Delineating urban functional use from points of interest data with neural network embedding: a case study in Greater London

Haifeng Niu[a,*], Elisabete A. Silva[a]

[a]*Lab of Interdisciplinary Spatial Analysis, Department of Land Economy, 19 Silver Street, Cambridge CB3 9EP*

Abstract

Delineating urban functional use plays a key role in understanding urban dynamics, evaluating planning strategies and supporting policymaking. In recent years, Points of Interest (POI) data, with precise geolocation and detailed attributes, have become the primary data source for exploring urban functional use from a bottom-up perspective, using local, highly disaggregated, big datasets. Previous studies using POI data have given insufficient consideration to the relationship among POI classes in the spatial context, and have failed to provide a straightforward means by which to classify urban functional areas. This study proposes an approach for delineating urban functional use at the scale of the Lower Layer Super Output Area (LSOA) in Greater London by integrating the Doc2Vec model, a neural network embedding method commonly used in natural language processing for vectoring words and documents from their context. In this study, the neural network vectorises both POI classes ('Word') and urban areas ('Document') based on their functional context by learning features from the spatial distribution of POIs in the city. Specifically, we first construct POI sequences based on the distribution of POI classes, and add their LSOA IDs as 'document' tags. By utilising these constructed POI–LSOA sequences, the Doc2Vec model trains the vectors of 574 POI classes (word vectors) and 4,836 LSOAs (document vectors). The vectors of POI classes are then used in calculating the functional similarity scores based on their cosine distance, with the vectors of LSOAs grouped into clusters (i.e., functional areas) via the k-means clustering algorithm. We also identify latent functions in each cluster of LSOAs by performing topic modelling and enrichment factor. Compared with TF–IDF, LDA and Word2Vec models, the Doc2Vec model obtains the highest accuracy when classifying functional areas. This study proposes a straightforward approach in which the model directly trains vectors for urban areas, subsequently using them to classify urban functional areas. By employing the enhanced neural network model with low-cost and ubiquitous POI datasets, this study provides a potential tool with which to monitor urban dynamics in a timely and adaptive manner, thereby providing enhanced, data-driven support to urban planning, development and management.

Keywords: Urban Functional Use, Points of Interest, Neural Network Embedding, Doc2Vec

Accepted by Computers, Environment and Urban Systems

1. Introduction

Scholars have studied the topic of urban functional use for decades, especially in the contexts of urban planning and geographical information science (Lynch, 1960; Crooks et al., 2015; Joshi et al., 2016). Urban functional use not only describes the configuration of the physical environment (e.g., buildings, spaces and facilities), but also reflects the socio-economic patterns of human activity at the collective level, which influences many urban processes from land-use regulations (e.g., the designation of permitted use of land) (Frias-Martinez & Frias-Martinez, 2014) to urban vibrancy (Yue et al., 2017). Understanding urban functional use is critical for planners and policymakers, especially in flexible planning systems where the uses of land and buildings change dynamically. For example, in England, the Use Classes Order (i.e., the legal framework that defines how land and buildings change from one class to another) was recently deregulated in order to introduce flexibility, which generated a surge in the conversion of uses without planning permission (Barton & Grimwood, 2019; Ferm et al., 2020). In practice within this context, delineating urban functional use offers a more nuanced understanding of how disaggregated land and building uses (e.g., offices, restaurants and so on) can be mapped by employing a bottom-up approach, and how the functional use of urban areas can be tracked. This approach allows for tracking, in near real-time, of what is happening with extended planning development rights, such as office-to-residential conversions or flexible use permission of commercial properties in cities such as London. It can also offer a tool to monitor the dynamic demand for urban facilities and services introduced by changes in the Use Classes Order, thereby assisting in solving planning-related issues such as refuse collection, parking supply and local taxation. The need for such tools is particularly important in flexible planning systems.

Traditionally, understanding urban functional use been heavily dependent on the identification of land-cover and land-use from remote sensing data (Forestier et al., 2012). However, this method only classifies urban land by its geographical features, neglecting the socio-economic characteristics of land use (Gao et al., 2017). It thus remains an insufficient method for exploring contemporary cities where urban land use is highly mixed and subject to rapid, dynamic changes. In the era of big data, new urban data sources such as social media check-ins, points of interest and mobile phone data provide disaggregated and fine-scale information about urban functional use (Liu et al., 2015; Crooks et al., 2015). In the urban context, POI data are a collection of location points such as commercial properties, offices, public spaces, transportation facilities, and so forth comprising detailed information including geospatial location, name, postcode, address, coordinates and so on (Elwood et al., 2012). POI data define the first-hand account of human activities and building uses, providing opportunities for researchers to explore the distribution of urban functions from a bottom-up perspective (Niu & Silva, 2020). As this type of data records accurate geolocation and specific uses of urban spaces and buildings at a very granular level, they are used to infer urban functional use by following a bottom-up approach to data mining. Particularly

*Corresponding author
Email address: hn303@cam.ac.uk (Haifeng Niu)

because there are many mixed-use buildings, POI data provide the finest level of building utilisation data that can be aggregated and categorised according to urban functions at a different scale. For instance, a commercial building might have a clothing shop on the ground floor and a nightclub on the upper floor. With POI data, we can identify the building's commercial use as two seperate functions, *Clothing* and *Nightclubs*, defined by their POI classes. Despite the disaggregated and fine-level information of functional use, POI data are also characterised by large volume and high dimensionality, presenting challenges in data processing and interpretation. Previous studies have introduced text mining techniques such as term frequency-inverse document frequency (TF–IDF) and latent Dirichlet allocation (LDA) to infer urban functional use from collections of POIs (Yuan et al., 2015; Liu et al., 2017; Gao et al., 2017). The drawback of these frequency-based methods is that they only take into account the number of POIs, without considering the spatial relationship among POIs in cities. To solve this issue, a group of studies have integrated neural word embedding models with POI data to classify urban areas by their functional use (Yao et al., 2017; Zhai et al., 2019; Hu et al., 2020).

The embedding method trains vectors of POI classes from their spatial context (i.e., the surrounding POI classes) and computes vector representations for urban areas based on the local combination of different POIs. Due to the fact that these low-dimensional vectors of urban areas extract information from POI classes (i.e., the finest-level of building utilisation data), the cosine distance (i.e., the similarity metric) between these vectors can be utilised for measuring the functional similarity between urban areas. However, there are gaps between the neural word embedding model and its applications in delineating urban functional use. First, few studies take full advantage of vector representations of POI classes when revealing the relationship between high-dimensional POI classes. Previous studies overlook the models potential for measuring functional similarities among hundreds of POI classes in the city, which is important if seeking to understand how different POI classes are configured as urban functions and how certain POI classes tend to cluster more easily than others. When classifying urban functional areas, it is important to note that as the Word2Vec model adopted by previous studies only trains vectors of POI classes, researchers have to calculate the vectors of urban areas using an average, or a weighted average, of the vectors of POIs in those areas. This rigorous compounding process fails to capture the spatial heterogeneity among urban areas because some areas may share the same ratios of different POI classes, but differ in terms of the spatial arrangement of the POIs.

To fill this gap, this study integrates the Doc2Vec model, an extension of Word2Vec developed by Le & Mikolov (2014), with POI data to train vectors of urban areas directly, as well as training the vectors of POI classes through the neural network. This proposed method considers spatial heterogeneity during the training process for vectors of urban areas, assisting in substantially improved identification of urban functional areas. Simultaneously, because the Doc2Vec model provides a direct means of obtaining vectors for urban areas, this study also contributes to the existing academic literature by proposing a more efficient way to delineate urban functional use, relative to previous studies.

Section 2 of this paper presents a literature review regarding POI data, previous methods, and the research trends in this area. Section 3 introduces the case study and its datasets.

Section 4 describes the methodology, including the pre-processing of POI data, training with the Doc2Vec model, analysis of the vectors of POIs and urban areas, and model evaluation. Section 5 presents the result for urban functional similarity between POI classes and urban functional areas in Greater London. Finally, Section 6 discusses the findings, implications and limitations of this study, before offering suggestions for future research.

2. Literature review

2.1. Urban functional use detection with crowdsourced data

Scholars typically conceptualise urban functional use based on the purposes of urban spaces by linking the preferences of human activities and the configurations of land or building use (Lynch, 1960; Crooks et al., 2015). In exploring urban functional use, researchers have traditionally employed remote sensing data to classify urban land use and monitor the dynamic change thereof (Joshi et al., 2016; Ma et al., 2019). Although remote sensing data provide a valuable source for extracting the physical characteristics of the land surface, they are less helpful in delineating urban functional use; namely, the collective activities in urban spaces and the socio-economic environment formed by these activities (Gao et al., 2017). In recent years, with the development of information and communications technology and the proliferation of location-based services (LBS), crowdsourced data (social media, points of interest and geotagged images) have demonstrated potential in understanding urban activities and the use (or multiple uses) of buildings, often with high levels of granularity and fine-temporal resolution (Goodchild, 2007; Niu & Silva, 2020). In exploring urban functional use, Crooks et al. (2015) highlights the notable contribution of crowdsourced data that provide primary accounts of urban form and function. In this vein, applications of crowdsourced data can be found in exploring urban function-related topics such as the detection of communities (Cranshaw et al., 2012; Hasan & Ukkusuri, 2015), identification of urban functional areas (Yuan et al., 2012; Chen et al., 2017; Liu et al., 2017), and in measurements of urban function mixture (Li et al., 2016; Yue et al., 2017).

Among the types of crowdsourced data, ubiquitous POI data are commonly utilised in studies of urban functional use. Previous studies have evaluated the capability of POI data for extracting urban functions from different sources (Jiang et al., 2015; Gao et al., 2017; Song et al., 2018; Chen et al., 2019). There are two primary sources of data. The first includes POI databases that provide a location-based directory, including national mapping agencies (e.g., Ordnance Survey UK), open-source platforms (e.g., OpenStreetMap[1]) and business platforms (e.g., Google Places[2] and Baidu Map[3]). These POI databases provide a wide range of types of POIs, including public services and privately-owned businesses, services and facilities. The second comes from location-based social networking data, or check-ins data, on platforms such as Foursquare, Yelp and Twitter. Although this type of data primarily links human activities with places, it has limited coverage in POI categories (mainly business-

[1] www.openstreetmap.org
[2] www.google.com/maps
[3] map.baidu.com

and leisure-related POIs) and spatial distribution (only covering particularly vibrant areas). Considering the diversity and complexity of POI sources, previous studies have developed several common requirements for POI data, especially for urban functional use exploration. First, the POI dataset should have a unified taxonomy that clearly defines the classification of POI classes. For instance, POI data from the Ordnance Survey (UK) have nine groups POI groups for 616 POI classes, including *Accommodation, Eating and Drinking*; *Commercial Services*; *Attractions* and so on (see Appendix A). It would be problematic to code different POIs into classes without a robust POI taxonomy, resulting in ambiguity when aggregating urban functional uses for urban areas (Jiang et al., 2015). POI data must also have broad spatial and categorical coverage in the study areas, so as to ensure it can accurately reflect heterogeneity within and among different areas. For example, the POI dataset derived from social media platforms (such as Yelp) provides more catering and leisure POIs in those vibrant areas, but fewer POIs of other types, such as businesses and industry, especially in the urban periphery. Therefore, extracting urban functions from this sampled POI dataset (i.e., limited coverage in both space and POI categories) may only reveal the patterns of commercial and entertainment functions in the central area of the city, thus impeding a comprehensive urban analysis.

2.2. Mining POI data with natural language processing techniques

Like many other types of new urban data, POI data are characterised by the large volume and high dimensionality that traditional methods mostly fail to incorporate and accommodate. Over recent decades, researchers have attempted to employ new techniques, such as natural language processing (NLP), in order to extract urban functional use. As with natural languages, the rank-frequency distribution of POI data follows Zipf's law (an empirical power law for rank and frequency), which states that the rank of a POI class is inversely proportional to its frequency, as demonstrated in Equation 1 (Gabaix, 1999; Soo, 2005). In other words, it reveals that only a small number of POI classes appear most frequently in cities, while others emerge only occasionally.

$$f \propto \frac{1}{r} \tag{1}$$

Yuan et al. (2012) introduced a topic-based inference model to identify urban functional areas with POI data using LDA topic modelling, similar to how researchers use the model to extract semantic topics from documents. Similarly, Gao et al. (2017) utilised LDA topic modelling to classify urban functional regions with POI and social media data. Xing & Meng (2018) made comparable attempts to extract semantics from POI data. Although LDA topic modelling helps to explore the hidden structure of functional semantics in POI data, this topic model is built on an approach known as the 'bag of words' model, which disregards grammar and word order, but maintains the word counts. In the case of POI data mining, LDA topic modelling retains the occurrences of POIs across the whole city, but disregards spatial relationships between POIs.

2.3. Neural network embedding: from Word2Vec to Doc2Vec

Neural network embedding offers a feasible solution to fill this gap, taking into account spatial relationships between POIs when delineating urban functional use. Neural network embedding, such as the Word2Vec model, refers to feature learning techniques based on a neural network in which words or terms in corpora (i.e., collections of text) are mapped to low-dimensional vectors (Mikolov et al., 2013b). The output vectors with real numbers represent words and their common context, which researchers can use to explore the semantic relations in a continuous vector space (Li & Yang, 2018). The integration of neural word embedding with POI data treats POI classes as 'words', POI sequences as 'sentences', regions/districts as 'paragraphs' and cities as the 'documents'. In this approach, the functions in urban areas hidden in the POI dataset are equivalent, to some extent, to semantic topics hidden in the text corpora. Most significantly, the spatial arrangement of POIs can be regarded as equivalent to the syntactic order of words in sentences, meaning that the spatial context of POIs in urban areas can be learned as a constituent element of vectors, an inclusion which is not able to be made with frequency-based methods such as LDA and TF–IDF models.

