Thesis Title

Thesis Subtitle

The-Huan Hoang
Supervised by Dr. Elsa Arcaute

A dissertation presented in partial fulfillment of the
requirements for the degree of
Master of Science in Urban Spatial Science

Bartlett Centre for Advanced Spatial Analytics (CASA)
University College London
United Kingdom
23 August 2024

Abstract

Abstract goes here...

Declaration

I, The-Huan Hoang, declare that the work presented in this dissertation is my own. Where claims, information or findings have been derived from other authors and sources, I confirm that this has herein been indicated in writing, attributing due credit. In the spirit of fostering reproducible research, all scripts used for analysis and their outputs are made available on GitHub at https://github.com/shaunhoang/casa-tfl-dissertation.

Acknowledgements

I am grateful for the guidance and insightful feedback from my supervisor, Dr. Elsa Arcaute, who has shaped the direction of my work from its inception. I would also like to acknowledge the invaluable support from the Public Transport Service Planning team at Transport for London. Their expert advice and real-world insights into the data science behind transport planning have greatly enriched my knowledge, as well as this research project.

Contents

List of Figures

List of Tables

1. Introduction

2. Literature Review

Estimating trip attraction is a crucial topic of the transport planning academic literature and profession, constituting an integral component of the demand forecast framework, termed the Four-Step Model (FSM), which has been in ubiquitous use since the 1950s (*Travel Forecasting Resource* 2024). More specifically, the FSM begins with the Trip Generation step, where the number of trips generated by each traffic analysis zone is estimated, and Trip Distribution allocates these trips to their destination zones before proceeding to mode choice determination and route assignment. Planners accomplish the task of Trip Distribution using a variety of spatial interaction models, most prominent among which is the classic *Gravity Model* and its derivatives inspired by Newton's law of universal gravitation (Erlander and Stewart, 1990). Originally formulated to estimate flows among cities in a region, it has been adapted to intraurban mobility and suggests that the number of trips between two locations is proportional to the difference in *mass* of each location and inversely proportional to the distance between them.

However, it is worth noting that the definition of *mass* in any gravity-based spatial interaction model has seen many interpretations and experimentations, depending on the types of travel demand of research interest. For example, a spatial interaction model for commuting trip forecasts may consider the emissivity of residential areas as origins and the attractiveness of employment centres as destinations. In fact, the estimation of trip attraction in transport planning literature has predominantly revolved around commutes (i.e., primary travel demand) from residential zones to zones where employment centres or education institutions are located. On the one hand, this choice is a practical one first and foremost, as commuting trips represent a large share of regular intracity mobility flows, whose high volume within short peak periods has important implications on road capacity management or public transit operations, among others. Due to its importance, the calibration of the spatial interaction model for commuting flows is further aided by nationwide flow datasets, such as that present in the UK Census Data[1], which provides detailed and explicit Origin-Destination data to serve as ground truth. In the absence of this declared data, researchers have also turned to spatial features to calibrate commute trip distribution, as explored by Yang et al., 2014.

On the other hand, a plethora of potential trips generated for other purposes is often underresearched, partially due to the complexity of non-commute trip purpose identification, classification and quantification: Non-commute trips can be defined as trips that are not carried out to fulfil time-bound and regular work or education obligations and can include trips for shopping, leisure, and social activities. Therefore, as opposed to commuting trips, Non-commute trips are highly irregular both spatially and temporally, i.e., they can vary in terms of distance, duration, and mode of transport from one to another, even for the same destination between different days. Nevertheless, Non-commute trips are an important component of urban travel behaviour and can have a significant impact on the overall transportation system. Advances in the field have brought about improvements to methodologies such as travel diary surveys[2], which collect data on personal and household travel behaviour, including Non-commute trips, to allow for comprehensive planning. By design, however, they are limited in scale and generalisability.

Overall, this reveals a gap in our understanding of the comprehensive travel behaviour of city dwellers on an aggregate scale. With the number of trips for shopping, leisure, and social activities increasing in recent years and the number of work-related trips decreasing as flexible

[1] An interactive visualisation of the UK Census 2021 origin-destination data can be found at https://www.ons.gov.uk/visualisations/censusorigindestination/

[2] An example of these is the UK National Travel Survey, accessible at https://www.gov.uk/government/collections/national-travel-survey-statistics

work arrangements become more prevalent (Wöhner, 2022), with recognised health benefits to workers (MacLeod, Cole, and Musselwhite, 2022), it is important to understand the factors that influence non-commute trip attraction and to incorporate them into transport planning models. Within the scope of this research, we hypothesise that a given urban area's non-commute trip attracting *mass* is associated with the built environments, which include their amenity and public transport accessibility profiles.

2.1 Urban amenities as attractors

Cervero and Kockelman (1997) was one of the foundational works to provide an econometric estimation of how three groups of factors pertaining to the built environments—namely Density, Diversity and Design—are associated with travel demand to a traffic analysis zone for different purposes (work/non-work) and with different mode choices (private vehicles/public transport). This study is novel for demonstrating that changes in the built environment—such as density, diversity, and walkability have a tangible impact on travel behaviour, which supports the idea that designing more compact, mixed-use, pedestrian-friendly neighbourhoods can significantly influence how people choose to travel. The components of density and diversity refer to the number and the variety of destinations available in a given area, respectively, while the design component refers to the physical layout of the area, including the presence of sidewalks, bike lanes, and other pedestrian-friendly infrastructure. It is worth noting that the study considered the distance as a control impedance factor—adhering to the classic Gravity Model paradigm—and did not investigate how these factors may interact with each other.

There have been many recent attempts to link the presence of urban amenities to trip attraction and calibrate how we define urban attractors. Beyond pure correlation observed between the density of points of interest (POIs) in a region and intensity of trips made to a given area (Melikov et al., 2021), the typology of the POIs differs between urban clusters that see distinct patterns of inflow trips made. For example, after Aaqib Javed et al. (2020) have clustered urban areas in one city into Global (e.g., airports), Downtown (d.g., central business district) and Residential attractors based on the volume, spatial dispersion and distance of the trips made using O-D data from mobile phone, they found that the differences in the types of POIs in each of these clusters were also statistically significant. The typology of amenity POIs explored in this study included not only Retail, as is common when thinking about non-commute trips, but also hospitals, public services, restaurants, religious institutions, and other categories. This suggests that the type of urban amenities available in an area can influence the volume of trips made to that area and that different types of amenities may attract different types of trip purposes.

Lastly, we must also consider the aeminity diversity factor, apart from the typology of the ameinities themselves, in the amenity profile of areas as a potential attractor, following the conclusion made by Cervero and Kockelman (1997). More specifically, for non-commute trips where locations are not necessarily prescribed, and the destination choice can be influenced by personal perferences and circumstantial convenience, people may be more likely to visit more vibrant areas that offer a variety of services and activities where multiple tasks could be carried out, as opposed to areas where one type of destinations is overrepresented. This subjective perception of quality of urban areas is consistent with the concept of 15-minute cities, which purports that mixed-use areas with a diversity of ameinities within a 15-minute walking distance are more attractive an area to live in, and by extension to visit, than single-use and exclusive ones. (Khavarian-Garmsir et al., 2023)

2.2 Transport accessibility as attractors

If the density and diversity of amenities were the only explanatory factors for non-commute trip demand, it would be tempting to conclude that the local configuration of destination points of interest is the sole determinant of why people go to one place and not another for their diverse needs. In reality, neighbourhoods do not exist in a vacuum but are intricately connected to one another by the transportation network. In simple terms, the consideration of distance as friction in the classic Gravity Model attests to the fact that the more accessible a destination is, the more likely it is to attract travel demand, especially trips with higher elasticity of choice such as non-commute (i.e., trips made to a specific destination not based on obligation, but mainly out of convenience). Public transport accessibility has been shown to contribute to the growth of intercity importance of certain areas as opposed to others and preserve travel demand variability over time (Zhong et al., 2015), which means that it reinforces the prominence of well-connected areas as urban attractors, presumably with similar density and diversity of destination points of interest.

