TRANSIT NETWORK CHANGES AND LONDON HOUSING MARKET DYNAMICS: A GRAPH-BASED ANALYSIS (2000–2023)

by

Andrey Zhuravlev, H.BSc, University of Toronto, 2021

A Major Research Paper
presented to Toronto Metropolitan University

in partial fulfillment of the requirements for the degree of

Master of Science
in the Program of
Data Science and Analytics

Toronto, Ontario, Canada, 2025

AUTHOR'S DECLARATION FOR ELECTRONIC SUBMISSION OF AN MRP

I hereby declare that I am the sole author of this MRP. This is a true copy of the MRP, including any required final revisions.

I authorize Toronto Metropolitan University to lend this MRP to other institutions or individuals for the purpose of scholarly research.

I further authorize Toronto Metropolitan University to reproduce this MRP by photocopying or by other means, in total or in part, at the request of other institutions or individuals for the purpose of scholarly research.

I understand that my MRP may be made electronically available to the public

Abstract

Major transport investments have fundamentally reshaped London's accessibility and real estate market, yet rigorous academic analysis specific to the city is limited. This research addresses this gap by analyzing how changes in the public transport network's structure influenced borough-level housing prices over a 25-year period (2000-2023). A dynamic temporal graph of the transit system is constructed by harmonizing TfL's RODS and NUMBAT datasets to create a consistent time series of passenger flows. Using graph mining, annual accessibility indicators such as centrality metrics are calculated for each borough. A fixed-effects panel regression model is then employed to quantify the relationship between these structural network metrics and changes in housing prices. The project provides novel, data-driven insights into the economic impacts of transport infrastructure, offering a new analytical perspective for urban planning and housing policy.

Acknowledgements

First and foremost, I would like to extend my profound gratitude to my supervisor, Dr. Pawel Pralat for his expert guidance and immense patience with me.

I am also deeply grateful to Dr. Ceni Babaoglu, whose courses provided a crucial framework for this research and whose seminars offered inspiring perspectives on future paths.

My sincere thanks also go to the entire faculty and staff at TMU for fostering a supportive and helpful learning environment.

Table of Contents

TRANSIT NETWORK CHANGES AND LONDON HOUSING MARKET DYNAMICS: A GRAPH-BASED ANALYSIS (2000–2023)

Author's Declaration for Electronic Submission of an MRP ii
Abstract iii
Acknowledgements iv
Table of Contents v
List of Tables vi
List of Figures vii

1. Introduction 1
    1.1 Literature Review 2
2. Data Description and Exploratory Data Analysis 5
    2.1 Datasets 5
        2.1.1 UK House Price Index 5
        2.1.2 TfL Rolling Origin-Destination Survey (RODS) 6
        2.1.3 TfL NUMBAT 7
        2.1.4 Station by Borough Data 8
    2.2 Data Preprocessing and Cleaning 8
    2.3 Graph Construction 13
        2.3.1 RODS Graph Generation (2000-2017) 13
        2.3.2 NUMBAT Graph Generation (2017-2023) 14
        2.3.3 Final Graph Structure and Validation 14
    2.4 Exploratory Data Analysis 15
    2.5 Network Accessibility Measures 16
3. Methodology and Experiments 22
    3.1 Methodology 22
    3.2 Model Estimation and Extensions 23
    3.3 Implementation 23
4. Results 24
    4.1 Overall Model Performance and Diagnostics 24
    4.2 The Impact of Accessibility on Housing Prices 24
    4.3 Borough Roles in the Network and Long-Term Dynamics 26
5. Conclusion and Future Works 28
6. Appendix (GitHub Repository) 29
7. References 30

List of Tables

Table	Page
Table 1: NLC Coverage Analysis Across Datasets	10
Table 2: A summary of key network statistics comparing the 2019 (pre-COVID) and 2022 (post-COVID) total weekday passenger flow graphs.	15
Table 3: Change in Mean Daily Passenger Arrivals by Borough (2019 vs. 2022)	18
Table 4: Fixed-Effects Panel Regression Results	24

List of Figures

Figure	Page
Figure 1: Venn diagrams showing overlaps between borough names from the different train station datasets and House Price Index dataset	12
Figure 2: Venn diagram showing overlap between borough names combined from all the different train station datasets and House Price Index dataset	12
Figure 3: Visualization of the 2019 network statistics, showing the top 10 boroughs by internal travel volume and the overall distribution of passenger flow weights.	15
Figure 4: Visualization of the 2022 network statistics.	16
Figure 5: Top 15 boroughs ranked by total passenger arrivals (Weighted In-Degree) for 2019	19
Figure 6: Percentage change in passenger arrivals by borough (2019 vs. 2022)	20
Figure 7: Correlation matrix of all calculated accessibility metrics	21
Figure 8: Regression coefficients and 95% confidence intervals for lagged accessibility metrics.	25
Figure 9: Scatter plot of Betweenness Centrality vs. Weighted Degree for 2019.	26
Figure 10: Heatmap of Betweenness Centrality by borough (2000-2023)	27

Introduction

The relationship between urban transportation networks and real estate markets is a foundational element of urban economics. As cities invest heavily in public transit, understanding how these changes in accessibility are reflected in property values becomes crucial for urban planning and policy. This Major Research Project investigates this dynamic within the unique context of London, a global city that has undergone significant transit evolution over the past twenty five years..

This paper begins by situating the research within the existing academic landscape through a literature review, identifying a gap in independent analysis specific to the London market. Following this, the report provides a detailed account of the extensive data preprocessing phase consisting of multi-stage process of cleaning, mapping, and merging several disparate data sources to construct a novel dataset covering the 25-year period. Key tasks included harmonizing station and borough names, mapping station location codes, and aggregating passenger flow data from TfL's RODS and NUMBAT systems with borough-level housing price indices. The methodology section then explains how this unified dataset was transformed into dynamic graphs, from which key network accessibility metrics were derived. Next, the paper outlines the regression model used to empirically test the relationship between these network metrics and housing prices. Finally, the results of this analysis are presented and discussed, offering new insights into the economic impacts of transit infrastructure and providing a foundation for future research.

Literature Review

The relationship between transportation infrastructure and real estate value is a cornerstone of urban economics, with a rich body of research demonstrating that enhanced accessibility is capitalized into property prices. Early models often relied on simple proximity-based metrics, such as distance to the nearest station. However, the academic literature has evolved significantly, now employing more sophisticated techniques to capture the complex, heterogeneous, and dynamic nature of this relationship. This review synthesizes key methodological trends and identifies a crucial gap in the academic analysis of London's housing market, which this research proposes to address.

A foundational approach in modern analysis involves modeling transit systems as networks and applying graph theory to quantify a location's importance. Scholars have consistently found that centrality indices are significant predictors of housing and land values across diverse urban contexts. Studies in Kolkata[1], Bangkok[2], numerous Chinese cities[3], and a bachelor's thesis on Munich[26] have shown that measures like closeness, eigenvector, and betweenness centrality serve as powerful proxies for the locational advantages conferred by a well-connected position in the transport network. For instance, Chakrabarti et al. (2022) found that higher closeness and eigenvector centrality were associated with price premiums of 9-40% in Kolkata[1], while Han and Wu (2023) demonstrated that a 10% rise in harmonic centrality (a variant of closeness) correlated with an 11% increase in land prices in China[3]. These studies validate the core premise of using graph-based metrics to analyze real estate markets.

While centrality metrics establish that a relationship exists, a significant body of research focuses on its spatial variability. The impact of transit access is rarely uniform across a city. To address this, many studies employ spatial econometric models like Geographically Weighted Regression (GWR) to account for heterogeneity. He (2020) used a Multiscale GWR (MGWR) to reveal that capitalization effects in Hong Kong varied significantly by sub-region[4], while others used GWR to map localized price premiums in Xiamen and Beijing[5], [6]. Research has also advanced beyond simple linear assumptions; Chen et al. (2022), for instance, used spline functions within a GWR framework to model the non-linear, "hump-shaped" effect of metro accessibility, where the price premium diminishes or even reverses at very close proximity to a station[7]. This precedent strongly supports this MRP's borough-level aggregation strategy, which is designed to uncover how the impacts of network changes vary across London's distinct sub-markets.

Moving beyond static, cross-sectional analyses, another research stream investigates the temporal and causal dimensions of transit investment. Quasi-experimental methods, particularly the Difference-in-Differences (DiD) framework, are crucial for isolating the impact of infrastructure changes from broader market trends. Comber and Arribas-Bel (2017) expertly applied a spatial DiD model to Ealing, demonstrating that the mere announcement of Crossrail in 2008 created a measurable, anticipatory uplift in house prices, long before the line became operational[19]. A more recent study on Shanghai's subway expansion used a Matching DiD design to track price dynamics

before and after station openings, finding that price increases began one month prior to launch and persisted for six months[8]. These studies underscore the importance of a long-term temporal analysis, as proposed in this MRP, to capture both the anticipation and realization of transit benefits.

The methodological frontier in this domain is increasingly characterized by the application of Graph Neural Networks (GNNs), a topic covered in depth by foundational surveys[11], [15]. GNNs can learn complex, non-linear relationships directly from graph-structured data, generating rich embeddings of nodes that encode both network topology and intrinsic features. Their versatility is demonstrated by applications ranging from predicting urban polycentricity from street networks[23] to serving as computationally efficient surrogate models for traditional four-step transport demand forecasting[14]. In the real estate domain, recent studies have successfully used GNNs to predict housing prices in Scotland[12], Santiago[13], and Melbourne[16], consistently outperforming conventional models[13]. The proposed GNN models in these papers, such as PD-TGCN with its attention mechanism[13] and the multipartite graph structure of GSNE[16], offer a sophisticated framework for integrating the diverse data sources available for this project.

A key advantage of recent GNN applications is the focus on explainability. The work by Karamanou et al. (2024), along with related conference proceedings[18], is particularly relevant, as it applies explainability techniques to a GNN model for house price prediction in Scotland. By using methods like GNNExplainer, the authors identify not only the most influential features but also the specific neighboring regions (subgraphs) that drive a price prediction for a given area[12]. This provides a powerful tool for moving beyond correlation to offer interpretable, policy-relevant insights, a goal central to this MRP.