Yao et al. (2017) first used the Word2Vec model, developed by Mikolov et al. (2013a), to train vectors of POI classes with POI sequences built by ascertaining the shortest path in traffic analysis zones (TAZs). With POI vectors, vectors for all TAZs can then be calculated and utilised to identify urban land-use types. Based on this work, Zhai et al. (2019) deployed the Place2Vec model and applied it to the neighbourhood level. The main difference in approach is that the Place2Vec model adopts the nearest neighbour method for POI sequences in neighbourhood areas, and constructs (POI_{centre}, $POI_{context}$) pairs. This method considers the first law of geography, namely that nearby POIs are more related than distant POIs. Apart from this distinction, the Place2Vec model essentially mirrors the Word2Vec model. Following a similar approach, this study aggregates POI vectors to generate vectors specific to urban areas, which then serve to identify urban functional areas by subsequent clustering. The recent study by Hu et al. (2020) also employed the Word2Vec model with POI sequences, constructed by POI centre-context pairs and classifying urban functional areas on a 1km-by-1km grid. However, these studies all build on the Word2Vec model that only trains vectors for POI classes, meaning that they had to calculate a compound vector to represent urban areas (i.e., spatial units such as TAZ, neighbourhood zones and grids) for later identification of urban functional areas through clustering. The problem is that researchers have calculated compound vectors by using either an average or a weighted average of all vectors of POIs that appear in each area. This rigorous compounding process ultimately fails to capture the spatial heterogeneity among urban areas because some areas may share a similar POI configuration (i.e., the same count in each POI category) but differ in the spatial arrangement of those POIs. By contrast, the Doc2Vec model, an extension of the Word2Vec model developed by Le & Mikolov (2014), uses the neural network to train additional vectors for paragraphs, which can be utilised to specifically avoid the aforementioned problem and in directly training vectors for urban areas directly. Although there have been some applications of the Doc2Vec model in the urban context, few of them focus on detecting urban functional use with POI data (Wang et al., 2017; Li et al., 2019).

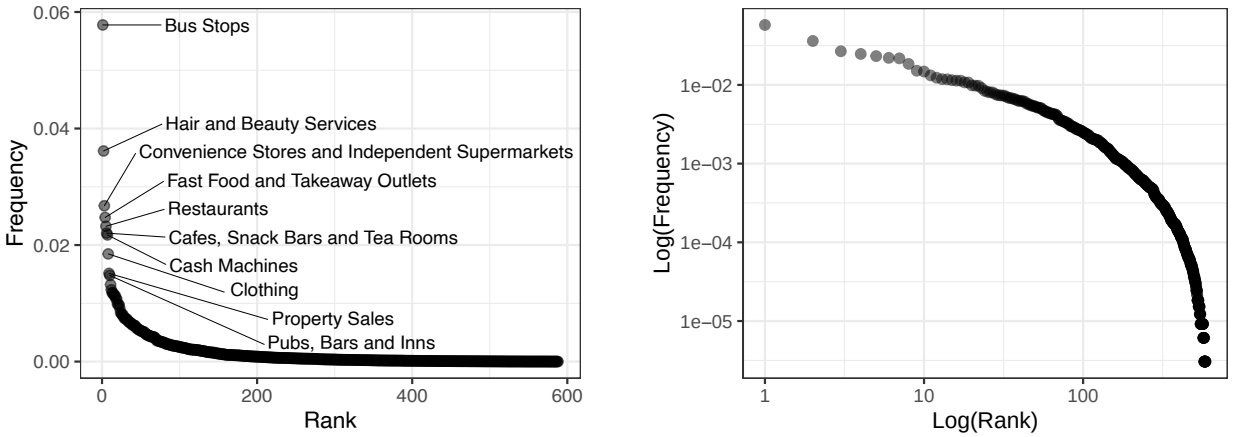
3. Study area and data

The study area covers Greater London as the extent is great enough to demonstrate the diversity of urban functional uses (see Figure 1). According to the Census Output Area population estimates (mid-2019) by the Office of National Statistics (ONS), 8.96 million people live in the region of Greater London, an area of 1572 km^2. As one of the largest megacities in the world, Greater London needs to monitor its urban dynamics, including urban functional uses, to accomplish smart and adaptive urban management.

Figure 1: The case study area: Greater London. The grey dots refer to locations of 420,559 POIs updated in March 2019 by the Ordnance Survey.

The POI dataset utilised for this research derives from the Ordnance Survey, the national mapping service for Great Britain. This dataset was last updated in March 2019, comprising 420,559 POIs for the London metropolitan area. Each record of a POI has numerous attributes, including a unique reference number, name, PointX classification code, geographic coordinates, address detail, street name, postcode, administrative boundary and other specific identifiers. The PointX classification code (POI classes), referring to the specific use of POI, contains an eight-digit number consisting of a two-digit group code, a two-digit category code and a four-digit class code, which assists with classification of POIs by their

functions at different levels. The POI classification scheme is a three-tier hierarchy, including nine groups, 52 categories and 616 classes with codes (See POI Groups and Categories in Appendix A). This study excludes some POI class codes, such as infrastructure features (electricity, gas and fire safety), industrial features (chimneys, pipelines, tanks, and so forth) and transport facilities that are unrelated to the functional use of a building or space in the city, leaving a total of 574 POI classes. Figure 2 illustrates the distribution of the dataset. The rank of POI classes and their frequency follows Zipf's law, in which a few POI classes account for a large proportion of all the POIs in Greater London. The ten most frequent POI categories are *Bus Stops*; *Hair and Beauty Services*; *Convenience Stores and Independent Supermarkets*; *Fast Food and Takeaway Outlets*; *Restaurants*; *Cafes, Snack Bars and Tea Rooms*; *Cash Machines*; *Clothing*; *Property Sales*; *Pubs, Bars and Inns*.



(a) Rank-Frequency plot

(b) Rank-Frequency log plot

Figure 2: Rank-Frequency plot for 620 POI classes in Greater London.

4. Methodology

This section first describes the preparation of the data for POI sequences with urban area tags. In order select the optimal unit of analysis, consideration is given to how scale and zone affect the spatial aggregation of POIs. Subsequently, this section introduces the Doc2Vec model for training vectors for POI classes and urban areas, followed by the method of calculating the functional similarity between POIs and the methods of classifying and annotating urban functional areas. The final part explains the overall evaluation of the model (Figure 3).

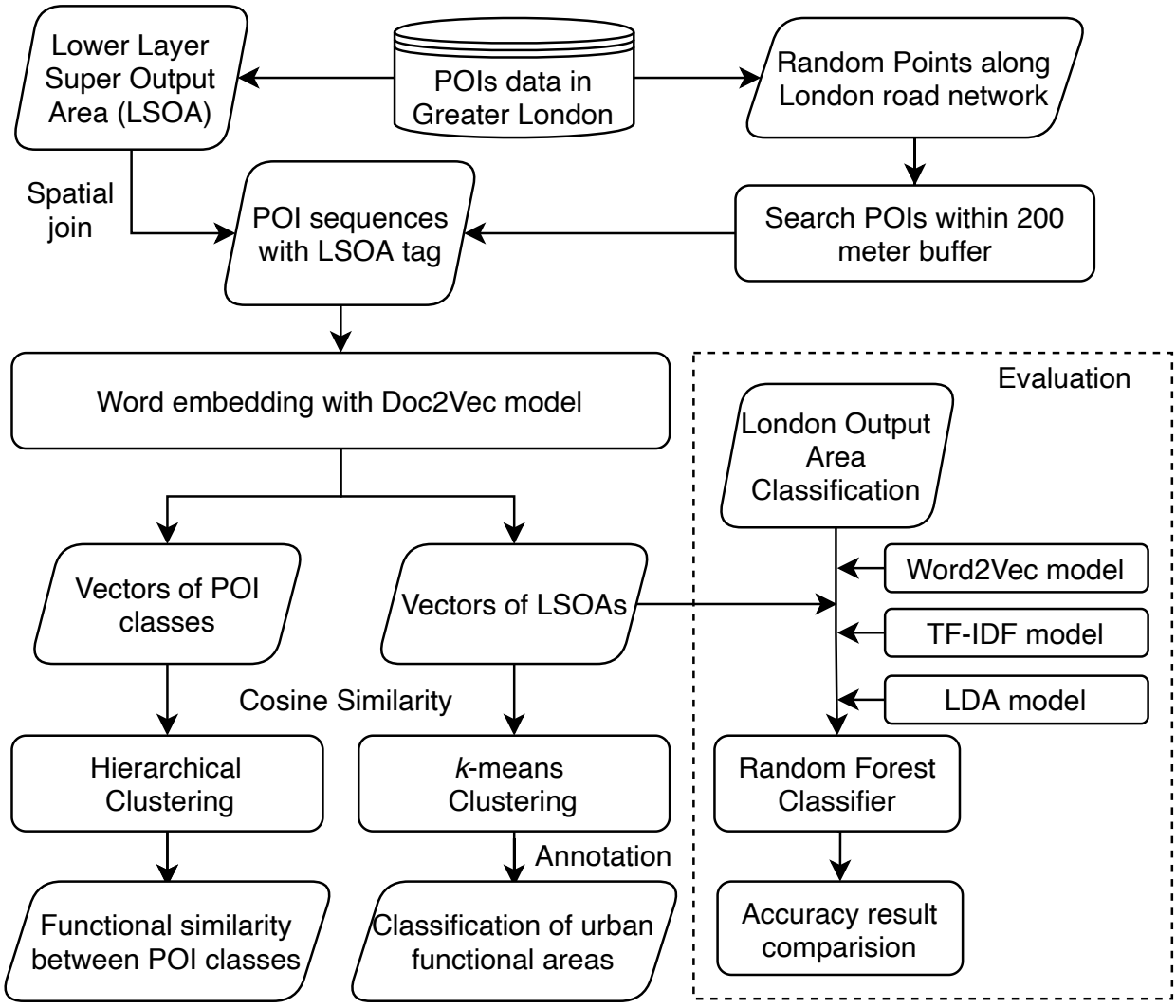


Figure 3: Flowchart of the methodology.

4.1. Constructing POI sequences tagged with urban areas

Unlike text corpora following a certain syntactic rule (i.e., how words are organised in a sentence), POIs in urban spaces essentially have no sequential order. This lack of order means that researchers must manually construct POI data as sequences, in order render them as valid inputs for the model. In practice, Yao et al. (2017) built a POI corpus based on traffic analysis zones and implemented the shortest path algorithm to construct POI sequential orders, where the length of path segments referred to the Euclidean distance within POI pairs. Zhai et al. (2019) later challenged this approach by arguing that sequences generated by the shortest path sorted all POIs in the traffic analysis zone with connecting points only once, which is not convincing when seeking to explore the spatial relationship between POIs. Zhai et al. (2019) adopted the nearest neighbour method for each POI to create (POI_{center}, $POI_{context}$) pairs, in which distance decay augmented the spatial context.

However, this method risks oversampling POI sequences in areas with a high density of POIs, resulting in an overfitting problem for the model.

To avoid the oversampling issue, this study utilises generated random points with a minimal interval to construct POI sequences in Greater London. To align with the POI density in the city, these random points are generated along with the road network (excluding unclassified roads). Setting the distance interval between any two random points at 50 metres produces a structured set of random points $R\{r_1, r_2, ..., r_i, ..., r_n\}$ with the maximum size being 201,000. Subsequently, for each random point r_i, we search its accessible POIs within 200 metres and thus obtain an accessible POI set S_i. By computing the distance for all of the pairs between the central random point and POIs within S_i, we can use the distance as a reference by which build a sequentially ordered POI list $L_i = [poi_1, poi_2, poi_3, ..poi_n]$, where i refers to the index of the random point, n is the number of accessible POIs around the random point r_i and elements in the list refer to the POI classes. The left part of Figure 4 illustrates all the lines between random points and their accessible POIs where colours are assigned by random point ID. The right part of Figure 4 shows a single POI sequence constructed for random point i with a 200-metre buffer. After constructing POI sequences, we retain only the POIs sequences with lengths which are greater than three; this is to avoid the sparse meaning and randomness produced by short POI sequences.

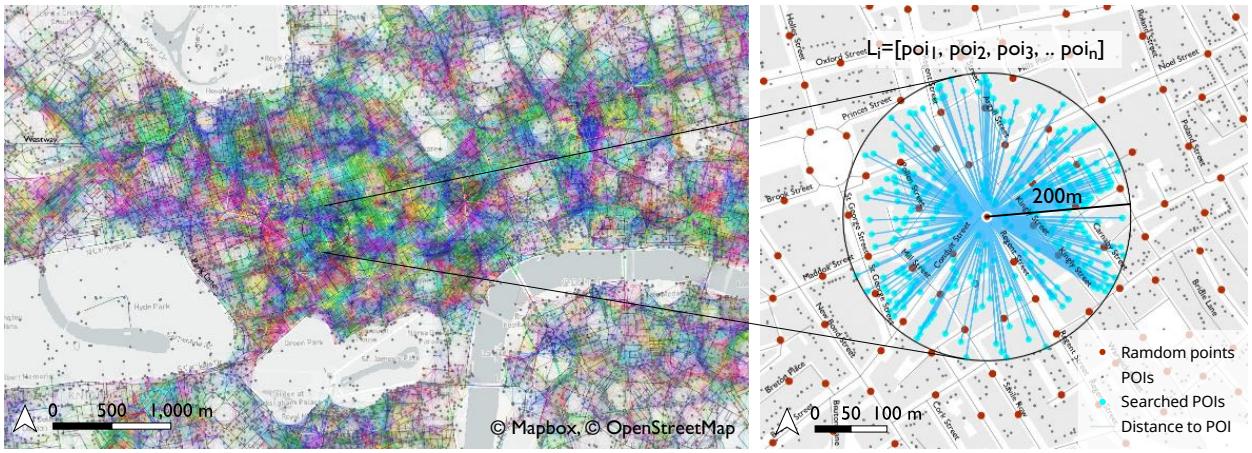


Figure 4: POI sequences construction in Greater London. Left: Constructed POI sequences with random points on the road network of Greater London. Right: Illustration of a single POI sequence constructed for random point r_i with a 200-metre buffer.

The most significant difference between Word2Vec and Doc2Vec is that, in the Doc2Vec model, each sentence can be tagged by its paragraph tag. This enables it to be later mapped into vectors representing the paragraphs. When applying the Doc2Vec model to POI sequences, the paragraph tag refer to an administrative district, a postcode area, part of a grid, a TAZ or any other spatial unit where POI sequences exist. During the training process, vectors of paragraph tags are generated from the semantics of POIs shared within their units, and can be served as the reference to classify urban functional areas.