Jayasinghe, Sano, and Rattanaporn (2017) went one step further to purport that the network centrality property of an urban area is enough to predict travel demand to a high degree of accuracy without the need to consider other attractors such as employment centres and amenities, thanks to its tight correlation. More specifically, the study looked at four centrality measures: Connectivity, Choice, Global Integration and Local Integration, which are analogous to the concept of degree centrality, betweenness centrality, closeness centrality, and closeness centrality in a confined radius, respectively. Among them, Global Integration (closeness centrality) was found to be the most significant predictor, suggesting that the role of the transport network in shaping urban travel demand at the aggregate scale is paramount, perhaps just as important as socio-economic factors (Convery and Williams, 2019)

2.3 Public transport and the walkable city

Walkability contributes to the quality of life in cities. The 15-minute city is one of the most recent urban planning concepts that has gained traction in recent years. The idea is to create more sustainable and resilient cities by reducing the need for long commutes and promoting active modes of transport such as walking and cycling. A resident who lives in a neighbourhood that has all the amenities they need within a 15-minute walk or cycle can ideally reduce their reliance on cars and public transport to a minimum. However, this does not mean that the role of public transport is diminished within the 15-minute city vision. On the contrary, public transport is an essential component of a sustainable system, which provides an efficient and affordable way for people to travel longer distances, effectively agglomerating a larger urban system made up of ideally 15-minute cities. In other words, public transport is the backbone of the 15-minute city, connecting neighbourhoods and enabling people to travel further than they can on foot or by bike. This, in turn, has multiple implications for the local communities:

- From the lens of urban planning and traffic management, the higher the volume of public transport trips made to an area, the higher the amount of pedestrian traffic it receives on top of the local population who patronises their local amenities. Understanding the trip attraction factors can help prioritise investments in public transport capacity planning, pedestrian or cycling infrastructures and services for these individuals, which can lead to a virtuous cycle of increased active travel and reduced car dependency.

- From the lens of fomenting local economies, increased pedestrian traffic injected by public transport can also be seen as an additional source of footfall, which is invaluable for businesses with storefronts. Findings on the economic benefits of walking and cycling done by Transport for London[3] showed that retail rents increased by 7%, and office space rents increased by 4% in areas that encouraged non-motorised access, suggesting that the street improvements translated into much more desirable spaces.

It is also worth mentioning that footfall estimation itself is a prolific field of research within retail analytics, which counts on a variety of technologies such as WiFi tracking, video analytics, and mobile phone data to estimate the number of people who pass by a given location. The insights derived from footfall data can be used to inform decisions on store location, opening hours, staffing levels, and marketing strategies, among others. Uncovering patterns of public transport trip attraction, which in turn is an indicator of pedestrian influx, could be used to complement footfall data and provide a more comprehensive understanding of the factors that influence the number of people in a given area.

2.4 Explainable Artificial Intelligence for spatial systems

Up to this point, our visited literature has shown that both urban amenities and transport accessibility are important factors in determining the attractiveness of an area to non-commute trips, albeit in isolation. In reality, the interaction between these two factors has not been well studied in conjunction. It is possible that the dense presence of diverse urban amenities is brought about by preexisting favourable transport accessibility or vice versa. In the interest of addressing the research questions, we are presented with a dilemma of interpretability versus prediction accuracy in the choice of modelling techniques.

Econometric methods, commonly used in policy analysis, can help us quantify and interpret the relationship between the dependent variables and independent variables in the form of explicit coefficients. However, even spatially aware methods such as Spatial Lag Model (SLM) or Geographically Weighted Regression (GWR) are limited in their linear assumptions of linearity, issues with multicollinearity, and thus may fail at attaining high prediction accuracy (Wheeler and Tiefelsdorf, 2005).

In the meantime, machine learning methods have gained popularity in the field of urban mobility research due to their ability to capture complex, non-linear relationships between spatial features with high prediction accuracy. However, the lack of interpretability of machine learning models has been a major drawback, especially in the context of urban planning, where the ability to understand the underlying factors that drive the model's predictions is crucial for decision-making. Among many opportunities to develop Machine Learning explanation techniques for causal inference in urban mobility is the need to evaluate causal effects of input features and derive actionable insights (Xin et al., 2022).

To address the interpretability challenge, local interpretation methods like SHAP (SHapley Additive exPlanations)—a method based on cooperative game theory based on the seminal work by Lundberg and Lee, 2017—to provide detailed attributions of input features to individual predictions made. Its introduction contributed to the growing body of literature on interpretable machine learning with spatial systems for regression with XGBoost (Li, 2022) or classification with Light GBM (Louhichi et al., 2023). One of the most notable recent examples of SHAP for

[3]https://tfl.gov.uk/corporate/publications-and-reports/economic-benefits-of-walking-and-cycling

urban mobility in action was the development of the Deep Gravity Model introduced by Simini et al., 2021, which used SHAP to explain the predictions of a deep learning model for spatial interaction and how the model's predictions were influenced by the input spatial features such as land-use mix, points of interest in the origin and destination zones.

As we approach this research with an eye on accurately modelling the real-world phenomenon and explaining the contributing factors, instead of GWR or SLM, we will use a machine learning model to predict non-commute trip attraction in urban areas and then leverage SHAP to explain the predictions in an attempt to quantify the relative importance and contribution of urban amenities and transport accessibility.

2.5 Open data and reproducibility

Many recent advances in the field of urban mobility research have been made possible by the availability of proprietary datasets owned by private companies, such as mobile phone data, GPS data, and social media data. Although the high data granularity has opened doors to new research opportunities and the development of new advanced techniques, its use also raises concerns about data privacy, data ownership, and data sharing. Moreover, it heightens the barrier to entry for independent researchers who would like to replicate the studies for their respective cities or regions. In contrast, open data not only democratises independent and participatory research in the public interest but also breaks down many reproducibility barriers when applying established methodologies to other localities (Yadav et al., 2017).

On one side, we have readily available open datasets maintained by local or federal government agencies in many countries. On the other, we have global platforms like **OpenStreetMap** and **Overture Maps Foundation** that provide a wealth of spatial data for cities around the world. The use of open data to foment reproducibility will be a key focus of this research to ensure that the findings can be replicated in other cities and regions, unburdened by commercial data availability and access barriers. It also enables flexible preprocessing of the amenity or transportation features to model the intracity mobility behaviour that is pertinent to the locality's needs, while adhering to the same model training and analysis methodology. These needs might include a different classification of the amenity points of interest that correspond to different trip purposes, or a different categorisation of the services in the public transport network.

Lastly, adhering to the spirit of openness and future reproducibility across different urban contexts, all of the datasets used in this research will be entirely sourced from open data hubs and platforms. The associated scripts used for data processing, model training and prediction analytics are available on GitHub.

2.6 Summary

In order to address the research questions that we have posed, informed by the literature review thus far, we intend to tune, evaluate and select a suitable machine learning model to predict non-commute trip attraction in urban areas, where we hypothesise that the built environment, including urban amenities and transport accessibility, are important factors. The use of a machine learning model to represent these realtionships, in conjunction with explainable machine learning techniques such as SHAP, allows us to capture complex, non-linear relationships between features and achieve high prediction accuracy of the target variable, as well as analyse attributions of input features to the predictions made in order to uncover underlying contributions.

3. Methodology

In this section, we will first describe the methodology used to preprocess data from various sources to engineer the necessary features for the predictive model. Second, we will outline the model selection and evaluation process, and finally, we will discuss the methodology used to interpret the model predictions using SHAP techniques. For ease of reference, the data preprocessing and feature engineering pipeline is summarised in Figure 3.1, and an overview of the model training and prediction interpretation methodology can be found in Figure 3.2.

