Proceedings of the 14th International Space Syntax Symposium

93

Exploring Pervasive Centres and Fuzzy Boundaries

A Spatial Configuration Analysis of Commercial Cognitive Regions in Shanghai

DEMIN HU, STEPHEN LAW, KIMON KRENZ & KAYVAN KARIMI
UNIVERSITY COLLEGE LONDON

ABSTRACT

This study is centred on the topic of cognitive regions and boundaries. Previous related research has improved the understanding of the human perception of urban 'places' and has been successful in informing urban planning practice and guiding real estate development. While numerous studies have explored human-perceived regions and boundaries using geographical data, there exists a gap in integrating spatial configuration models with social perceptions of cognitive regions derived from social media data. This study seeks to refine popular identification methods by employing space syntax measures, which have successfully linked spatial morphology to human behaviour, particularly pedestrian footfall. The study focus is on the named boundaries of commercial cognitive regions in the city of Shanghai.

The study's primary novelty lies in the methodological proposal of space-syntax-adjusted clustering methods, named as ssx_AggCluster. This approach builds on space syntax theory, particularly the concepts of 'pervasive centres' and 'fuzzy boundaries' and hypothesizes that space syntax measures may reflect the solidarity and evolution of commercial regional boundaries. The study incorporates commercial Points of Interest (POI) data from social media into a street network model from Open Street Map data and tests three clustering algorithms (DBSCAN, K-means, and Agglomerative Clustering) using network distance. The calculated configuration boundaries of commercial regions in Shanghai are then compared with owner-generated cognitive boundaries from social media data. A space syntax adjusted clustering algorithm successfully classified over 60% of POI points for Shanghai. This finding suggests that spatial configuration plays a significant role in shaping human perception of named places in an urban context.

KEYWORDS

PERVASIVE CENTRES, FUZZY BOUNDARIES, COMMERCIAL REGIONS, SPACE SYNTAX, CLUSTERING ALGORITHM

1 INTRODUCTION

“Place”, which Tuan (1977) defined as “spatial locations that have been given meaning by human experiences”, plays an important role in the organisation of urban life. Various urban places serve as attractors of destination, influencing human perception and behaviour. The seminal work of Kevin Lynch in 1964 shows that the perception of a “place” is related to the physical environment, road network, buildings and natural landscapes such as rivers. However, collecting data on human perception of place is traditionally often small-scale in nature and fuzzy in defining a place’s boundary. As in Lynch and other’s work, the boundary of place can be categorized into two types: functional boundary/region which are often administrative boundaries and perceptual boundary/region which is non-consistent within different groups of people at different time periods.

Our research focuses on the latter – the fuzzy perceptual boundary of place (Wu et al. 2019). To initiate, it is imperative to acknowledge that the city, viewed as a system, embodies a complex network of interactions and structures. Hillier (2007) emphasizes that any model aiming to effectively encapsulate this complexity must transcend traditional, linear narratives, adopting a non-discursive approach to truly reflect the multifaceted nature of urban systems. The boundary of a “place” in human perception is fuzzy in nature and is related to multiple elements in the urban context which can never be constructed as definite as, e.g., in administrative boundaries. However, defining these vaguely defined places in higher resolution is important in urban design as it relates to how we understand, associate, and navigate in a city.

Looking back on the development of space syntax theories, the concept of ‘pervasive centres’ and ‘fuzzy boundaries’ has been developed as a pure typological study of centres and boundaries of neighbourhoods (Hillier 2007). Pervasive centres are “the function of centrality in cities that pervades the urban grid in a more intricate way than has been thought, and that multi-scale centrality should be seen as a pervasive function in cities, with clear spatial correlates, and not simply as a hierarchy of locations”. Fuzzy boundaries (Yang and Hillier 2007) refer to “the area boundaries arising from the way space is structured internally and how this relates to the external structure of space, and so maintaining inter-accessibility between the areas.” These two concepts are derived through an Angular Segment Model (Turner 2004) and are directly linked to integration (to-movement) and choice (through-movement) of the public urban realm. These two concepts have demonstrated a possible link between cognitive regions and the physical urban network. This study hypothesises that human perception of a 'place' could represent a form of solidarity based on the perceived boundaries and centres of a specific 'space'. In contemporary urban environments, commercial areas can be categorized as 'spaces of organic solidarity', as defined by Hillier and Hanson, which exhibit weak control over the boundaries of the region. Prior to this, the contribution of spatial configuration to the delineation of neighbourhoods has been discussed linking space syntax models to Lynch's maps (Dalton 2006).

Although these two concepts have been qualitatively examined in previous research (Davis and Griffiths 2022), the concept of 'place' and 'boundary' has mainly focused on the study of neighbourhoods, particularly residential or administrative neighbourhoods (Law et al. 2017, 2019; Dalton and Hurrell 2022). For example, Law (2017) applied community detection methods to identify street-based local area in relation to housing submarkets. There is generally a lack of research that relates syntactic neighbourhoods and actual cognitive regions defined by a user. One exception is Dalton and Mark's (2023) work which related user-drawn cognitive regions to point intelligibility measures. However, these studies are limited in scale.

With the advent of social media and map services, there has been an increased number of studies applying user-generated data to the study of 'place' (Goodchild 2011; Liu et al. 2015; Wu et al. 2019). This geo-tagged information has been a substitute for the questionnaires and interviews originally used. Among these, the use of points of interest (POIs) and check-in data (McKenzie et al. 2015) has been widely explored to identify the fuzzy boundary of cognitive places using geo-parsing techniques in quantitative geography (Gao et al. 2017). Although these studies have proven that user-generated data can successfully identify the hierarchical and fuzzy nature of neighbourhoods, they do not investigate the extent these data-driven neighbourhoods relate to the spatial configuration of the urban road network. As such, this research aims to utilise only the spatial configuration of the street network in defining street-based cognitive regions which will then be associated with the ground truth user-generated commercial cognitive region from social media data.

The research conjectures that 1) according to prior studies, the morphological typologies of road networks significantly influence human perception of space and boundaries; 2) space syntax measures from segment models have great potential in uncovering and substantiating these hidden typologies, as the concept of Pervasive Centres and Fuzzy Boundaries has demonstrated qualitatively.

2 RELATED WORK

2.1 The human perceptions of place

There have been constant efforts in the field of geography and urban study of modelling the human perception of place. In the 1960s, Kevin Lynch is among the first to investigate this issue. In his book, ‘The Image of the City’, Lynch conducted empirical studies in three American cities: Boston, Jersey City, and Los Angeles through extensive interviews and surveys to understand how people perceive and organize urban information (Lynch 1960). Lynch emphasizes how an individual's understanding and mental mapping of their urban environment shape their experience of the city. Among his five elements of the urban environment, “paths, edges, districts, nodes and landmarks”, he described Districts as *Medium* to large sections of the city that are recognized as having a common character which inspired later study on cognitive places.

Since this research is closely related to human perception, the data collection has been a major effort and challenging task. Studies have been conducted using surveys including interviews and questionnaires to collect human perception of places (Montello 2003, 2014; Nasar 1997). Moreover, Quercia et al. conducted an online survey of 700,000 streets to collect data on the perception of the street in terms of safety and beauty (Quercia et al. 2014). These studies have been focused on the merits of using user-generated content to address the challenge of collecting and summarising people’s perceptions.

