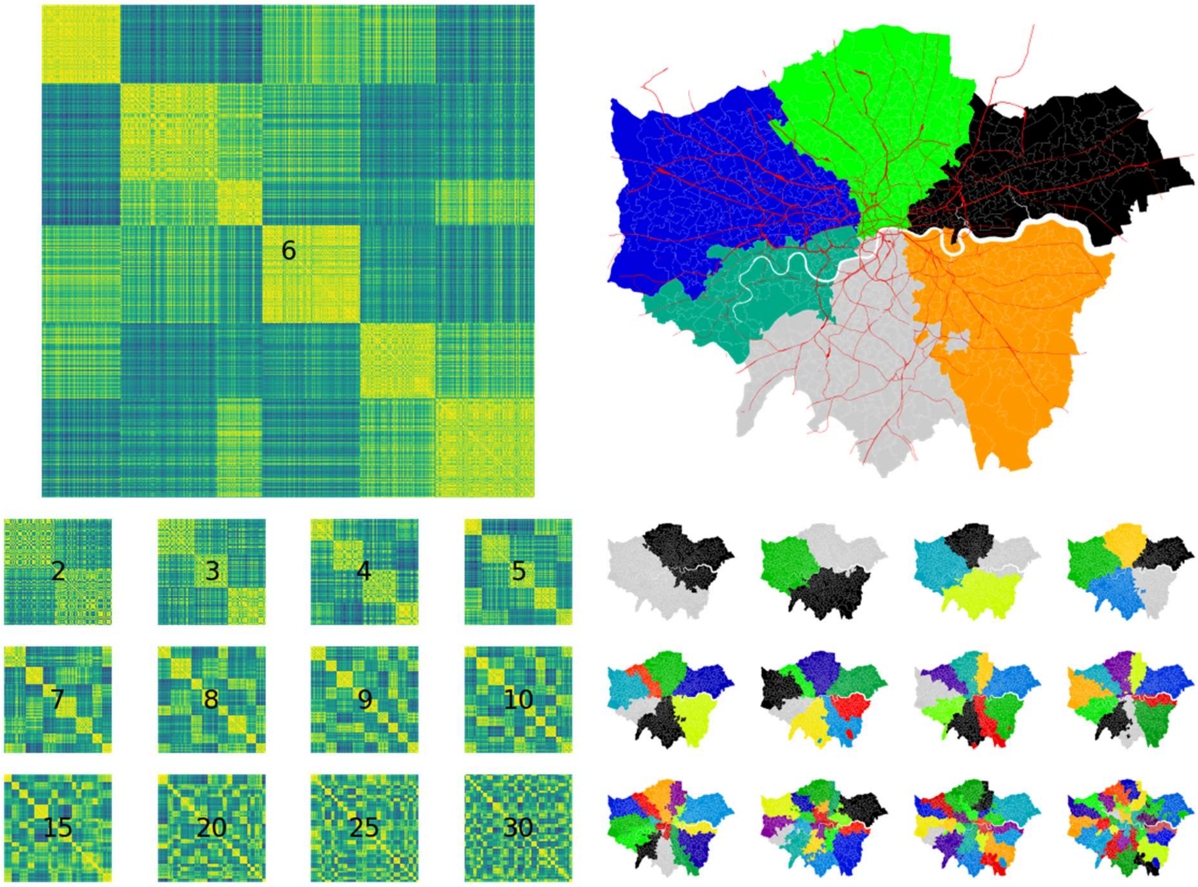
The Journey Planner tends to systematically underestimate the travel time actually needed for a trip, because it does not account for all possible sources of delay during a trip. This can be partially addressed in future studies by combining multiple sources of TfL open data. For example, once the route for a trip is obtained from the Journey Planner, the stations where a traveller needs to board and alight can be identified, and separate scripts can then repeatedly request the live arrivals for each station from a separate API endpoint [31] in order to track the vehicles that the traveller would ride on in real time. Actual waiting and in-vehicle times for each leg of the trip can then be obtained and updated into the database of trip times. This would enable travel times to reflect the effects of en-route congestion that results in slower in-vehicle speeds and delayed arrivals.

