Computers, Environment and Urban Systems 85 (2021) 101539

Contents lists available at ScienceDirect

Computers, Environment and Urban Systems

journal homepage: www.elsevier.com/locate/ceus

A geographic data science framework for the functional and contextual analysis of human dynamics within global cities

Alessia Calafiore*, Gregory Palmer, Sam Comber, Daniel Arribas-Bel, Alex Singleton

Geographic Data Science Lab, Department of Geography and Planning University of Liverpool, UK

ARTICLE INFO

Keywords:
Foursquare
Geographic data science
Urban analytics

ABSTRACT

This study develops a Geographic Data Science framework that transforms the Foursquare check-in locations and user origin-destination flows data into knowledge about the emerging forms and characteristics of cities' neighbourhoods. We employ a longitudinal mobility dataset describing human interactions with Foursquare venues in ten global cities: Chicago, Istanbul, Jakarta, London, Los Angeles, New York, Paris, Seoul, Singapore, Tokyo. This social media data provides spatio-temporally referenced digital traces left by human use of urban environments, giving us access to the intangible aspects of urban life, such as people behaviours and preferences. Our framework capitalizes on these new data sources, bringing about a novel Geographic Data Science and human-centered methodological approach. Combining network science – a study area with great promise for the analysis of cities and their structure – with geospatial analysis methods, we model cities as a series of global urban networks. Through a spatially weighted community detection algorithm, we uncover functional neighbourhoods for the ten global cities. Each neighbourhood is linked to hyper-local characterisations of their built environment for the Foursquare venues that compose them, and complemented with a range of measures describing their diversity, morphology and mobility. This information is used in a clustering exercise that uncovers a set of four functional neighbourhood types. Our results enable the profiling and comparison of functional neighbourhoods, based on human dynamics and their contexts, across the sample of global cities. The framework is portable to other geographic contexts where interaction data are available to bind different localities into functional agglomerations, and provide insight into their contextual and human dynamics.

1. Introduction

Cities are complex entities at the heart of the main human challenges in the present century. Cities have been framed both as engines of innovation, productivity and sustainability (Glaeser, 2011), but also as sources of concentrated pollution and crime (Bettencourt, Lobo, Helbing, Kuhnert, & West, 2007). Given that the majority of humanity now live within cities (DESA, 2018), it has never been more important to develop better understanding of their main strengths and weaknesses, as well as those mechanisms underpinning their functional structure and dynamics.

An important concept when articulating urban structure is the notion of the neighbourhood. Neighbourhoods can be understood as the building blocks of cities, coherent and meaningful entities that, put together, make up a city and can also be leveraged for policy making (PCAST, 2016). There is a long tradition within the social sciences to frame the neighbourhood as a social construct (Sampson, 2019) that evolves over time, both in nature and extent (Knaap, Wolf, Rey, Kang, & Han, 2019). Equally, neighbourhoods also have a very physical dimension. At their core, a neighbourhood is the geographical representation and delimitation of a place that, while being part of a larger more diverse city, shares similar character, composition, and/or activity patterns. The concept thus plays at the intersection and interaction of the built and lived or experienced environment (Sennett, 2018). In this sense, it relates to that of activity space, although it differs in some crucial aspects. While activity spaces thread together a series of locations based on whether an individual has been in contact with them in their everyday life (Patterson & Farber, 2015), neighbourhoods bring together the individuals that are in contact with a particular set of locations, and build a place from this collection. This paper brings both concepts in conversation.

Scarcity of relevant data is a clear barrier to furthering our understanding of social systems (Lazer et al., 2009), such as cities and neighbourhoods (Arribas-Bel, 2014; Shelton & Poorthuis, 2019). Traditional analyses of neighbourhoods have been limited by the availability of data that meaningfully represent their extent and accurately capture their nature.