It is important to note that a number of papers reviewed here were published in Sustainability, an open-access journal by MDPI[2], [6], [7]. While widely cited, the journal has faced scrutiny regarding its academic rigor. Notably, it was removed from the official lists of approved academic journals by the Norwegian and Finnish national publication committees in 2023. Concerns have been raised about its rapid peer-review timelines and a business model reliant on article processing charges. While this does not invalidate the findings of individual papers, it highlights the need to critically assess their methodologies, a principle applied throughout this review.

Furthermore, several studies referenced in this review, particularly those focused on inference rather than prediction, do not employ a traditional training-testing data split. Instead, models are applied to the full dataset to explore spatial patterns and estimate the significance of relationships. This is a well-established practice in spatial econometrics where the primary goal is understanding coefficients across the entire study area, rather than optimizing a model for out-of-sample prediction. This methodological precedent aligns with the exploratory and inferential goals of this MRP.

While the global literature provides a robust methodological toolkit, a critical analysis reveals a significant gap concerning the London market. Most of the rigorous, peer-reviewed academic studies are focused on other global cities. Conversely, the most detailed, long-term analyses of London's

transit-property value link come from reports commissioned by public or private sector entities. These include economic impact assessments of TfL's supply chain spending[22], analyses of future funding scenarios[21], and reports on specific infrastructure projects.

These reports offer invaluable data and context, confirming significant property value uplift around the Elizabeth Line[10], Crossrail[9], and the London Overground[20]. For instance, the TfL-commissioned report on Crossrail found a 7-10% price uplift near stations even before opening[9], and CBRE's analysis showed an 80% price increase near Elizabeth Line stations between 2008 and 2023[10]. However, as these studies are commissioned by organizations with a vested interest in demonstrating positive outcomes, they may carry an implicit promotional bias and typically lack the methodological transparency and peer-reviewed validation of academic research. They often rely on descriptive comparisons or simpler hedonic models rather than the advanced spatial, causal, or graph-based techniques found in scholarly journals.

This identifies a clear and compelling research opportunity: there is a lack of independent, academic research that applies state-of-the-art graph mining and spatial-temporal modeling techniques to analyze the impact of London's transit network evolution on its housing market. While Zhang et al. (2021) provided a landmark academic study using London's NUMBAT data to classify station roles through community detection[17], its focus was not on housing market impacts. The novelty of this MRP, therefore, lies in bridging this gap. It will synthesize advanced, academically-rigorous methods—validated by the international literature and drawing on concepts from transport geography and graph mining[25]—and apply them to a unique, 25-year dataset for London. By modeling the transit network based on observed passenger flows from RODS and NUMBAT, a practice supported by recent mobility data benchmarks[24], this project will capture a more functionally accurate measure of accessibility, allowing for a deeper and more nuanced understanding of how network structure shapes urban economies.

Data Description and Exploratory Data Analysis

Datasets

(i) UK House Price Index - Monthly average house prices for each London borough from 1995 to the present, including sales counts and property type breakdowns.

Dataset Description and Link: https://data.london.gov.uk/dataset/uk-house-price-index

The dataset is available under UK Open Government License (v3)

It contains 5 sheets:

- Metadata
- By Type

  Monthly house prices and corresponding index values for London and the United Kingdom, broken down by property type (Detached, Semi Detached, Terraced, and Flat). The data is organized with a multi-level header: the first two columns indicate year and month, followed by eight columns for London (four for prices and four for indices) and eight for the UK (also split into prices and indices by property type). Each row represents a specific month, starting in 1995.

- Average price

  Monthly average house prices across all 33 London boroughs, grouped English regions, and national aggregates from January 1995 onward. The top row shows borough and region names (e.g., Camden, Westminster, South East, England), and the second row includes their official administrative codes (e.g., E09000007 for Camden). Each subsequent row begins with a date (e.g., Jan-95) followed by numeric values representing average sale prices in GBP for that month and geography. Data is granular at the borough level for London and aggregated by region elsewhere in England, with Inner/Outer London summaries and a national average ("England") included.

- Index Price

  This sheet follows the same structure as the previous one but contains index prices instead of average sale prices

- Sales Volume

  The structure is the same as the one of the Average Price and Index Price sheets, but it presents Sales Volume data, indicating the number of residential property

transactions recorded each month in a specific area (33 London boroughs, Inner and Outer London, English regions, and the entire country).

Prices are nominal and not adjusted for inflation. Excludes non-market sales (e.g., right-to-buy, gifts, leases under 7 years). As of February 2025, the index was re-referenced to January 2023, and historical data was revised. A 3-month moving average is applied at the borough level.

(ii) TfL Rolling Origin-Destination Survey (RODS) - Annual survey capturing typical weekday entries, exits, and OD flows for London Underground stations, covering 2000–2017. Provides station-to-station trip matrices, boarding and alighting patterns, and station-level demand profiles.

Dataset Description (including License information):
https://data.london.gov.uk/dataset/tfl-rolling-origin-and-destination-survey

Dataset Link:
https://tfl.gov.uk/corporate/transparency/freedom-of-information/foi-request-detail?referenceId=FOI-1386-2021

The dataset is available under UK Open Government License (v2)

There is a separate .xls file for each year and each file consists of 2 sheets:

- Matrix

  Origin-Destination (OD) flow matrix, showing weekday passenger journey counts between London Underground stations, broken down by time bands. Each row records the number of trips from a specific origin station (Station Name in column 2, with National Location Code in column 1) to a specific destination station (Station Name in column 4, with its own NLC in column 3). The remaining columns capture journey volumes across six time periods: Early (before 7am), AM Peak (7am–10am), Midday (10am-4pm), PM Peak (4pm-7pm), Evening (7pm-10pm), and Late (after 10pm), along with a final Weekday Total column summing across these bands. This format enables temporal analysis of commuter flows, revealing not only station-to-station connections but also the distribution of travel activity throughout the day.

- Zone

  Presents a zone-to-zone OD flow matrix from the RODS report, summarizing estimated journey volumes between London fare zones based on survey data reconciled to Autumn counts. The table is structured into multiple matrices for the same 6 time bands, with each matrix showing trip counts and percentage distributions

across zones 1 through 7. For each time band, the rows represent origin zones and the columns represent destination zones, with totals and proportional flows included. Below each matrix, a compressed summary matrix aggregates flows into four broader groups: Zone 1, Zone 2, Zone 3, and combined outer zones (Zones 4–7).

(iii) TfL NUMBAT

Detailed smartcard-based estimates of rail demand across London Underground, Overground, DLR, and TfL Rail, covering 2016–2023. According to TfL, compared to RODS, NUMBAT is a much larger sample size as it uses ticketing data, oyster taps, train loadweigh and passenger counts, as opposed to the manual survey method of RODS which required a lot of scaling.

Dataset Link: https://crowding.data.tfl.gov.uk/

The dataset is available under UK Open Government License (v3)

NUMBAT Origin-Destination Data consists of multiple files per year, each corresponding to a specific day type in the autumn period. For every year, separate files are provided for Fridays, Saturdays, and Sundays, while in more recent years (e.g., 2022), Monday flows are also split out from the midweek group (Tuesday to Thursday). Though filenames are inconsistent and not standardized, each one represents a typical autumn day for that day type, with major disruptions and anomalous days excluded to ensure data stability.

Each file contains a single sheet structured as an Origin-Destination (OD) matrix. Every row corresponds to a unique journey from an origin station (`mnlc_o`) to a destination station (`mnlc_d`), identified by their Master National Location Codes (NLCs), unique identifiers assigned to each station across the TfL network. The remaining columns (starting at column 3) represent the number of trips in each 15-minute interval throughout the traffic day, spanning from 05:00 to 04:59 the next morning. These intervals are indexed using quarter-hour slot numbers, starting at 21 (05:00-05:15) and cycling through to 96 (23:45-00:00), followed by 97-116 to capture post-midnight flows.

Trip counts in NUMBAT are typically non-integer values, reflecting modelled estimates derived from smartcard data and gateline counts, adjusted using timetable-based assignment. This contrasts with the RODS OD matrices, where all values are integers representing actual or scaled passenger counts from survey samples. It is designed for detailed analysis of travel demand, service provision, and customer experience, and for use in service planning and performance evaluation.

Unlike OD matrices in RODS, which are based on survey data, aggregated by fare zone and coarse time bands, NUMBAT provides station-to-station OD flows at 15-minute granularity across all TfL-operated services. This makes it a more precise and modern tool for analyzing how, when, and where passengers travel across the network.

NUMBAT Outputs are other measures representing typical flows, they are also given by the same 15 minute bends described earlier and there are also multiple files per year, with the same structure as NUMBAT Origin-Destination Data. Each file consists of 9 sheets:

- VersionControl, Note, Cover, Cover Page, _Cover: Metadata and documentation (title, publishing info, definitions), varies slightly in name year-to-year.
- Link Load: Number of passengers on the train between two different consecutive stations on a line e.g. from Lambeth North to Waterloo on the Bakerloo line.
- Link frequency (supply): Number of scheduled trains per quarter hour between two consecutive stations.
- Station Boarders and Station Alighters: Number of passengers boarding or alighting at a specific platform in a specific station.
- Line Boarders: Total boarders across a rail line, derived from summing station boarders. This gives a better picture than Entry/Exits for the utilisation of a line as it includes interchangers.
- Station Flow: Number of passenger movements inside a station between its entrances and its different platforms. This includes boarding/alighting and interchange flows.
- Station Entries and Station Exits: number of passengers entering/exiting the station through its gatelines.

(iv) Tube, Overground, Docklands Railway and Elizabeth Line Stations by Borough Data

London Underground by Borough:
https://en.wikipedia.org/wiki/Category:Tube_stations_in_London_by_borough

London Overground:
https://en.wikipedia.org/wiki/Category:Railway_stations_served_by_London_Overground

Elizabeth Line:
https://en.wikipedia.org/wiki/Category:Railway_stations_served_by_the_Elizabeth_line

Docklands Railway: https://en.wikipedia.org/wiki/List_of_Docklands_Light_Railway_stations

Data Preprocessing and Cleaning

One of the first challenges that I faced is a lack of publicly available dataset linking each Underground station to its corresponding Borough. To obtain that information, I created *london_tube_station_borough_scraper.py* script designed to go to every Subcategory on "Category:Tube stations in London by borough" Wikipedia page, scape the corresponding Subcategory page for the list of all the Tube stations in any given borough, and then save that information in *london_tube_stations_by_borough.csv*. The challenge I encountered was the

inconsistency in naming conventions across datasets obtained from different sources. The *UK_House_price_index.xlsx* file used one set of borough names, while the *london_tube_stations_by_borough.csv* file used another, leading to mismatches that prevented direct merging; additionally, tube station names in the *OD_matrix* files sometimes differed from those in the borough mapping file. It took some adjustment and cleaning for boroughs (for instance removing "London Borough of" or "Royal Borough of" from the beginning of borough names, or substituting "and" in boroughs' names for "&"), as well as station names (removing " tube station" from the end of the names, fixing minor punctuation differences like apostrophes, removing additional clarifying geographical information in parentheses (e.g., '(London)'), matching older terminal numbering to newer designations, etc.) to make all those names match. I used Venn diagrams to compare and visualize overlaps and mismatches in both borough and station names and listing the non-overlapping entries to guide manual mapping.