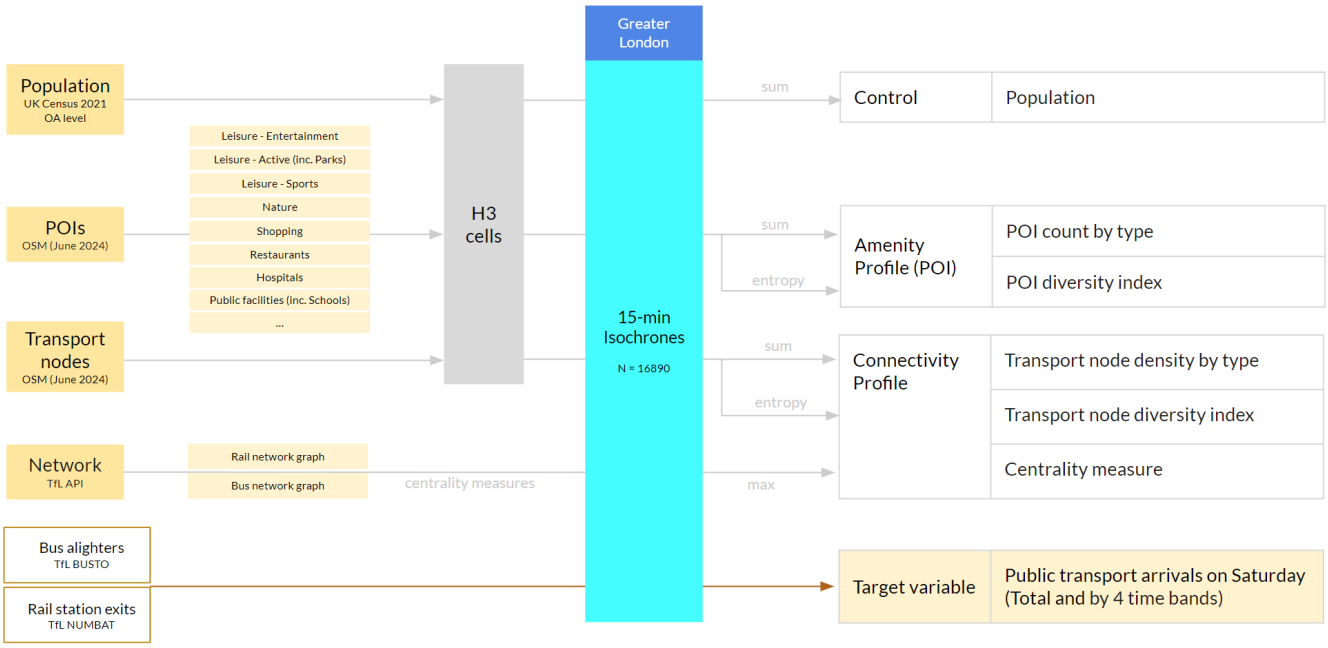

Figure 3.1: Overview of feature engineering pipeline

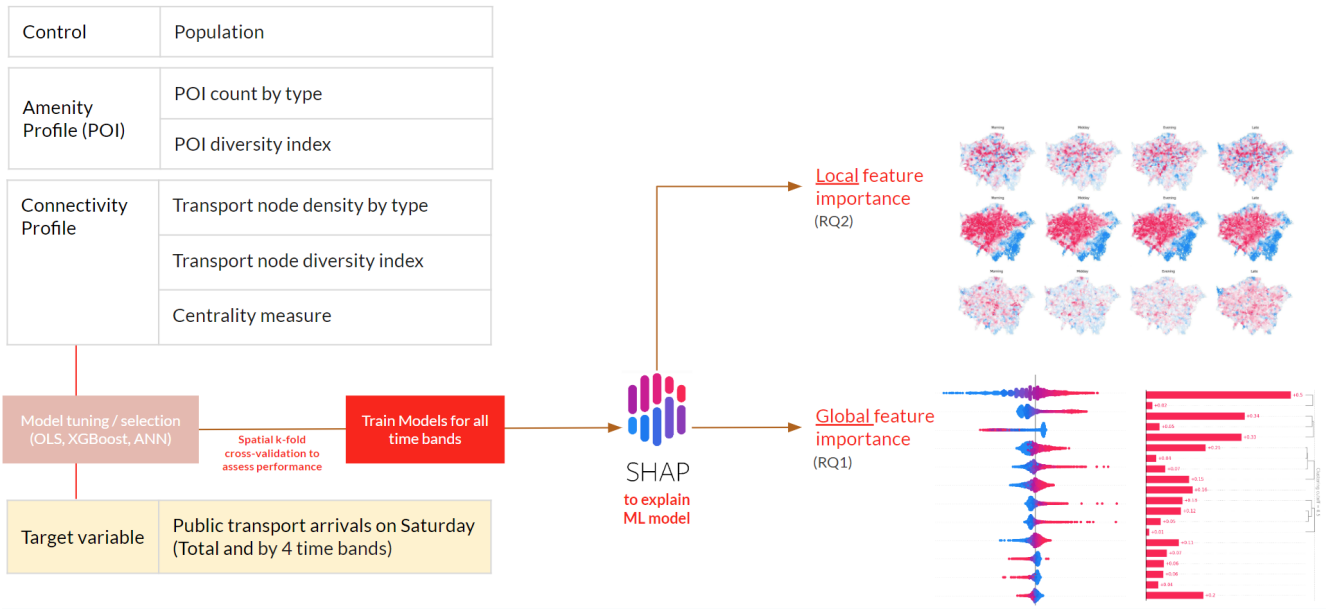

Figure 3.2: Overview of model training and interpretation

3.1 Open data sources

The analysis carried out in this dissertation will utilise the Greater London area as the working case study, serving as an example for further replications of the analysis for other localities. The Office for National Statistics (ONS) defines the Greater London metropolitan area to include the City of London and the 32 London boroughs, with a total area of 1,572 km^2 and a population of 9.4 million people, making it the largest in the United Kingdom. The choice of the Greater London metropolitan area is also motivated by the availability of open data sources from which necessary features could be derived and the complexity of the public transport network that is coupled with rich usage data made available to the public. The data sources used in this analysis are summarised in Table 3.1 and detailed below.

	Variables	**Sources**
Public transport arrivals (target)	Tube and Rail station exits	TfL network demand (NUMBAT)
	Bus alightings	TfL network demand (BUSTO)
Population	Population count	UK Census 2021
Amenity Profile	POI count by type	OSM Amenity POIs
	POI diversity index	
Public Transport (PT) Connectivity Profile	PT node count by type	OSM Transport POIs and TfL
	PT node diversity index	
	PT network centrality	

Table 3.1: Summary of data used in the analysis and their sources

Transport for London (TfL) Open Data

We use the following GIS datasets from the TfL GIS Open Data Hub and TfL API to derive transport-related features for the study area including density of transport facilities, diversity of transport options, and accessibility in the form of centrality measures in the Rail and Bus networks separately. The networks of interest are defined as follows:

- *Rail network* includes the London Underground, London Overground, the Docklands Light Railway, and the Elizabeth Line, excluding stations and segments that are not managed by TfL or fall outside the Greater London area.
- *Bus network* includes stop locations and route data of all bus and tram services, excluding stops and route segments that are not managed by TfL or within the Greater London area.