* Corresponding author.
E-mail addresses: A.Calafiore@liverpool.ac.uk (A. Calafiore), G.J.Palmer@liverpool.ac.uk (G. Palmer), S.Comber@liverpool.ac.uk (S. Comber), D.Arribas-Bel@liverpool.ac.uk (D. Arribas-Bel), Alex.Singleton@liverpool.ac.uk (A. Singleton).

https://doi.org/10.1016/j.compenvurbsys.2020.101539
Received 20 November 2019; Received in revised form 14 July 2020; Accepted 31 July 2020
Available online 11 November 2020

For the most part, researchers have been limited to the use of official area estimates such as censuses, or bespoke surveys, each of them with their own frequency, coverage and bias issues (Spielman, Folch, & Nagle, 2014). The former provides a representative picture at the expense of low frequency (e.g. every ten years) and spatial aggregations to administrative boundaries that ignore the social nature of neighbourhoods; while the latter provides more detail and fine-grain scale at the cost of limited coverage and rare repeated collection over time.

Advances in Information and Communication Technology such as location-aware technology, sensor technology and mobile technology, have enabled our capability of collecting detailed data about human dynamics (Shaw & Sui, 2018). The traditional data landscape in urban and neighbourhood research is currently being redefined by new forms of data. Characterized by some as a revolution (Kitchin, 2014), the rise of new data sources are making possible analyses on cities and social systems that just a few years ago were unthinkable (Lazer & Radford, 2017). New opportunities to explore life in cities have been enabled by the volunteering of spatio-temporally referenced digital traces left by human use of urban environments (Campbell et al., 2008; Crooks et al., 2015). One recent source of particularly interesting data for understanding cities and the places that make them up are location-based services (LBSs). These are online applications that, thanks to the geolocation and connectivity technology embedded in smartphones, allow their users to broadcast their location at a given point in time to their social network. Through these "check-ins", users contribute to building databases of locations (or "venues", in the LBS jargon) with detailed information not only of their characteristics but also of the type of people that frequent them. Of all LBS services currently available, the most prominent, standalone one is Foursquare[1]. Existing research using LBSs, and Foursquare more specifically, has focused on studying user behaviour (e.g. Noulas, Scellato, Mascolo, & Pontil, 2011), the motivations behind checking in (e.g. Lindqvist, Cranshaw, Wiese, Hong, & Zimmerman, 2011), global mobility patterns (e.g. Noulas, Scellato, Lambiotte, Pontil, & Mascolo, 2012), or on exploring the extent, coverage and implicit biases recorded in these datasets (e.g. Arribas-Bel & Bakens, 2019; Hecht & Stephens, 2014). On their own, data such as Foursquare check-ins provide insight into both where and when activities are taking place within cities. But, as we demonstrate in this study, through their linkage, augmentation and analysis they also provide a great opportunity to model the functional form and characteristics of neighbourhoods within cities.

New urban data require new urban analytics (Batty, 2019; Singleton, Spielman, & Folch, 2017). Developing an understanding of contemporary human mobility, behaviour, context and outcome poses great challenges to many existing instruments that urban scholars have traditionally relied upon for the empirical study of cities (Arribas-Bel, 2014). Because new forms of data do not represent more of the same nature as traditional sources, but qualitatively different typologies and characteristics (e.g. more granular, different sampling strategies and geographical representations), it is important that methods used to fully leverage and unlock their potential recognize it and be tailored to their unique nature. In other disciplines, advances in this direction are being made in the nascent field of Data Science (Donoho, 2017) and, within the discipline of Geography, there are also calls for a Geographic and/or Urban Data Science (Arribas-Bel & Reades, 2018; Organizers et al., 2019; Singleton & Arribas-Bel, 2019).