After all of this pre-processing, there were still some discrepancies between the datasets, though there are explanations for those:

The house price index contained six boroughs that were not found in the tube station dataset: Bexley, Bromley, Croydon, Kingston upon Thames, Lewisham, and Sutton. The reason for this discrepancy is that these are London boroughs that are not served by the London Underground network, relying on other transport systems like National Rail instead.

The comparison between the 2017 Origin-Destination (OD) Matrix and the scraped tube stations names showed mismatches in both directions:

Stations only in OD Matrix 2017 (16 total): Amersham, Buckhurst Hill, Chalfont & Latimer, Chesham, Chigwell, Chorleywood, Croxley, Debden, Epping, Grange Hill, Loughton, Moor Park, Rickmansworth, Roding Valley, Theydon Bois and Watford.

These stations exist in the operational OD matrix because they are part of the London Underground network, but they were not captured by the borough-based scrape as they are located outside the official Greater London boundaries.

Stations only in *london_tube_stations_by_borough* CSV (18 total):Battersea Power, Beckton, City Road, Emlyn Road, Heathfield Terrace, Hounslow Town, King William Street, London Victoria, Lothbury, Ludgate Circus, Mark Lane, Millwall, Nine Elms, Paddenswick Road, Rylett Road, St Katharine Docks, Surrey Docks North and The Grove.

This list is a composite of several categories. It includes stations that opened after 2017 (Battersea Power, Nine Elms), permanently closed "ghost" stations captured from historical Wikipedia lists (City Road, Mark Lane, King William Street, Hounslow Town), stations on other transit systems like the DLR (Beckton), and fully operational multimodal interchange stations that are not exclusively part of the Underground (London Victoria). It also includes a substantial number of authorised but never built Underground stations, which were included in historical or planning-related Wikipedia categories and

thus scraped as valid entries despite never having been constructed. These are: Emlyn Road, Lothbury, Paddenswick Road, Heathfield Terrace, Rylett Road, The Grove, Surrey Docks North, St Katharine Docks, Millwall, and Ludgate Circus. These planned stations were associated with early Central London Railway proposals or unbuilt extensions of the Jubilee line, particularly the intended line to Woolwich Arsenal.

Since RODS was replaced with NUMBAT, analysis of the stations in NUMBAT files was also required. What made things even more challenging was that NUMBAT Origin-Destination matrix data only contained National Location Codes (NLCs), so the mapping of the stations to Location Codes was required. I created *create_station_nlc_mapping.py* that systematically processes all NUMBAT model output files from 2016 to 2023. Each of these files contains a Station_Entries sheet, which consistently stores the NLC in the first column and the station name in the third column. The script was designed to extract all unique pairs of NLCs and station names from each file, ensuring that any station present in any scenario or year would be included in the final mapping. After extracting the data from all 32 output files, the script combined and deduplicated the results, resulting in a master mapping that contains 472 unique station-NLC pairs. This mapping, saved as comprehensive_station_nlc_mapping.csv, provides a robust and up-to-date reference for all stations modeled in NUMBAT across the studied years.

(Table 1)

415	NLCs in all three datasets
39	NLCs only in NLC Mapping
0	NLCs only in NUMBAT 2019
18	NLCs only in NUMBAT 2022
1	NLCs in NLC Mapping and NUMBAT 2019 only
0	NLCs in NLC Mapping and NUMBAT 2022 only
241	NLCs in NUMBAT 2019 and NUMBAT 2022 only

However, a detailed investigation revealed an important distinction between different NUMBAT datasets. While the comprehensive mapping successfully captured all 472 stations listed in the NUMBAT output files' Station_Entries sheets, the 2022 NUMBAT OD matrix files contain 674 unique NLC codes, which is 202 more than the output files. This discrepancy highlights that the OD matrices include additional nodes beyond the passenger stations listed in the output files. Among these additional codes, three specific NLCs (6070, 6073, and 8204) were found to have significant

passenger flows in the 2022 OD matrix but do not appear in any of the NUMBAT output files, or the *StationNodesDescription.xls* file provided by TfL themselves.

This finding suggests that these codes likely represent special network nodes such as interchange points, junction stations, or virtual nodes used for routing purposes rather than traditional passenger stations. For borough-based analysis, these codes can be safely excluded since they do not represent actual passenger stations with geographic locations that can be assigned to specific boroughs. Including them would introduce artificial nodes that could distort the spatial distribution of passenger flows and complicate the borough-level analysis without providing meaningful geographic insights.

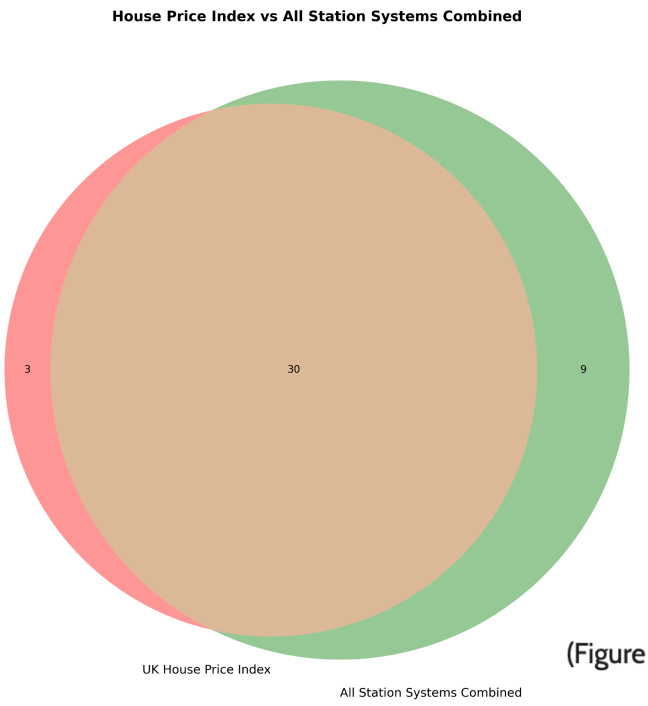
Since initially NUMBAT Data on TfL's Open Data Portal was only available for years 2019 and 2022, I submitted a Freedom of Information request to TfL to obtain data for years 2017, 2018, 2020, 2021 and 2023. After obtaining the data I verified (verification code is in *check_NUMBAT_OD_NLC_Codes.py* script) that in all NUMBAT Origin_destination matrices files among the entire 2017-2023 period there are exactly 674 unique NLC codes. Additional script *Comprehensive_Mapping_Coverage_Over_NUMBAT.py* also verifies that all stations in those files related to London Underground, Overground, Docklands Light Railway (DLR) and Elizabeth Line a covered by *comprehensive_station_nlc_mapping.csv*.

Since NUMBAT files contain Overground, Docklands Light Railway and Elizabeth Railway Stations in addition to the Underground ones, the mapping of the names of those stations to borough names was also required. To address this need, I created three additional web scraping scripts following the same methodology and cleaning logic as the original tube station scraper. The first script, (*london_overground_station_borough_scraper.py)*, scrapes the Wikipedia category page "Railway stations served by London Overground" to extract all 113 Overground stations and their corresponding boroughs by visiting each individual station page and extracting the "Local authority" information from the infobox. The second script (*london_ELZline_station_borough_scraper.py)* performs the same process for the Elizabeth line, scraping the "Railway stations served by the Elizabeth line" category page to obtain borough information for all 41 Elizabeth line stations. The third script, (*london_DLR_station_borough_scraper.py*) takes a different approach by scraping the single comprehensive Wikipedia page "List of Docklands Light Railway stations" which contains all DLR stations and their borough information in tabular format, eliminating the need to visit individual station pages. All three scripts apply the same standardized cleaning procedures for both station names and borough names as the original *london_tube_station_borough_scraper.py* script.

I then investigated the overlays of the borough names between the newly obtained *london_tube_stations_by_borough.csv, LO_stations_by_borough.csv*, *ELZ_stations_by_borough.csv, DLR_stations_by_borough.csv* and the Home Price Index dataset:

(Figure 1)

There remaining mismatches were resolved through two targeted corrections. First, compound borough names for stations serving multiple boroughs were standardized to their primary administrative borough (e.g., changing "Tower Hamlets & Hackney" to "Tower Hamlets", "Bexley & Royal Borough of Greenwich" to just "Greenwich"). Second, all stations located outside the Greater London boundaries were categorized as "Out of London" to maintain geographic accuracy, including Elizabeth line stations in Berkshire and Buckinghamshire (Reading, Slough, Maidenhead, Twyford, Burnham, Taplow, Iver, Langley), London Overground stations in Hertfordshire (Borough of Watford, Borough of



(Figure 2)

Broxbourne, District of Three Rivers), and other non-London locations (Borough of Brentwood in Essex). This standardization process resulted in a clean dataset with 33 London administrative areas plus an "Out of London" category, ensuring consistent geographic attribution across all transport systems for subsequent borough-level analysis.

The final step was to align the station names in my aggregated scraped borough table (*all_stations_by_borough_standardized.csv*) to the canonical names in *comprehensive_station_nlc_mapping.csv.* There were 362 (76.7% of NLC mapping stations) overlapping stations, 110 only in NLC mapping and 65 only in borough mapping.
Firstly, I removed 39 Tramlink stations, as the scope of this research is specifically focused on London's core rapid transit rail networks (Underground, Overground, DLR, and Elizabeth line). Including the Tramlink system, which has different operational characteristics, would be inconsistent with the study's defined boundaries.
The next step was to systematically map the remaining stations using a sequential mapping strategy, executed by the *create_station_borough_nlc_mapping.py* script. The process was designed to handle the data in stages, moving from broad, automated matches to more specific, manual corrections. Initially, the script performed a direct match for stations with identical names, followed by a second stage that stripped common service suffixes (e.g., 'LU', 'DLR') to match base names. To address the final 11 stations with missing boroughs, a dictionary was manually constructed to resolve the remaining complex inconsistencies. Finally, stations outside Greater London were assigned a distinct 'Out of London' category to ensure geographic accuracy. For the few stations situated directly on administrative boundaries, such as Manor House (Hackney/Haringey) or Anerley (Bromley/Croydon), a single primary borough was assigned based on the station's main entrance or predominant location to maintain analytical consistency. The final result was saved in *station_borough_nlc_mapping.csv* file.