TfL also provides annually-updated network demand datasets, which will be used to derive the target variable of our analysis. The datasets are special interest to our analysis are namely *Rail demand data (NUMBAT)* and the *Bus demand data (BUSTO)*, from which we will extract estimated counts of station exits and bus alightings at the station or stop level, respectively. The dataset's temporal granularity is also an important aspect to consider. To address non-commute trips for the analysis, we will specifically use **Saturday average demand**, representing

the demand in a typical period without any major citywide disruptions or events[1]. To examine possible variations between different times of day, besides the total daily counts of exits and bus alightings, we will aggregate the data into four larger time bands[2]:

- *Morning* (05:00 - 10:00): Limited mobility expected due to early hours
- *Midday* (10:00 - 19:00): High daytime mobility expected
- *Evening* (19:00 - 00:00): High evening mobility expected
- *Late* (00:00 - 05:00): Minimal mobility expected due to sparser night PT services

It is worth noting that despite the existence of the Oyster Card integrated fare system, TfL does not provide origin-destination data for individuals as open data. Therefore, station exits and bus alightings are not directly associated with the number of passengers traversing through the system but are treated as discrete events. For example, a person who takes a bus to reach a Tube station for onward travel and then exits to arrive at their destination would be counted as two separate data points, one alighting for the bus and one station exit for the Tube. Nevertheless, our model will take this into account when making poreDiction and allows us to investigate patterns and the contribution of modal interchange behaviour in our interpretation.

Lastly, we need to acknowledge the particularity of the TfL bus demand dataset (BUSTO). Although passengers do not 'tap off' when alighting from busses, TfL has developed a methodology to estimate the number of passengers alighting at each bus stop based on individuals' next actions and other statistical methods. The methodology is described in detail in the BUSTO User Guide and Data Dictionary

UK Census Data 2021

For the estimated population datasets, we rely on the most recent United Kingdom Census data administered in 2021. The data is ingested at the Output Area (OA) level, which is the smallest geographical unit for which census data is available. Further processing will aggregate the data to the desired spatial unit of analysis.

OpenStreetMap POIs

OpenStreetMap (OSM) is a collaborative mapping project that provides a free and open-source map of the world. The version of OSM data used in this analysis is packaged by an OpenStreetMap community project Geofabrik and provides a snapshot of the OSM database as of June 2024. Geofabrik also standardises the categorisation of POIs' amenity attributes in the OSM database into larger classes[3]. Our analysis will use this default POI classification, extracting 12 amenity POI types. Lastly, apart from amenity POIs, we also make use of the OSM transport POIs point data to derive non-TfL transport-related features, such as national rail stations, ferry piers, long-distance bus stations and airports.

Combined with transport node data acquired from TfL open data sources, all extracted Amenity and Transport POI types of interest are summarised in Table 3.2.

[1]This Saturday travel demand may still include work trips for those working on Saturdays, albeit not concentrated into peak periods as with weekday travel demand. For the scope of this analysis, we will treat these potential work trips as non-commute.

[2]The choice of time band delineation is a compromise between different granularity levels afforded by the TfL rail and the bus demand datasets. There are opportunities to deepen the analysis by further segmenting the time bands into finer intervals, if data permits.

[3]More information about Geofabrik's OSM data product can be found in Geofabrik OSM data dictionary

	POI Type	Examples
Amenity	*Public Facilities*	Post offices, schools, universities, libraries,...
	Medical	Hospitals, clinics, pharmacies,...
	Entertainment	Theatres, Night clubs, cinemas,...
	Outdoors	Parks, playgrounds,...
	Active	Sports centres, swimming pools, stadiums,...
	Restaurants	Restaurants, cafes, pubs, bars,...
	Hotels	Hotels, motels, hostels,...
	Shopping	Supermarkets, bakeries, florists, bookshops, malls,...
	Banking	Banks, ATMs,...
	Tourism	Tourist attractions, museums, monuments, viewpoints,...
	Religious	Churches, mosques, temples,...
	Nature	Riverbanks, beaches, hills,...
Transport	*Bus stops*	Bus stops and stations (TfL)
	Rail stations	TfL rail stations (TfL) and Non-TfL rail stations (OSM)
	Other transport hubs	Airports, ferry piers, long-distance bus stations (OSM)

Table 3.2: Summary of Amenity and Transport POI types used in the analysis

3.2 Defining the spatial unit of analysis

As features will be extracted and engineered from different data sources, the spatial unit is an important consideration, at which the features are aggregated, and the Machine Learning model is trained. In this analysis, we will use 15-minute walking isochrones—i.e., the area that can be reached within a 15-minute walk from a given point—as the spatial unit of analysis. The isochrones were generated based on initial seeds of 16,890 points across the Greater London area, using the pedestrian profile of the OpenStreetMap Network API (osmnx), which takes into account pedestrian infrastructure such as footpaths, pedestrian crossings, and pedestrianised areas. The choice of using isochrones as the spatial unit of analysis for our model is motivated by the following reasons:

- First, the spatial unit of aggregation and analysis for the model must reflect how individuals interact with transport nodes and amenity POIs that are accessible using the existing pedestrian network. Proximity using the Euclidean distance is inaccurate when there are natural barriers such as rivers, or private estates that prevent through access.
- Second, human mobility in cities is not bound by administrative boundaries such as Output Areas or boroughs, and the use of overlapping isochrones allows us to capture the spatial heterogeneity of the urban environment in our dataset.
- Third, the 15-minute specification of the isochrones aims to reflect a uniformity among the spatial units of analysis while following a predetermined standard set out by the 15-minute city concept that characterises the extent of desirable mobility in a neighbourhood.
- Lastly, the density of the isochrones produced is a compromise between the granularity of the spatial unit and the computational resources required to process the data. The initial seeds of 16,890 points are originally centroids of the same number of H3 cells at resolution 9 needed to cover the Greater London area.

The detailed methodology to generate the isochrones is illustrated in Figure 3.3, and two example isochrones in Central London can be found in Figure 3.4. It is worth noting that depending on where the spatial unit is located in the Greater London area, the isochrones may vary in shape and size due to the underlying pedestrian network and the distribution of transport nodes and amenity POIs. However, these variances are expected and even desirable as they more closely reflect pedestrian accessibility to different types of amenities and transport nodes. All raw data extracted from various sources will be aggregated and processed at the isochrone level to generate the necessary features and target variable for the model training (See Figure 3.1). In the next section, we will describe the data transformation and feature engineering process in more detail. Note that the terms spatial units (of analysis), isochrones, or simply areas may be used interchangeable from this point on in our discussion.

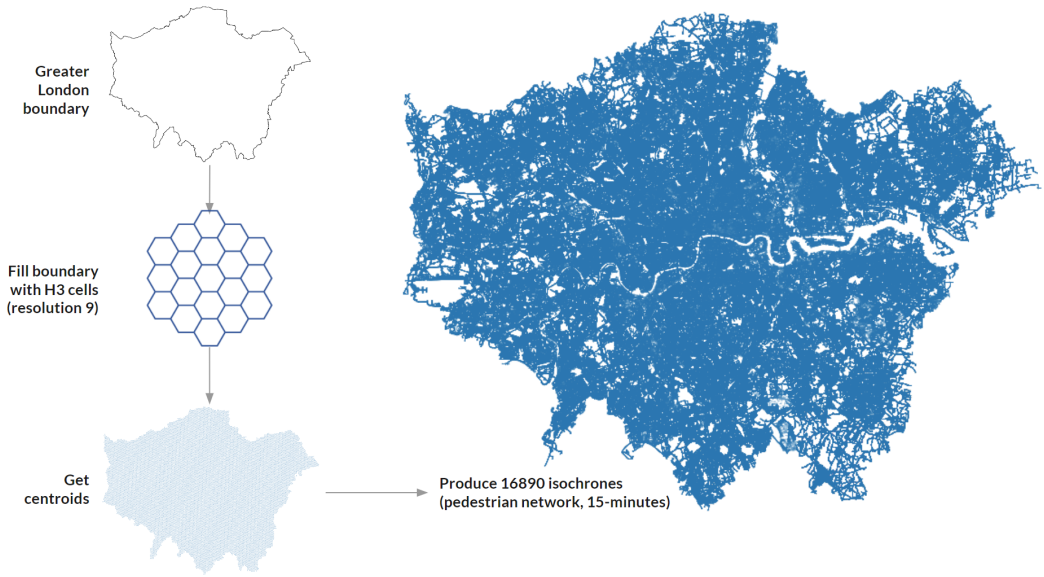

Figure 3.3: Methodology to generate 15-minute walking isochrones as the spatial unit of analysis