A methodological area with great promise for the analysis of cities and their structure is network science. Networks are an increasingly important conceptual and methodological tool in contemporary urban theory. They are used to both represent and model various types of interaction, flow or relation (e.g. movement, finance, communication, friendships etc), and to elucidate the hidden structure manifest through the agglomeration of human connections (Nelson & Rae, 2016; Ratti, 2004). Many applications of networks focus on human dynamics across a range of temporal scales: from mapping patterns of global migration to daily commuting patterns (Campbell et al., 2008). Network analysis has also been applied to new data sources, such as LBS. Noulas et al. (2012) use Foursquare data to uncover universal pattern in human urban mobility. In Jiang and Miao (2015), spatial traces from a former location social media, have been used to build an Irregular Triangulated Network and analyse the evolution of natural cities. Jiang and Miao (2015)'s work has demonstrated the potential of these new data sources to generate realistic geographic units bottom-up and, as they put it, social media such as "Fourquare can act as a proxy for studying and understanding evolving mechanism of cities" Jiang and Miao (2015). To obtain geographic units from network data, scholars have increasingly combined the use of community detection algorithms, a traditional network science approach, with the spatial element Ratti et al. (2010), Chen, Xu, and Xu (2015), Gao, Liu, Wang, and Ma (2013), Guo, Jin, Gao, and Zhu (2018).

In the present study, we couple information from Foursquare with (Geographic) Data Science methods to provide a framework that explores the functional structure of cities, how their populations' preferences underpin them, and the relationship between activity spaces and their contexts. We propose and operationalize the notion of functional neighbourhoods, which intersects that of neighbourhood and activity space.

Differently from the traditional conceptualization of neighbourhood – which also accounts for the identity of people inhabiting it (Kallus & Law-Yone, 2000) – we only focus on spaces that originate from people movements. Such movements are quantified via Foursquare origin destination flows – the number of trips that took place by any user from an origin venue to a destination venue – providing knowledge about the functional role of neighbourhoods. By functional here we refer to areas that display a significant degree of spatial interdependency between venues (Dunford, 2009). At the same time, we expand the idea of activity space, usually centered on individuals' day to day experience (Horton & Reynolds, 1971), by disclosing geographic forms derived from collective behaviours. Then we characterise the hyper-local built environment of each venue; and explore how built and experienced environments interact across a wide range of global cities by building a global typology of functional neighbourhood.

Through this work we introduce a framework to generate functional neighbourhoods and provide insights into the human dynamics characterizing them. Taking advantage of the global coverage of crowdsourced geographic data, our framework also implements a set of methods to identify functional neighbourhoods sharing similar properties – in terms of diversity, morphology, people preferences and mobility – across different cities.

The remainder of the paper is structured as follows: Section 2 describes the unique dataset we rely upon to build functional neighbourhoods; Section 3 presents the methodological approach we assemble, including how we delineate neighbourhoods, how they are characterized, and how a clustering exercise groups them in similar types; in Section 4 we discuss our results; and Section 5 concludes with a discussion and ideas for future work in this area.

2. The data

This work concerns data that were acquired for the Foursquare Future Cities Challenge (FCC) [2], which provide a set of longitudinal mobility data describing check in activity (movement) between different venues (POIs) in Chicago, Istanbul, Jakarta, London, Los Angeles, New York, Paris, Seoul, Singapore and Tokyo. The following data are provided for each city:

1. A venue information table, providing, the id, coordinates and category for each venue (See Sub-Fig. 1a)

2. A file listing movements between venue pairs aggregated for month-year and period of the day. Each row consisted of venue1, venue2, month-year, period and flows (See Sub-Fig. 1b)

[1] LBS services are also integrated in larger social media platforms such as Facebook or Twitter.

[2] https://www.futurecitieschallenge.com

	id	name	lat	lng	category
0	4adcda12f964a520643621e3	Les Grandes Marches	48.853009	2.370073	French Restaurants
1	4b9fb106f964a520563537e3	Sofa Café	48.873162	2.333964	Cafés
2	4bfa6835b182c9b625397a5a	Square Paul Langevin	48.847940	2.350379	Parks
3	51007c8ae4b05ffc2bee51a7	Chloé	48.867258	2.326858	Boutiques
4	4bc8648a14d79521fded68e9	Le Bailli de Suffren	48.856703	2.292298	Bistros
...	...	...	...	...	...