2.2 User-generated content and urban places

User-generated content (UGC) has been widely applied in the modelling of places. Points of interest (POIs) and social media check-in data have been the major types of data used in this field of research (McKenzie et al.2015; Li and Goodchild 2012). Li and Goodchild (2012) applied the method of kernel density estimation (KDE) to generate travel patterns from Flickr data as a hint for possible places in the perception of tourists. The significance of building fuzzy boundaries has also been explored in models like the egg-yolk model (Cohn and Gotts 1996) and the 9-intersection model (Clementini and Di Felice 1996). The point-set-based region (PSBR) model (Liu et al. 2010) also proposes a model of approximating regions using KDE. However, few attempts have been made to identify the fuzziness based on road networks.

2.3 Community detection methods

Community detection refers to the process of identifying groups or clusters within a larger network. In urban studies, this often applies to identifying subgroups within urban areas based on various criteria like social interactions, economic activities, and physical characteristics. Community detection methods (CDM) define a set of subgraphs that maximises internal ties and minimises external ties using the topology of the graph (Girvan and Newman 2002; Caschili et al. 2009). The spatial elements including roadwork, street images and building information have been applied in different CDM Workflows (Law 2017).

3 DATASET AND DATA PROCESSING

3.1 Data set

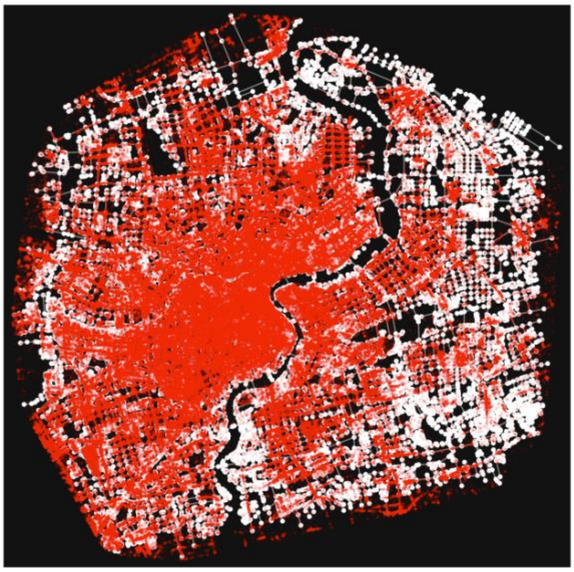


Figure 1: Segment Map of Shanghai & Retail POI Points (Red Points) in 2018

3.1.1 Open Street Map data

This study utilizes the approach outlined in 'Utilizing Volunteered Geographic Information for Space Syntax Analysis' to convert road centreline data of Shanghai into segment models (Krenz 2017). The distinction between a segment model and a road centreline model has been examined in previous research by Turner (2005) concerning road centreline maps. Given the chosen Chinese cities' scale, the disparity between these two models can be considered negligible.

3.1.2 POI data

The primary dataset for Points of Interest (POI) is sourced from publicly available data on the Chinese social media platform dianping.com, specifically from the year 2018 (figure 1). This dataset encompasses a total of 917,186 retail and commercial establishments within the range of Shanghai. The dataset includes various attributes related to each retail POI, including 'shop_id,' 'name,' 'city,' 'province,' 'area_code,' 'phone,' 'region,' 'address,' 'avg_price,' 'cross_road,' 'big_cate,' 'small_cate,' 'stars,' 'review_count,' 'good_remarks,' 'bad_remarks,' 'bookable,' 'take-away,' 'alias,' 'product_rating,' 'environment_rating,' 'service_rating,' 'longitude,' 'latitude,' 'default_pic,' and 'dishes.'

POI data includes, besides locational information, a series of user-generated metadata. This includes a 'region' classification. Shop owners relate their establishments to specific commercial regions, which can be categorized into three main types: administrative regions, commercial centres, and institutional landmarks such as universities or hospitals. In the case of Shanghai, there are a total of 480 unique regions identified in the dataset which are voluntarily provided by the shop owners.

In addition, metadata includes reviews, remarks, and ratings generated by customers who have frequented these establishments and are publicly available. Finally, a 'small_cate' variable describes the commercial land-use type as one of 16 categories, i.e., Food, Service, Shopping, Leisure, Training, Beauty, Kid, Furniture, Wedding, Fitness, Health, Tourism, Vehicle, Pet, Hotel, and Other.

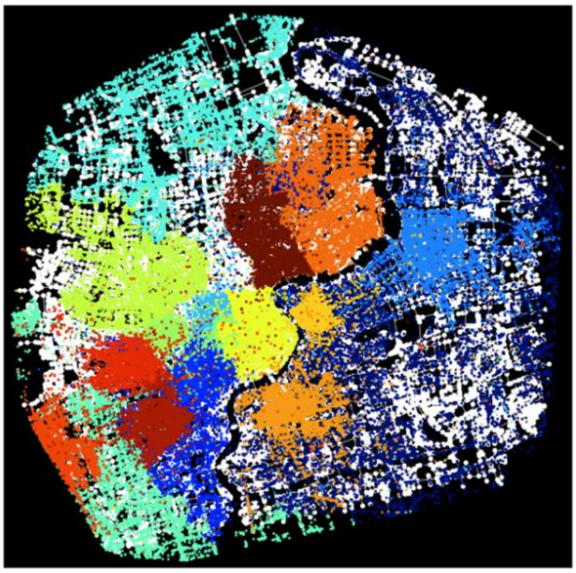


Figure 2: POI Points of 20 Different Regions Represented by Colours

3.2 Data Cleaning for Commercial cognitive region detection

Given that the POI data is user-generated, it's susceptible to containing erroneous data resulting from human errors, biases or outdated information. The initial stage involves data cleansing for the POI dataset, particularly focusing on the 'region' attribute. the section aims to reduce erroneous information. These methods include Kernel density estimation (KDE), standard scores (Z-Scores) and Density-Based Spatial Clustering of Applications with Noise (DBSCAN). These three methods are applied to pinpoint and handle outliers within the POI dataset, with a specific focus on the 'region' attribute.

The Z-score, also referred to as the standard score or standardized score, serves as a statistical metric employed to detect outliers within a dataset. It aids in the identification of data points that exhibit significant differences from the dataset's mean. Kernel Density Estimation (KDE), on the other hand, is a non-parametric statistical technique utilized for the estimation of the probability density function (PDF) of a continuous random variable (Sheather and Michael 1991). The estimation is based on a finite set of data points. When applied to Points of Interest (POI) data, one of the primary challenges encountered is the selection of an appropriate bandwidth parameter. In this study, a user-selected bandwidth is used. DBSCAN, which stands for Density-Based Spatial Clustering of Applications with Noise, is primarily employed for density-based clustering (Liu et al. 2022). It groups data points that are near to each other into clusters based on their density. Nevertheless, data points that do not belong to any of these clusters, meaning they are not sufficiently close to other data points, are considered outliers.

Following the execution of these three different outlier detection methods, the results are assessed using the Jaccard similarity measure (Yue and Murray 2005). Jaccard similarity quantifies the similarity between two sets by comparing the intersection of the sets with their union.

In this research, road network distance measurements are utilized instead of the traditional Euclidean distance commonly employed in geographical studies of Points of Interest (POIs) (Wu et al. 2019). The network distance is posited to reflect real-world urban contexts more closely, particularly in terms of pedestrian movement at a neighborhood scale. This approach potentially offers a more precise depiction of spatial relationships between POI points.

3.2.1 Babaiban example

Figure 3(a) shows the segment model of Shanghai and highlights 14,500 POI in the 'Babaiban' commercial area, a well-known area for shopping for over 30 years. The main feature is the 'Babaiban' shopping mall, located near the main subway station and central to the district. Visual inspection of the map reveals that many POIs stretch along the road network, some even located over 5 kilometers from the shopping mall, the area's focal point. While considering these points as correctly tagged by the owner, there are some points further away from the center which are considered as outliers. These outliers are likely due to human input errors which is detected using three outlier detection methods.