Appendix 4: Selection of 7 Key Destinations for Travel Time Reliability Analysis

As part of coursework for ST445 Managing and Visualising Data, I collected an earlier version of the baseline travel time between the population-weighted centroids of each MSOA and the jobs-weighted centroids of every MSOA in the GLA area for trips ending at 0830 on 28 Nov 2018. I then carried out a clustering analysis on the travel time matrix. For each origin, I had a vector of 983 travel times to every destination. Then, for every pair of origins, I calculated the cosine similarity of their travel time vectors, yielding a 983 by 983 matrix of similarity measurements. This captures the extent to which travel times to every destination vary in similar ways between any two origins. I then carried out spectral co-clustering using the similarity matrix, which grouped origins together by how similar their travel time vectors were.

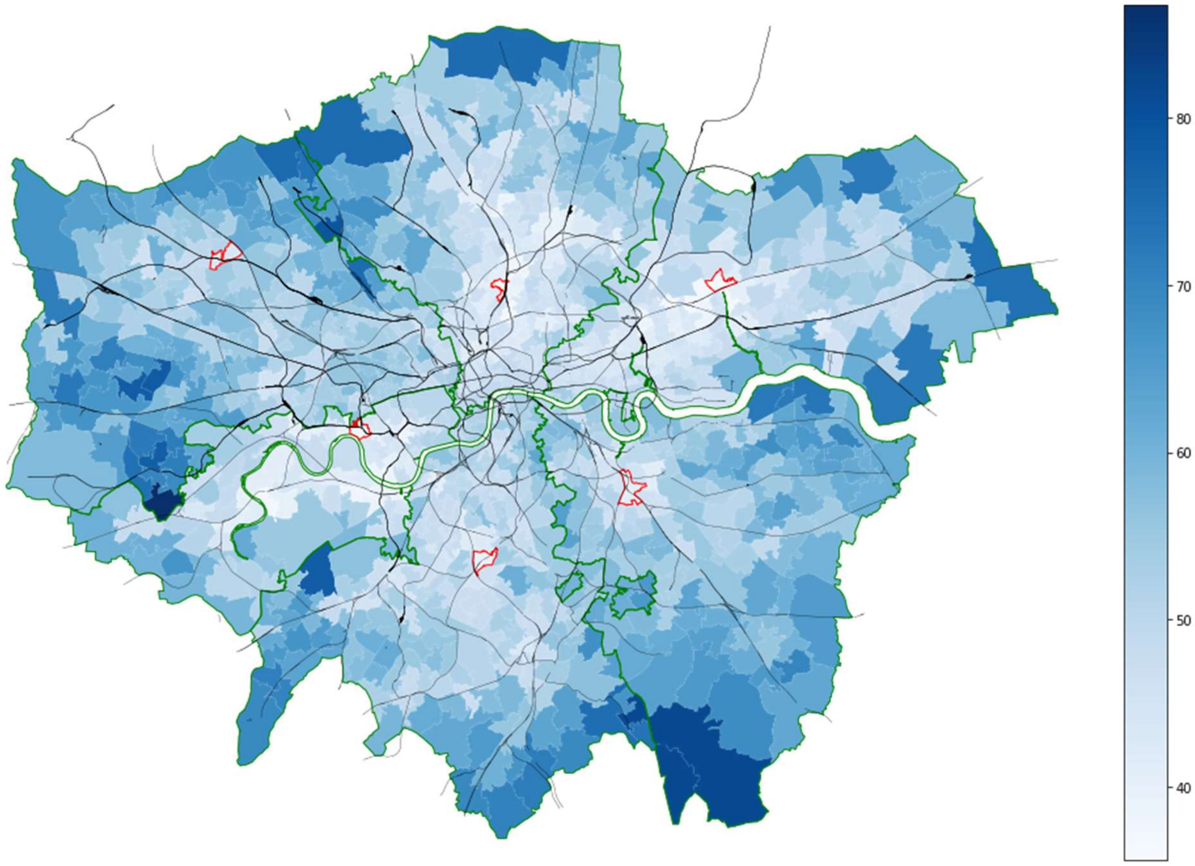
Figure 17 shows the spectral co-clustering results for a range of different numbers of clusters. In the left section, each square displays the similarity matrix after clustering into a certain number of clusters (superimposed on the square). If any given cell (indexed by the two origins o1,o2) has a stronger yellow (green) colour, this indicates that the travel time vectors for o1 and o2 are more (less) similar. In the right section, the clustering is visualised spatially in a corresponding map. We can see that there is a very strong radial pattern in the clustering, persisting up to 20 clusters. This reflects both how MSOAs that are near each other tend to be connected similarly to the larger transport network (and thus to have similar travel times to all destinations), and how the transport network strongly emphasises fast connections to the City. Each cluster resembles a “travel-shed” of areas that rely on key radial rail lines. Overall, a clustering with 6 clusters maximises both high intra-cluster similarity and high inter-cluster dissimilarity.