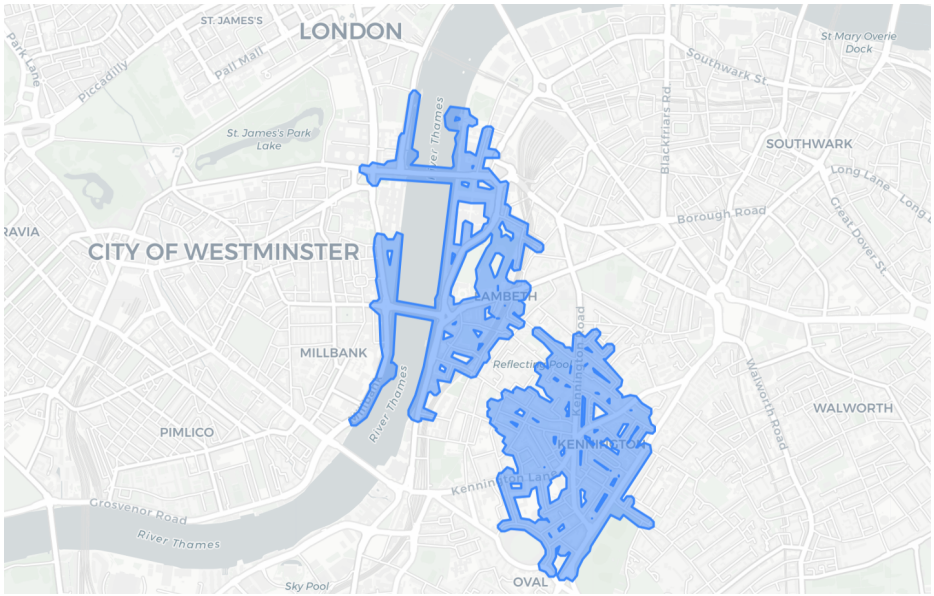

Figure 3.4: Two example spatial units of analysis located in Central London

3.3 Data Preprocessing

3.3.1 Aggregating target variables

The total numbers of passengers who arrive at a transport node or alight from a bus at a stop are the target variables for our model, used to train the model to predict the number of passengers arriving at a transport node or alighting from a bus at a given isochrone. The aggregation results are illustrated in Figure 3.5, which shows the total daily figures as well as the figures aggregated into the four time bands described earlier in order to capture the temporal variations. The total daily figures show a clear concentration of passenger arrivals in the Central London area, with the highest number of passengers arriving at transport nodes in the morning and evening time bands. The time band figures show a similar pattern, with the highest number of passengers arriving during the Midday time band (10:00-19:00). The Evening time band also shows a significant number of passengers arriving at transport nodes, while the Morning and Late time bands show lower numbers of passengers using the system, due to early hours in the day on weekends and sparser night services, respectively.

Lastly, since the target variable has a highly right-skewed distribution, we will apply a log transformation to the target variable to normalise the distribution and improve the model's performance. The log transformation of the target variable is shown in Figure 3.6, which shows a

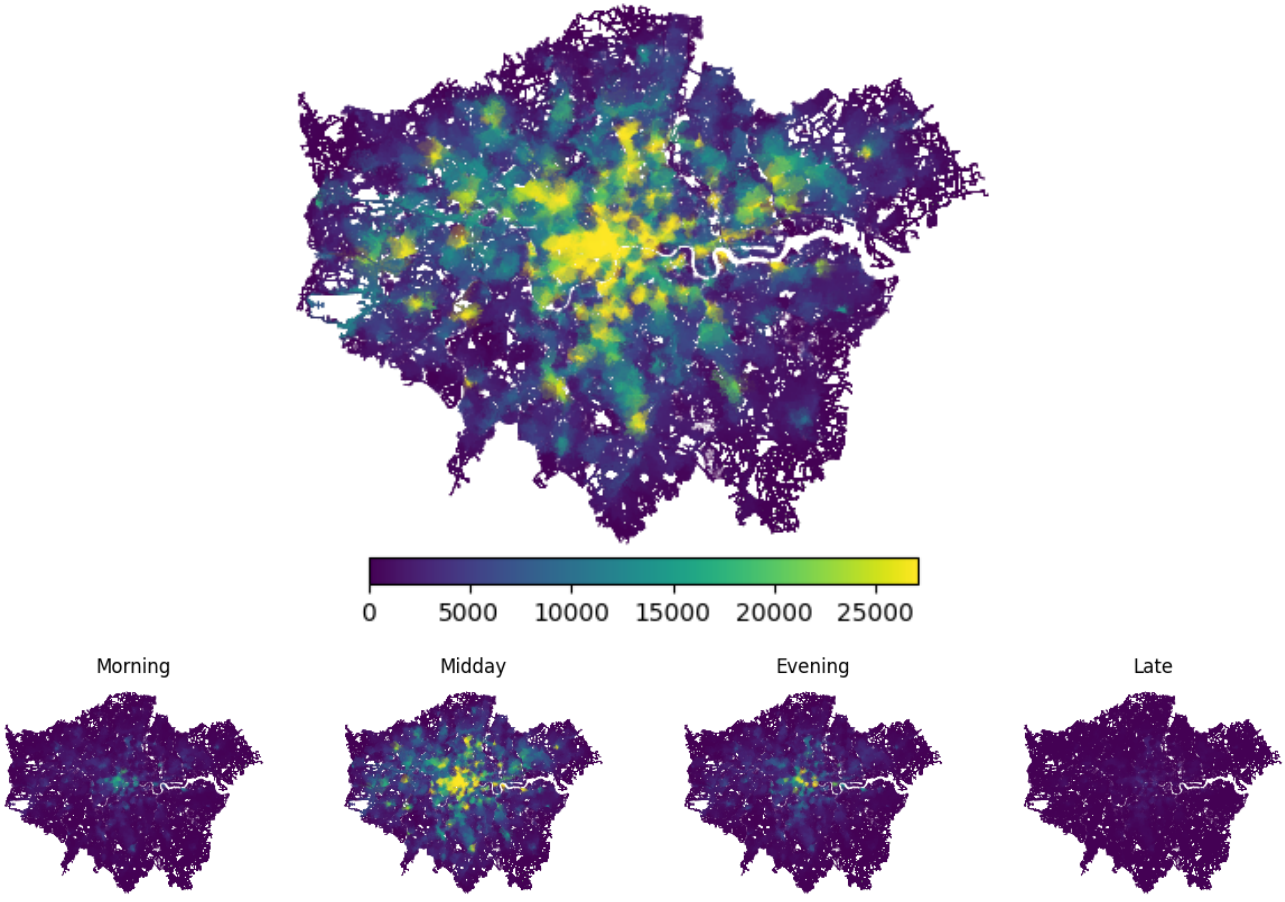

Figure 3.5: Aggregated arrivals at the chosen spatial unit level
by total daily (top) and by time band (bottom)

more normal distribution of the target variable after the transformation. The log-transformed target variable will be used in the model training and evaluation process, and will be implicit from this point on.