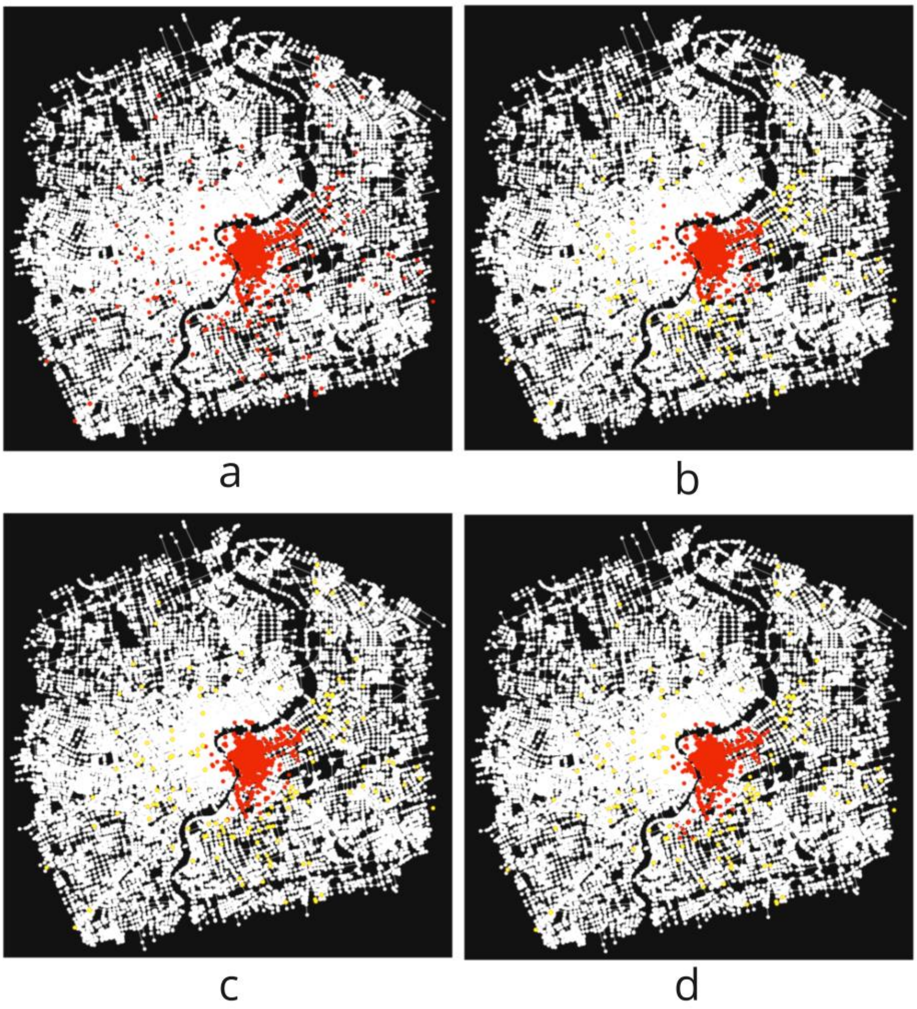


Figure 3: POI Points of 'Babaiban' Region & Noise Detection (Core as Red Points & Periphery as Yellow Points)

Figure 3(b) illustrates the application of the Z-score method for outlier detection, where points with Z-scores exceeding the 5% threshold are filtered out. This method effectively identifies and removes the outliers, aligning with the initial visual assessments.

Figure 3(c) displays the filtering outcome achieved by manually setting the bandwidth of KDE to 800 meters. This choice of bandwidth is informed by the specific conditions in Shanghai, where the minimum distance between two subway stations is approximately 400 meters. As a result, the minimum network distance between these two subway stations typically falls within the range of 800 to 1000 meters.

Figure 3(d) displays the outcome of a DBSCAN analysis with an epsilon (neighborhood distance) set to 800 meters, which aligns with the choice made in the KDE method due to the subway station proximity. Additionally, the minimum number of data points required to form a cluster is manually set to 12 points. In this analysis, data points that do not belong to any cluster are classed as outliers and labelled “-1.”

The Jaccard Similarity values obtained between Z-score and KDE, Z-score and DBSCAN, and KDE and DBSCAN are **0.96**, **0.98**, and **0.98**, respectively. These high similarity scores indicate that nearly identical outliers are identified by these different methods. It can be concluded that the data cleaning result is robust and can be used as a ground truth representation of human perception regarding a commercial region.

Moreover, the calculated center of KDE closely aligns with the predefined center, thus further validating the effectiveness of this outlier removal method. This data processing method is employed to identify commercial cognitive regions which serve as the ground truth for the comparative experiment.

4 METHODOLOGY

Various clustering methods are applied to partition the segment model into distinct sub-networks. The clustering results are then assessed in comparison to the ground truth commercial cognitive region identified in the last section.

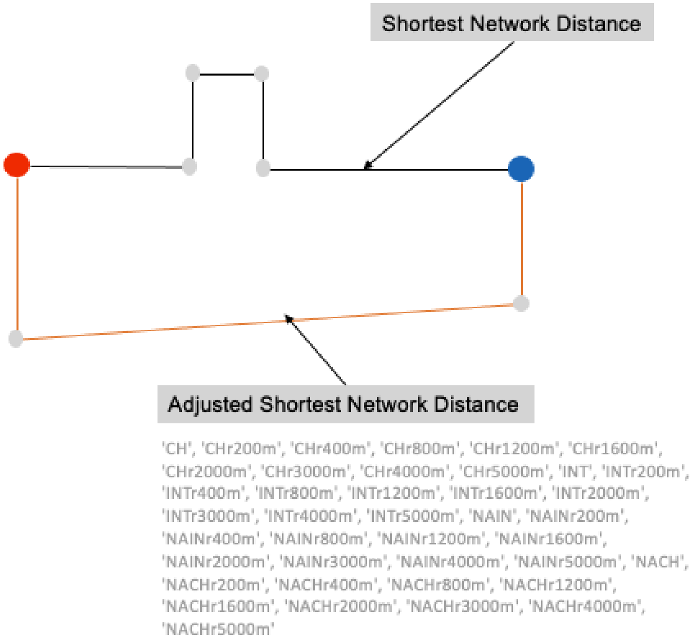
4.1 Density-Based Spatial Clustering of Applications with Noise (ssx-DBSCAN)

DBSCAN, short for Density-Based Spatial Clustering of Applications with Noise, is an unsupervised clustering algorithm employed in machine learning and data science. It was initially introduced by Ester et al. in 1996. DBSCAN's primary function is to group data points that are in proximity within high-density regions while identifying data points in low-density areas as noise or outliers. One of its strengths is its ability to handle clusters of varying sizes without requiring prior knowledge of the number of clusters. The method is sensitive to the setting of two hyper-parameters: epsilon (ε) a parameter that specifies the radius of a neighbourhood around a point and minimum points (MinPts) the minimum number of points to form a dense region. The choice of these parameters significantly impacts the final clustering result (Schubert et al. 2017). While DBSCAN has been widely applied in clustering geographical data, particularly POI data using Euclidean distance (Wu et al. 2019), its application in urban networks (segment models) with network distance has not been extensively explored.