4.2. Selecting the optimal aggregation level

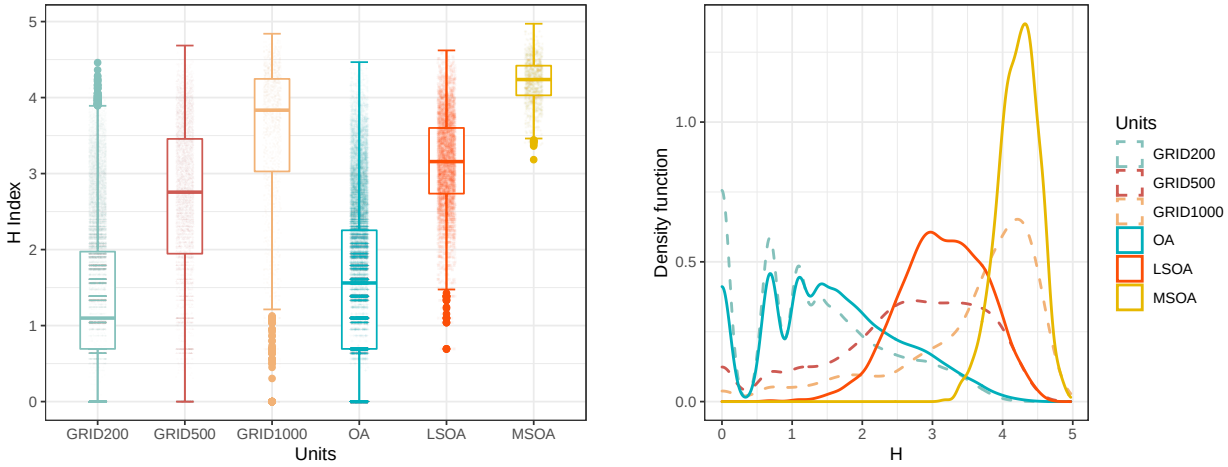
As the model extracts information about urban functional use from point-based data (POIs), the effects of the modifiable areal unit problem (MAUP) must be considered. When researchers aggregate spatial phenomena from points data into areas, MAUP can occur because results can be impacted both by the shape (zoning effects) and the size (scale effects) of the analysis units (Openshaw, 1983; Viegas et al., 2009). To select the optimal aggregation level for examining urban function from POI data, we compared two types of zoning methods (UK census units and grid) at three levels of scale. The UK census units, including OA (output area), LSOA (lower layer super output area) and MSOA (middle layer super output area), are geospatial statistical units built from postcodes, with standardised populations and household sizes. For example, OA units typically contain 100 people, while LSOAs have an average population of around 1,500. We also applied three sizes of grid (200m, 500m, 1000m) corresponding to the median areas of OA, LSOA and MSOA, respectively. We used three indicators, including POI class diversity, richness, and POI density, to compare the six analytical units. Table 1 presents the description of indicators and the related effects.

To examine the scale effect, we measured POI class richness and diversity for both grid-based and census units at three levels. Figure 5 shows that small units in scales GRID200 and OA have a greater number of low outliers (See the number of NaN values shown in Table 2) and a low mean of the diversity indices (1.33 and 1.57, respectively). When the scale size increases, the mean and median of the diversity and richness indexes (GRID500 and LSOA) are greater, because the wider extent includes more areas with varying functions. However, the density function distribution for GRID1000 and MSOA is left-skewed, with a notably high mean and median of diversity index. This would increase the difficulty of extracting urban functional use because the units include too much POI information (high entropy in information theory). For zoning effects, we compared the POI density of units between grids and census units (See Figure 6). Table 2 demonstrates that grid-based units have more NaN values than census units, especially at the smallest scale. Compared with larger scales, GRID500 and GRID1000 had lower POI densities than LSOA and MSOA, respectively. This is because the grid partition method generates more units than the geostatistical method in outer London, where there are fewer POIs. By synthesising both scale and zoning effects of the different analytical units with POIs, we chose LSOA (the census units at the middle scale) as the optimal unit of analysis to aggregate urban functional use from POI data.

All POI sequences of L_i generated in the last step by LSOA units were tagged, marked as a tag set $T\{t_1, t_2, ..., t_j, ..., t_m\}$. By spatially combining random points and the LSOA layer, each POI sequence receives an LSOA ID for the area in which the central random point is located. This process produces a tagged POI sequence $ST_i = [t_j, poi_1, poi_2, poi_3, ...poi_n]$ where i refers to the ID of the random point and j refers to the LSOA ID.

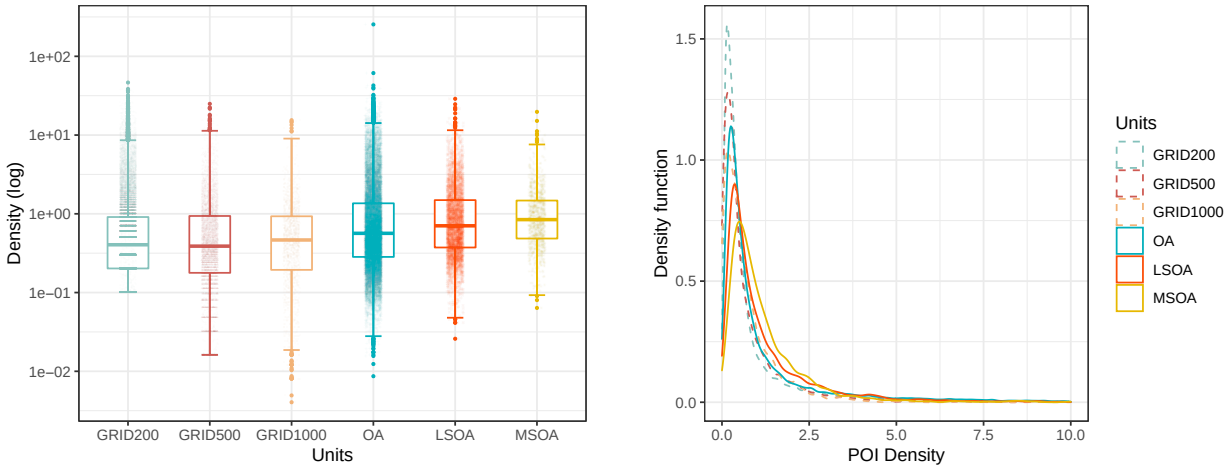
Table 1: Indicators for examining scale and zoning effects when aggregating POIs.

Indicator	Description	Related Effects
POI density	Number of POIs per square metre: $D = n/a$, where n refers to the number of POIs and a refers to the area of unit of analysis.	Zoning effects: indicates the intensity of urban functional use
POI class richness	Number of POI classes present in a unit: $R = m$, where m refers to the number of POI classes.	Scale effects: indicates the degree of mixed-use
POI class diversity	The weighted geometric mean of the POI classes' proportional abundances. The index is calculated by Shannon's Diversity Index: SDI $= -\sum_{i=1}^{m}(P_i \ln P_i)$, where p_i refers to the proportion of type i_{th} POI class	Scale effects: indicates the diversity of urban functions in units.



(a) Boxplot (b) Density plot

Figure 5: Plots of POI class diversity for different analysis units (NaN values excluded).



(a) Boxplot (b) Density plot

Figure 6: Plots of POI density for different analysis units (NaN values excluded).

Table 2: Results of indicators measured for six analysis units.

Units	N	N_{NaN}	POI Class Diversity			POI Class Richness			POI Density		
			Mean	St. Dev.	Median	Mean	St. Dev.	Median	Mean	St. Dev.	Median
GRID200	40637	9524	1.33	1.01	1.10	8.17	3.00	0.11	0.80	1.83	0.30
GRID500	6672	423	2.59	1.11	21.47	18.40	16.00	0.42	0.78	1.40	0.36
GRID1000	1731	52	3.47	1.08	3.83	46.59	29.23	46.00	0.76	1.18	0.45
OA	25053	1498	1.57	0.98	1.56	7.92	9.19	5.00	1.37	2.97	0.51
LSOA	4836	0	3.15	0.61	3.16	27.70	16.42	24.00	1.28	1.74	0.70
MSOA	983	0	4.22	0.28	4.24	70.44	19.25	69.00	1.25	1.44	0.84

4.3. Training vectors for POI classes and urban areas with Doc2Vec model

The neural network architectures of the Doc2Vec model, and their training process, are similar to those in the Word2Vec model. The Doc2Vec model has two neural network architectures: the Bag of Words version of Paragraph Vector (PV-DBOW) and the Distributed Memory version of Paragraph Vector (PV-DM) (Le & Mikolov, 2014). The PV-DBOW model predicts POI classes using a spatial tag (the paragraph tag), while the PV-DM model uses context POI classes to predict the central POI class. In this study, we employ the PV-DM model that takes the tagged POI sequences ST as the input. Two types of vectors can be learned in the model. The first is the 'word vector' type, which represents vectors for POI classes in the POI sequences ST. The second is 'paragraph vectors' type, which represents urban areas tagged in set T. The training process of PV-DM for POI classes (word vector) is based on the CBOW model in Word2Vec, which Goldberg & Levy (2014) and Rong (2016) thoroughly explain. In brief, the objective of the CBOW model is to predict a word, based on the surrounding words in a given context. The aim is to maximise the average of the log probability for predicting the central word (Equation 2).

$$\text{minimize } J = -\log P\left(w_c | w_{c-m}, \ldots, w_{c-1}, w_{c+1}, \ldots, w_{c+m}\right) \tag{2}$$

w_c is the POI class of the central POI and m is the window size of the context. Equation 3 calculates the probability of the central word by the context $P\left(w_c|w_{c-m},\ldots,w_{c+m}\right)$ where $\hat{v}$ is the mean embedding vector over all context words, u_c is the output vector of word w_c, and V refers to the size of vocabulary (the number of POI classes). In the above equation, a score vector is initially generated by $u_c^T\hat{v}$, which is then mapped to a probability via the softmax function. With the probability estimate $P\left(w_c|w_{c-m},\ldots,w_{c+m}\right)$, the model uses cross-entropy to calculate the loss against the true distribution (i.e., one-hot word representation of the central word) (See Figure 7, left). Stochastic gradient descent is used to minimise J, while updating input and output word representations.

$$P\left(w_c|w_{c-m},\ldots,w_{c-1},w_{c+1},\ldots,w_{c+m}\right)=\frac{\exp\left(u_c^T\hat{v}\right)}{\sum_{j=1}^{|V|}\exp\left(u_j^T\hat{v}\right)} \tag{3}$$

Based on the CBOW model in Word2Vec, the PV-DM model additionally output a new embedding for the whole context (See Figure 7, right). The variation of the input data is the additional document token. In the process of averaging all context words, the method uses this token to predict the subsequent word. For all POI sequences from the same LSOA, the represented document vector can perform as a shared memory of the functional context in this area. Considering the total number of POI classes (574), we chose a dimension size of 20 for vectors trained in the model.

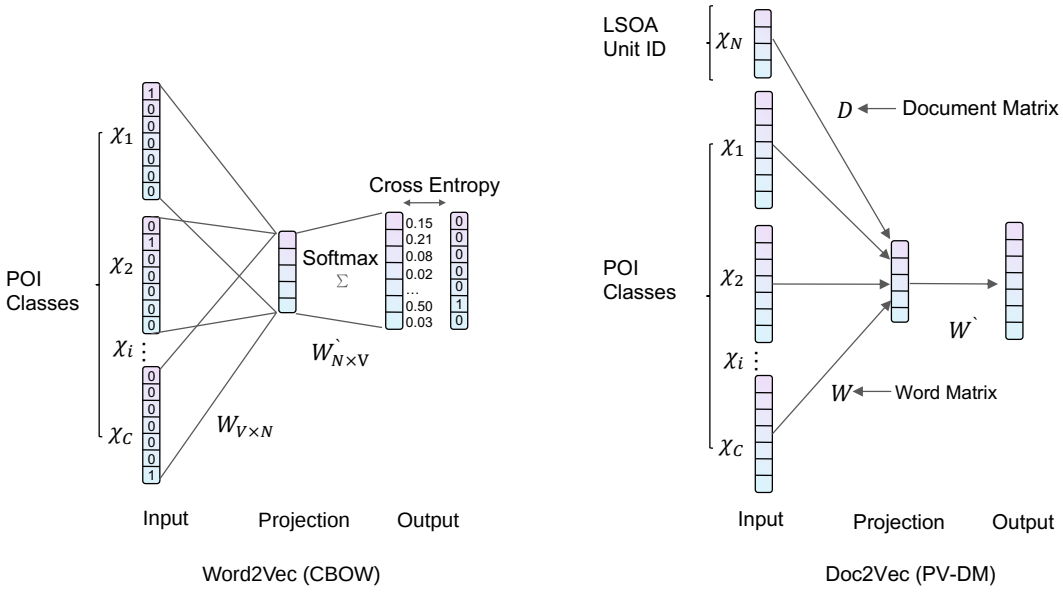


Figure 7: Comparison of the training processes using Word2Vec (CBOW) and Doc2Vec (PV-DM).

4.4. Functional similarity between POI classes with hierarchical clustering

When the training process is complete, each POI class and LSOA is represented by a 20-dimensional vector. For any two POI classes, the similarity between them can be measured by the cosine similarity in the vector space using Equation 4 where A_i and B_i are components of vectors A and B, respectively. A lower cosine distance indicates a higher similarity. The similarity score range is from -1 to 1, where 1 indicates the same and -1 signifies the opposite.

$$sim(A, B) = \frac{A \cdot B}{||A||||B||} = \frac{\sum_{i=1}^{N} A_i B_i}{\sqrt{\left(\sum_{i=1}^{N} A_i^2\right)\left(\sum_{i=1}^{N} B_i^2\right)}} \tag{4}$$

With similarity scores computed for all pairs of POI classes, we can generate a similarity matrix for all POI classes. The similarity between the vectors of POI classes from the Doc2Vec model refers to their semantic similarity – they have a similar context in urban space (Mikolov et al., 2013a). The similar pairs of POI classes mean that they tend to appear in the city as associated functional use. Based on this understanding, we conducted agglomerative hierarchical clustering to further explore the similarity scores between all pairs of 574 POI classes in Greater London. We chose the hierarchical clustering algorithm because this non-flat algorithm provides a sequence of nested relationships between POI classes and may, therefore, produce meaningful taxonomies. In hierarchical clustering, we chose complete linkage (i.e., furthest-neighbour linkage) as the metric by which to calculate all pairwise dissimilarities between observations in any two groups. As complete linkage emphasises the maximal inter-cluster dissimilarity, assisting the classification of classify POI groups with different functions (James et al., 2013b). The result of the clustering can be visualised by way of a dendrogram, a tree-based representation.