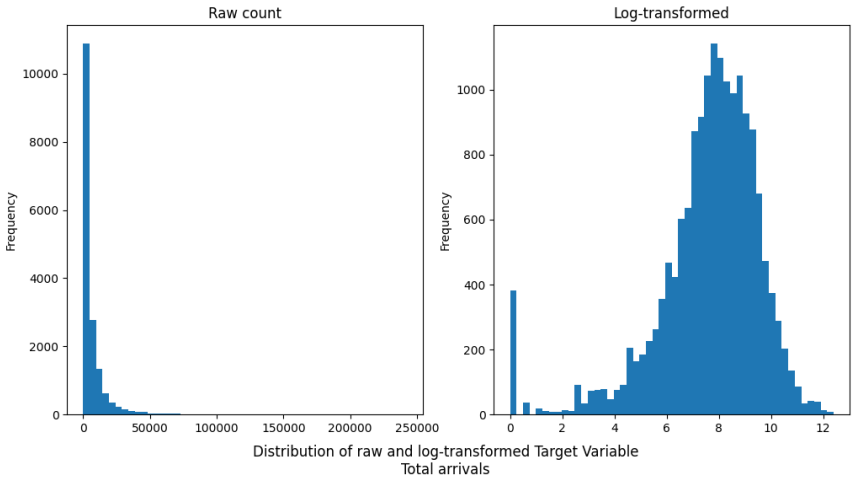

Figure 3.6: Distribution of the Total arrivals as raw count (left) and log-transformed (right). Similar transformations carried out for arrivals by each of the four time bands

3.3.2 Feature engineering

Density

Density of POIs is one of the first feature type of interest for our model (Cervero and Kockelman, 1997), which represents the presence intensity of amenities or transport nodes of each type in an area that would attract visitors (in our case, coming by public transport). We expect a positive correlation between the density of POIs of all types to the total number of arrivals. An area with a high concetration of retail within a 15-minute walk is expected be more attractive than a similar-sized area with a low concetration of retail. Similarly, an area with a high concentration of transport facilities such as stations, bus stops are expected to be accommodating a higher inflow volume of visitors. However, feature importance and spatial nonstationarity (different impact on the target variable depending the location) in this regard will be factors of interest as we interpret the model predictions post-training.

All density-related features—i.e., population, amenity POI density and transport node POI density—are formulated by aggregating their values for each spatial unit (isochrone) by type, resulting in a set of **16 features** (See Figure A.2 in the Appendix)

Diversity

Diversity of Amenity POIs according to Cervero and Kockelman (1997) is also an important factor that may influence attractiveness of an urban area for certain type of non-commute travel demand. The implication from their conclusion is that between two areas with all else being equal, the area that is more diverse in the types of amenities it provides will be more attractive on average, because of their vibrancy and the possibility it can affords visitors to accomplish different tasks in one trip.

- Transport POI diversity

Cervero and Kockelman (1997) conceptualised diversity in the form of *entropy*. A concept originated in information theory, entropy is the average amount of information conveyed by an event when considering all possible outcomes. Batty et al. (2014) further adapted it to convey complexity in a spatial system. Interestingly, they also considered the increasing spatial entropy to mean a trade-off between density of amenity types and the unique types of amenity present, which stands in competition with other density-based features created. Therefore, the inclusion of entropy as a measure of POI diversity may improve the model prediction accuracy.

Although there are many formulation for the calculation of entropy, we intend to replicate the methodology used by both of the cited works above and adopt the formulation of entropy introduced by Shannon (1948) as follows:

- formula

Mapping

- map

Centrality

- Which 3 centrality measures are used
  - Why not multi-Layered network

3.3.3 Incorporating spatial lags as feature

When working with machine learning models with data exist in a closed spatial system such as the Greater London area, it is important to consider the spatial autocorrelation in the target variables. Spatial autocorrelation refers to the phenomenon where the value of a variable at one location is correlated with the value of the same variable at nearby locations. In the context of our analysis, this means that the number of passengers arriving at a transport node or alighting from a bus at one isochrone is likely to be correlated with the number of passengers arriving at nearby isochrones. Although deep learning models' parameters and hyperparameters can be made more complex and tuned to reach a high degree of prediction accuracy without considering the spatial aspect of the dataset and can practically be considered non-parametric (i.e., not assuming the variables to have a specific distributions), accounting for potential spatial autocorrelation could reduce probability of biased outcomes even with a low-complexity and high-interpretability atchitecture, e.g. tree-based (Meyer et al., 2019).

The existence of spatial autocorrelation in the target variables is confirmed by the Local Indicators of Spatial Association (LISA) analysis, which shows the presence of spatial clusters in the target variables (Figure 3.7), using Local Moran's I indicators. High-High and Low-Low clusters of the total daily arrivals with public transport are respectively present in the inner core and the outer edge. Unless spatial relational information among these spatial units are made apparent to the model during tuning and evaluation, we are bound to see low accuracy espeically in these localities that make up more than 50% of the study area.

To address this issue, Liu, Kounadi, and Zurita-Milla (2022) has proposed incoporating spatial lag into the model as a feature. More specifically, the authors have concluded that when spatial features were induced in the random forest models, the global spatial autocorrelation among the predictions' residuals was significantly reduced, also leading to higher validation accuracy. For our analysis, the spatial lag feature of each spatial unit is calculated as the average number of passengers arriving at neighbouring isochrones[4]. The creation of a **Spatial Lag features** is replicated

[4]The choice of the distance band of 750m was calibrated with the objective to ensure all spatial units have

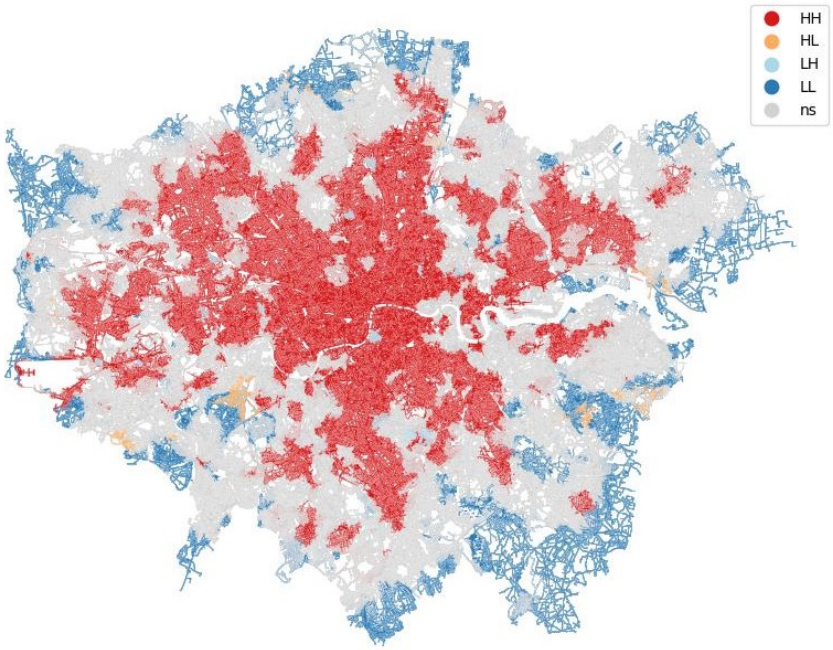

Figure 3.7: LISA cluster map of PT arrival volume (Total daily)
Global Moran's I: 0.742 (High spatial autocorrelation)

for the total daily arrival counts, as well as for each time band, resulting in a set of additional features that will be included for the training of the model in predicting the corresponding target variable,

No.	Target variable	Main Features		Spatial Lag Feature
1	Total arrivals	Population, amenity and transport profile features (25 features)	+	Lag - Total arrivals
2	Morning arrivals			Lag - Morning arrivals
3	Midday arrivals			Lag - Midday arrivals
4	Evening arrivals			Lag - Evening arrivals
5	Late arrivals			Lag - Late arrivals

Table 3.3: Mapping spatial lag features to corresponding models to predict target variables

neighbours, i.e., there are no islands.