(a) Venues

	venue1	venue2	month	period	flows
0	4adcda08f964a5206f3321e3	4adcda05f964a5208d3221e3	2019-03	NIGHT	1
1	4adcda10f964a520af3521e3	4adcda05f964a5208d3221e3	2018-10	OVERNIGHT	1
2	4c9f8907d3c2b60cd468d5bc	4adcda05f964a5208d3221e3	2018-05	OVERNIGHT	1
3	4b52428bf964a5205e7327e3	4adcda05f964a5208d3221e3	2018-10	MIDDAY	1
4	4b52428bf964a5205e7327e3	4adcda05f964a5208d3221e3	2018-11	OVERNIGHT	1
...	...	...	...	...	...

(b) Movements

Fig. 1. Example rows extracted from the venues and movements tables for Paris.

The data extract refers to April 2017 to March 2019, with each entry assigned into a time category determined by the time of arrival at venue2: overnight (00:00:00–05:59:59); morning (06:00:00–09:59:59); midday (10:00:00–14:59:59); afternoon (15:00:00–18:59:59); and night (19:00:00–23:59:59). A flows column provides a count of the number of times any user is recorded as travelling between venue pairs during a given month-year and time category [3]. From this column the number of check-ins at a venue can be calculated through summing the number of incoming flows arriving at venue2. The data aggregations are implemented by Foursquare to mitigate privacy concerns. Beyond the necessary aggregations, there are two further limitations of the FCC dataset: (i) no data are provided for Brooklyn, USA despite the rest of NYC having coverage; (ii) very granular venue categories are provided (e.g. 215 food outlet types, and 813 categories overall).

While Foursquare also provides a hierarchical venue type schema, it consists of multiple hierarchical levels. This categorical branching is unsuitable for our evaluation, as we only require one additional level of coarser categories. In addition, Foursquare's top level categories are too general for our current analysis. Therefore, to address the second limitation we manually define our own aggregate groups to simplify the activity types. While the coarse categories are similar to those defined within Foursquare's own hierarchical venue type scheme, there are a number of key differences that make the current set of categories more suitable for our work. For instance, we consider that spiritual venues, e.g., Buddhist Temples, Churches and Mosques, are worthy of their own category. In contrast Spiritual Centres is a sub-category under Professional & Other Places within Foursquare's hierarchy. In addition we separate Professional & Other Places categories, with our other category consisting of Lines / Queues and Public Bathrooms. We also identify more suitable coarse categories for a number sub-categories. For example, we consider the category Shops & Services too general, and as a result create the category coffee to capture a group of sub-categories for which a large number of venues exists within the data-set. Finally, given that there are no inter-city flows, and the fact that interesting flows exist within airports (between Airport Food Courts, Airport Gates, Airport Lounges, Airport Services, etc), we define airport as category independent from the more general travel category. Fig. 2 provides a bar chart illustrating the category frequencies across all cities by the new classification.

3. Methods

The overarching objective of this paper is to provide insight into emergent urban structures and dynamics for ten global cities. The framework designed for this study is illustrated in Fig. 3. As reported above, two input data are employed: 1) Foursquare's venues; 2) people movements between venue pairs. The latter enables the development of a set of urban mobility networks to which a spatially augmented community detection algorithm is applied (Section 3.1). The resulting geographic units – functional neighbourhoods – are cohesive zones of interconnected agglomerations of activity emerging from human mobility dynamics. Along with the analysis of people movement, Foursquare's venues are contextualized within ten minutes walk catchment areas through several metrics describing the built environment and behavioural patterns (Section 3.2). To compare the form and functions of neighbourhoods we posit three drivers of differentiation that include: context, mobility and diversity (Section 3.3). While contextual metrics are ascribed to each venue and based on its catchment area (see left hand side of the diagram in Fig. 3), analytics on mobility and diversity are derived from the urban networks underlying each functional

[3] We visualize the number of flows per month-period combination for each city in Appendix B.