Graph Construction

Following the extensive data preprocessing and cleaning, the next critical phase of the research involved transforming the harmonized origin-destination (OD) matrices into a series of dynamic temporal graphs. This process converts the tabular passenger flow data into a rich network representation, where London's boroughs are the nodes and the passenger volumes between them constitute the weighted, directed edges. This foundational step is essential for applying graph mining techniques to derive accessibility metrics. The entire process was automated using a series of Python scripts leveraging the igraph library for graph creation and manipulation. All resulting graphs were saved in the standard GraphML format for portability and future analysis.

RODS Graph Generation (2000-2017)
To process the historical RODS data, a script systematically iterates through the 18 annual .xls files from 2000 to 2017. For each year, the script reads the 'matrix' sheet, extracting the origin and

destination National Location Codes (NLCs) and the corresponding passenger counts for the six defined time bands (Early, AM Peak, Midday, PM Peak, Evening, Late).

The core of this process was the aggregation from the station level to the borough level. Using the station_borough_nlc_mapping.csv file, each station's NLC was mapped to its corresponding London borough. The script then aggregated all passenger flows using a groupby operation, summing the trips between every pair of boroughs. From these aggregated matrices, seven distinct igraph graphs were generated for each year: one for the total weekday flow and one for each of the six time bands. Edges with zero passenger flow were pruned from the final graphs.

NUMBAT Graph Generation (2017-2023)

The construction of graphs from the NUMBAT dataset was more complex due to the data's higher granularity and variability in file structure. The goal was to create a consistent annual series of graphs that aggregated flows across all transport modes (London Underground, Overground, DLR, and the Elizabeth Line).

An automated script was designed to handle this complexity. For each year, the script first scans the filenames to automatically identify the distinct day-type groups available (e.g., MTT for Tuesday-Thursday, FRI, SAT, SUN). The processing logic then adapts based on the year's file format, even accommodating inconsistencies in column naming conventions across different years (e.g., mnlc_o vs. mode_mnlc_o).

For years with mode-specific files, the script first sums the OD flows from all individual mode files, and for years with network-level files, it uses the provided data directly.

Where coarse time band (_tb_) files were not provided, the script automatically derived them by summing the flows from the appropriate 15-minute quarter-hour (_qhr_) slots according to a predefined mapping, ensuring consistency across all years.

For each year, this process generates a comprehensive hierarchy of graphs, including an overall annual graph, day-type specific graphs, and graphs for each of the six time bands and 96 quarter-hour slots.

Final Graph Structure and Validation

The output of the construction phase is a complete, harmonized time series of passenger flow graphs from 2000 to 2023. Each graph is directed, with edges pointing from the origin borough to the destination, and weighted, with the edge weight representing the total number of passenger journeys. The nodes consist of the 33 London boroughs plus a single 'Out of London' category. A key feature is the presence of self-loops (e.g., an edge from Westminster to Westminster), which are retained as they represent meaningful intra-borough travel.

To ensure integrity, a series of sanity checks were performed on sample outputs. This validation was crucial, as it uncovered and allowed me to correct an initial column-mapping error in the RODS

processing script. The corrected scripts were verified to produce graphs with logical node and edge counts and realistic passenger flow distributions, confirming that the resulting graph dataset provides a robust foundation for the subsequent network analysis.

Here are visual representations of the examples graphs:

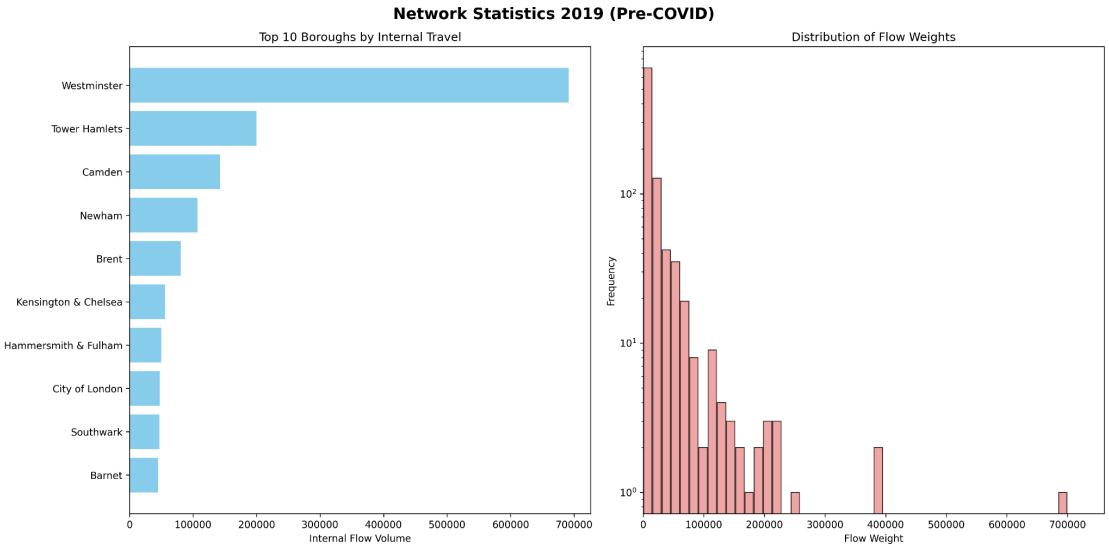
Exploratory Data Analysis

Table 2 compares key network metrics for the 2019 (pre-COVID) and 2022 (post-COVID) weekday passenger flow graphs. The analysis reveals a significant redistribution of travel patterns despite the Total Flow of passengers remaining nearly static with only a 0.8% increase.

(Table 2)

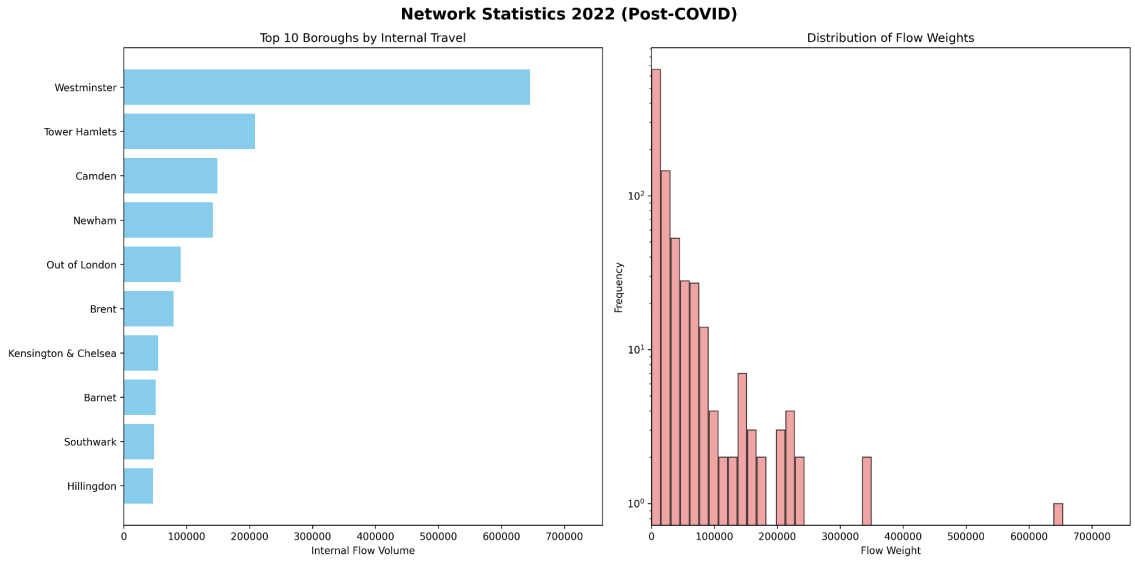
Metric	2019	2022	Change (%)
Nodes (Boroughs)	31	31	0
Edges (Flows)	961	961	0
Total Flow	17079375.4	18452022.1	8
Average Flow	17772.5	19200.9	8
Maximum Flow	691347.4	645654.8	-6.6
Minimum Flow	2.9	23.5	706.8
Total Internal Flow	1829557.4	1914058.9	4.6
Average Internal Flow per Borough	59018	61743.8	4.6
Network Density	1	1	0
Average Degree	62	62	0

The data shows a trend towards intensified use of major corridors and more localized travel. The Maximum Flow on the network's single busiest connection grew by 6.6%, while Total Internal Flow (journeys within the same borough) rose by 4.6%. In stark contrast, the Minimum Flow on the least-used connection plummeted by over 700%. Taken together, these metrics point to a post-pandemic network that relies more heavily on its core arteries and localized trips, while connectivity between peripheral boroughs has weakened considerably.



(Figure 3)

(Figure 4)



Figures 3 and 4 offer a comparative visual summary of the network characteristics for 2019 (pre-COVID) and 2022 (post-COVID), respectively. Each figure presents two key visualizations: a bar chart on the left ranking the top boroughs by their internal travel volume (the number of journeys starting and ending in the same borough) and a histogram on the right showing the distribution of passenger flow weights across all connections.

A direct comparison of the bar charts reveals the network's structural resilience. In both 2019 and 2022, Westminster is clearly established as the dominant hub for internal activity, followed by a consistent cohort of central boroughs like Tower Hamlets and Camden. The most significant difference is the emergence of the 'Out of London' category as a top-ranked traffic generator in the 2022 chart, highlighting a post-pandemic shift where external commuting plays a more prominent role. The histograms in both figures reinforce this theme of stability, each depicting a nearly identical, heavily right-skewed distribution. This visual pattern confirms that the network's fundamental topology—a high frequency of low-volume connections complemented by a few major, high-volume corridors—has remained unchanged, even as the nature of some journeys has evolved.