4.5. Urban functional areas classification and annotation

The vectors of LSOAs are used for exploring urban functional areas in Greater London. By clustering these vectors referring to the functional context in LSOAs, we classify all LSOAs into different groups, which are urban functional areas. Here, we use the k-means method to cluster LSOAs and choose an external evaluation method, Adjusted Rand Index (ARI), to determine the optimal k for the clustering. Researchers often utilise ARI to compare the similarity between two clustering types (Santos & Embrechts, 2009). This study employed ARI to compare the predicted clustering by the k-means algorithm and the ground truth that we derived from London Output Area Classification (LOAC). This classification system for 340 statistical units in Greater London, is founded on 60 attributes recorded in the 2011 census at the OA level, indicating the social, economic and demographic characteristics of local communities (Singleton & Longley, 2015). Specifically, we calculated the Rand Index (RI) with Equation 5.

$$RI = \frac{TP + TN}{TP + FP + TF + FN} = \frac{TP + TN}{C_N^2} \tag{5}$$

- True positive (TP): two similar POI classes are in the same cluster

- True negative (TN): two dissimilar POI classes are in different clusters
- False positive (FP): two dissimilar POI classes are in the same cluster
- False negative (FN): two similar POI classes are in different clusters

We then computed ARI with Equations 6-8, where n_{ij} denotes the number of times a POI class occurs in cluster i of the POI classification scheme and cluster j of k-means clustering; $\binom{n}{2}$ refers to unordered pairs in a set of n elements. The bounded range of the ARI score is [-1, 1], where a negative score means independent labelling and a positive score represents similar labelling. The random label assignments produce an ARI score of close to 0.

$$\mathrm{ARI} = \frac{\mathrm{RI} - E[\mathrm{RI}]}{\max(\mathrm{RI}) - E[\mathrm{RI}]} \tag{6}$$

$$E(RI) = E\left(\sum_{i,j}\binom{n_{ij}}{2}\right) = \left[\sum_{i}\binom{n_i}{2}\sum_{j}\binom{n_j}{2}\right] / \binom{n}{2} \tag{7}$$

$$\max(RI) = \frac{1}{2}\left[\sum_{i}\binom{n_{i.}}{2} + \sum_{j}\binom{n_{,j}}{2}\right] \tag{8}$$

After classifying LSOAs into several groups of functional areas, the next step is to interpret these groups. This is necessary because the clustering result gives no guarantee regarding on the interpretation of functional use in each cluster. Therefore, it is necessary to annotate the clusters with latent functions. In this study, we implemented LDA topic modelling for clusters separately, in order to extract topics from POI sequences. LDA topic modelling is one of the most powerful text mining tools for topic extraction from a collection of texts. The LDA model regards documents as mixtures of several topics, in which words characterise each topic with certain probabilities (Blei et al., 2003). Based on this conception, LDA topic modelling can be used for extracting the latent functions (topics) with probabilities of POI classes (words) in all the clusters identified by k-means.

We also applied the enrichment factor (EF) to supplement the annotation for urban functional areas. EF explores the difference in expression of elements between the sample data and background data; it is an approach that has been widely adopted in previous studies (Chen et al., 2017; Zhai et al., 2019). The metric allows us to aggregate the configuration of 574 POI classes at the category level (52 POI categories), so as to better interpret the urban functions in each cluster. This study used EF to calculate the ratio of the proportion of a POI category in a cluster to the proportion in the whole city, as shown in Equation 9:

$$EF_i^q = (P^q)_{cluster} / (P^q)_{context} = (N_i^q / N_i) \,/\, (N^q / N) \tag{9}$$

where N_i^q denotes the number of POIs in category q in the cluster i, N_i refers to the number of POIs in cluster i, N^q represents the total number of POIs in category q in the city, and N refers to the total number of POIs.

4.6. Model evaluation

To better comprehend the performance of the Doc2Vec model in delineating urban functional areas, we compared it with other three open-source models: TF–IDF, LDA and Word2Vec. We used the features output from the above models to train a random forest classifier, and compared their accuracy by the out-of-bag scores (James et al., 2013a). Random forest is a supervised learning algorithm that consists of multiple decision trees as an ensemble. The random forest algorithm is a popular classification tool due to its advantages in reducing the overfitting problem, responding to high-dimensional features, and exploring the non-linear correlation between independent variables (Biau, 2012). The true labels used to train the random forest classifier were the super groups identified in the LOAC. Table 3 lists the classification scheme comprising eight super groups and 19 groups.

Table 3: London Output Area Classification (LOAC): Super Groups and Groups.

Super Group	Group
A: Intermediate Lifestyles	A1: Struggling suburbs
	A2: Suburban localities
B: High Density and High-Rise Flats	B1: Disadvantaged diaspora
	B2: Bangladeshi enclaves
	B3: Students and minority mix
C: Settled Asians	C1: Asian owner-occupiers
	C2: Transport service workers
	C3: East-End Asians
	C4: Elderly Asians
D: Urban Elites	D1: Educational advantage
	D2: City central
E: City Vibe	E1: City and student fringe
	E2: Graduation occupation
F: London Life-Cycle	F1: City enclaves
	F2: Affluent suburbs
G: Multi-Ethnic Suburbs	G1: Affordable transitions
	G2: Public sector and service employees
H: Ageing City Fringe	H1: Detached retirement
	H2: Not quite Home Counties

Source: Singleton & Longley (2015)

5. Results

This study implemented the methodology using Python version 3.7 and PostGIS version 3.0 on Windows 10 (x64). PostGIS was used to manage geospatial data and run geoprocessing. The Gensim library for Python (`https://radimrehurek.com/gensim/`) was used to conduct Word2Vec, Doc2Vec and LDA models. Scikit-learn library for Python

(https://scikit-learn.org/) provided codes for machine learning algorithms, including TF–IDF, k-means clustering, agglomerative hierarchical clustering and random forest classification.

5.1. Functional similarity between POI classes in London

The Doc2Vec model returned 20-dimensional vectors for 574 POI classes and 4,836 LSOAs. By calculating the similarity between all pairs of POI classes in Greater London, we produced the 574-by-574 similarity matrix shown in Figure 8. Cells in the cluster map are coloured with the RdBu ramp, based on their similarity score, with a range from -1 to 1; red cells indicate high similarity and blue cells denote dissimilarity. Along the diagonal, POI classes are grouped into several clusters according to the functional similarity. The details of annotation for POI classes in each cluster can be found in Supplementary File 1. The dendrogram shown at the left of Figure 8 illustrates the underlying hierarchical structure of these clusters. We cut off the hierarchical tree with eight clusters (coloured in the dendrogram) in order to enable subsequent comparison with the nine groups in the original classification scheme. The results show that POI classes are not necessarily from the same POI category or group. This finding is not surprising, as a mix of functional uses is common in metropolitan cities such as London. From top to bottom, the eight clusters regarding functional use are: Leisure and domestic services; Industrial production, farming and municipal facilities; Food & household retail and personal services; Media, financial and business services; High-end retail, arts and real estate; Entertainment and transport; Public facilities, schools and public areas; Educational, government and institutes.

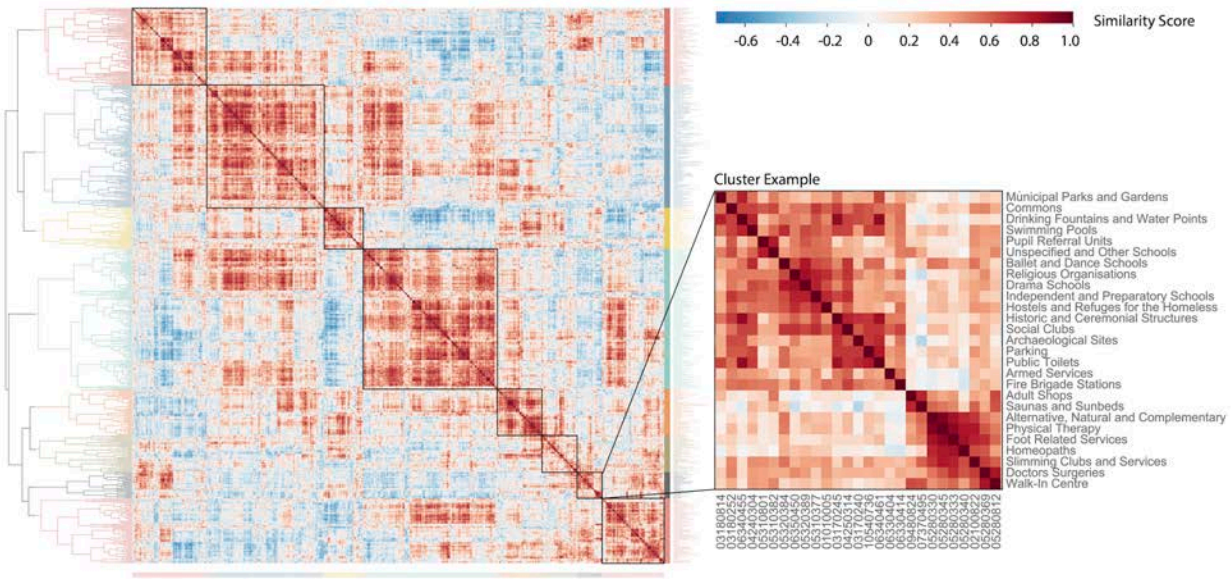


Figure 8: The functional similarity cluster map of 574 POI classes in Greater London. Each cell in the cluster map refers to the similarity value $sim(POI_i, POI_j)$ calculated by the cosine distance between the characteristic vector trained in the Doc2Vec model. Matrix elements are coloured with the RdBu ramp according to the similarity value with a range from -1 to 1, where red elements indicate high similarity and blue elements show low similarity. The similarity scores are clustered by the hierarchical clustering shown as the coloured dendrogram on the left-hand side. The square clusters around the diagonal indicate the clusters of POI classes with similar functional use. Names and codes for POI classes are labelled in the axes.

To further explore the difference between hierarchical clusters based on functional similarity and POI classification taxonomy, we reordered the index of the same matrix by their PointX classification code defined in Appendix A. In Figure 9, POI classes are sequentially ordered from Group 01 to Group 10 (See classification code for POI classes on the x-axis). The details regarding annotation of the POI classes can be found in Supplementary File 2. Along the diagonal, POI groups can be identified by the black frames. We found that POI classes in Groups 02 (*Commercial Services*), 07 (*Manufacturing and Production*) and 09 (*Retail*) shared a functional similarity within their groups, while POI classes in other groups had lower similarity scores within their groups. POI classes in both Group 07 (*Manufacturing and Production*) and Group 09 (*Retail*) demonstrated an overall functional similarity, while those in other groups were only partially congregated. Figure 9 also shows the functional similarity between different groups in Greater London. For example, one would not expect to see a high similarity score between Group 02 (*Commercial Services*) and Group 07 (*Manufacturing and Production*). The results show that only specific commercial services such as construction, contract, and engineering services tend to appear around manufacturing and production POIs. Similarities also appear in group pairs between Groups 07 and 09 (*Retail*), and between Groups 03 (*Attractions*) and 04 (*Sport and Entertainment*).

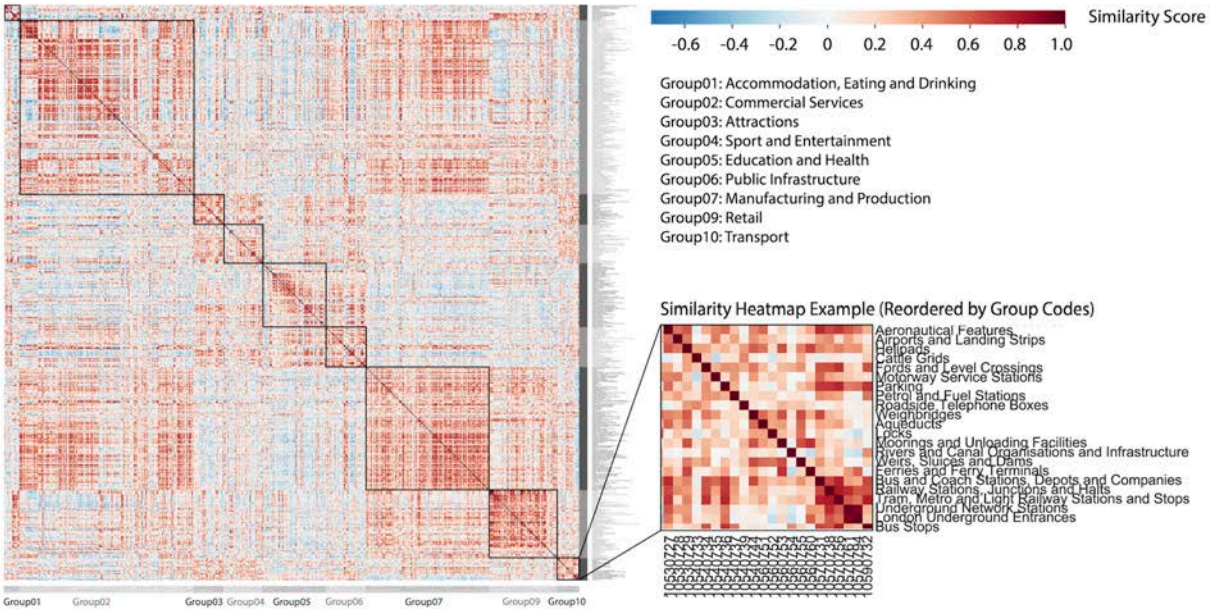


Figure 9: The functional similarity heat map of 574 POI classes in Greater London (re-indexed by classification codes). Each element in the matrix refers to the similarity score $sim(POI_i, POI_j)$ calculated by the cosine distance between the characteristic vector trained in the Doc2Vec model. Matrix elements are coloured with the RdBu ramp according to their similarity score with a range from -1 to 1, where red elements denote high similarity and blue elements indicate low similarity. The matrix is sequentially re-indexed according to POI codes from Group 01 to Group 10. Black frames around the diagonal refer to the similarity matrix of POI classes from the same POI group. Names and codes for POI classes are labelled in the axes.