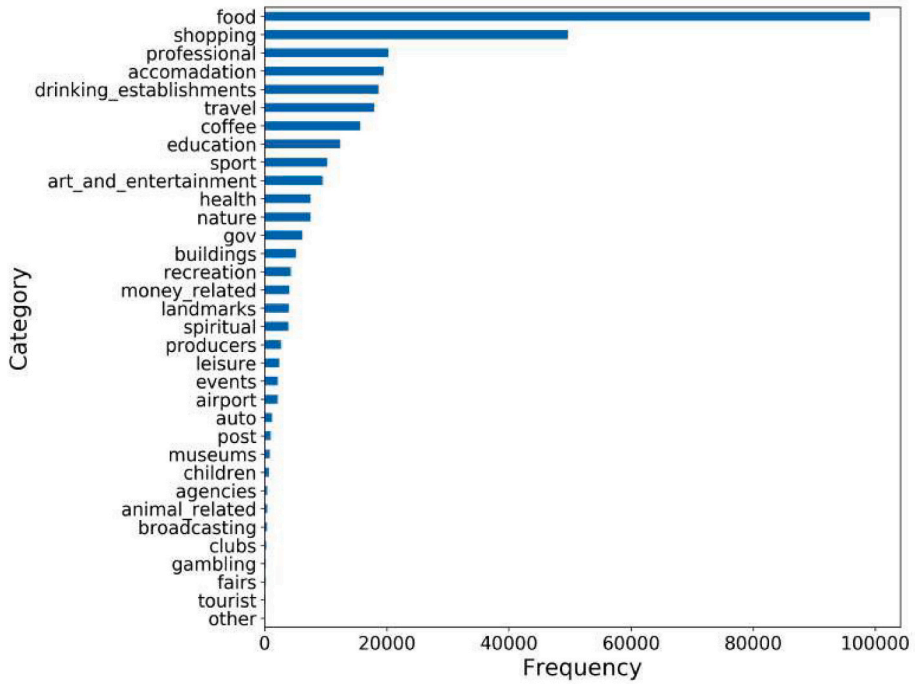


Fig. 2. The frequency of venue category types across all cities.

neighbourhood. To upscale the former at a functional neighbourhood level, all measures linked to venues belonging to the same neighbourhood are averaged and then reduced through Principal Component Analysis. Finally, the functional neighbourhoods are clustered to highlight global similarities across cities. Results are discussed in Section 4.

3.1. Urban networks & functional neighbourhoods

The first stage of our analysis is to detect areas of functional structure that emerge across different global cities by introducing and operationalizing the notion of functional neighbourhoods.

Such geographic unit can be seen from two different angles: 1) a typological perspective where the term functional directly refers to the variety of activity types that certain areas can afford (Assem, Xu, Buda, & O'Sullivan, 2016; Gao, Janowicz, & Couclelis, 2017); 2) an organizational perspective which considers functional those areas made up of places that display a significant degree of spatial interdependency (Dunford, 2009). The definition of neighbourhood developed in this study is based on the latter perspective. We capitalize on the Foursquare flows data and employ community detection to delineate neighbourhoods where venues are strongly interconnected through flows of people, therefore suggesting a high level of interdependency. The use of community detection on mobility networks is not new (Chen et al., 2015; Gao et al., 2013; Guo et al., 2018; He, Glasser, Pritchard, Bhamidi, & Kaza, 2019), which are more representative of people's activity space (Patterson & Farber, 2015). However, one problem of these derived geometries is that they tend to have a much higher degree of overlapping than statistical unit. As Nelson (2020) puts it "the key role of proximity in the organization and structuring of regions is challenged by the inside and outside criss-crossing of several flows".

In this study, we have approached community detection to identify functional neighbourhoods and balance interdependence between venues, even when spatially dispersed, and proximity.

Below we first discuss the notion of spatially weighted community detection, then proceed on our partition selection process, before discussing the resulting communities.