Network Accessibility Measures

The goal of this stage was to calculate a set of centrality and community structure metrics for all graphs, creating a complete time-series dataset for the final analysis.
The analysis was performed using three separate Python scripts. The first script (*calculate_centrality.py*) calculated all centrality measures, the second (*aggregate_results.py*) handled

community detection, and the third (*calculate_community_metrics.py*) merged their outputs into a single file. This separation kept the process organized and easy to manage.

The main challenge was community detection on directed graphs, since the community detection algorithm required an undirected graph (where connections have no direction), but the passenger flow data is directed (from an origin to a destination). Therefore, for the community detection step only, the graph's directed edges were temporarily converted to undirected edges by summing the passenger flows in both directions. This was done because community detection focuses on the total strength of a connection, for which the combined two-way flow is the best measure. The original directed graph was still used for all other calculations.

Metrics Calculated:

Five centrality measures were calculated for each borough in every graph to measure its importance in the network:

Weighted In-Degree: Total passenger arrivals.

Weighted Out-Degree: Total passenger departures.

Betweenness Centrality: Measures a borough's role as a "pass-through" corridor. Because this metric is calculated based on shortest paths, the passenger flow weights were inverted (distance = 1 / flow) before calculation. This transformation ensures that high-volume routes are correctly treated as the functionally "shortest" paths for passenger journeys.

Closeness Centrality: Measures how easily a borough can reach all others. This metric also relies on shortest paths, so the same weight inversion (distance = 1 / flow) was applied. By treating high-volume routes as 'shorter' paths, the metric accurately reflects network connectivity based on travel intensity.

Eigenvector Centrality: Measures a borough's connection to other important hubs.

For detecting community structure, **t**he **Leiden** algorithm was used to find travel clusters in the network. It was selected as it ensures more accurate communities compared to Louvain method, while remaining computationally efficient enough for a large-scale analysis of numerous graphs compared to the ECG algorithm which would be too slow.

The process resulted in the *creation of all_metrics_timeseries.csv*, a single dataset containing all 57,319 records. This file includes the metadata for each observation (year, day-type, etc.) along with all the calculated accessibility metrics, ready for the final regression analysis.
To provide a clear overview of the network's evolution, Table 3 summarizes the change in mean daily passenger arrivals for a selection of representative boroughs between the pre-pandemic (2019) and

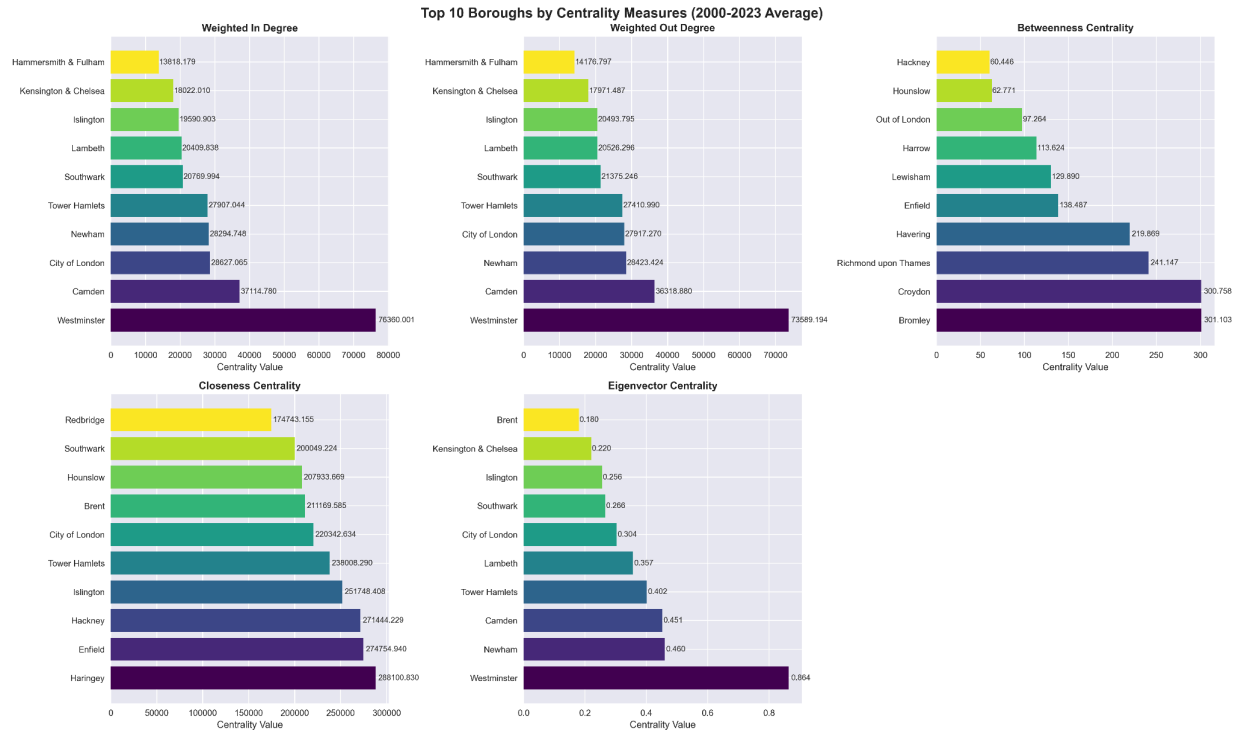
post-pandemic (2022) periods. The table highlights a key spatial trend: while major central hubs like Westminster and the City of London saw a relative decline in passenger traffic, many outer, residential boroughs experienced significant growth. This provides initial evidence of the shift in travel patterns away from the traditional city core that will be explored in the main analysis.

(Table 3)

Category	Borough	Mean Arrivals 2019	Mean Arrivals 2022	Change (%)
Top Central Boroughs	Westminster	350,850	340,801	-2.90%
Top Central Boroughs	Camden	117,768	115,278	-2.10%
Top Central Boroughs	City of London	88,421	80,832	-8.60%
Outer Boroughs	Havering	13,810	16,174	+17.1%
Outer Boroughs	Sutton	1,846	2,751	+49.0%
Outer Boroughs	Barking & Dagenham	22,298	26,013	+16.7%
All Boroughs Average	All Boroughs Average	59,018	61,744	+4.6%

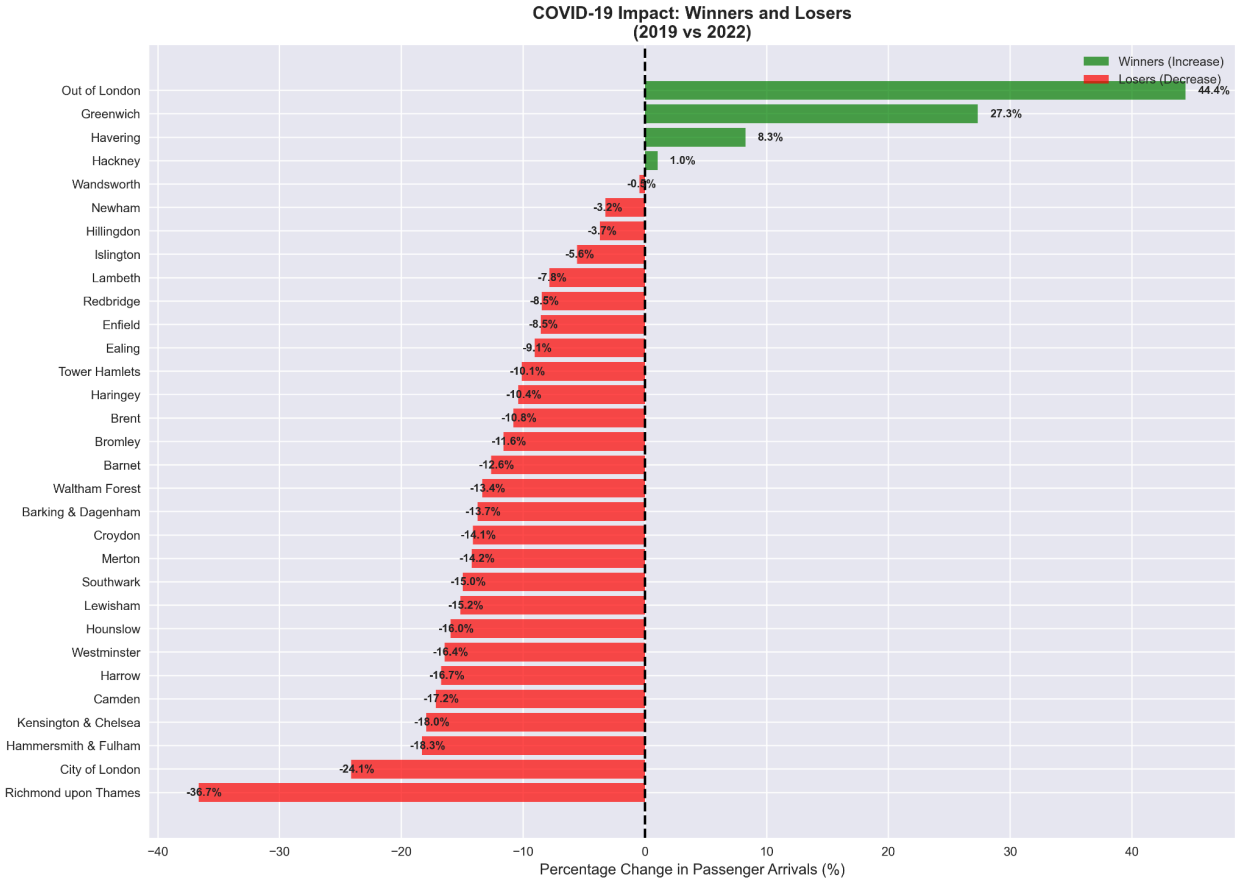
The clear hierarchy of the network is visualized in Figure 5, which ranks boroughs by their total passenger traffic (in-degree) for a representative year. The chart shows the overwhelming dominance of a few key central boroughs, particularly Westminster, establishing the highly monocentric nature of London's public transport network.