3.4 Model selection and evaluation

- Comparing Linear, XGBoost, ANN
    - Preference for XGBoost, to be evaluated in Results

3.5 Explainable machine learning with SHAP

- What is SHAP
    - SHAP use cases for different machine learning models
    - SHAP for spatial
    - True to Data, True to Model
    - Limitations and comparisons to econometric models

4. Results and Discussion

4.1 Model evaluation results

4.2 Model interpretation

4.2.1 Global feature importance

4.2.2 Local feature importance

4.3 Hot spots

4.4 Discussions

4.5 Limitations

Figure 4.1: Example

Look at Figure 4.1 for an example.

If Casello and Smith, 2006 we want a page with no headers or footers except for a simple page number at the bottom we would use the keyword plain. However you need to be aware that using this command changes the page style for all the pages following the command. Therefore we need to turn the page style back to fancy as soon as we want the headers back.

If (Casello and Smith, 2006) we want a page with no headers or footers except for a simple page number at the bottom we would use the keyword plain. However you need to be aware that using this command changes the page style for all the pages following the command. Therefore we need to turn the page style back to fancy as soon as we want the headers back.

If Casello and Smith (2006) we want a page with no headers or footers except for a simple page number at the bottom we would use the keyword plain. However you need to be aware that using this command changes the page style for all the pages following the command. Therefore we need to turn the page style back to fancy as soon as we want the headers back.

This concludes our discussion on page layout. In the next post we'll look at using images and tables.If we want a page with no headers or footers except for a simple page number at the bottom we would use the keyword plain. However you need to be aware that using this command changes the page style for all the pages following the command. Therefore we need to turn the page style back to fancy as soon as we want the headers back.

This concludes our discussion on page layout. In the next post we'll look at using images and tables.If we want a page with no headers or footers except for a simple page number at the bottom we would use the keyword plain. However you need to be aware that using this command changes the page style for all the pages following the command. Therefore we need to turn the page style back to fancy as soon as we want the headers back.

This concludes our discussion on page layout. In the next post we'll look at using images and tables.If we want a page with no headers or footers except for a simple page number at the bottom we would use the keyword plain. However you need to be aware that using this command changes the page style for all the pages following the command. Therefore we need to turn the page style back to fancy as soon as we want the headers back.

This concludes our discussion on page layout. In the next post we'll look at using images and tables.If we want a page with no headers or footers except for a simple page number at the bottom we would use the keyword plain. However you need to be aware that using this command changes the page style for all the pages following the command. Therefore we need to turn the page style back to fancy as soon as we want the headers back.

This concludes our discussion on page layout. In the next post we'll look at using images and tables.If we want a page with no headers or footers except for a simple page number at the bottom we would use the keyword plain. However you need to be aware that using this command changes the page style for all the pages following the command. Therefore we need to turn the page style back to fancy as soon as we want the headers back.

This concludes our discussion on page layout. In the next post we'll look at using images and tables.If we want a page with no headers or footers except for a simple page number at the bottom we would use the keyword plain. However you need to be aware that using this command changes the page style for all the pages following the command. Therefore we need to turn the page style back to fancy as soon as we want the headers back.

This concludes our discussion on page layout. In the next post we'll look at using images and tables.If we want a page with no headers or footers except for a simple page number at the bottom we would use the keyword plain. However you need to be aware that using this command changes the page style for all the pages following the command. Therefore we need to turn the page style back to fancy as soon as we want the headers back.

This concludes our discussion on page layout. In the next post we'll look at using images and tables.If we want a page with no headers or footers except for a simple page number at the bottom we would use the keyword plain. However you need to be aware that using this command changes the page style for all the pages following the command. Therefore we need to turn the page style back to fancy as soon as we want the headers back.

This concludes our discussion on page layout. In the next post we'll look at using images and tables.If we want a page with no headers or footers except for a simple page number at the bottom we would use the keyword plain. However you need to be aware that using this command changes the page style for all the pages following the command. Therefore we need to turn the page style back to fancy as soon as we want the headers back.

This concludes our discussion on page layout. In the next post we'll look at using images and tables.If we want a page with no headers or footers except for a simple page number at the bottom we would use the keyword plain. However you need to be aware that using this command changes the page style for all the pages following the command. Therefore we need to turn the page style back to fancy as soon as we want the headers back.

This concludes our discussion on page layout. In the next post we'll look at using images and tables.If we want a page with no headers or footers except for a simple page number at the bottom we would use the keyword plain. However you need to be aware that using this command changes the page style for all the pages following the command. Therefore we need to turn the page style back to fancy as soon as we want the headers back.

This concludes our discussion on page layout. In the next post we'll look at using images and tables.If we want a page with no headers or footers except for a simple page number at the bottom we would use the keyword plain. However you need to be aware that using this command changes

the page style for all the pages following the command. Therefore we need to turn the page style back to fancy as soon as we want the headers back.

This concludes our discussion on page layout. In the next post we'll look at using images and tables.If we want a page with no headers or footers except for a simple page number at the bottom we would use the keyword plain. However you need to be aware that using this command changes the page style for all the pages following the command. Therefore we need to turn the page style back to fancy as soon as we want the headers back.

This concludes our discussion on page layout. In the next post we'll look at using images and tables.If we want a page with no headers or footers except for a simple page number at the bottom we would use the keyword plain. However you need to be aware that using this command changes the page style for all the pages following the command. Therefore we need to turn the page style back to fancy as soon as we want the headers back.

This concludes our discussion on page layout. In the next post we'll look at using images and tables.If we want a page with no headers or footers except for a simple page number at the bottom we would use the keyword plain. However you need to be aware that using this command changes the page style for all the pages following the command. Therefore we need to turn the page style back to fancy as soon as we want the headers back.

This concludes our discussion on page layout. In the next post we'll look at using images and tables.If we want a page with no headers or footers except for a simple page number at the bottom we would use the keyword plain. However you need to be aware that using this command changes the page style for all the pages following the command. Therefore we need to turn the page style back to fancy as soon as we want the headers back.

This concludes our discussion on page layout. In the next post we'll look at using images and tables.

5. Conclusion

A. Appendix

Population and POI Density Maps

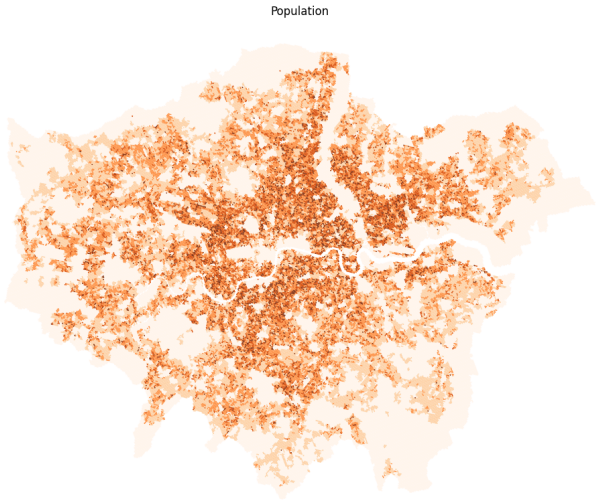

Figure A.1: Population density (UK Census 2011)

Figure A.2: Amenity and Transport POI density by type (Source: OSM)

Bibliography

Aaqib Javed, Syed et al. (Apr. 6, 2020). "Estimation of Trip Attraction Rates and Models for Shopping Centers in Dhaka City". In: DOI: 10.5281/zenodo.3733088.

Batty, Michael et al. (Oct. 1, 2014). "Entropy, Complexity, and Spatial Information". In: *Journal of Geographical Systems* 16.4, pp. 363–385. ISSN: 1435-5949. DOI: 10.1007/s10109-014-0202-2. URL: https://doi.org/10.1007/s10109-014-0202-2 (visited on 08/01/2024).