The DBSCAN algorithm can be directly applied to the nodes or edges of a segment model. However, in this study, we propose a new space syntax-adjusted variant here termed ssx-DBSCAN. In space syntax, the concept of movement economy reveals how the layout of physical space influences people's movement and perception (Hillier 2009). As a result, network distance in human perception may be influenced by its underlying space syntax values integration or choice values at both the local and global scales. Consequently, the modified distances between nodes are represented as:

$$dist_{ssx} = \frac{dist}{ssx\_measures}$$

This alteration implies that the identification of the shortest path connecting two nodes within the network may change to include more segments with higher space syntax measures of global integration or choice (refer to Figure 4 for illustration).



'CH', 'CHr200m', 'CHr400m', 'CHr800m', 'CHr1200m', 'CHr1600m', 'CHr2000m', 'CHr3000m', 'CHr4000m', 'CHr5000m', 'INT', 'INTr200m', 'INTr400m', 'INTr800m', 'INTr1200m', 'INTr1600m', 'INTr2000m', 'INTr3000m', 'INTr4000m', 'INTr5000m', 'NAIN', 'NAINr200m', 'NAINr400m', 'NAINr800m', 'NAINr1200m', 'NAINr1600m', 'NAINr2000m', 'NAINr3000m', 'NAINr4000m', 'NAINr5000m', 'NACH', 'NACHr200m', 'NACHr400m', 'NACHr800m', 'NACHr1200m', 'NACHr1600m', 'NACHr2000m', 'NACHr3000m', 'NACHr4000m', 'NACHr5000m'

Figure 4: Adjusted Shortest Distance

In addition to the Space-syntax-adjusted DBSCAN method, this study explores other variations of DBSCAN, for example, Space-syntax-adjusted HDBSCAN hierarchical clustering which we term here ssx-HDBSCAN.

The rationale behind the selection of these two methods is rooted in the observation that in the ground truth data commercial regions often overlap. For instance, the 'Nanjing East Road' and ‘Renminguangchang' regions overlap with each other around the junction of Nanjing East Road and Renmin Square. It is hypothesized that these overlapping regions can be effectively captured through the application of soft clusters and hierarchical clustering. HDBSCAN, introduced by Campello, M., Moulavi, D., & Sander, J. (2015) in their paper titled "Density-Based Clustering Based on Hierarchical Density Estimates," provides a framework that offers both hierarchical clustering and the ability to create soft clusters.

4.2 K-means clustering (ssx-Kmeans)

K-means is a popular unsupervised clustering technique employed in this research which separates data into K distinct, non-overlapping clusters based on K number of randomly chosen initial centroids (MacQueen 1967).

Algorithmically, we first apply Principal Component Analysis (PCA) on space syntax measures namely Choice, Integration, Normalised Choice and Normalised Integration at radii of 200, 400, 800, 1200, 1600, 2000, 3000, 4000, 5000 and radii n, to reduce its dimension before clustering.

Principal Component Analysis is a widely employed dimensionality reduction technique in statistics and machine learning whose objective is to transform a dataset with a high number of dimensions into a lower-dimensional representation, all while retaining as much of the original data's variance as possible. PCA accomplishes this by identifying new orthogonal axes, known as principal components, along which the data exhibits the greatest variance (as initially described by Pearson, K. (1901)).

Subsequently, K-means clustering is applied to the segment model utilizing the PCA components. Various numbers of clusters are tested to explore the optimal clustering solution on the dataset of Shanghai.

4.3 Hierarchical or Agglomerative Clustering (ssx-AggCluster)

Hierarchical or Agglomerative clustering is another popular unsupervised clustering algorithm employed in this research. It constructs clusters by progressively merging individual data points or clusters, driven by a specified linkage criterion (e.g., single linkage, complete linkage, or average linkage). The Agglomerative clustering implementation follows a bottom-up approach, commencing with each data point as a separate cluster and subsequently combining clusters iteratively until a predetermined stopping condition is satisfied (Ackermann et al. 2014).

One notable advantage of Agglomerative clustering is its ability to produce a hierarchical representation of the data, enabling the exploration of different levels of granularity within the clustering results. Additionally, it doesn't require a predefined number of clusters, which can be advantageous in various applications (Ackermann et al. 2014). We will employ a similar space syntax adjustment illustrated in section 4.1 for Agglomerative Clustering. All three clustering methods will be compared against the commercial cognitive region experiment described in the next section.

4.4 Experiment comparing street clusters and commercial cognitive regions

In order to test the efficacy of the three ssx-adjusted clustering methods, a comparative experiment is proposed. The aim of the experiment is to study the extent the three street network clustering methods are coherent with commercial cognitive regions defined by the commercial named-area tagged POI as defined in section 3.2. Figure 5 illustrates a simple example of the ground truth cognitive region data in relation to the three clustering methodologies. Figure 5(a) shows an example of the segment model with retail POI tagged as two different commercial regions (in red and blue) which serves as the ground truth in the experiment. DBSCAN/HDBSCAN/Agglomerative Clustering is conducted on the intersections of segments as shown in Figure 5(b) where the intersections are in two clusters. The POI points are getting a new label of the cluster number by linking to the nearest intersection. Finally, the label of the original 'region' and the 'cluster number' are evaluated through Homogeneity, Completeness and V-measure (Figure 5(d)). Homogeneity, completeness, and V-measure (Rosenberg and Hirschberg 2007) are three commonly used metrics for evaluating the performance of clustering algorithms[1]. These metrics help quantify the effectiveness and accuracy of the clustering algorithms used in the study, providing valuable insights into how well the proposed street-based clustering methods capture the fuzzy boundary of commercial regions. On the other hand, K-means clustering would be applied to the segments and as Figure 5(c) shows, the segments are categorised into two clusters. Then POI points would be assigned to the nearest segment and the 'region' label and 'cluster' label would be compared in the same way.

[1] Homogeneity measures the extent to which all data points within the same cluster belong to the same true class or category. Completeness measures the extent to which all data points that belong to the same true class or category are assigned to the same cluster. V-measure is a balanced metric that combines both homogeneity and completeness to provide an overall assessment of the clustering quality.

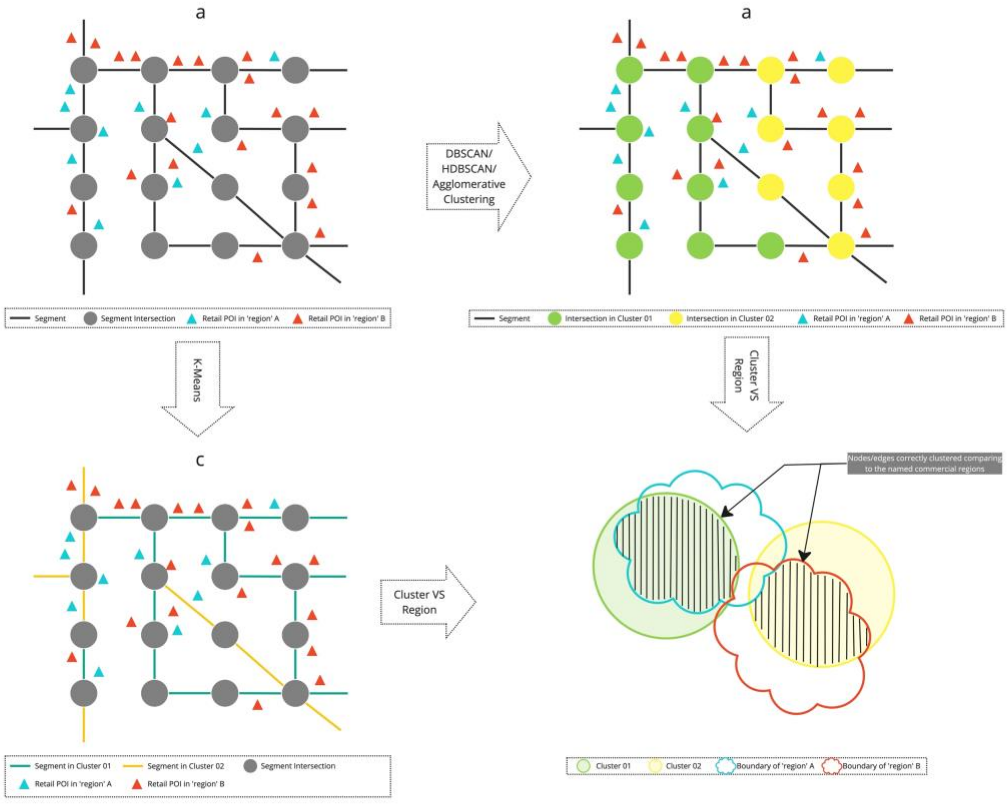


Figure 5: Clusters of road network and named POI points.