(Figure 5: Top 15 boroughs ranked by total passenger arrivals (Weighted In-Degree) for 2019)



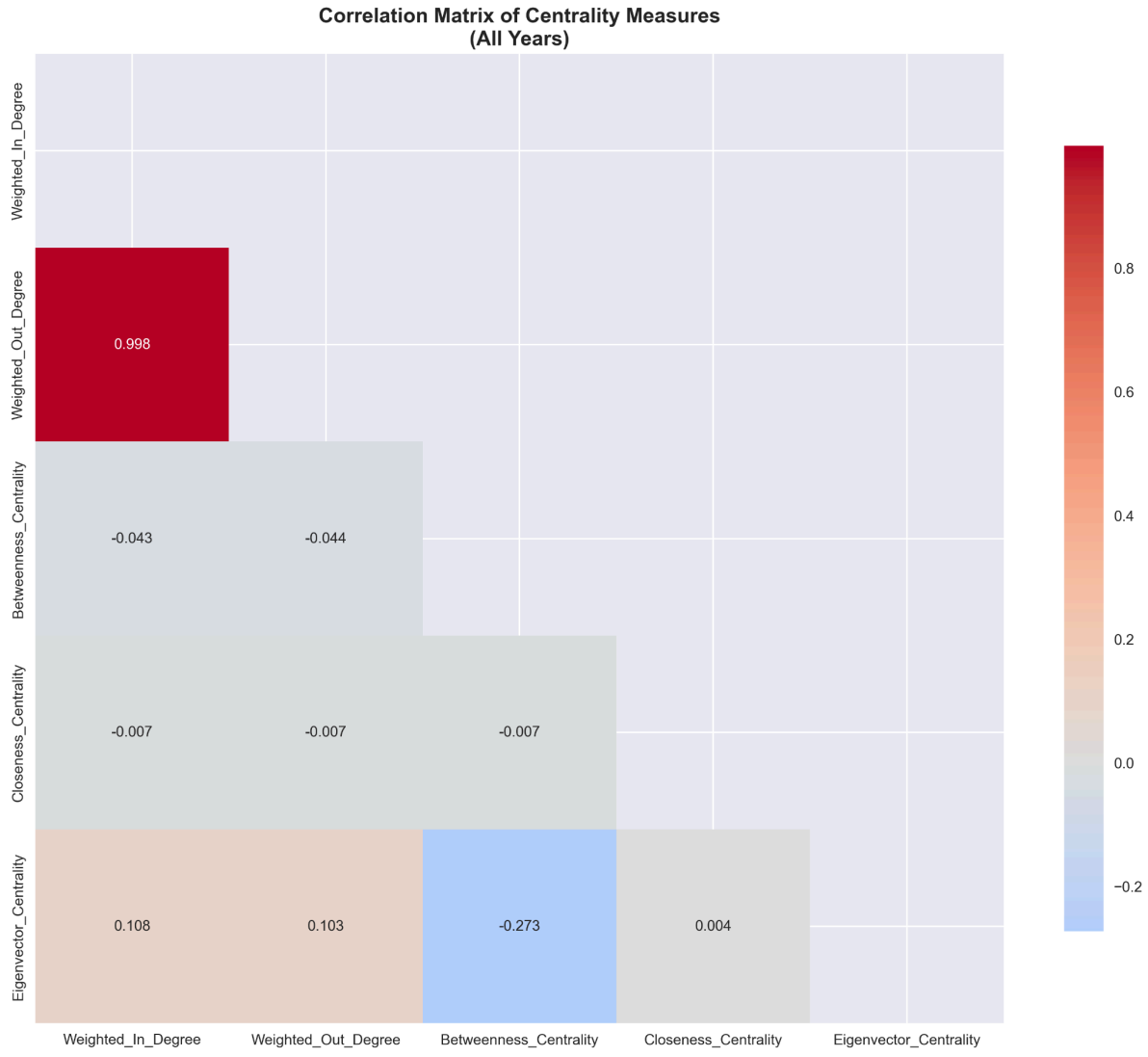
To visualize the impact of the COVID-19 pandemic, the percentage change in passenger arrivals between 2019 and 2022 was calculated for each borough. As shown in Figure 6, this analysis reveals a distinct spatial pattern of change. Central, business-focused boroughs such as the City of London and Westminster saw a relative decline in traffic, while many outer, more residential boroughs like Sutton, Havering, and Barking & Dagenham saw a relative increase. This illustrates a significant post-pandemic shift in travel patterns away from the traditional city core.

(Figure 6: Percentage change in passenger arrivals by borough (2019 vs. 2022))



Finally, to prepare for the regression modeling stage, a correlation matrix of all the calculated accessibility metrics was generated (Figure 7). This step is crucial for understanding the relationships between the variables and checking for multicollinearity. The heatmap confirms a near-perfect correlation between in-degree and out-degree, as expected. More importantly, it shows that the other centrality measures have low correlations with each other, indicating that they capture distinct and independent aspects of network accessibility and are therefore suitable for inclusion as separate variables in the final model.

(Figure 7: Correlation matrix of all calculated accessibility metrics)



Methodology and Experiments

Methodology

To analyze the relationship between transport accessibility and housing prices, a fixed-effects panel regression model was employed. The model is specified as follows:

$$ln(HousingPrice_{it}) = \beta_0 + \beta_1(X_{i,t-1} - 1) + \alpha_i + \delta_t + \varepsilon_{it}$$

Where $HousingPrice_{it}$ is the average house price in borough *i* in year *t*. $X_{i,t-1}$ the vector of lagged and standardized accessibility metrics from the prior year. α_i represents borough fixed effects, controlling for time-invariant local characteristics, while δ_t represents year fixed effects, controlling for macro-economic trends.

Dependent Variable: The annual average housing price for each London borough. To account for the exponential nature of price growth and to ensure the residuals are more normally distributed, the natural logarithm of the average house price was used.

Independent Variables: The set of five accessibility metrics (Weighted Degree, Betweenness, etc.). To allow for direct comparison of the effect sizes of these different metrics, all independent variables were standardized using a Z-score transformation.

The choice of a fixed-effects model is directly informed by the methodological precedents identified in the literature review. Drawing on the Difference-in-Differences (DiD) frameworks used in studies like Comber and Arribas-Bel (2017)[19] on Crossrail, this model structure is adept at isolating the impact of network changes from both macro-economic trends (year fixed effects) and unique local characteristics (borough fixed effects). This directly addresses the spatial heterogeneity noted in the Geographically Weighted Regression (GWR) literature by controlling for the unique, underlying attributes of each borough.

Furthermore, the decision to use lagged accessibility metrics as the primary independent variables is a direct response to findings from the Shanghai and Crossrail studies, which demonstrate that property value impacts are not instantaneous but can be both anticipatory and delayed. The graph-based centrality measures themselves are validated as powerful proxies for locational advantage by a wide body of international research cited in the review, from Kolkata[1] to various Chinese cities[3]. By synthesizing these established spatial-temporal and graph-based techniques, this model provides a rigorous, academically-grounded framework to analyze the transit-property value link in London.

Model Estimation and Extensions

The model was estimated using Ordinary Least Squares (OLS) with clustered standard errors at the borough level. This was done to account for potential autocorrelation (the likelihood that observations within the same borough are correlated over time), ensuring that the statistical inference is reliable.

To specifically test for structural changes during the pandemic, an extended model including an interaction term was also estimated. The specification for this model is as follows:

$$ln(HousingPrice_{it}) = \beta_0 + \beta_1 X_{i,t-1} + \beta_2 COVID_t + \beta_3 X_{i,t-1} \times COVID_t + \alpha_i + \delta_t + \epsilon_{it}$$

$COVID_t$: A binary indicator variable that equals 1 for years 2020 and later, and 0 otherwise.

$\beta_2 COVID_t$ captures the average baseline change in housing prices during the pandemic period, independent of any changes in accessibility.

β_3 measures the additional effect of the accessibility metric specifically during the COVID period. A statistically significant β_3 would suggest that the fundamental relationship between transport accessibility and housing prices was structurally different after 2020.

Implementation

The model (data preparation and analysis) is implemented in three Python scripts located in the modeling/scripts/ directory.

01_prepare_final_dataset.py: This script handles all data preprocessing. Its primary responsibility is to merge the calculated all_metrics_timeseries.csv file with the annual housing price data. It then created the crucial one-year lagged variables for all accessibility metrics before saving the final, model-ready dataset.

02_run_panel_regression.py: Loads the prepared data and uses the statsmodels library to estimate both the primary fixed-effects panel model and the extended model with COVID-19 interaction terms. A key implementation detail is the application of clustered standard errors at the borough level to ensure statistical validity. The full regression summary for each model is saved to a text file for later use.

03_visualize_results.py: This final script is responsible for reporting the results. It provides a clear visual interpretation of the model's findings by parsing the saved model results to generate the final regression tables and create a coefficient plot.

Results

Overall Model Performance and Diagnostics

The primary fixed-effects panel model demonstrated an exceptionally strong fit to the data, explaining 98.5% of the variance in the natural logarithm of average housing prices (Adjusted R^2 = 0.984). The overall model was highly statistically significant (F-statistic p < 0.001), confirming that the variables are collectively powerful predictors of housing prices. The high R-squared is largely attributable to the inclusion of borough and year fixed effects, which effectively control for unobserved heterogeneity.

Diagnostic tests confirmed the robustness of the model. A Variance Inflation Factor (VIF) analysis was conducted, and after correcting for initial multicollinearity, all VIF scores for the core accessibility metrics were found to be within an acceptable range. The Breusch-Pagan test indicated the presence of heteroskedasticity, which was addressed through the use of clustered standard errors to ensure all reported p-values are reliable.

The Impact of Accessibility on Housing Prices

The main findings of the regression analysis are presented in Table 4. The table displays the estimated coefficients for each lagged accessibility metric, which can be interpreted as the percentage change in housing prices for a one standard deviation increase in the metric from the previous year.