5.2. Urban functional areas in Greater London

To identify functional areas, we applied k-means clustering to classify vectors of all 4,836 LSOAs in Greater London. In selecting parameter k for the algorithm, we used the Adjusted Rand Index to evaluate the clustering results by comparing them with super groups in the LOAC (James et al., 2013b). Figure 10 presents the ARI with different numbers of clusters k, from 2 to 15. Although $k = 6$ has the highest ARI, it was necessary to examine the optimal number of clusters for the task of detecting functional areas. Therefore, we selected another two k with high ARI, $k = 4$ and $k = 7$, as comparisons.

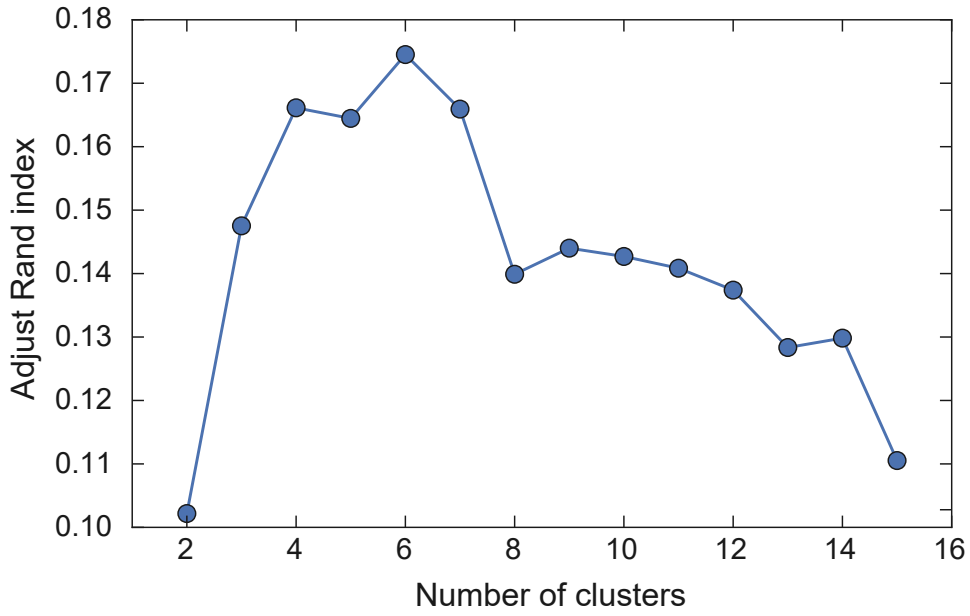
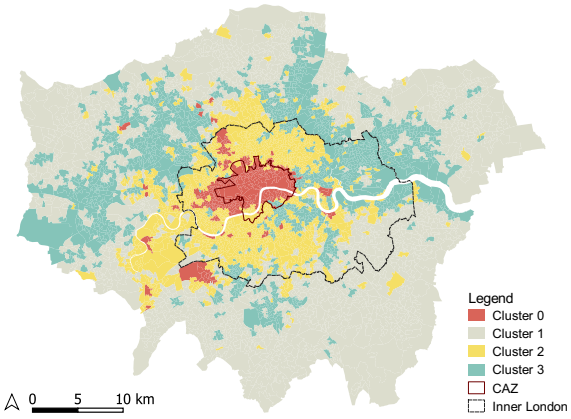
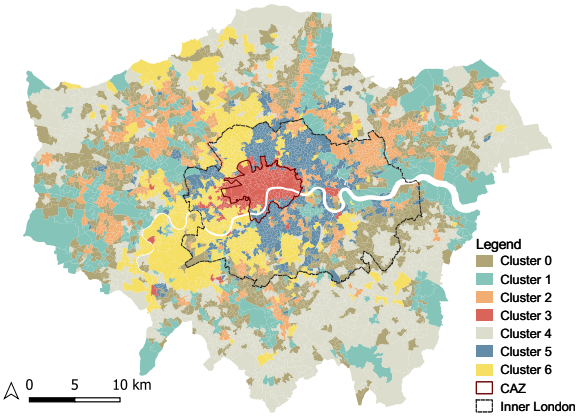


Figure 10: Adjusted Rand Index against the number of clusters using k-means.

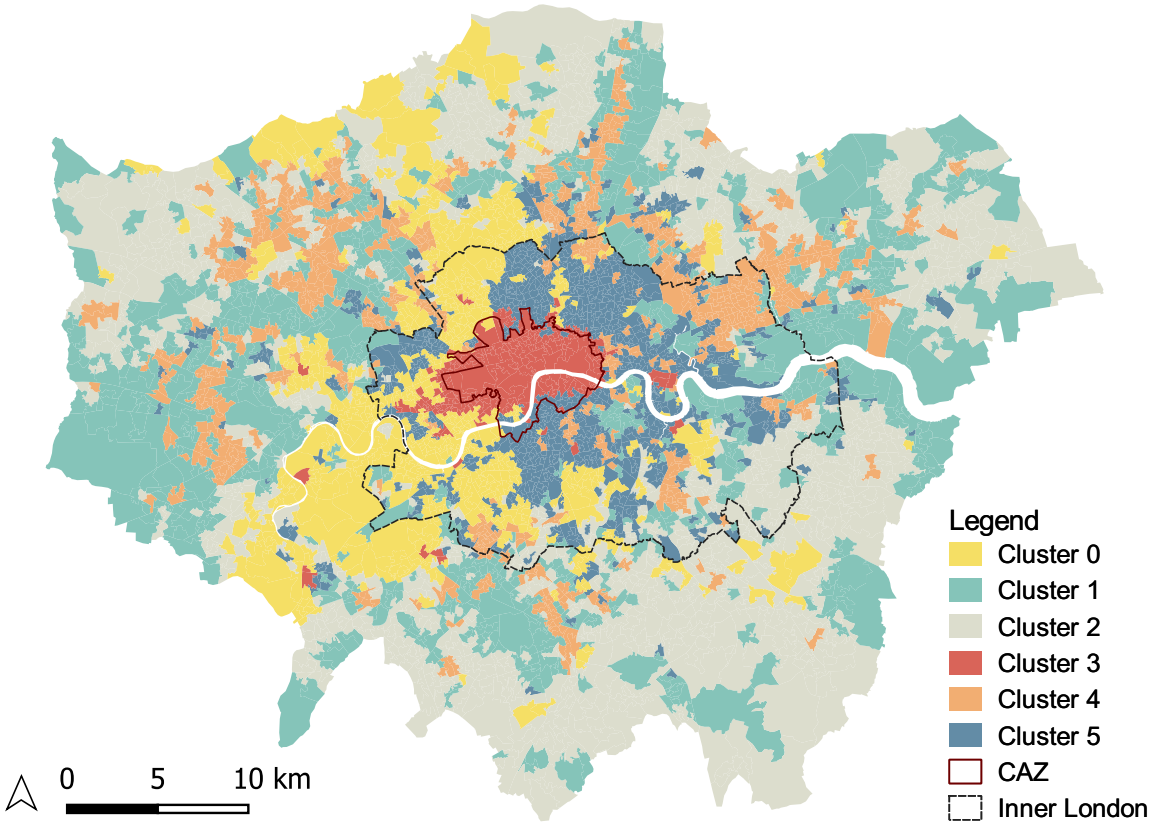
Figure 11 illustrates the clustering result when $k = 4$, 6 and 7. It is apparent that three classifications were successful in identifying the central areas of London, which overlap with the Central Activities Zone (CAZ) in the city. It is a reasonable result, as the CAZ is the vibrant centre hub of London, which is very different from the outer areas in most aspects of urban function. However, in the classification with $k = 4$ (Figure 11a), Cluster 0 not only covers the CAZ, but also includes large areas such as Wimbledon and Hampstead. Comparing this result with the other two, it clearly fails to subdivide the functional areas both in inner and outer London. For Figure 11b ($k = 7$) and Figure 11c ($k = 6$), the clustering results were similar. Both classifications subdivide inner London (except for the CAZ) into two main functional areas shown as yellow- and blue-coloured regions, implying the spatial division between western and eastern areas. The difference between them is that the result for $k = 7$ (Figure 11b) further classifies the areas in outer London, which is challenging to interpret with the random distribution. Ultimately, we chose six as the optimal k for k-means clustering to aggregate all LSOA vectors.



(a) $k = 4$



(b) $k = 7$



(c) $k = 6$

Figure 11: Urban functional areas classification with k-means clustering when $k = 4, 6, and 7$.

After classifying the urban functional areas with k-means ($k = 6$), we used LDA topic modelling and enrichment factor explained in section 4.5 to annotate these clusters. In order to find the optimal number of topics for interpreting each cluster, we utilised the coherence score as the indicator. Figure 12 shows the results of coherence scores against

the number of topics. Due to the fact that most of the clusters shared one or two common topics (i.e., basic urban functions), including the most common POI classes such as *Hair and Beauty Services*, *Convenience Stores and Independent Supermarkets*, *Restaurants* and *Cafes*, we chose at least three topics so as include more latent topics in the functional areas. By looking at the average score of coherence in Figure 12, we found that Cluster 3 has the lowest average score, indicating a heterogeneous collection of POI classes. Cluster 1 and Cluster 2 both exhibit a higher coherence than other clusters. The table of EFs for POI categories (Table 4) also served as another reference for the annotation of urban functional areas.

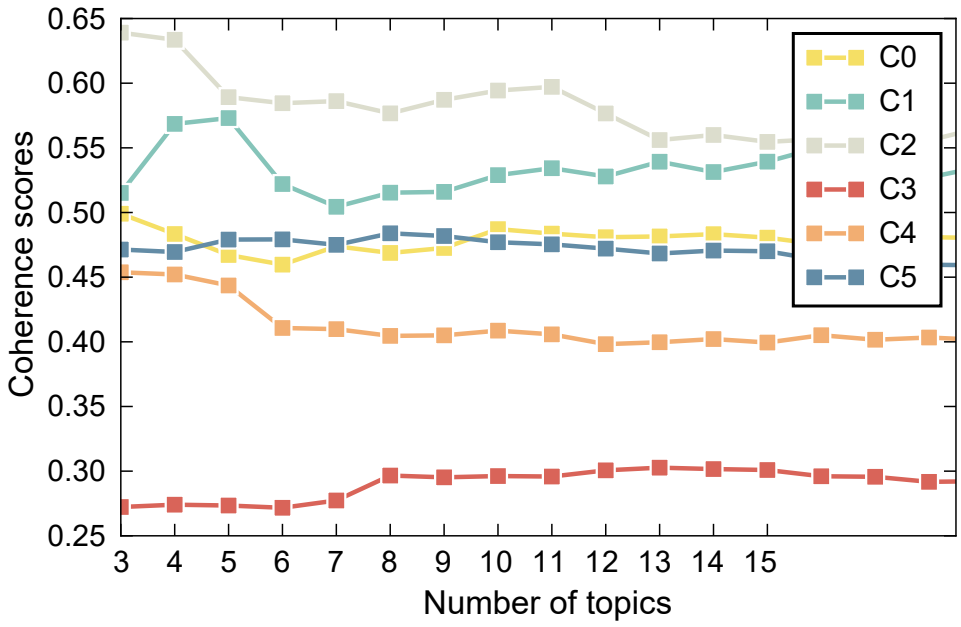


Figure 12: Coherence scores against the number of topics in each cluster.

Table 4: Enrichment factor for POI category in each cluster.

POI Category		Enrichment Factor					
Code	Names	C0	C1	C2	C3	C4	C5
01	Accommodation	1.067	0.5	0.391	**2.499**	0.689	0.666
02	Eating and Drinking	0.968	0.467	0.787	1.434	1.158	1.137
03	Construction Services	0.718	1.536	**2.064**	0.228	0.649	0.72
04	Consultancies	1.157	0.688	0.751	1.883	0.556	0.905
05	Employment and Career Agencies	0.707	0.646	0.718	**2.126**	0.92	0.612
06	Engineering Services	0.655	1.611	1.315	1.223	0.519	0.502
07	Contract Services	0.686	1.577	1.023	0.968	0.767	0.951
08	IT, Advertising, Marketing and Media Services	1.023	0.766	0.655	1.881	0.533	1.114
09	Legal and Financial	0.708	0.646	0.887	1.481	**1.278**	0.815
10	Personal, Consumer and Other Services	1.122	0.599	1.127	0.889	1.176	1.101
11	Property and Development Services	1.508	0.353	0.797	1.275	1.17	0.931
12	Recycling Services	0.604	2.308	1.191	0.284	0.728	0.951

Table 4: Enrichment factor for POI category in each cluster.