Spatially weighted community detection: The structure of a network is typically measured using modularity, with dense connections between vertices within individual communities and sparse connections between communities having a higher modularity. A widely used community detection algorithm based on modularity maximization is proposed in Clauset, Newman, and Moore (2004). It identifies partitions of a network characterized by a high modularity value. However, this approach does not take into consideration the spatial dimension, which significantly limits the analysis of networks representing relations among geographical objects. In recent years, some scholars have introduced different methods to include space in the community detection process.

Gao et al. (2013) adopt a hierarchical agglomerative clustering algorithm based on a Newman-Girvan modularity metric and an alternative modularity function incorporating a gravity model to study the dynamics of spatial interaction communities. Such modularity measure compares the real number of edges within communities with the same estimated value under a random model. This approach favours communities where the number of edges is higher than expected based on the gravity model. While it increases the probability of generating disconnected communities, we aim at partitioning interdependent but possibly adjacent venues to generate non-overlapping, and therefore more usable, geographical units.

Another method to include the spatial element into community detection is described in (Chen et al., 2015). A distance decay function $P(d) \sim d^{-n}$ is employed to measure the likelihood of a connection between two nodes and weight the network accordingly, where d is the Euclidean distance between nodes and n would depend on the size and compactness of the network. Following this approach the optimal value of n – to maximize the modularity of each city's network – has to be selected.

An extension of (Chen et al., 2015)'s approach to obtain a more stable geographic whole is presented in Guo et al. (2018). Such a method successfully encloses nodes into proximal communities by implementing a spatially contiguity constraint. However, constraining the partitioning process risks to hide relevant underlying movement patterns between venues located far away. Differently, this work's objective is to balance the need of proximity required to obtain usable geographic units with the possibility of relevant, although spatially distant, relations.

To achieve such objective we propose a selection of the optimal partitioning by tuning the distance function exponent n – employed to weight edges – and the resolution parameter as described below.