5 RESULTS

5.1 Baseline results

Prior to applying the street-based clustering methods as described in the method section, this study conducts a random clustering experiment using only the segment model. The results of 100 random attempts to separate the segment model of Shanghai yield average metrics as follows: **Average Homogeneity: 0.147; Average Completeness: 0.170; Average V-measure: 0.158**. These measures serve as a benchmark or null model for evaluating the performance of the clustering methods. By comparing the results of the clustering methods to these random clustering averages, the study can assess whether the applied methodologies offer meaningful and improved clustering outcomes.

5.2 Space Syntax adjusted Density-Based Spatial Clustering of Applications with Noise (DBSCAN)

The original DBSCAN method was first tested on the Shanghai dataset with minimum sample numbers as 10 and the epsilon (eps) as 400 meters.

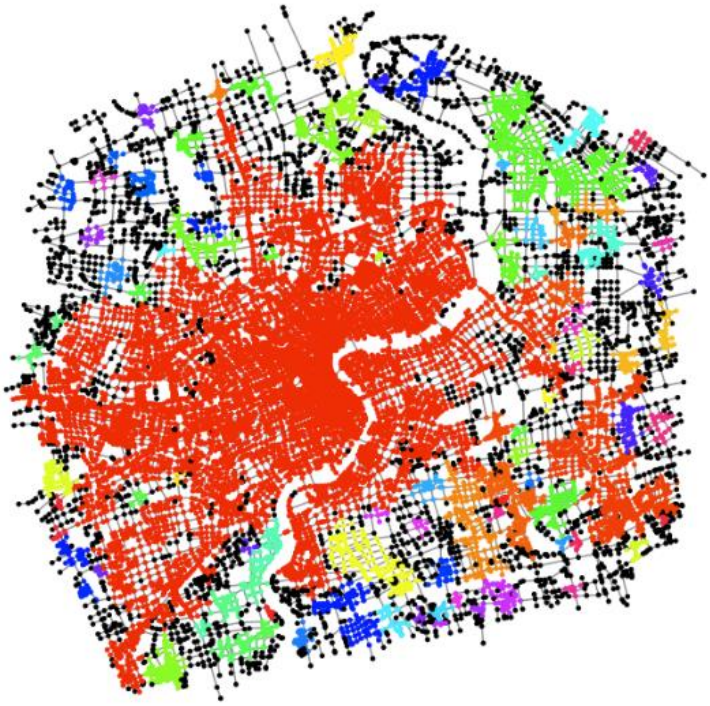


Figure 6: DBSCAN at (400, 10) with Clusters of Different Colours

Figure 6 is showing many noise points as black and the massive central cluster as red. This is due to the calculation of DBSCAN clusters, where the selection of minimum sample numbers and the epsilon (eps) will affect the size of cluster and the global density of cluster. This result indicating two hypotheses: 1) Typology of Road Network Structure: big portion of the road network structure in Shanghai does not exhibit a distinct density characteristic; instead, it appears to be evenly distributed in terms of network distances. 2) Dilemma in Eps Selection: The choice of the minimum distance parameter (eps) has a substantial impact on clustering results. There exists a trade-off between grouping the city centre with higher density and the city border with lowe r density which cannot be achieved at the same time. The reported clustering metrics of network clusters and POI commercial regions (**Homogeneity: 0.145, Completeness: 0.537, V-measure: 0.228**) do not show a significant improvement over the random clustering results. In conclusion, DBSCAN is not an appropriate approach for identifying cognitive commercial regions.

As an improvement, space syntax measures are incorporated as weights for network distances. The segment length is weighted by NAINr2000m and NACHr200m to calculate Weighted Length_IN using the formula: Weighted Length_IN = Segment Length / (NAINr2000m * NACHr200m). DBSCAN is then conducted with varying parameter settings (eps = 200, 400, 800, and min_samples = 10, 20, 30). Figure 7 illustrates that this weighted adjusted DBSCAN approach yields improved results compared to the original algorithm.