POI Category		Enrichment Factor					
Code	Names	C0	C1	C2	C3	C4	C5
13	Repair and Servicing	0.724	1.822	1.177	0.272	1.091	0.959
14	Research and Design	**1.814**	0.766	0.922	1.325	0.336	1.006
15	Transport, Storage and Delivery	0.677	2.392	0.573	0.554	0.956	0.953
16	Botanical and Zoological	**1.79**	1.568	1.305	0.459	0.292	0.829
17	Historical and Cultural	1.416	0.679	0.973	1.599	0.506	0.816
18	Recreational	0.786	1.242	0.711	0.249	0.645	**2.88**
19	Landscape Features	0.997	1.999	**1.703**	0.279	0.592	0.399
20	Tourism	**1.782**	0.171	0.362	**2.597**	0.296	0.774
21	Sport and Entertainment Support Services	1.281	0.883	0.779	1.285	0.475	**1.47**
22	Gambling	0.684	0.514	0.893	0.75	**1.909**	1.182
23	Outdoor Pursuits	**1.929**	1.47	1.616	0.358	0.126	0.738
24	Sports Complex	**1.7**	1.14	**1.688**	0.3	0.615	0.684
25	Venues, Stage and Screen	0.79	0.455	0.756	**2.011**	0.8	1.031
26	Animal Welfare	1.527	1.01	**2.239**	0.143	0.54	0.563
27	Education Support Services	0.757	0.64	0.609	1.541	1.079	**1.344**
28	Health Practitioners and Establishments	1.32	0.57	1.266	0.807	1.106	0.953
29	Health Support Services	1.145	0.683	1.081	0.965	1.027	1.135
31	Primary, Secondary and Tertiary Education	1.081	0.975	1.208	0.498	0.954	**1.435**
32	Recreational and Vocational Education	1.203	0.877	1.297	0.593	0.927	1.207
33	Central and Local Government	0.661	0.842	0.76	1.345	**1.313**	0.951
34	Infrastructure and Facilities	1.029	1.127	1.251	0.437	1.02	1.243
35	Organisations	0.871	0.635	0.798	1.522	0.836	**1.323**
37	Consumer Products	1.024	1.288	0.971	0.946	0.84	0.965
38	Extractive Industries	0.457	**3.169**	0.45	1.279	0.235	0.351
39	Farming	0.576	**3.62**	0.8	0.134	0.285	0.768
40	Foodstuffs	0.555	2.288	0.472	0.757	0.892	1.138
41	Industrial Features	0.522	**3.297**	0.558	0.163	0.611	1.069
42	Industrial Products	0.671	1.976	0.87	0.997	0.649	0.822
46	Clothing and Accessories	0.986	0.434	0.367	**1.96**	**1.44**	0.617
47	Food, Drink and Multi-Item Retail	0.858	0.738	0.863	0.601	**1.689**	1.289
48	Household, Office, Leisure and Garden	1.13	1.04	0.87	0.989	1.143	0.829
49	Motoring	0.46	**2.413**	1.225	0.183	1.114	0.545
53	Air	0.127	**5.525**	0.453	NA	NA	NA
54	Road and Rail	0.844	1.643	1.422	0.507	0.967	0.541
55	Walking	1.254	1.523	1.494	0.439	0.64	0.713
56	Water	1.492	2.056	0.709	0.473	0.355	1.273
57	Public Transport, Stations and Infrastructure	1.085	0.866	0.734	1.304	0.994	1.025
58	Bodies of Water	1.651	1.683	**2.117**	0.099	0.222	0.296
59	Bus Transport	0.97	1.353	1.431	0.381	0.853	1.083
60	Hire Services	0.645	2.35	0.993	0.495	0.793	0.747

Cluster 0 (Affluent neighbourhoods and leisure spaces) The spatial distribution of LSOAs in Cluster 0 (shown in Figure 13a) covers the affluent suburbs of Kensington and Chelsea, Hammersmith and Fulham, Hampstead, Richmond Upon Thames, Wandsworth

and so on. Table 5 provides an example of four topics extracted by LDA topic modelling where each topic is represented as a latent function by a list of weighted POI classes. Topic 1 and Topic 3, both group the POI classes, related to essential shops and services for daily life, identifying the residential function as the primary purpose in this cluster. Topic 2 includes a high number of property-related POI classes and associated facilities, including *Design Services*, *Architectural and Building-Related Consultants*, *Estate and Property Management*, *Tennis Facilities* and *Playgrounds*. Topic 4 includes a series of facilities and services in these affluent neighbourhoods, including *Tennis Facilities*, *Nursery Schools and Pre and After School Care*, *Accountants and Auditors*, *Sports Grounds, Stadia and Pitches* and *Entertainment Services*. Table 4 displays a high EF for both *Outdoor Pursuit* and *Sports Complex* in Cluster 0, indicating that leisure and outdoor activities play an important role in LSOAs within this cluster. The high EFs for *Botanical and Zoological* and *Tourism* reveal the popularity of these affluent areas among tourists.

Cluster 1 (Municipal facilities and industrial areas) As Figure 13b illustrates, Cluster 1 covers large areas in outer London. The topic extraction identified five topics that can be summarised as industrial areas and municipal facilities. One of the topics includes a group of municipal facilities such as *Sports Grounds, Stadia and Pitches*, *Cemetreies and Crematoria* and *Waste Storage, Processing and Disposal* which typically occupy large areas, requiring idiosyncratic site conditions. The presence of industrial areas is evident from the high weights for POI classes such as *Distribution and Haulage*, *Container and Storage*, *Import and Export Services*, *Construction and Tool Hire* and *Courier, Delivery and Messenger*. The high EFs for a range of manufacturing and production POIs (POI category codes 37 to 42) in Table 4 further reveals that most of the industrial production is located in this cluster. Air transport also emerges as a key function, because the EF for *Air*, in this specific cluster, is the highest among all other scores. Airports such as London Heathrow Airport, London City Airport and Royal Air Force Northolt are included in this cluster.

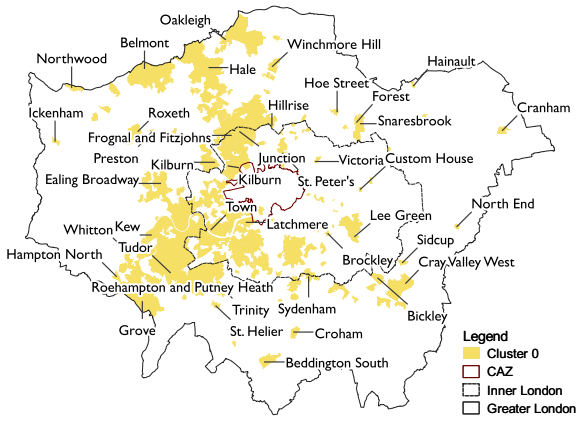
Cluster 2 (Suburbs and developing areas) LSOAs in Cluster 2 are distributed along the urban fringe, including Hillingdon, Enfield, Havering, Bexley, Bromley, Sutton and so on (Figure 13c). The result of topic extraction shows that LSOAs in this cluster function as residential areas. In neighbourhoods within this cluster, there are more facilities such as *Nursing and Residential Care Homes*, *Tennis Facilities* and *Places of Worship*. The results also demonstrate that building services such as *Building Contractors*, *Electrical Contractors* and so on are another significant function within the cluster. The high EF (2.064) of *Construction Services* further confirms this finding. Table 4 also shows that *Body of Waters*, *Animal Welfare* and *Landscape Features* are uniquely high POI classes in this cluster, compared with the other five clusters. This result is not surprising, because the urban fringe in Greater London is surrounded by the Green Belt.

Cluster 3 (Vibrant city centre) Cluster 3 is predominantly comprises the CAZ, which is the most vibrant area of London. The cluster also identifies similar areas outside the CAZ, such as Canary Wharf, Richmond and Wimbledon town centre. With the lowest average coherence score among all the clusters (Figure 13d), Cluster 3 has a highly mixed use of urban functions, therefore, requires more topics to interpret. We selected eight topics with which to annotate the cluster and to summarise the main functions of tourism, business, commercial

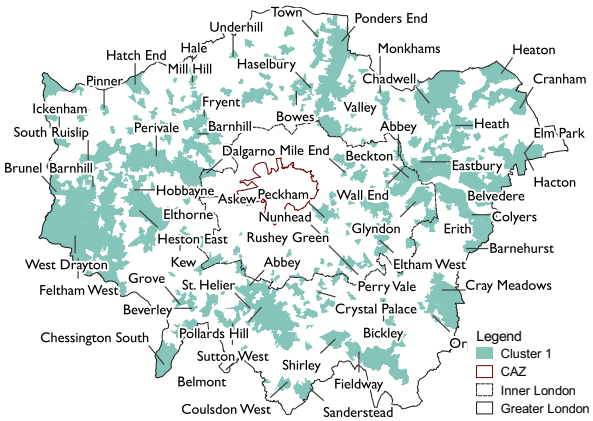
and cultural activities. For example, the topic of tourism includes POI classes such as *Hotels, Motels, Country Houses and Inns*, *B&B Accommodation*, and *Youth Accommodation*. The commercial function can be identified from POIs related to high-end retail and fashion such as *Jewellery, Gems, Clocks and Watches* and *Jewellery and Fashion Accessories*. The business function in this cluster is diverse, involving the media industry, financial services, legal consultancy, property sales, health services and others. The EF result also reveals that POIs in categories such as *Accommodation*, *Employment and Career Agencies*, *Tourism* and *Clothing and Accessories* are representative functions in this cluster. The high EF for *Venues, Stage and Screen* indicates that LSOAs in Cluster 3 also operate as entertainment centres.

Cluster 4 (Local service centres) Most of the LSOAs in Cluster 4 are distributed in suburban areas surrounded by Cluster 1 and Cluster 2 (see Figure 11c). The functions in this cluster are predominantly local services, including convenience stores, retail businesses, and professional support. The high EFs for both *Clothing and Accessories* and *Food, Drink and Multi-Item Retail* show that these LSOAs comprise local retail centres for surrounding neighbourhoods. The EF for *Central and Local Government* in this cluster also indicates that most of the public services and local government offices are within these LSOAs.

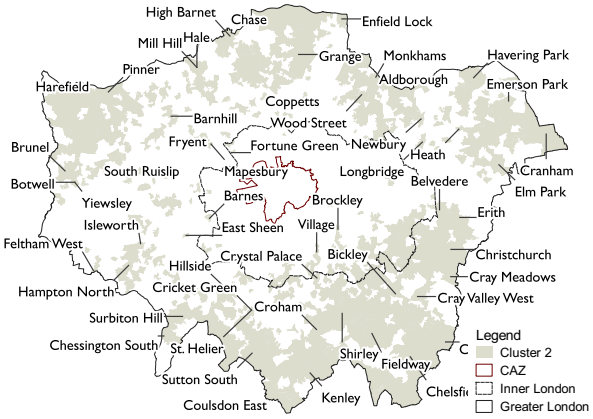
Cluster 5 (Recreation, education, and organisations) The LSOAs in Cluster 5 are mainly distributed in inner London and cover regenerated areas such as Hackney, Islington, Wandsworth, Lambeth and so on(See Figure 13f). The results of the topic extraction show that recreation and residency are the main functions in this cluster. The EF result suggests that *Recreational* and *Sports and Entertainment Support Services* are representative POI categories in LSOAs from Cluster 5. Education-related POIs such as *Education Support Services* and *Primary, Secondary and Tertiary Education* also play a notable role in these neighbourhoods. *Charitable Organisations* and *Headquarters, Administration and Central Offices* also have high weightings in this cluster.



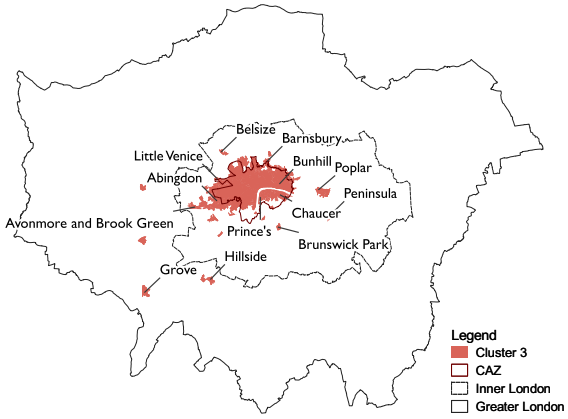
(a) Cluster 0: Affluent neighbourhoods and leisure spaces



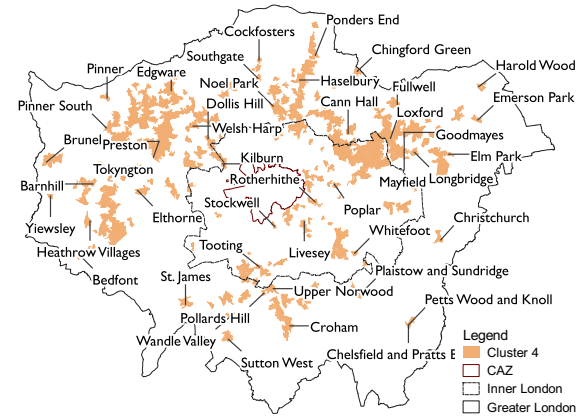
(b) Cluster 1: Municipal facilities and industrial areas



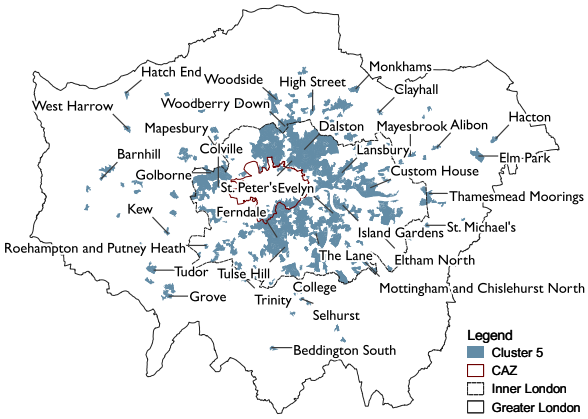
(c) Cluster 2: Suburbs and developing areas



(d) Cluster 3: Vibrant city centre



(e) Cluster 4: Local service centres



(f) Cluster 5: Recreation, education, and organisations

Figure 13: Urban functional areas in Greater London (partial labelled by wards name).