Figure 17: Spectral Co-Clustering of Origin MSOAs by Cosine Similarity of Travel Times
Squares visualise the matrix of cosine similarity values between pairs of origins after clustering
Labels in squares indicate number of clusters requested
Corresponding maps indicate the locations of MSOAs in each cluster



For this study, I extended this analysis. For each destination MSOA, I calculated the mean time it takes to travel from every other MSOA in the same cluster, weighted by the number of potential workers in the origin MSOAs. I then found the destination MSOA where this weighted mean travel time was minimised. This represent a network centroid which is on average the closest to every potential worker in the same cluster in terms of travel time. Figure 18 shows for every MSOA the weighted mean travel time from every MSOA within the same cluster, with the 6 cluster centroids outlined in red. As it turns out, each of the 6 centroids contain major rail interchanges, providing easy access to that centroid from anywhere else in the cluster. Then, I added the City of London as a 7th destination due to its importance as the most job-rich area in London.

Figure 18: Weighted Mean Travel Time (min) from all MSOAs in the Same Cluster
6 cluster boundaries outlined in green
6 cluster centroids outlined in red



Appendix 5: Diagnostics for Locally Weighted Regression

Figure 19 shows the adjusted R^2 for each regression fitted for each MSOA considering only its 30 closest neighbours. The median adjusted R^2 is around the adjusted R^2 value for the global spatially autoregressive model in Column 1 of Table 1, but ranges from a low of 0.05 in the outskirts to a maximum of 0.69 around Westminster. Apparently, in areas where the rail network is denser, the covariates included in the model account for a significant proportion of the local variance in RBT, but in the outskirts, local variance in RBT is due more to factors not included in the model (such as mean service frequencies at nearby rail stations, or the distance travelled on the services of different operators).