Casello, Jeffrey and Tony Smith (Dec. 1, 2006). "Transportation Activity Centers for Urban Transportation Analysis". In: *Journal of Urban Planning and Development-asce - J URBAN PLAN DEV-ASCE* 132. DOI: 10.1061/(ASCE)0733-9488(2006)132:4(247).

Cervero, Robert and Kara Kockelman (Sept. 1, 1997). "Travel Demand and the 3Ds: Density, Diversity, and Design". In: *Transportation Research Part D: Transport and Environment* 2.3, pp. 199–219. ISSN: 1361-9209. DOI: 10.1016/S1361-9209(97)00009-6. URL: https://www.sciencedirect.com/science/article/pii/S1361920997000096 (visited on 06/14/2024).

Convery, Sheila and Brendan Williams (Sept. 2019). "Determinants of Transport Mode Choice for Non-Commuting Trips: The Roles of Transport, Land Use and Socio-Demographic Characteristics". In: *Urban Science* 3.3 (3), p. 82. ISSN: 2413-8851. DOI: 10.3390/urbansci3030082. URL: https://www.mdpi.com/2413-8851/3/3/82 (visited on 06/18/2024).

Erlander, Sven and Neil F. Stewart (Dec. 1990). *The Gravity Model in Transportation Analysis: Theory and Extensions*. VSP. 252 pp. ISBN: 978-90-6764-089-3. Google Books: tId3PU1leR8C.

Jayasinghe, Amila, Kazushi Sano, and Kasemsri Rattanaporn (Oct. 1, 2017). "Application for Developing Countries: Estimating Trip Attraction in Urban Zones Based on Centrality". In: *Journal of Traffic and Transportation Engineering (English Edition)* 4.5, pp. 464–476. ISSN: 2095-7564. DOI: 10.1016/j.jtte.2017.05.011. URL: https://www.sciencedirect.com/science/article/pii/S2095756416302458 (visited on 05/17/2024).

Khavarian-Garmsir, Amir Reza et al. (Feb. 2023). "From Garden City to 15-Minute City: A Historical Perspective and Critical Assessment". In: *Land* 12.2 (2), p. 512. ISSN: 2073-445X. DOI: 10.3390/land12020512. URL: https://www.mdpi.com/2073-445X/12/2/512 (visited on 07/29/2024).

Li, Ziqi (Sept. 1, 2022). "Extracting Spatial Effects from Machine Learning Model Using Local Interpretation Method: An Example of SHAP and XGBoost". In: *Computers, Environment and Urban Systems* 96, p. 101845. ISSN: 0198-9715. DOI: 10.1016/j.compenvurbsys.2022.101845. URL: https://www.sciencedirect.com/science/article/pii/S0198971522000898 (visited on 07/07/2024).

Liu, Xiaojian, Ourania Kounadi, and Raul Zurita-Milla (Apr. 2022). "Incorporating Spatial Autocorrelation in Machine Learning Models Using Spatial Lag and Eigenvector Spatial Filtering Features". In: *ISPRS International Journal of Geo-Information* 11.4 (4), p. 242. ISSN: 2220-9964. DOI: 10.3390/ijgi11040242. URL: https://www.mdpi.com/2220-9964/11/4/242 (visited on 07/25/2024).

Louhichi, Mouad et al. (Jan. 1, 2023). "Shapley Values for Explaining the Black Box Nature of Machine Learning Model Clustering". In: *Procedia Computer Science*. The 14th International

Conference on Ambient Systems, Networks and Technologies Networks (ANT) and The 6th International Conference on Emerging Data and Industry 4.0 (EDI40) 220, pp. 806–811. ISSN: 1877-0509. DOI: 10.1016/j.procs.2023.03.107. URL: https://www.sciencedirect.com/science/article/pii/S1877050923006427 (visited on 07/24/2024).

Lundberg, Scott M and Su-In Lee (2017). "A Unified Approach to Interpreting Model Predictions". In: *Advances in Neural Information Processing Systems*. Vol. 30. Curran Associates, Inc. URL: https://proceedings.neurips.cc/paper/2017/hash/8a20a8621978632d76c43dfd28b67767-Abstract.html (visited on 07/24/2024).

MacLeod, Kara E., Brian L. Cole, and Charles Musselwhite (June 2022). "Commuting to Work Post-Pandemic: Opportunities for Health?" In: *Journal of Transport & Health* 25, p. 101381. ISSN: 2214-1405. DOI: 10.1016/j.jth.2022.101381. pmid: 35540370. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC9074865/ (visited on 07/24/2024).

Melikov, Pierre et al. (Apr. 1, 2021). "Characterizing Urban Mobility Patterns: A Case Study of Mexico City". In: pp. 153–170. ISBN: 9789811589829. DOI: 10.1007/978-981-15-8983-6_11.

Meyer, Hanna et al. (Nov. 2019). "Importance of Spatial Predictor Variable Selection in Machine Learning Applications – Moving from Data Reproduction to Spatial Prediction". In: *Ecological Modelling* 411, p. 108815. ISSN: 03043800. DOI: 10.1016/j.ecolmodel.2019.108815. arXiv: 1908.07805 [cs, stat]. URL: http://arxiv.org/abs/1908.07805 (visited on 08/01/2024).

Shannon, C. E. (July 1948). "A Mathematical Theory of Communication". In: *The Bell System Technical Journal* 27.3, pp. 379–423. ISSN: 0005-8580. DOI: 10.1002/j.1538-7305.1948.tb01338.x. URL: https://ieeexplore.ieee.org/document/6773024 (visited on 08/01/2024).

Simini, Filippo et al. (Nov. 12, 2021). "A Deep Gravity Model for Mobility Flows Generation". In: *Nature Communications* 12.1, p. 6576. ISSN: 2041-1723. DOI: 10.1038/s41467-021-26752-4. URL: https://www.nature.com/articles/s41467-021-26752-4 (visited on 05/02/2024).

Travel Forecasting Resource (2024). URL: https://tfresource.org (visited on 05/09/2024).

Wheeler, David and Michael Tiefelsdorf (June 1, 2005). "Multicollinearity and Correlation among Local Regression Coefficients in Geographically Weighted Regression". In: *Journal of Geographical Systems* 7.2, pp. 161–187. ISSN: 1435-5949. DOI: 10.1007/s10109-005-0155-6. URL: https://doi.org/10.1007/s10109-005-0155-6 (visited on 07/25/2024).

Wöhner, Fabienne (June 1, 2022). "Work Flexibly, Travel Less? The Impact of Telework and Flextime on Mobility Behavior in Switzerland". In: *Journal of Transport Geography* 102, p. 103390. ISSN: 0966-6923. DOI: 10.1016/j.jtrangeo.2022.103390. URL: https://www.sciencedirect.com/science/article/pii/S0966692322001132 (visited on 07/24/2024).

Xin, Yanan et al. (Oct. 18, 2022). *Vision Paper: Causal Inference for Interpretable and Robust Machine Learning in Mobility Analysis*. DOI: 10.48550/arXiv.2210.10010.

Yadav, Piyush et al. (June 1, 2017). *The Role of Open Data in Driving Sustainable Mobility in Nine Smart Cities*.

Yang, Yingxiang et al. (July 11, 2014). "Limits of Predictability in Commuting Flows in the Absence of Data for Calibration". In: *Scientific reports* 4, p. 5662. DOI: 10.1038/srep05662.

Zhong, Chen et al. (July 1, 2015). "Measuring Variability of Mobility Patterns from Multiday Smart-Card Data". In: *Journal of Computational Science*. Computational Science at the Gates of Nature 9, pp. 125–130. ISSN: 1877-7503. DOI: 10.1016/j.jocs.2015.04.021. URL: https://www.sciencedirect.com/science/article/pii/S1877750315000599 (visited on 06/18/2024).