(Table 4)

Variable	Coefficient	Std. Error	p-value	Sig.
Weighted In-Degree (Lag 1)	0.0128	0.003	< 0.001	***
Betweenness Centrality (Lag 1)	-0.0069	0.004	0.052	*
Closeness Centrality (Lag 1)	0.0022	0.002	0.354	
Eigenvector Centrality (Lag 1)	-0.0067	0.014	0.632	

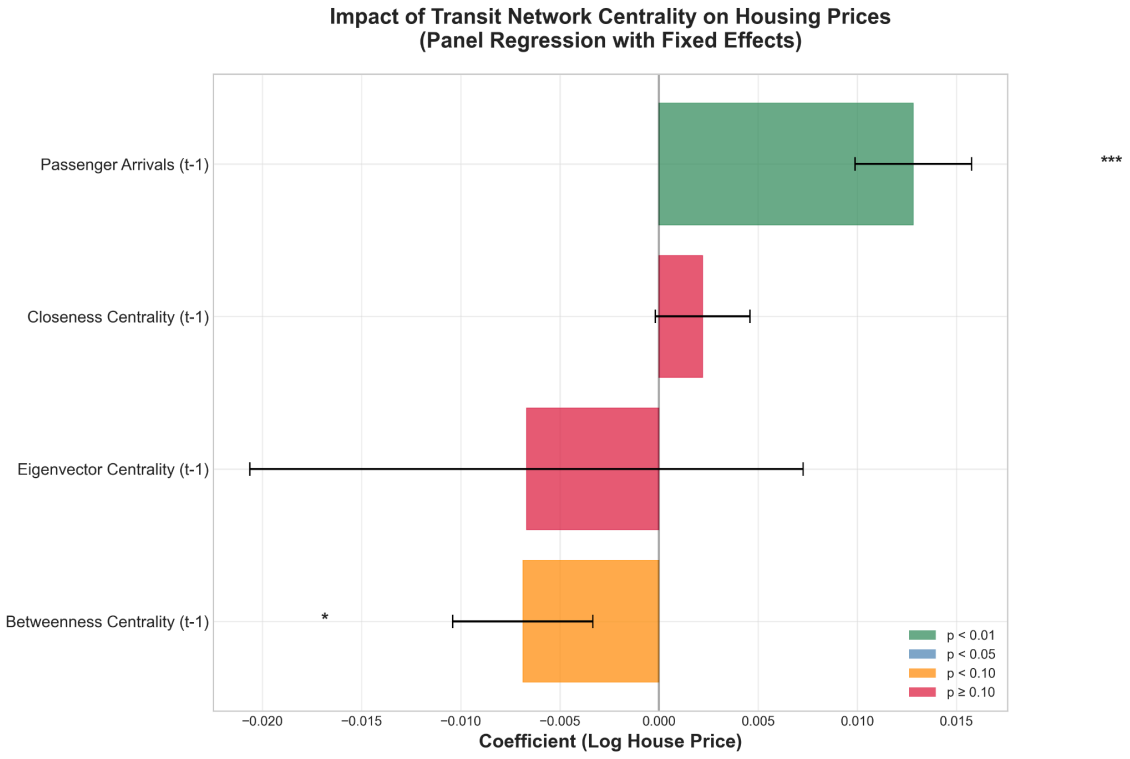
Significance levels: * p < 0.1, ** p < 0.05, *** p < 0.01

The results reveal two statistically significant relationships. First, Weighted In-Degree, representing the total passenger flow into a borough, has a highly significant positive effect on housing prices. The coefficient of 0.0128 indicates that a one standard deviation increase in a borough's passenger arrivals in the previous year is associated with a 1.28% increase in its average housing prices, holding all other factors constant. This finding aligns with the economic theory that areas with higher accessibility and activity command higher property values.

Second, Betweenness Centrality shows a marginally significant negative relationship with housing prices (p=0.052). This suggests that a one standard deviation increase in a borough's role as a "pass-through" corridor is associated with a 0.69% decrease in housing prices. This counter-intuitive result may suggest that the disamenities that come with being a major transport corridor—such as traffic, noise, and transient populations—could slightly offset the capitalization of pure connectivity benefits. The other centrality measures were not found to have a statistically significant effect.

Figure 8 provides a visual summary of these findings. The coefficient plot displays the estimated effect of each accessibility metric on housing prices, along with its 95% confidence interval. The plot clearly shows the positive and statistically significant effect of passenger flow, as its confidence interval is entirely above zero. It also visualizes the marginally significant negative effect of Betweenness Centrality, with its confidence interval just touching the zero line, while the intervals for all other metrics clearly overlap with zero, confirming their lack of statistical significance.

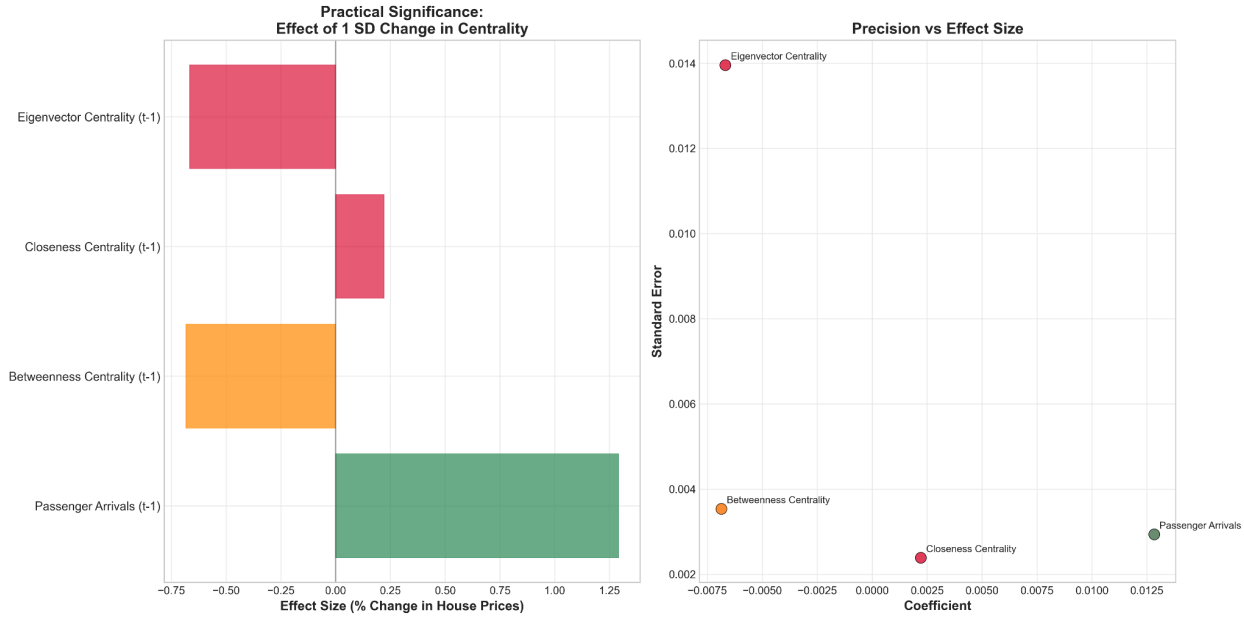
(Figure 8)



Borough Roles in the Network and Long-Term Dynamics

To better understand the roles that different boroughs play, the relationship between a borough's total passenger volume (Weighted Degree) and its function as a transport corridor (Betweenness Centrality) was analyzed. Figure 9 plots these two metrics against each other for a representative year.

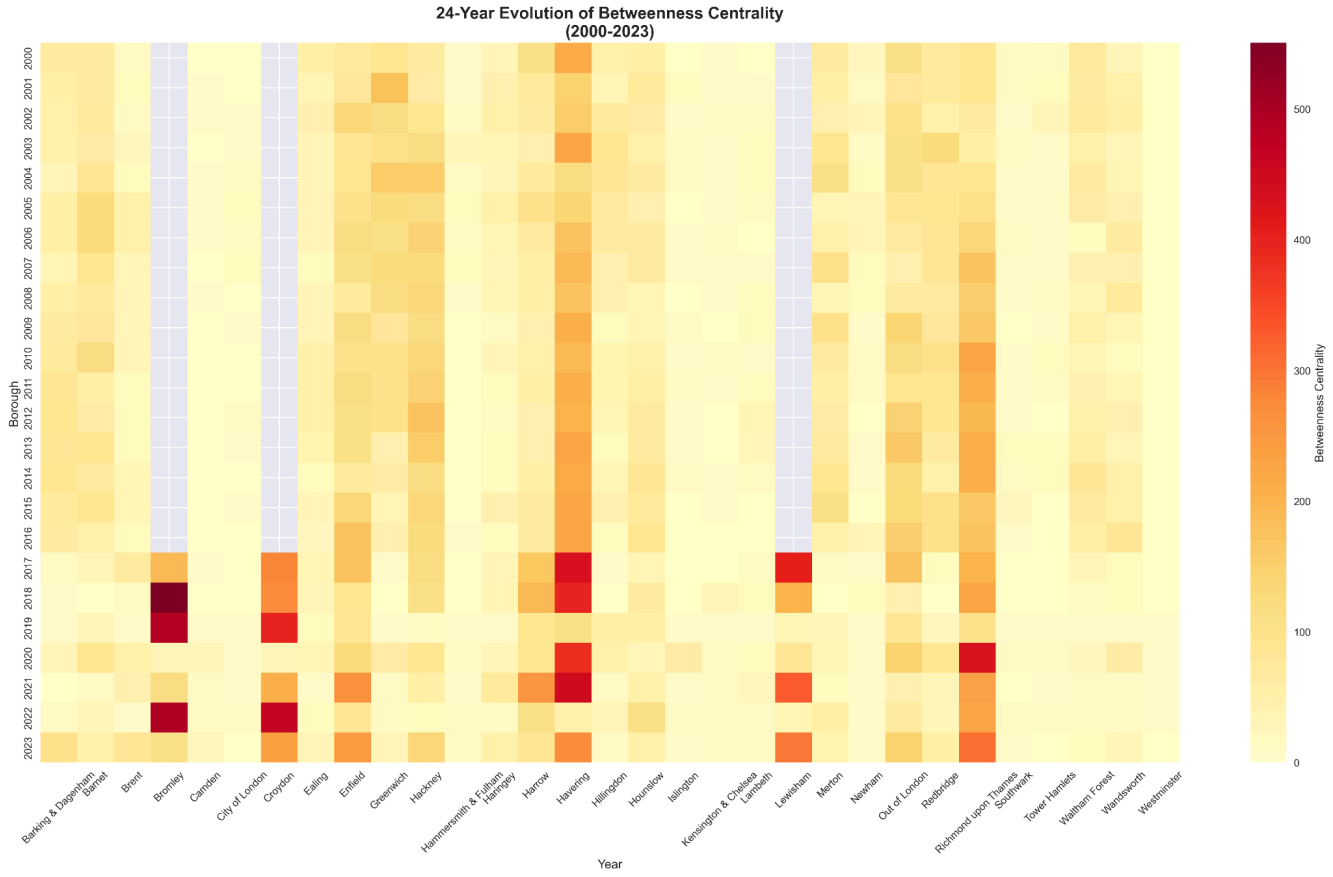
(Figure 9)



The plot reveals distinct borough types. Boroughs like Westminster and Camden exhibit both high degree and high betweenness, cementing their status as London's primary hubs that are both major destinations and critical interchanges. In contrast, boroughs like Tower Hamlets (home to Canary Wharf) show high degree but relatively lower betweenness, characterizing them as major "destination zones" that are not primary corridors for cross-city travel. This distinction helps explain the regression results: while being a major destination (high degree) is positively capitalized into house prices, the specific function of being a "pass-through" corridor (high betweenness) may carry negative externalities that temper property values.