Figure 19: Adjusted R^2 Values for Local Fit of Regression Model

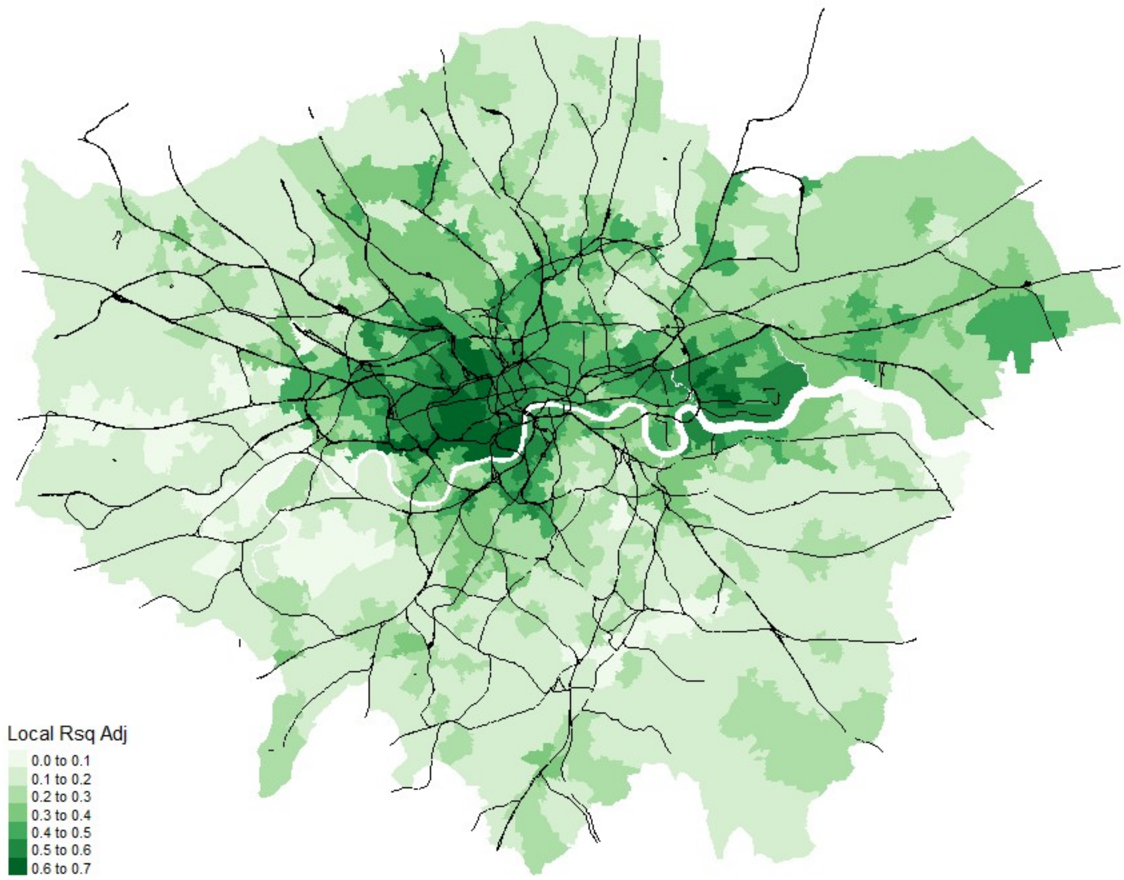
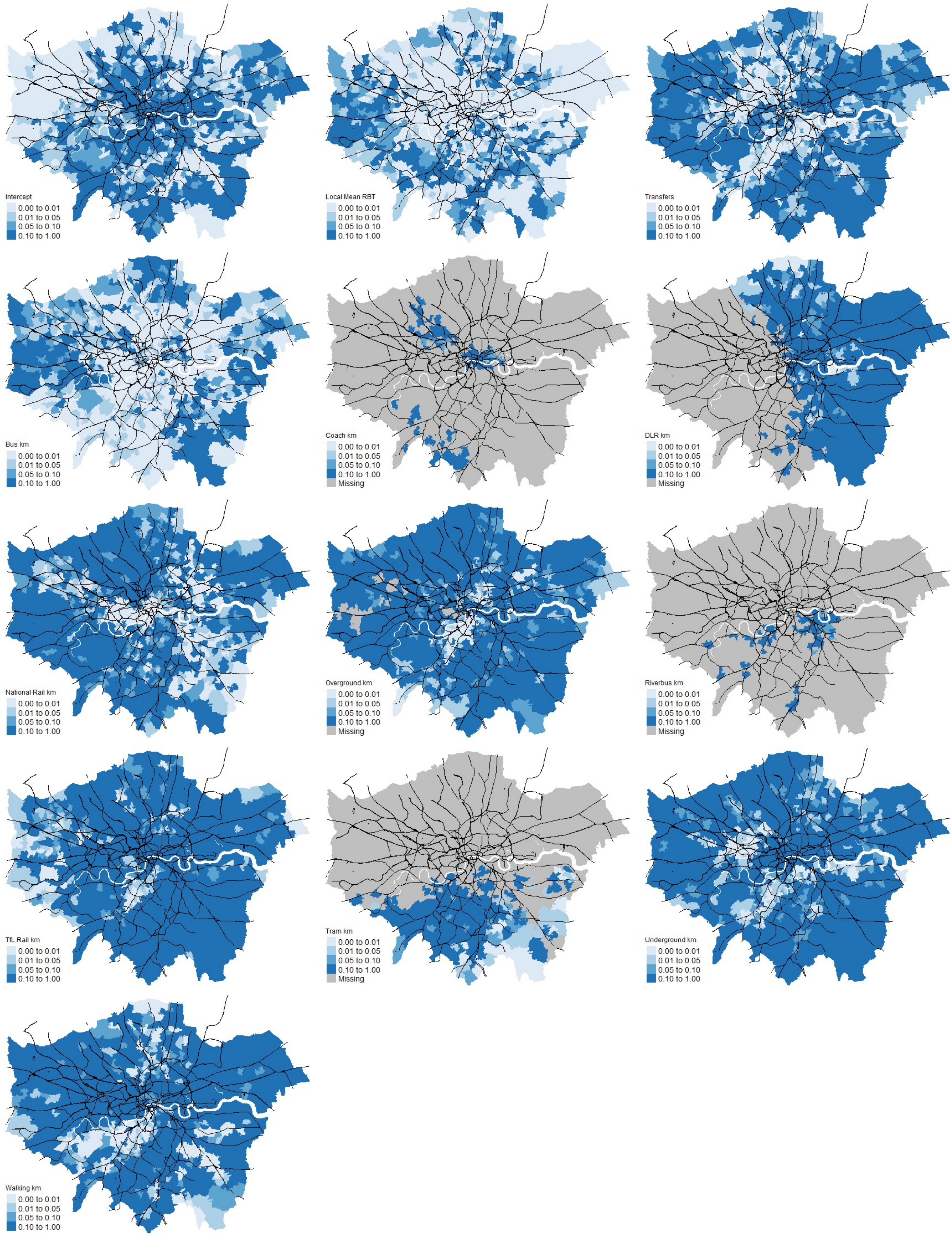