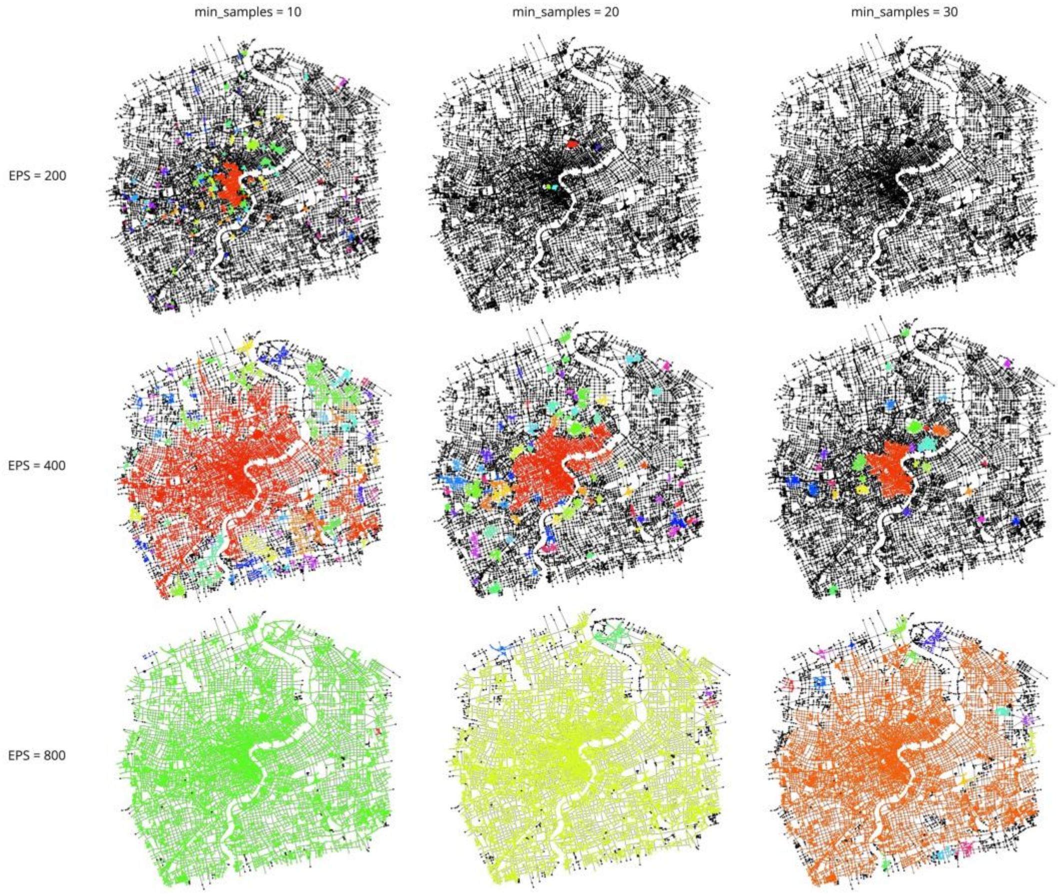


Figure 7: Weighted DBSCAN

The results indicate a notable improvement with the space syntax adjusted DBSCAN method, which we term here ssx-DBSCAN (**Homogeneity: 0.250, Completeness: 0.628, and V-measure: 0.357)**. These improvements suggest that the ssx-DBSCAN more effectively captures the centers of density within the data. However, the issue of noise or outliers becomes severe, largely due to the overlapping nature of the ground truth dataset. To address this challenge, the study then tested the HDBSCAN version, which we term ssx-HDBSCAN at a radius of 400 meters, shown in Figure 8. The results indicate further improvements in **Homogeneity: 0.608, Completeness: 0.524, and V-measure: 0.564**. This outcome signifies that over 50% of the human perception of commercial regions aligns with the clusters generated based on the shortest network distances, weighted by space syntax measures. In conclusion, ssx-HDBSCAN appears to be a promising approach for addressing the overlapping nature of commercial regions and enhancing the clustering performance in this context.

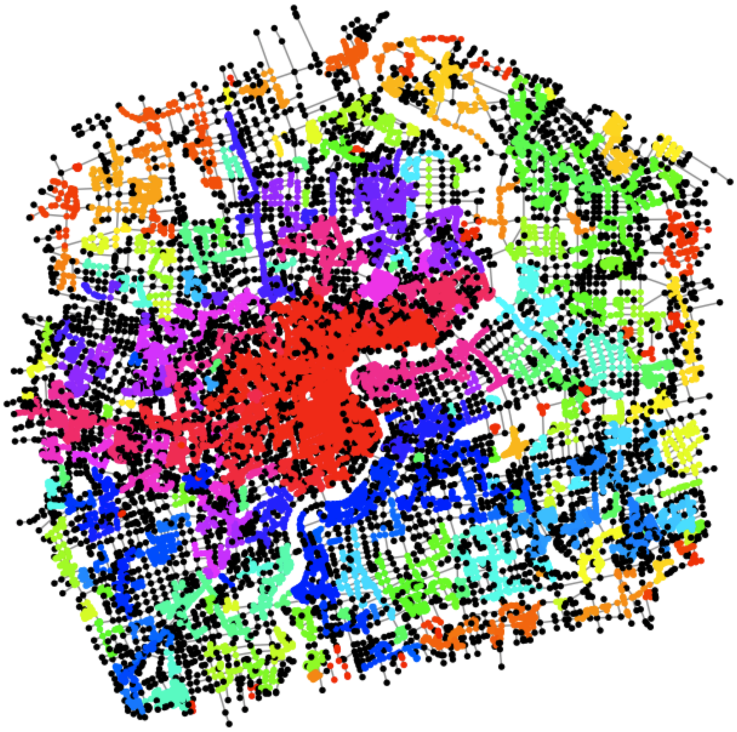


Figure 8: HDBSCAN at (400, 10) with Clusters of Different Colours, Shanghai

5.3 Space syntax adjusted K-means clustering

The next stage is to apply K-means clustering with space syntax measures. We first applied Principal Component Analysis (PCA) on the space syntax measures including Choice, Integration, Normalised Choice, and Normalised Integration at radii of 200, 400, 800, 1200, 1600, 2000, 3000, 4000, 5000 and radii n to identify a sparse set of components that captures the majority of the data variance. By calculating the cumulative variance explained by the principal components for a range of components, the study chooses to retain 14 principal components, as this selection accounts for over 95% of the total variance in the data.

The K-means clustering method is then applied on the projected principal components, varying the number of clusters from 2 to 100. Figure 9 depicts the changes in Homogeneity, Completeness, and V-measure based on the number of clusters (For specific clustering result, please see appendix 01). We term this measure ssx-Kmeans.

The results indicate that these metrics generally increase as the number of clusters rises, especially from 2 to around 60 clusters. However, when the number of clusters reaches 60, the clustering performance reaches a plateau with **Homogeneity: 0.238, Completeness: 0.291, and V-measure: 0.262**. This outcome suggests that ssx-Kmeans, as a clustering method, struggles to effectively partition the segment model in a manner that aligns well with the commercial regions based on space syntax measures. The analysis demonstrates that K-means may not be suitable for this specific clustering task.

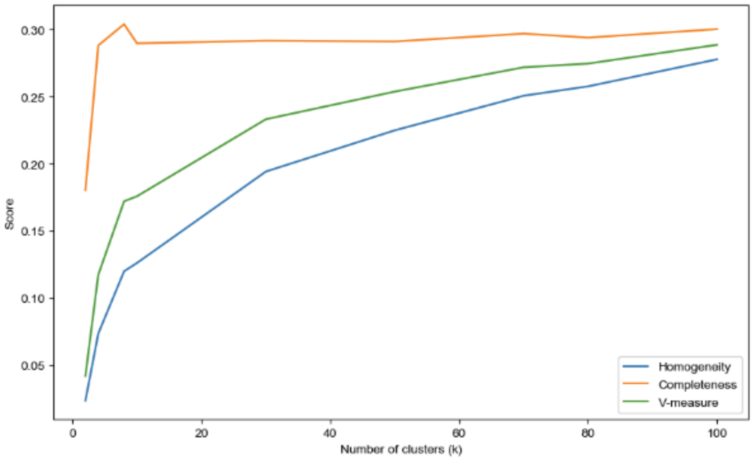


Figure 9: K-means Evaluation Based on Cluster Numbers

The observed results can be explained by the underlying algorithm of K-means. K-means tends to group segments that are distant from each other in physical space but share similarity in terms of space syntax measures. For instance, K-means might place different high street segments with high integration and choice values into the same cluster, even if they belong to different commercial regions that are far apart from each other geographically. This phenomenon is inherent to K-means, as it primarily relies on minimizing the high dimensional distance between data points in the data space, which might not align well with the actual geographical proximity or the human perception of commercial regions.

5.4 Space syntax adjusted Agglomerative Clustering

Figure 10 demonstrates that the Agglomerative Clustering method weighted by space syntax measures, termed ssx-AggCluster, when applied with the number of clusters set to 100 is effective in dividing the segment map into clusters without the presence of noise or outliers.