Finally, to visualize the long-term dynamics of the network, the evolution of Betweenness Centrality was plotted for all boroughs over the entire 24-year study period. Figure 10 presents this data as a heatmap, providing a powerful overview of structural changes in London's transport network.

(Figure 10)



The heatmap clearly shows the consistent dominance of central boroughs like Westminster, Camden, and Southwark as the primary transport corridors over the last two decades. It also highlights the transformative impact of major infrastructure projects. For example, the steady increase in the centrality of Tower Hamlets and Newham in the mid-to-late 2000s coincides with the expansion of the DLR and the development leading up to the 2012 Olympics. This visualization confirms that while the core structure of the network is stable, it is not static, and targeted investments can significantly reshape the functional roles of different parts of the city over time.

Conclusion and Future Works

The analysis yielded several key findings. The primary modeling approach revealed a statistically significant positive relationship between a borough's total passenger flow (Weighted In-Degree) and its subsequent housing prices, confirming that higher levels of activity and accessibility are capitalized into the housing market. On the other hand, a negative relationship was found for Betweenness Centrality, suggesting that the disamenities that come with being a major "pass-through" corridor may offset the positive benefits of pure connectivity.

Moreover, a long-term analysis of the network's structure confirmed its stability while also showing its capacity for change. Central boroughs like Westminster and Camden have consistently dominated as the primary transport corridors over the last two decades. At the same time, the network has shown to be dynamic, not static, meaning there was a noticeable transformative impact of major infrastructure projects. The centrality of Tower Hamlets and Newham steadily increased in the mid-to-late 2000s, a period coinciding with the DLR expansion and development for the 2012 Olympics. This shows how targeted investments can significantly reshape the functional roles of different parts of the city over time.

The findings carry significant implications for urban planning and transport economics. The positive relationship between passenger flow and housing prices provides quantitative support for infrastructure investment as a mechanism for property value growth. However, the negative effect of 'pass-through' centrality suggests a critical distinction: the quality and function of accessibility are as important as its quantity. This implies that to maximize economic benefits, infrastructure projects should be designed to cultivate a borough's role as a destination, as the advantages of simply being a transport corridor may be offset by negative externalities like congestion and noise.

This study is not without its limitations. The analysis was conducted at the borough level, which may obscure more granular, neighborhood-level effects. The model also does not account for all potential factors influencing house prices, such as local school quality or zoning policies.

Future research could build on this work by using more advanced models, such as the Graph Neural Networks (GNNs) identified in the literature review, to capture more complex, non-linear relationships. The model could also benefit by incorporating additional borough-level variables to control for other factors known to influence house prices, such as socio-economic data (e.g., crime rates, average income), built-environment features (e.g., green space availability, school quality ratings), or policy information (e.g., designated regeneration zones). Finally, moving to a more granular spatial scale, such as analyzing individual station areas instead of entire boroughs, would provide a deeper and more precise understanding of the spatial dynamics. Overall, this project successfully demonstrated the value of applying network science techniques to understand the complex interplay between transport infrastructure and urban economics.

Appendix

GitHub Repository

https://github.com/azhuravlev1/MRP_LondonTransit_RealEstate

References

1. Chakrabarti, S., Kushari, T., & Mazumder, T. (2022). *Does transportation network centrality determine housing price? Journal of Transport Geography, 103*, 103397. https://doi.org/10.1016/j.jtrangeo.2022.103397
2. Vichiensan, V., Wasuntarasook, V., Prakayaphun, T., et al. (2023). *Influence of Urban Railway Network Centrality on Residential Property Values in Bangkok. Sustainability, 15*(22), 16013. https://doi.org/10.3390/su152216013
3. Han, D., & Wu, S. (2023). *The capitalization and urbanization effect of subway stations: A network centrality perspective. Transportation Research Part A: Policy and Practice, 176*, 103815. https://doi.org/10.1016/j.tra.2023.103815
4. He, S. Y. (2020). *Regional impact of rail network accessibility on residential property price: Spatially heterogeneous capitalization effects in Hong Kong. Transportation Research Part A: Policy and Practice, 135*, 244–263. https://doi.org/10.1016/j.tra.2020.01.025]
5. Yang, L., Chu, X., Gou, Z., et al. (2020). *Accessibility and proximity effects of bus rapid transit on housing prices: Heterogeneity across price quantiles and space. Journal of Transport Geography, 88*, 102850. https://doi.org/10.1016/j.jtrangeo.2020.102850
6. Zhou, Y., Tian, Y., Jim, C. Y., et al. (2022). *Effects of Public Transport Accessibility and Property Attributes on Housing Prices in Polycentric Beijing. Sustainability, 14*(22), 14743. https://doi.org/10.3390/su142214743
7. Chen, K., Lin, H., Liao, L., et al. (2022). *Nonlinear Rail Accessibility and Road Spatial Pattern Effects on House Prices. Sustainability, 14*(8), 4700. https://doi.org/10.3390/su14084700
8. Wang, J., & Tang, H. (2025). *Subway Expansion and Housing Price Dynamics: Accessibility Improvements and Spatial Heterogeneity in Shanghai. SSRN Working Paper*. https://doi.org/10.2139/ssrn.5187633
9. Transport for London, & Department for Transport. (2022, April 27). *Crossrail baseline evaluation: Evaluation of Crossrail pre-opening property impacts*. https://content.tfl.gov.uk/property-impacts-report-acc.pdf
10. CBRE. (2024). *The Elizabeth Line: The impact on London's housing market*. Adaptive Spaces. https://www.cbre.co.uk/insights/reports/the-elizabeth-line-the-impact-on-london-s-housing-market
11. NYU Marron Institute of Urban Management. (2021). *Access to Opportunity: Housing Affordability and Transportation in New York City*. ArcGIS StoryMaps. https://storymaps.arcgis.com/stories/6c15fbaed4ff49e3bdaecddcfdbf4e9b
12. Karamanou, A., Brimos, P., Kalampokis, E., & Tarabanis, K. (2024). Explainable Graph Neural Networks: An Application to Open Statistics Knowledge Graphs for Estimating House Prices. *Technologies, 12*(8), 128. https://doi.org/10.3390/technologies12080128
13. Riveros, E., et al. (2023). *Scalable Property Valuation Models via Graph-based Deep Learning. Preprint under review.* https://arxiv.org/abs/2405.06553

14. Narayanan, S., Makarov, N., & Antoniou, C. (2024). *Graph neural networks as strategic transport modelling alternative — A proof of concept for a surrogate*. *IET Intelligent Transport Systems, 18*(11), 2059–2077. https://doi.org/10.1049/itr2.12551
15. Li, H., Zhao, Y., Mao, Z., Qin, Y., Xiao, Z., Feng, J., Gu, Y., Ju, W., Luo, X., & Zhang, M. (2024). *A Survey on Graph Neural Networks in Intelligent Transportation Systems*. arXiv preprint arXiv:2401.00713. https://arxiv.org/abs/2401.00713
16. Das, S. S. S., Ali, M. E., Li, Y.-F., Kang, Y.-B., & Sellis, T. (2021). *Boosting House Price Predictions using Geo-Spatial Network Embedding.* arXiv preprint arXiv:2009.00254. https://doi.org/10.48550/arXiv.2009.00254
17. Zhang, Y., Marshall, S., & Manley, E. (2021). Understanding the roles of rail stations: Insights from network approaches in the London metropolitan area. Journal of Transport Geography, 94, 103110. https://doi.org/10.1016/j.jtrangeo.2021.103110
18. Brimos, P., Karamanou, A., Kalampokis, E., Mamalis, M. E., & Tarabanis, K. (2023). *Explainable Graph Neural Networks on Linked Statistical Data for Predicting Scottish House Prices.* In *Proceedings of the 27th Pan-Hellenic Conference on Informatics (PCI 2023)* (pp. 48–56). ACM. https://doi.org/10.1145/3635059.3635065
19. Comber, S., & Arribas-Bel, D. (2017). *"Waiting on the train": The anticipatory (causal) effects of Crossrail in Ealing.*Journal of Transport Geography, 62, 92–104. https://doi.org/10.1016/j.jtrangeo.2017.08.004
20. Transport for London. (2012, February). *London Overground: Impact Study.* London: TfL. Retrieved from https://content.tfl.gov.uk/Item08-020212-Board-London-Overground-Impact-Study.pdf
21. Greater London Authority. (2022). *Economic impacts under future funding scenarios for TfL.* London: GLA Economics. Retrieved from https://www.london.gov.uk/business-and-economy-publications/economic-impacts-under-future-funding-scenarios-tfl
22. Hatch. (2024, November). *Transport for London Supply Chain: Economic Impact Assessment 2023/24*. Transport for London. Retrieved from https://content.tfl.gov.uk/tfl-supply-chain-economic-impact-assessment-2023-24.pdf
23. Ma, D., He, F., Yue, Y., Guo, R., Zhao, T., & Wang, M. (2024). *Graph convolutional networks for street network analysis with a case study of urban polycentricity in Chinese cities*. *International Journal of Geographical Information Science*. https://doi.org/10.1080/13658816.2024.2321229
24. Na, J., Nam, Y., Yoon, S., Song, H., Lee, B. S., & Lee, J.-G. (2025). Mobility networked time-series forecasting benchmark datasets. In Proceedings of the International AAAI Conference on Web and Social Media (Vol. 19, pp. 2539–2549). Association for the Advancement of Artificial Intelligence. https://doi.org/10.1609/icwsm.v19i1.35955
25. Beecham, R. (n.d.). *Class 5 — Commutes, connectivity and flows* [Web page]. Retrieved June 18, 2025, from https://www.roger-beecham.com/vis-for-gds/class/05-class/
26. Schwarzbart, J. (2021, May 27). *Analysis of public transport connectivity and different places of residence in Munich: Described by rental prices and the Isar* (Bachelor's thesis, Ludwig-Maximilians-Universität München). https://doi.org/10.5282/ubm/epub.77437