Figure 20 shows the p-values associated with the coefficient estimates for each covariate in each MSOA. Dark blue areas indicate MSOAs where the coefficient estimates for a particular covariate are not statistically significantly different from zero. Grey areas indicate MSOAs where coefficient estimates were unavailable, because there was no variance in the covariate among the MSOAs considered when fitting the model locally. For instance, there are no coefficient estimates for distance travelled by tram for MSOAs north of the Thames, because all trips from these MSOAs to the 7 destinations do not involve any tram travel.

Figure 20: P-values of Coefficient Estimates from Local Fit of Regression Model



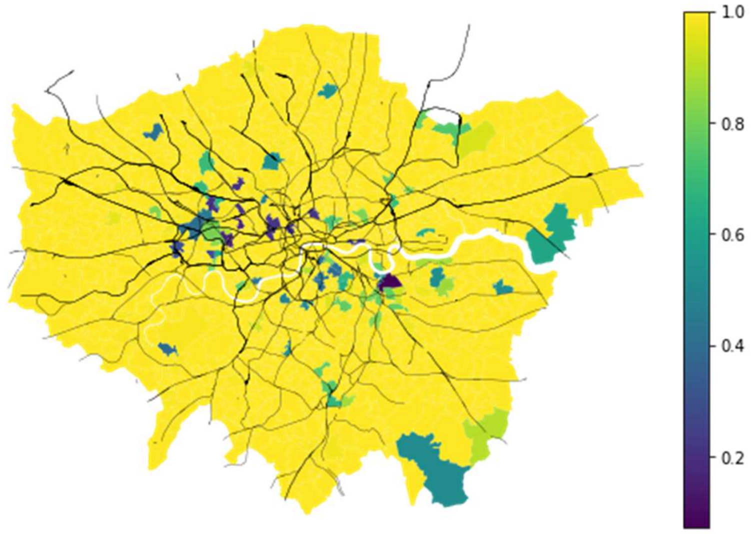
Appendix 6: Stability of K-Means Clustering Results