Table 5: An example of topic extraction with LDA for Cluster 0 (20 POI classes in each topic)

Topics	Result (Top 20 POI Classes with weight)
Topic 1	0.065**Hair and Beauty Services* + 0.049**Restaurants* + 0.046**Property Sales* + 0.037**Cafes, Snack Bars and Tea Rooms* + 0.029**Convenience Stores and Independent Supermarkets* + 0.029**Bus Stops* + 0.026**Clothing* + 0.025**Pubs, Bars and Inns* + 0.023**Fast Food and Takeaway Outlets* + 0.022**Cleaning Services* + 0.022**Cash Machines* + 0.019**Design Services* + 0.013**Alternative, Natural and Complementary* + 0.011**Property Letting* + 0.010**Architectural and Building-Related Consultants* + 0.010**Building Contractors* + 0.009**Accountants and Auditors* + 0.009**DIY and Home Improvement* + 0.009**Unspecified and Other Attractions* + 0.009**Places of Worship*)
Topic 2	0.061**Bus Stops* + 0.047**Design Services* + 0.024**Places of Worship* + 0.019**Architectural and Building-Related Consultants* + 0.018**Photographic Services* + 0.016**Marketing Services* + 0.016**Entertainment Services* + 0.016**Computer Systems Services* + 0.015**Estate and Property Management* + 0.015**Tennis Facilities* + 0.014**Alternative, Natural and Complementary* + 0.014**Building Contractors* + 0.013**Nursery Schools and Pre and After School Care* + 0.013**Playgrounds* + 0.013**Mental Health Centres and Practitioners* + 0.012**Plumbing and Heating Services* + 0.012**Unspecified and Other Attractions* + 0.012**Business-Related Consultants* + 0.012**Property Development Services* + 0.012**First, Primary and Infant Schools*
Topic 3	0.096**Bus Stops* + 0.040**Hair and Beauty Services* + 0.040**Convenience Stores and Independent Supermarkets* + 0.034**Unspecified and Other Attractions* + 0.021**Pubs, Bars and Inns* + 0.020**Places of Worship* + 0.020**Cleaning Services* + 0.019**Property Sales* + 0.018**Cash Machines* + 0.015**Cafes, Snack Bars and Tea Rooms* + 0.014**Restaurants* + 0.014**Fast Food and Takeaway Outlets* + 0.014**Design Services* + 0.013**PayPoint Locations* + 0.013**Ponds* + 0.012**Nursing and Residential Care Homes* + 0.011**Nursery Schools and Pre and After School Care* + 0.011**Estate and Property Management* + 0.010**Alternative, Natural and Complementary* + 0.010**Building Contractors*
Topic 4	0.093**Bus Stops* + 0.045**Tennis Facilities* + 0.028**Nursery Schools and Pre and After School Care* + 0.023**Accountants and Auditors* + 0.022**Hair and Beauty Services* + 0.019**Places of Worship* + 0.018**Unspecified and Other Attractions* + 0.018**Cafes, Snack Bars and Tea Rooms* + 0.018**Cleaning Services* + 0.017**Convenience Stores and Independent Supermarkets* + 0.016**Design Services* + 0.015**Pubs, Bars and Inns* + 0.014**Building Contractors* + 0.013**Sports Grounds, Stadia and Pitches* + 0.013**Cash Machines* + 0.013**Business-Related Consultants* + 0.013**Vehicle Repair, Testing and Servicing* + 0.011**Playgrounds* + 0.011**Entertainment Services* + 0.010**Historic and Ceremonial Structures*

5.3. The evaluation results

Table 6 provides the accuracy results of random forest classifiers with vectors of LSOAs trained by TF–IDF, LDA, Word2Vec and the model proposed herein (Doc2Vec). The TF–IDF model transformed POI classes into a sparse matrix of N-gram counts, producing 574-dimensional features of documents. The LDA model provided a distribution of 10 topics for all LSOAs and represented each LSOA as a vector of topic proportions. The alpha parameter was set as 0.5. The Word2Vec model used the same corpus (without LSOA tags) constructed for the Doc2Vec model. We ascertained vectors for LSOAs by averaging vectors of POI classes within LSOAs, which is the method most commonly utilised in previous studies (Yao et al., 2017; Zhai et al., 2019). The dimension of vectors in the Word2Vec model was the same as that of the Doc2Vec model (20-dimensional). To increase the reliability of the results, we iterated the evaluation process 100 times. In each iteration, the dataset was randomly split

into a training set (70%) and a testing set (30%). The result includes out-of-bag (OOB) scores during the training process, and Overall Accuracy (OA) and Kappa scores during the predicting process. The OA of the proposed model (training process) had a 9.3% higher value than that of the TF–IDF model, an 11.5% higher value than that of the Word2Vec model, and a 20% higher value than that of the LDA model. In the prediction process, the Kappa score of the Doc2Vec model was more than 10% higher than those of the other models. The result thus demonstrates that the Doc2Vec model outperforms frequency-based models, such as TF–IDF and LDA.

Table 6: Accuracy assessment of functional area classifications with different language models

Models	Training process		Prediction process
	OOB score	Overall Accuracy	Kappa score
Doc2Vec	**0.486** ± 0.0106	**0.496** ± 0.0086	**0.419** ± 0.0100
Word2Vec	0.376 ± 0.0097	0.381 ± 0.0084	0.286 ± 0.0097
TF–IDF	0.370 ± 0.0083	0.403 ± 0.0086	0.310 ± 0.0098
LDA	0.292 ± 0.0090	0.297 ± 0.0067	0.189 ± 0.0077

6. Discussion

The results of calculating the functional similarity between POI classes demonstrates the diverse and mixed configuration of urban functional use, as well as revealing the correlation between POI classes in Greater London (See Figure 8). For example, we found that entertainment-related POIs have a strong functional similarity with transport POIs such as bus stops, subway stations and taxi ranks, and that real estate-related POIs are often grouped with high-end retail and art shop POIs. By comparing the functional similarity matrix and predefined POI groups in the POI taxonomy, we discovered that only POIs in certain groups such as *Commercial Services*, *Manufacturing and Production* and *Retail* exhibit high functional similarities within their groups, meaning POIs in these groups tend to agglomerate more in the city, compared to those in other groups. Differing from the POI categorisation that classifies POIs by types, the functional similarity matrix measured the spatial context of 574 POI classes and revealed how different POIs appear in London, alongside local characteristics.

The classification results illustrated in Figure 13 identify the city's vibrant areas, not only in the commonly known CAZ, but also in a number of central London, such as Canary Wharf, Richmond, and Wimbledon, where the results reveal similar functional uses, implying that these areas have a similar function as the city's vibrant centre; however, contrary to common perceptions, they are outside of the city's core, which points to a polycentric structure of Greater London. Surrounding central London, there is a spatial division of urban functions between Cluster 0 (affluent neighbourhoods and leisure spaces) and Cluster 5 (regenerated areas of inner London). Eastern areas of inner London predominantly provide functions such

as recreation, education, and workplaces for organisations, while western London areas such as Kensington and Chelsea, Hammersmith and Fulham, Richmond Upon Thames and so on contain more affluent residential areas and corresponding leisure facilities. In outer London areas with low POI densities, the results also identify three clusters (Clusters 1, 2 and 4) with different functions, one of which (Cluster 4) functions as the local service centre. The clustering results presented in Table 6 show that the proposed Doc2Vec model with POI data can better delineate urban functional use, compared with frequency-based models (e.g., TF–IDF and LDA) or other neural network embedding models, such as Word2Vec. The Doc2Vec model performs better because it directly generates vector representations for urban areas from POIs in different classes, whilst establishing the spatial context as a part of a vector.

The results of this study also demonstrate that this enhanced neural network embedding model can be used in exploring the functional similarity between POI classes, outperforming other commonly used language models which utilise POI data to delineate urban functional areas across an urban setting. The functional similarity matrix for POI classes in Greater London (Figure 8) illustrates how 574 POI classes connect with each other, based on their spatial context. The hierarchical clustering result for vectors of POI classes proves that urban function is highly mixed in Greater London, which can be intuitively identified by comparison with the POI classification taxonomy (Appendix A). The vectorised POIs provide an opportunity for exploring the functional similarity between POI classes, while considering the pattern of their spatial proximity. It proves the capability of neural network embedding to reveal the spatial relationship better than frequency-based models can. Although exploring the functional similarity for each pair of POI classes is beyond the scope of this study, this result provides a firm reference from which to explore how different POI classes group together in cities. The other principal result (See Figure 11c) gives an interpretable and reasonable classification of urban functional areas in Greater London.

6.1. Linking previous studies

In comparison to the existing literature on detecting urban function, this study is distinct in two ways. First, it uses POI data to extract urban functional use, because this type of data provides the most granular level of functional use. The results suggest that POI data reflect more socio-economic characteristics of urban land use, rather than the remote sensing data used in previous studies (Hu & Wang, 2013; Joshi et al., 2016; Liu et al., 2017). Regarding studies using social media check-in data (Cranshaw et al., 2012; Gao et al., 2017; Liu et al., 2017; Chen et al., 2017; Zhang et al., 2019), although they consider the intensity of human activity in urban areas, are nonetheless biased due to the digital gaps inherent in the datasets used. Furthermore, they are not suitable for identifying urban functional use in metropolitan areas, where outer areas normally have fewer check-in data. Moreover, as POI data are ubiquitous in cities around the world, the transferability and applicability of this present studys method to other urban cases increases accordingly. Second, the methodology proposed in this study yielded results consistent with other studies (Yao et al., 2017; Zhai et al., 2019; Hu et al., 2020) that have argued that the neural word embedding method offers greater advantages in learning the spatial contexts of POIs and extracting urban functional

use than frequency-based methods, such as TF–IDF and LDA (Yuan et al., 2015; Liu et al., 2017; Gao et al., 2017). This study also demonstrates that the LDA topic model can be useful in interpreting urban functions from clustered areas with disaggregated POI classes.

6.2. Potential in urban planning and management

From the perspective of urban planning and management, the proposed framework has significant potential benefits for different stakeholders, including both the public and private sectors. First, this study proposes a tool for local authorities (i.e., the Greater London Authority) to classify urban functional areas efficiently and at different scales, while maintaining the capability of sensing functional changes in urban areas in a timely manner. For example, in the recently published planning reform white paper (Ministry of Housing, Communities & Local Government, 2020), the UK government proposed that new local plans are required to identify three types of land use for urban development in the future. As the proposed model in this study uses disaggregated POI data with detailed uses of buildings/land, it thereby supports planning authorities in extricating functional uses of urban areas at various scales (e.g., blocks, grids, census units or TAZs), furnishing them with an up-to-date reference when developing new land-use plans.

The UK also recently further deregulated planning regulations by allowing the conversion of some buildings from one class of use to another (e.g., office-to-residential change) without planning permission. Although this decision increases the flexibility and adaptability of the planning process, thereby helping to resolve issues such as housing shortages and a high office-vacancy rate, it undoubtedly makes the ability of local authorities to monitor the changes at the micro-level more difficult. The lack of planning permission records and the unanticipated cumulative effects thereof, may lead to some negative unintentional consequences, exacerbated by a lack of granular. For example, large numbers of office-to-residential conversions, in some communities in London, bring with them substantially increased demand for education, health care provision, parking, and waste disposal, and other associated services; this is because, concordant with the change of occupants, the functional use of these buildings has dramatically changed (Ferm et al., 2020). However, it is difficult for local authorities to plan for these additional demands without applications, at least not until the demand exceeds the capability of local public services to meet them. To sense the dynamic change of urban functional use, planning authorities can apply the Doc2Vec model to London POI datasets in two successive quarters, allowing observations of changes in urban functions by comparing the classification results shown in Figure 11. For example, urban areas that are identified as changing from Cluster 3 (vibrant city centre, mainly referring to commercial and business functions) to Cluster 4 (local service centres providing services for residential areas) suggest areas which are being the areas impacted by the office-to-residential policy. Accordingly, local authorities can begin to promptly plan and implement local policies, such as increasing parking price rates or local council taxes, before seeing a sudden increase in the public service burden.

For private sector groups such as developers and real estate agents, this framework also has great potential for providing building/land use recommendations and selecting locations in terms of functional need. As the Doc2Vec model trains vector representations of POI

classes from its functional context (i.e., co-occurring POI classes in close proximity), in practice, it can be utilised to determine the suggested POI classes from the context POIs in urban areas, such as high streets, commercial centres, communities and so on. For example, by identifying many POIs (including cafes, restaurants, shoe shops and grocery stores) on a segment of a high-street in London, developers could calculate a compound vector by averaging the vectors of those POI classes and comparing them with all vectors of POI classes to find functionally similar POI classes, which could potentially be added to the high-street. This POI recommendation can help developers and real estate managers easily recognise suitable POI classes for assigning a new developments or new uses of property, in concordance with its functional context.

The above opportunities in urban planning and management, either for the public sector or the private sector, can be realised not just in London, but also in other metropolitan cities with highly mixed land use (e.g., Shanghai and Hong Kong), and other regions with flexible planning policies, such as in Australia and New Zealand (Ferm et al., 2020). It can help to identify urban functional use from disaggregated and fine-scale POIs, and provide a tool with which to track functional changes for evaluating and adjusting planning policies. Given the ubiquity and accessibility of POI datasets, the framework could be easily applied to help reveal the urban functional use and track its dynamics, enabling a more enhanced, data-driven support of planning policies.

6.3. Limitations and future work

There are some limitations to the application of this model, data sets and method. First, this study does not include urban activities as part of exploring urban functions (Crooks et al., 2015). The underlying idea of using POI data, as with many previous studies, is that the functional use derived from POI classes represents the preference of urban activities; thus, urban function can be delineated from the POI dataset. This assumption risks overlooking the role of human activities in shaping urban function. Second, there exists no perfect urban functional use dataset for Greater London, by which to evaluate the Doc2Vec model. The absence of such data explains why the land-use classification (Generalised Land Use Database) in the UK provides only simplified classifications (e.g., domestic buildings, non-domestic buildings, roads, green space, water and so on) at the regional level, which is insufficient as a true indication of granular urban functions, especially in a metropolitan area. We eventually selected a compound classification (LOAC) as the alternative to evaluate the classification of urban functional areas in Greater London. The limitations of this dataset go some way to explain the overall low accuracy of results for all of the models shown in Table 6, albeit the model proposed herein demonstrates an unambiguous advantage over the others.

It is worth noting that the reliability of the results of the Doc2Vec model is dependent on the quality of the POI datasets. Those used in this present study are derived from the Ordnance Survey, the national mapping service for Great Britain. Previous studies have used datasets from map services (e.g., Google Place, Baidu Map and Gaode Map), location-based services (e.g., Foursquare and Yelp) or volunteered geographic databases (e.g., OpenStreetMap). This dataset provides POIs with larger coverage, more precise information

and less inaccurate data, which helps to comprehensively reveal the different urban functions in Greater London. For instance, it records POI classes such as industries, farming and infrastructures, which are not normally included in the other POI data sources. Therefore, this study could identify functions such as municipal facilities and industrial areas in the results for urban functional areas, which rendering the classification even more robust.

The proposed methodology provides remarkable potential for exploring cross-sectional and longitudinal urban functions. For developing a better understanding of the similarities among POI classes, it can be used with selected POI datasets for specific locations or local characteristics. For example, by training POIs only from deprived or affluent communities (e.g., in accordance with census data) in the city, researchers can evaluate the relationship between POI classes and socio-economic characteristics. The results of the classification of urban functional areas could also be used to investigate their correlation with socio-demographic characteristics, by combining them with census data. With updated POI datasets, it would also be possible to develop a tool, utilising the Doc2Vec model, to monitor changes of functional use in cities.

7. Conclusion

This study proposes a framework for implementing the Doc2Vec model on POI data. The Doc2Vec model not only vectorises POI classes by considering the spatial relationship among POIs in cities, but also directly generates vectors representing urban areas; this is something which previous studies have not provided. The chosen case study of Greater London demonstrates that the vectors of POI classes can be used to calculate the functional similarity among POIs, and that Doc2Vec model outperforms the TF–IDF, LDA and Word2Vec models in classifying spatial units as functional areas. The framework established in this study provides a bottom-up analysis by efficiently inferring urban functions from the fine-scale POI uses provided by POI data, with assistance from neural network embedding. It allows authorities and policymakers to monitor the urban dynamics, especially of metropolitan areas with higher mixed land use, which hitherto has been a challenging task. Moreover, as ubiquitous POI data becomes more accessible because of the diversity of the data sources, the methods employed in this study can be applied to a broader array of cases.