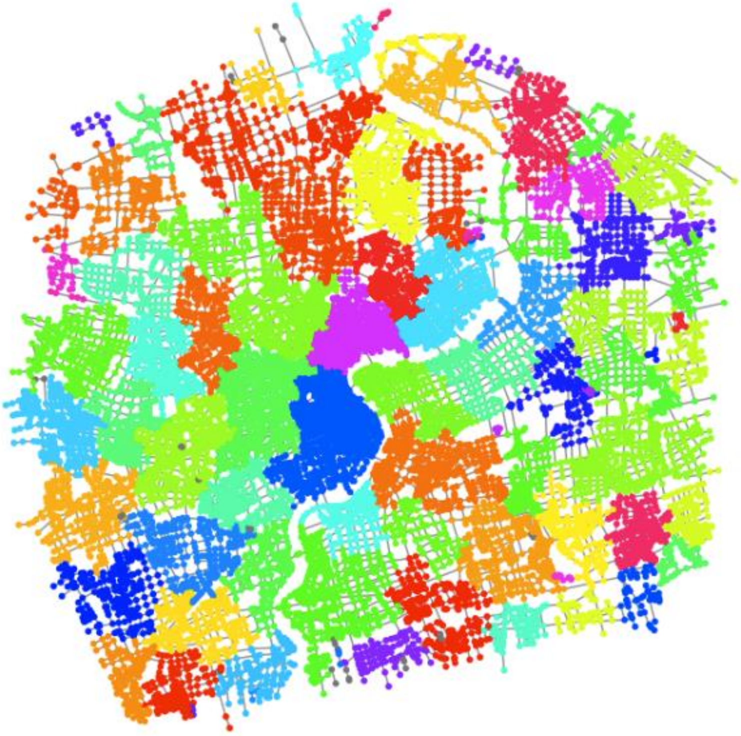


Figure 10: Agglomerative Clustering with 100 Clusters in Colours, Shanghai

The results of the network distance-based clustering using ssx-AggCluster are promising, with metrics of **Homogeneity: 0.583, Completeness: 0.739, and V-measure: 0.652**. This clustering approach has substantially improved the clustering performance, correctly clustering over 65% of the POI points. These results represent a significant improvement over the outcomes achieved with ssx-HDBSCAN, further highlighting the effectiveness of the network distance-based clustering method in capturing the commercial regions and aligning with human perception. For specific regions, the best results are achieved in middle-sized commercial regions, while larger and small regions exhibit less coherence between cluster assignments and regional tags which would be a question for future research.

6 CONCLUSION AND DISCUSSION

In conclusion, this study explored the application of clustering methodologies on urban street segment models and their relevance in successfully identifying commercial cognitive regions for the city of Shanghai. The initial analysis revealed challenges in accurate clustering due to the uneven distribution of the road network structure and the sensitivity of clustering results to parameter settings, such as the minimum distance (eps) in DBSCAN. To address these challenges, the study introduced a novel approach: Space-syntax-adjusted Clustering which incorporated space syntax measures to enhance existing unsupervised clustering methodologies. We tested the adjustment for three methods namely, DBSCAN, Kmeans and Agglomerative Clustering.

The first one we tested is ssx-DBSCAN which resulted in improved clustering results, although noise and overlapping regions remained problematic. Subsequently, the study adopted ssx-HDBSCAN, which further improved clustering performance by accounting for overlapping commercial regions and achieved a high degree of alignment with human perception. The study also tested a space syntax adjusted version of K-means which we termed ssx-Kmeans. However, ssx-Kmeans struggled to accurately partition the segment model based on space syntax measures, highlighting its limitations in this context. The study then tested a space syntax adjusted version of Agglomerative clustering (ssx-AggCluster) which performed significantly better than both ssx-Kmeans and ssx-HDBSCAN. There are two hypotheses for this high performance: a) The street network likely maintains a relatively consistent density across adjacent commercial cognitive regions. At this level, ssx-AggCluster outperforms ssx-HDBSCAN in identifying typological differences and reducing extraneous noise. b) The hierarchical data structure of agglomerative clustering might more accurately reflect the spatial typology in road networks. This approach, building each cluster starting from a single node, may align with human perception in understanding a commercial area from recognition of the region's central activity hub.

The favorable results from ssx-HDBSCAN and ssx-AggCluster confirm the link between the typology of the urban road network and human cognition of commercial regions. Over 60% of the cognitive regions can be explained by the urban road network alone, providing evidence for the pervasive center and fuzzy boundaries. The space syntax measures, which not only link physical space to human movement, significantly contribute to the formation of human perception of commercial regions.

Several limitations remain as the method of other network clustering methods such as percolation and other community detection have not been tested for the same dataset. Using the owner-generated tags might be a biased ground truth to be used. For the next step of this study, other Chinese cities and cities around the world could be compared with the results of Shanghai.

In summary, this research demonstrated that adapting clustering algorithms to account for the urban network structure and incorporating space syntax measures can significantly enhance the representation of commercial regions within a city. The Space syntax adjusted Agglomerative Clustering approach, leveraging network distance and space syntax, emerged as a powerful tool for capturing the underlying structure of the data and aligning with human perception.

REFERENCES

Ackermann, M.R., Blömer, J., Kuntze, D. and Sohler, C. (2014). Analysis of agglomerative clustering. Algorithmica, 69, pp.184-215.

Campello, R.J., Moulavi, D., Zimek, A. and Sander, J. (2015). Hierarchical density estimates for data clustering, visualization, and outlier detection. ACM Transactions on Knowledge Discovery from Data (TKDD), 10(1), pp.1-51.

Caschili, S., De Montis, A., Chessa, A. and Deplano, G. (2009). Weighted networks and community detection: planning productive districts in sardinia.

Clementini, E. and Di Felice, P. (1996). A model for representing topological relationships between complex geometric features in spatial databases. Information sciences, 90(1-4), pp.121-136.

Cohn, A.G. and Gotts, N.M. (1996). Representing spatial vagueness: a mereological approach. KR, 96, pp.230-241.

Dalton, N.S. and Hurrell, M. (2023). Methods for neighbourhood Mapping, boundary agreement. Environment and Planning B: Urban Analytics and City Science, 50(2), pp.401-415.

Dalton, N.S.C. (2006), September. Configuration and neighbourhood: Is place measurable?. In Space Syntax and Spatial Cognition Workshop of the Spatial Cognition'06 conference. Bremen, Germany.

Davis, H. and Griffiths, S. (2022). Bill Hillier, Christopher Alexander and the representation of urban complexity: their concepts of 'pervasive centrality'and 'field of centres' brought into dialogue. In Proceedings of the 13th International Space Syntax Symposium (pp. 1-19). Western Norway University of Applied Sciences (HVL).

Ester, M., Kriegel, H.P., Sander, J. and Xu, X. (1996), August. A density-based algorithm for discovering clusters in large spatial databases with noise. In kdd (Vol. 96, No. 34, pp. 226-231).

Gao, S., Janowicz, K., Montello, D.R., Hu, Y., Yang, J.A., McKenzie, G., Ju, Y., Gong, L., Adams, B. and Yan, B. (2017). A data-synthesis-driven method for detecting and extracting vague cognitive regions. International Journal of Geographical Information Science, 31(6), pp.1245-1271.

Girvan, M. and Newman, M.E. (2002). Community structure in social and biological networks. Proceedings of the national academy of sciences, 99(12), pp.7821-7826.

Goodchild, M.F. (2010). Formalizing place in geographic information systems. In Communities, neighborhoods, and health: Expanding the boundaries of place (pp. 21-33). New York, NY: Springer New York.

Hillier, B. and Hanson, J. (1989). The social logic of space. Cambridge university press.

Hillier, B. (1999). Centrality as a process: accounting for attraction inequalities in deformed grids. Urban design international, 4, pp.107-127.

Hillier, B. (2007). Space is the machine: a configurational theory of architecture. Space Syntax.

Krenz, K. (2017), . Employing volunteered geographic information in space syntax analysis. In Proceedings-11th International Space Syntax Symposium, SSS 2017 (Vol. 11, pp. 150-1). Instituto Superior Técnico, Departamento de Engenharia Civil.

Law, S. and Neira, M. (2019), November. An unsupervised approach to geographical knowledge discovery using street level and street network images. In Proceedings of the 3rd ACM SIGSPATIAL International Workshop on AI for Geographic Knowledge Discovery (pp. 56-65).