The clusters identified by k-means clustering can be unstable because the algorithm initialises by randomly selecting an MSOA as an initial cluster centroid, and then selecting other MSOAs as initial centroids to maximise the inter-centroid distances. Depending on how similar each MSOA is to other MSOAs and on which 8 MSOAs are initially selected, each run of the clustering algorithm may group each MSOA into different clusters. Thus, each MSOA's cluster membership and the centroids of the clusters identified could both vary significantly from run to run [37]. If cluster centroid locations and MSOAs' cluster memberships are unstable, this could qualitatively change the characteristics that define each cluster and undermine the validity of interpretations specific to each MSOA.

To assess the extent to which my findings from k-means clustering may be vulnerable to instability, I first identified which 8 MSOAs were closest to the centroids of the 8 clusters identified. Then, I reran the clustering algorithm 1000 times and tracked for each replication whether each MSOA was grouped with the same cluster of MSOAs. For instance, if in the original clustering results, MSOA 1 was one of the 8 MSOAs closest to the centroids of the 8 clusters identified and MSOA 2 was clustered with MSOA 1, I tracked over 1000 replications of the clustering algorithm how often MSOA 2 was clustered with MSOA 1.

Figure 21 shows for each MSOA the proportion of the replications where each MSOA's clustering membership was the same as the original clustering membership. As we can see, most MSOAs do no change clustering memberships, regardless of which 8 MSOAs were initially chosen as cluster centroids.

Figure 21: Stability of MSOA Membership in 8 Clusters
MSOA colours reflect proportion of 1000 replications of the k-means clustering algorithm where MSOA cluster membership is the same as the original k-means clustering results



I also tracked the locations of each of the 8 cluster centroids from each run, and calculated the variance of the coordinates of each cluster centroid for each dimension across all 1000 replications. The variance of each centroid's location along each dimension is shown in Table 3. Recall that the values in all dimensions were standardised to have mean 0 and variance 1 before carrying out k-means clustering. As we can see, the centroid coordinates identified in each dimension over 1000 replications are very stable, with variances the coordinates in each dimension that are no more than 10% of the overall variance in that dimension. Taken together, these results give us confidence that the characteristics of the 8 clusters identified reflect real and distinct types of job accessibility experienced by different areas in London, and that location-specific interpretations of the k-means clustering results are valid.

Table 3: Variance of Coordinates of each Cluster Centroid per Dimension across 1000 Replications

Centroid	Index of Multiple Deprivation (IMD) Rank	Overall Job Provision Ratio	Proportion of Job Availability from High-End Services	RBT for Trip to City	RBT for Trip to Local Destination	Bus km Effect on RBT	National Rail km Effect on RBT	Underground km Effect on RBT	Walking km Effect on RBT
0	0.0107	0.0014	0.0021	0.0001	0.0035	0.0003	0.0020	0.0036	0.0030
1	0.0000	0.0000	0.0001	0.0001	0.0001	0.0001	0.0001	0.0000	0.0002
2	0.0007	0.0000	0.0002	0.0014	0.0010	0.0005	0.0001	0.0001	0.0004
3	0.0010	0.0443	0.0131	0.0038	0.0054	0.0084	0.0415	0.0060	0.0018
4	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
5	0.0002	0.0036	0.0007	0.0007	0.0019	0.0002	0.0007	0.0121	0.0002
6	0.0001	0.0001	0.0002	0.0000	0.0012	0.0001	0.0001	0.0005	0.0001
7	0.0029	0.0094	0.0026	0.0008	0.0145	0.0085	0.0042	0.0012	0.0010

References

[1] L. Bettencourt and G. West, "A unified theory of urban living," *Nature,* vol. 467, pp. 912-913, 10 2010.