Appendix A. POI Classification Scheme used by the Ordnance Survey

POI Groups	POI Categories
01 Accommodation, eating and drinking	01 Accommodation
	02 Eating and drinking
02 Commercial services	03 Construction services
	04 Consultancies
	07 Contract services
	05 Employment and career agencies
	06 Engineering services
	60 Hire services
	08 IT, advertising, marketing and media services
	09 Legal and financial
	10 Personal, consumer and other services
	11 Property and development services
	12 Recycling services
	13 Repair and servicing
	14 Research and design
	15 Transport, storage and delivery
03 Attractions	58 Bodies of water
	16 Botanical and zoological
	17 Historical and cultural
	19 Landscape features
	18 Recreational
	20 Tourism
04 Sport and entertainment	22 Gambling
	23 Outdoor pursuits
	21 Sport and entertainment support services
	24 Sports complex
	25 Venues, stage and screen
05 Education and health	26 Animal welfare
	27 Education support services
	28 Health practitioners and establishments
	29 Health support services
	31 Primary, secondary and tertiary education
	32 Recreational and vocational education
06 Public infrastructure	33 Central and local government
	34 Infrastructure and facilities
	35 Organisations
07 Manufacturing and production	37 Consumer products
	38 Extractive industries
	39 Farming
	40 Foodstuffs
	41 Industrial features
	42 Industrial products
09 Retail	46 Clothing and accessories
	47 Food, drink and multi-item retail
	48 Household, office, leisure and garden
	49 Motoring
10 Transport	53 Air

	59 Bus transport
	57 Public transport, stations and infrastructure
	54 Road and rail
	55 Walking
	56 Water

Acknowledgement

This research is supported by a scholarship from the China Scholarship Council (CSC No. 201808060346). We thank the anonymous reviewers for their many insightful comments and suggestions.

References

Barton, C., & Grimwood, G. G. (2019). *Planning: Change of Use*. Briefing Report House of Commons Library.

Biau, G. (2012). Analysis of a random forests model. *The Journal of Machine Learning Research*, *13*, 1063–1095.

Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent Dirichlet Allocation. *Journal of Machine Learning Research*, *3*, 993–1022.

Chen, M., Arribas-Bel, D., & Singleton, A. (2019). Understanding the dynamics of urban areas of interest through volunteered geographic information. *Journal of Geographical Systems*, *21*, 89–109. doi:10.1007/s10109-018-0284-3.

Chen, Y., Liu, X., Li, X., Liu, X., Yao, Y., Hu, G., Xu, X., & Pei, F. (2017). Delineating urban functional areas with building-level social media data: A dynamic time warping (DTW) distance based k-medoids method. *Landscape and Urban Planning*, *160*, 48–60. doi:10.1016/j.landurbplan.2016.12.001.

Cranshaw, J., Schwartz, R., Hong, J. I., & Sadeh, N. (2012). The Livehoods Project: Utilizing Social Media to Understand the Dynamics of a City. In *Sixth International AAAI Conference on Weblogs and Social Media*. California, United States: AAAI.

Crooks, A., Pfoser, D., Jenkins, A., Croitoru, A., Stefanidis, A., Smith, D., Karagiorgou, S., Efentakis, A., & Lamprianidis, G. (2015). Crowdsourcing urban form and function. *International Journal of Geographical Information Science*, *29*, 720–741. doi:10.1080/13658816.2014.977905.

Elwood, S., Goodchild, M. F., & Sui, D. Z. (2012). Researching Volunteered Geographic Information: Spatial Data, Geographic Research, and New Social Practice. *Annals of the Association of American Geographers*, *102*, 571–590. doi:10.1080/00045608.2011.595657.

Ferm, J., Clifford, B., Canelas, P., & Livingstone, N. (2020). Emerging problematics of deregulating the urban: The case of permitted development in England. *Urban Studies*, (p. 0042098020936966). doi:10.1177/0042098020936966.

Forestier, G., Puissant, A., Wemmert, C., & Gançarski, P. (2012). Knowledge-based region labeling for remote sensing image interpretation. *Computers, Environment and Urban Systems*, *36*, 470–480. doi:10.1016/j.compenvurbsys.2012.01.003.

Frias-Martinez, V., & Frias-Martinez, E. (2014). Spectral clustering for sensing urban land use using Twitter activity. *Engineering Applications of Artificial Intelligence*, *35*, 237–245. doi:10.1016/j.engappai.2014.06.019.

Gabaix, X. (1999). Zipf's Law for Cities: An Explanation. *The Quarterly Journal of Economics*, *114*, 739–767. doi:10.1162/003355399556133.

Gao, S., Janowicz, K., & Couclelis, H. (2017). Extracting urban functional regions from points of interest and human activities on location-based social networks. *Transactions in GIS*, *21*, 446–467. doi:10.1111/tgis.12289.

Goldberg, Y., & Levy, O. (2014). Word2vec Explained: Deriving Mikolov et al.'s negative-sampling word-embedding method. *arXiv:1402.3722 [cs, stat]*, . arXiv:1402.3722.

Goodchild, M. F. (2007). Citizens as Sensors: The World of Volunteered Geography. *GeoJournal*, *69*, 211–221. doi:10.1007/s10708-007-9111-y.

Hasan, S., & Ukkusuri, S. V. (2015). Location Contexts of User Check-Ins to Model Urban Geo Life-Style Patterns. *PLOS ONE*, *10*, e0124819. doi:10.1371/journal.pone.0124819.

Hu, S., He, Z., Wu, L., Yin, L., Xu, Y., & Cui, H. (2020). A framework for extracting urban functional regions based on multiprototype word embeddings using points-of-interest data. *Computers, Environment and Urban Systems*, *80*, 101442. doi:10.1016/j.compenvurbsys.2019.101442.

Hu, S., & Wang, L. (2013). Automated urban land-use classification with remote sensing. *International Journal of Remote Sensing*, *34*, 790–803. doi:10.1080/01431161.2012.714510.

James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013a). Tree-Based Methods. In G. James, D. Witten, T. Hastie, & R. Tibshirani (Eds.), *An Introduction to Statistical Learning: With Applications in R* Springer Texts in Statistics (pp. 303–335). New York: Springer. doi:10.1007/978-1-4614-7138-7_8.

James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013b). Unsupervised Learning. In G. James, D. Witten,

T. Hastie, & R. Tibshirani (Eds.), *An Introduction to Statistical Learning: With Applications in R* Springer Texts in Statistics (pp. 373–418). New York: Springer. doi:10.1007/978-1-4614-7138-7_10.

Jiang, S., Alves, A., Rodrigues, F., Ferreira, J., & Pereira, F. C. (2015). Mining point-of-interest data from social networks for urban land use classification and disaggregation. *Computers Environment and Urban Systems*, *53*, 36–46. doi:10.1016/j.compenvurbsys.2014.12.001.

Joshi, N., Baumann, M., Ehammer, A., Fensholt, R., Grogan, K., Hostert, P., Jepsen, M., Kuemmerle, T., Meyfroidt, P., Mitchard, E., Reiche, J., Ryan, C., & Waske, B. (2016). A Review of the Application of Optical and Radar Remote Sensing Data Fusion to Land Use Mapping and Monitoring. *Remote Sensing*, *8*, 70. doi:10.3390/rs8010070.

Le, Q. V., & Mikolov, T. (2014). Distributed Representations of Sentences and Documents. *arXiv:1405.4053 [cs]*, . arXiv:1405.4053.

Li, M., Shen, Z., & Hao, X. (2016). Revealing the relationship between spatio-temporal distribution of population and urban function with social media data. *GeoJournal*, *81*, 919–935. doi:10.1007/s10708-016-9738-7.

Li, W., Jin, G., & Dong, Y. (2019). Scene classification based on the bag-of-visual-words and Doc2Vec models for high-spatial resolution remote-sensing imagery. *Journal of Applied Remote Sensing*, *13*, 026506. doi:10.1117/1.JRS.13.026506.

Li, Y., & Yang, T. (2018). Word Embedding for Understanding Natural Language: A Survey. In S. Srinivasan (Ed.), *Guide to Big Data Applications* (pp. 83–104). Cham: Springer International Publishing volume 26. doi:10.1007/978-3-319-53817-4_4.

Liu, X., He, J., Yao, Y., Zhang, J., Liang, H., Wang, H., & Hong, Y. (2017). Classifying urban land use by integrating remote sensing and social media data. *International Journal of Geographical Information Science*, *31*, 1675–1696. doi:10.1080/13658816.2017.1324976.

Liu, Y., Liu, X., Gao, S., Gong, L., Kang, C., Zhi, Y., Chi, G., & Shi, L. (2015). Social Sensing: A New Approach to Understanding Our Socioeconomic Environments. *Annals of the Association of American Geographers*, *105*, 512–530. doi:10.1080/00045608.2015.1018773.

Lynch, K. (1960). *The Image of the City*. MIT Press.

Ma, L., Liu, Y., Zhang, X., Ye, Y., Yin, G., & Johnson, B. A. (2019). Deep learning in remote sensing applications: A meta-analysis and review. *ISPRS Journal of Photogrammetry and Remote Sensing*, *152*, 166–177. doi:10.1016/j.isprsjprs.2019.04.015.

Mikolov, T., Chen, K., Corrado, G., & Dean, J. (2013a). Efficient Estimation of Word Representations in Vector Space. *arXiv:1301.3781 [cs]*, . arXiv:1301.3781.

Mikolov, T., Sutskever, I., Chen, K., Corrado, G., & Dean, J. (2013b). Distributed Representations of Words and Phrases and their Compositionality. *arXiv:1310.4546 [cs, stat]*, . arXiv:1310.4546.

Ministry of Housing, Communities & Local Government (2020). Planning for the future. https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/958420/MHCLG-Planning-Consultation.pdf.

Niu, H., & Silva, E. A. (2020). Crowdsourced Data Mining for Urban Activity: Review of Data Sources, Applications, and Methods. *Journal of Urban Planning and Development*, *146*, 04020007. doi:10.1061/(ASCE)UP.1943-5444.0000566.

Openshaw, S. (1983). *The Modifiable Areal Unit Problem*. Norwick: Geo Books.

Rong, X. (2016). Word2vec Parameter Learning Explained. *arXiv:1411.2738 [cs]*, . arXiv:1411.2738.

Santos, J. M., & Embrechts, M. (2009). On the Use of the Adjusted Rand Index as a Metric for Evaluating Supervised Classification. In C. Alippi, M. Polycarpou, C. Panayiotou, & G. Ellinas (Eds.), *Artificial Neural Networks – ICANN 2009* Lecture Notes in Computer Science (pp. 175–184). Berlin, Heidelberg: Springer. doi:10.1007/978-3-642-04277-5_18.

Singleton, A. D., & Longley, P. (2015). The internal structure of Greater London: A comparison of national and regional geodemographic models. *Geo: Geography and Environment*, *2*, 69–87. doi:10.1002/geo2.7.

Song, J., Lin, T., Li, X., & Prishchepov, A. V. (2018). Mapping Urban Functional Zones by Integrating Very High Spatial Resolution Remote Sensing Imagery and Points of Interest: A Case Study of Xiamen, China. *Remote Sensing*, *10*, 1737. doi:10.3390/rs10111737.

Soo, K. T. (2005). Zipf's Law for cities: A cross-country investigation. *Regional Science and Urban Economics*, *35*, 239–263. doi:10.1016/j.regsciurbeco.2004.04.004.

Viegas, J. M., Martinez, L. M., & Silva, E. A. (2009). Effects of the Modifiable Areal Unit Problem on the Delineation of Traffic Analysis Zones. *Environment and Planning B: Planning and Design*, *36*, 625–643. doi:10.1068/b34033.

Wang, J., Kong, X., Rahim, A., Xia, F., Tolba, A., & Al-Makhadmeh, Z. (2017). IS2Fun: Identification of Subway Station Functions Using Massive Urban Data. *IEEE Access*, *5*, 27103–27113. doi:10.1109/ACCESS.2017.2766237.

Xing, H., & Meng, Y. (2018). Integrating landscape metrics and socioeconomic features for urban functional region classification. *Computers Environment and Urban Systems*, *72*, 134–145. doi:10.1016/j.compenvurbsys.2018.06.005.

Yao, Y., Li, X., Liu, X., Liu, P., Liang, Z., Zhang, J., & Mai, K. (2017). Sensing spatial distribution of urban land use by integrating points-of-interest and Google Word2Vec model. *International Journal of Geographical Information Science*, *31*, 825–848. doi:10.1080/13658816.2016.1244608.

Yuan, J., Zheng, Y., & Xie, X. (2012). Discovering Regions of Different Functions in a City Using Human Mobility and Pois. In *Proceedings of the 18th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (pp. 186–194). New York, United States: ACM. doi:10.1145/2339530.2339561.

Yuan, N. J., Zheng, Y., Xie, X., Wang, Y., Zheng, K., & Xiong, H. (2015). Discovering Urban Functional Zones Using Latent Activity Trajectories. *IEEE Transactions on Knowledge and Data Engineering*, *27*, 712–725. doi:10.1109/TKDE.2014.2345405.

Yue, Y., Zhuang, Y., Yeh, A. G. O., Xie, J.-Y., Ma, C.-L., & Li, Q.-Q. (2017). Measurements of POI-based mixed use and their relationships with neighbourhood vibrancy. *International Journal of Geographical Information Science*, *31*, 658–675. doi:10.1080/13658816.2016.1220561.

Zhai, W., Bai, X., Shi, Y., Han, Y., Peng, Z.-R., & Gu, C. (2019). Beyond Word2vec: An approach for urban functional region extraction and identification by combining Place2vec and POIs. *Computers, Environment and Urban Systems*, *74*, 1–12. doi:10.1016/j.compenvurbsys.2018.11.008.

Zhang, Y., Li, Q., Tu, W., Mai, K., Yao, Y., & Chen, Y. (2019). Functional urban land use recognition integrating multi-source geospatial data and cross-correlations. *Computers, Environment and Urban Systems*, *78*, 101374. doi:10.1016/j.compenvurbsys.2019.101374.