Law, S. (2017). Defining Street-based Local Area and measuring its effect on house price using a hedonic price approach: The case study of Metropolitan London. Cities, 60, pp.166-179.

Li, L. and Goodchild, M.F. (2012), November. Constructing places from spatial footprints. In Proceedings of the 1st ACM SIGSPATIAL international workshop on crowdsourced and volunteered geographic information (pp. 15-21).

Liu, J., Qin, H., Liu, Z., Wang, S., Zhang, Q. and He, Z. (2022). A Density-Based Spatial Clustering of Application with Noise Algorithm and its Empirical Research. Highlights in Science, Engineering and Technology, 7, pp.174-179.

Liu, Y., Yuan, Y., Xiao, D., Zhang, Y. and Hu, J. (2010). A point-set-based approximation for areal objects: A case study of representing localities. Computers, Environment and Urban Systems, 34(1), pp.28-39.

Liu, Y., Liu, X., Gao, S., Gong, L., Kang, C., Zhi, Y., Chi, G. and Shi, L. (2015). Social sensing: A new approach to understanding our socioeconomic environments. Annals of the Association of American Geographers, 105(3), pp.512-530.

Liu, Y., Yuan, Y. and Zhang, Y. (2008). A cognitive approach to modeling vague geographical features: a case study of Zhongguancun. JOURNAL OF REMOTE SENSING-BEIJING-, 12(2), p.376.

MacQueen, J. (1967), June. Some methods for classification and analysis of multivariate observations. In Proceedings of the fifth Berkeley symposium on mathematical statistics and probability (Vol. 1, No. 14, pp. 281-297).

McKenzie, G., Janowicz, K., Gao, S. and Gong, L. (2015). How where is when? On the regional variability and resolution of geosocial temporal signatures for points of interest. Computers, Environment and Urban Systems, 54, pp.336-346.

Montello, D.R., Goodchild, M.F., Gottsegen, J. and Fohl, P. (2017). Where's downtown?: Behavioral methods for determining referents of vague spatial queries. In Spatial Vagueness, Uncertainty, Granularity (pp. 185-204). Psychology Press.

Montello, D.R., Friedman, A. and Phillips, D.W. (2014). Vague cognitive regions in geography and geographic information science. International Journal of Geographical Information Science, 28(9), pp.1802-1820.

Nasar, J.L. (1997). New developments in aesthetics for urban design. In Toward the integration of theory, methods, research, and utilization (pp. 149-193). Boston, MA: Springer US.

Pearson, K. (1901). LIII. On lines and planes of closest fit to systems of points in space. The London, Edinburgh, and Dublin philosophical magazine and journal of science, 2(11), pp.559-572.

Purves, R.S., Winter, S. and Kuhn, W. (2019). Places in information science. Journal of the Association for Information Science and Technology, 70(11), pp.1173-1182.

Quercia, D., Schifanella, R. and Aiello, L.M. (2014), September. The shortest path to happiness: Recommending beautiful, quiet, and happy routes in the city. In Proceedings of the 25th ACM conference on Hypertext and social media (pp. 116-125).

Rosenberg, A. and Hirschberg, J. (2007), June. V-measure: A conditional entropy-based external cluster evaluation measure. In Proceedings of the 2007 joint conference on empirical methods in natural language processing and computational natural language learning (EMNLP-CoNLL) (pp. 410-420).

Schubert, E., Sander, J., Ester, M., Kriegel, H.P. and Xu, X. (2017). DBSCAN revisited, revisited: why and how you should (still) use DBSCAN. ACM Transactions on Database Systems (TODS), 42(3), pp.1-21.

Sheather, S.J. and Jones, M.C. (1991). A reliable data‐based bandwidth selection method for kernel density estimation. Journal of the Royal Statistical Society: Series B (Methodological), 53(3), pp.683-690.

Tuan, Y.F. (1977). Space and place: The perspective of experience. U of Minnesota Press.

Turner, A. (2004). Depthmap 4: a researcher's handbook.

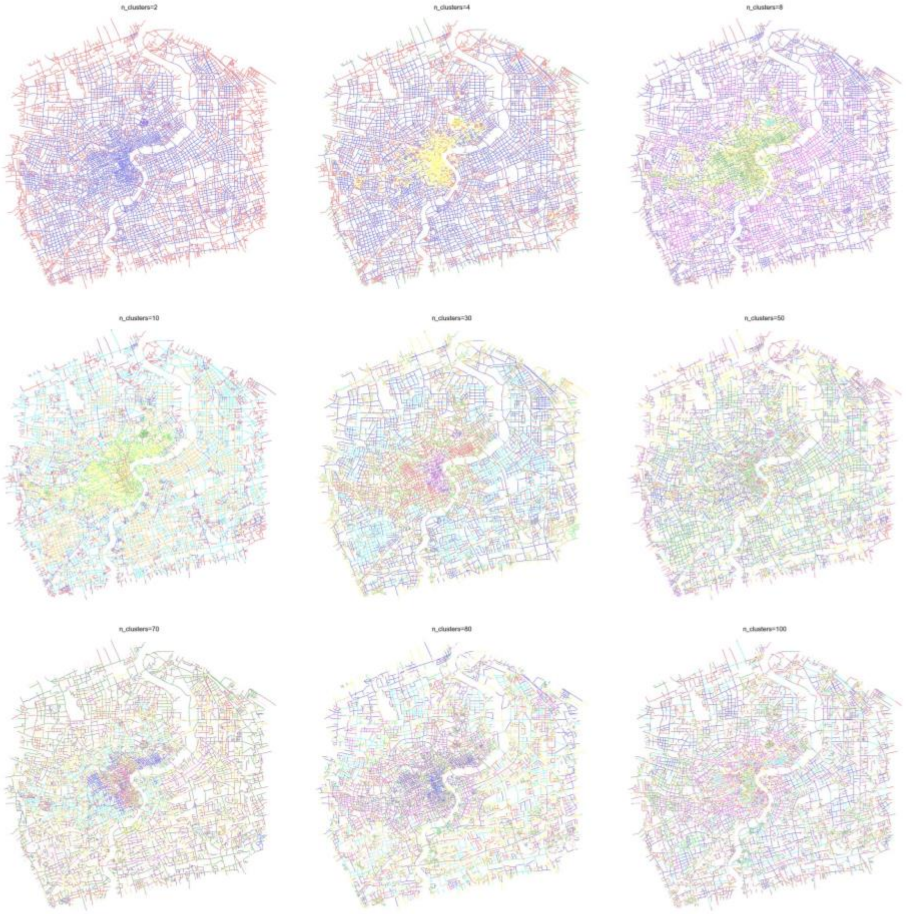
Turner, A. (2005), June. Could a road-centre line be an axial line in disguise. In Proceedings of the 5th International Symposium on Space Syntax (Vol. 1, No. 4, pp. 145-159). Delft: TU Delft.

Wu, X., Wang, J., Shi, L., Gao, Y. and Liu, Y. (2019). A fuzzy formal concept analysis-based approach to uncovering spatial hierarchies among vague places extracted from user-generated data. International Journal of Geographical Information Science, 33(5), pp.991-1016.

Yang, T. and Hillier, B. (2007). The fuzzy boundary: the spatial definition of urban areas. In Proceedings, 6th International Space Syntax Symposium, İstanbul, 2007 (pp. 091-01). Istanbul Technical University.

Yue, J.C. and Clayton, M.K. (2005). A similarity measure based on species proportions. Communications in Statistics-theory and Methods, 34(11), pp.2123-2131.

7 APPENDIX



Appendix 01: K-means Evaluation Based on Cluster Numbers