[2] W. G. Hansen, "How accessibility shapes land use," *Journal of the American Institute of Planners,* vol. 25, pp. 73-76, 5 1959.

[3] A. Ford, S. Barr, R. Dawson and P. James, "Transport accessibility analysis using GIS: assessing sustainable transport in London," *ISPRS International Journal of Geo-Information,* vol. 4, pp. 124-149, 1 2015.

[4] Y. Shen and M. Batty, "Ripples and undulations in the perceived supply–demand mismatch surfaces of London's job market," *Regional Studies, Regional Science,* vol. 5, pp. 263-266, 1 2018.

[5] W. Luo and F. Wang, "Measures of spatial accessibility to health care in a GIS environment: synthesis and a case study in the Chicago region," *Environment and Planning B: Planning and Design,* vol. 30, pp. 865-884, 12 2003.

[6] K. Lucas, "Transport and social exclusion: where are we now?," *Transport Policy,* vol. 20, pp. 105-113, 3 2012.

[7] F. Wang, "Measurement, optimization, and impact of health care accessibility: a methodological review," *Annals of the Association of American Geographers,* vol. 102, pp. 1104-1112, 9 2012.

[8] J. Cheng and L. Bertolini, "Measuring urban job accessibility with distance decay, competition and diversity," *Journal of Transport Geography,* vol. 30, pp. 100-109, 6 2013.

[9] T.-q. Dai, Z.-b. Liu, C. Liao and H.-y. Cai, "Incorporating job diversity preference into measuring job accessibility," *Cities,* vol. 78, pp. 108-115, 8 2018.

[10] Social Exclusion Unit, Making the connections: final report on transport and social exclusion, London: Office of the Deputy Prime Minister, 2003.

[11] D. L. Uniman, J. Attanucci, R. G. Mishalani and N. H. M. Wilson, "Service reliability measurement using automated fare card data," *Transportation Research Record: Journal of the Transportation Research Board,* vol. 2143, pp. 92-99, 1 2010.

[12] Greater London Authority, Mayor's transport strategy: March 2018, London: Greater London Authority, 2018.

[13] Q. Shen, "Location characteristics of inner-city neighborhoods and employment accessibility of low-wage workers," *Environment and Planning B: Planning and Design,* vol. 25, pp. 345-365, 1998.

[14] T. W. Sanchez, "The connection between public transit and employment: the cases of Portland and Atlanta," *Journal of the American Planning Association,* vol. 65, pp. 284-296, 9 1999.

[15] A. El-Geneidy, D. Levinson, E. Diab, G. Boisjoly, D. Verbich and C. Loong, "The cost of equity: assessing transit accessibility and social disparity using total travel cost," *Transportation Research Part A: Policy and Practice,* vol. 91, pp. 302-316, 9 2016.

[16] M. Neihaus, P. Galilea and R. Hurtubia, "Accessibility and equity: an approach for wider transport project assessment in Chile," *Research in Transportation Economics,* 2016.

[17] L. A. Merlin and L. Hu, "Does competition matter in measures of job accessibility? Explaining employment in Los Angeles," *Journal of Transport Geography,* vol. 64, pp. 77-88, 10 2017.

[18] L. A. Guzman, D. Oviedo and C. Rivera, "Assessing equity in transport accessibility to work and study: the Bogotá region," *Journal of Transport Geography,* vol. 58, pp. 236-246, 1 2017.

[19] L. A. Guzman and D. Oviedo, "Accessibility, affordability and equity: assessing `pro-poor' public transport subsidies in Bogotá," *Transport Policy,* vol. 68, pp. 37-51, 9 2018.

[20] K. T. Geurs and B. Wee, "Accessibility evaluation of land-use and transport strategies: review and research directions," *Journal of Transport Geography,* vol. 12, pp. 127-140, 6 2004.

[21] Department for Housing, Communities and Local Government, "The English Indices of Deprivation 2015: Research Report," Department for Communities and Local Government, London, 2015.

[22] R. Deboosere and A. El-Geneidy, "Evaluating equity and accessibility to jobs by public transport across Canada," *Journal of Transport Geography,* vol. 73, pp. 54-63, 12 2018.

[23] O. Kotavaara, H. Antikainen, M. Marmion and J. Rusanen, "Scale in the effect of accessibility on population change: GIS and a statistical approach to road, air and rail accessibility in Finland, 1990-2008," *The Geographical Journal,* vol. 178, pp. 366-382, 3 2012.

[24] H. Tenkanen, Saarsalmi, Perttu, O. Järv, M. Salonen and T. Toivonen, "Health research needs more comprehensive accessibility measures: integrating time and transport modes from open data," *International Journal of Health Geographics,* vol. 15, no. 1, 2016.

[25] O. Järv, H. Tenkanen, M. Salonen, R. Ahas and T. Toivonen, "Dynamic cities: location-based accessibility modelling as a function of time," *Applied Geography,* 2018.

[26] Y. Hu and J. Downs, "Measuring and visualizing place-based space-time job accessibility," *Journal of Transport Geography,* vol. 74, pp. 278-288, 1 2019.

[27] J. Katz and K. Quealy, "How Unpredictable Is Your Subway Commute? We'll Show You," 2019. [Online]. Available: https://www.nytimes.com/interactive/2019/07/08/upshot/nyc-subway-variability-calculator.html. [Accessed 12 July 2019].

[28] X. Chen and P. Jia, "A comparative analysis of accessibility measures by the two-step floating catchment area (2SFCA) method," *International Journal of Geographical Information Science,* vol. 33, pp. 1739-1758, 3 2019.

[29] Transport for London, Travel in London: report 11, London: Transport for London, 2018.

[30] Transport for London, "Plan a journey," Transport for London, London, 2019.

[31] Transport for London, "Transport for London unified API," Transport for London, London, 2019.

[32] L. Anselin, "Under the hood: issues in the specification and interpretation of spatial regression models," *Agricultural Economics,* vol. 27, pp. 247-267, 11 2002.

[33] J. Micklethwaite, "Southern and Southeastern trains branded worst in the UK by disgruntled commuters," 2016. [Online]. Available: https://www.standard.co.uk/news/london/southern-and-southeastern-branded-worst-rail-networks-in-the-country-by-disgruntled-passengers-a3284666.html. [Accessed 1 August 2019].

[34] J. Nurden, "Southeastern and Southern trains voted among worst services in Britain," 2018. [Online]. Available: https://www.kentonline.co.uk/kent/news/commuters-vent-their-anger-on-158556/. [Accessed 1 August 2019].

[35] E. Wills, "Southeastern passengers' despair after report reveals service will not improve until 2024," 2018. [Online]. Available: https://www.standard.co.uk/news/transport/southeastern-passengers-despair-after-report-reveals-service-will-not-improve-until-2024-a3818876.html. [Accessed 1 August 2019].

[36] S. A. Fotheringham, C. Brundson and M. Charlton, Geographically weighted regression: the analysis of spatially varying relationships, John Wiley & Sons, 2002.

[37] U. Luxburg, "Clustering stability: an overview," *Foundations and Trends in Machine Learning,* vol. 2, pp. 235-274, 2010.

[38] Crossrail 2, "Crossrail 2: supporting growth in the South East," 2019. [Online]. Available: https://crossrail2.co.uk. [Accessed 1 August 2019].

[39] G. Papadogeorgou, C. Choirat and C. M. Zigler, "Adjusting for unmeasured spatial confounding with distance adjusted propensity score matching," *Biostatistics,* vol. 20, pp. 256-272, 1 2018.