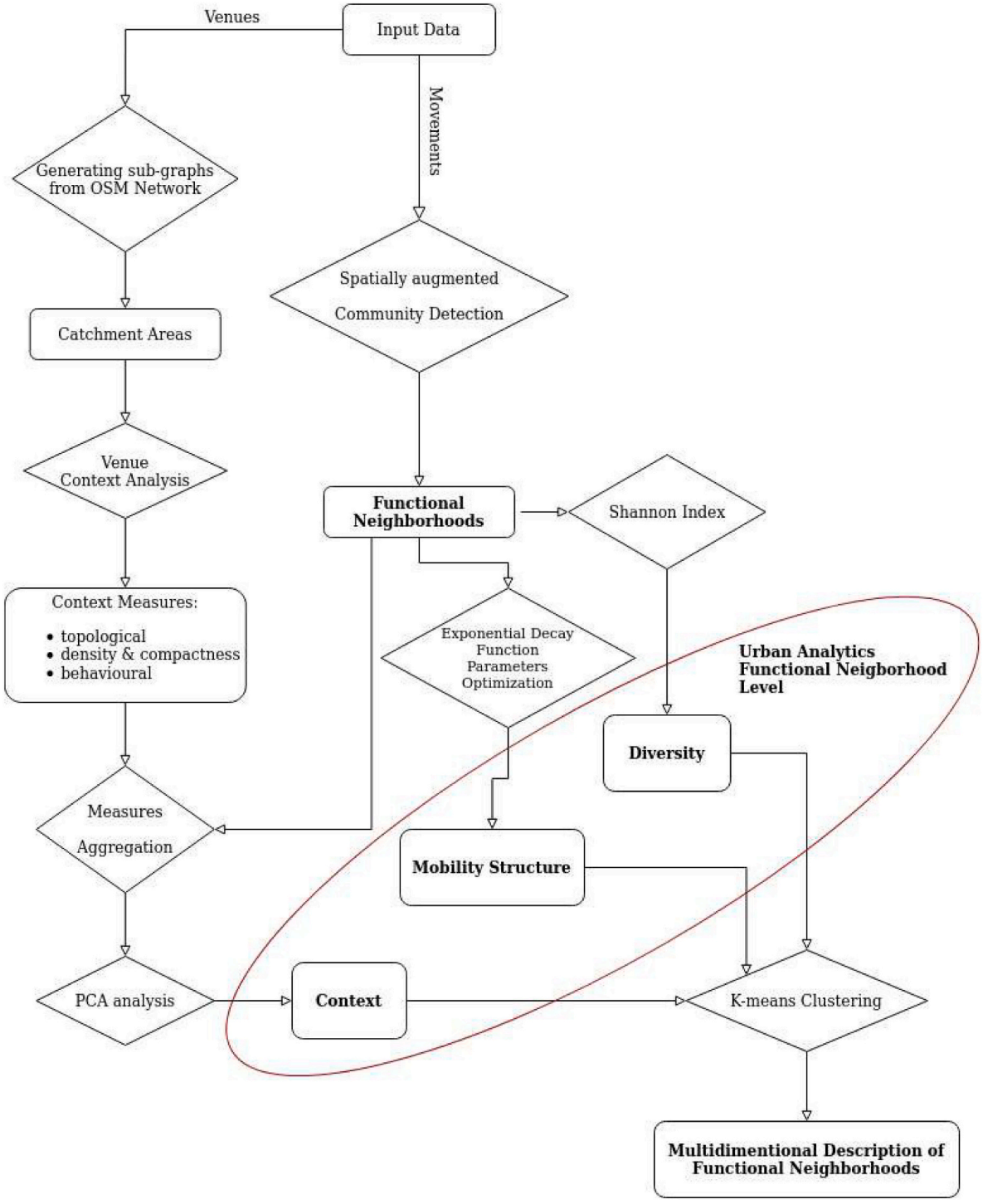


Fig. 3. Illustration of the geographic data science framework for the functional and contextual analysis of human dynamics. Methods implemented are in diamonds. A detailed description of the methods can be found in Section 3. Input and output data are in squares, while the main results are highlighted in bold. The two input datasets (Venues and Movements) are detailed in Section 2. Results are discussed in Section 4.

Partition selection process: We create cities' networks using Python's NetworkX library (Hagberg, Swart, & S Chult, 2008), with nodes representing each venue and edges determined by the existence of users flows between them. Edges' weights are therefore calculated as follow: $w \sim x * d^{-n}$, where x is the number of flows.

Community detection is implemented using Python's community package's best partition method that searches for network partitions by maximizing the modularity using the Louvain heuristics.

To identify the best partitioning for each city's network, we try to maximize modularity while minimizing the number of nodes where all spatially adjacent venues belong to a different community. We refer to such cases as "outliers". Therefore, we conduct a hyperparameter sweep using exponents $n \in \{1.0, 1.5, 2.0, \ldots, 3.5\}$ and the Louvain algorithm's resolution parameter, $res \in \{0.5, 1.0, 1.5, \ldots, 6.0\}$. The resolution determines the time-sale of the community detection algorithm, whereby increasing the resolution produces a larger number of smaller communities (Lambiotte, Delvenne, & Barahona, 2008). We identify the outliers for each partition by looking at the Triangulated Irregular Network obtained through the

Delaunay triangulation algorithm, and select those venues where all spatial neighbours belong to a different community.

For the majority of the cities (London, Paris, New York Seoul, Los Angeles, Singapore, Tokyo and Chicago), the highest modularity is achieved using either a decay exponent $n = 2.0$ or $n = 2.5$, while the highest modularity resulted from $n = 1.5$ for Jakarta [4]. In contrast, for Istanbul a modularity of 0.99 is achieved for all of the listed exponents $n \geq 2.0$ (See heatmaps in Appendix C for an overview). We observe that as the exponent n increases so does the number of outliers (See Fig. D.18 in the Appendix). To identify which partitioning parameters (n and res) minimize the number of outliers while maintaining the highest possible modularity, we compute an evaluation metric:

$$s_i = z_{mi} - z_{oi} \quad (1)$$

In the above equation we compute and subtract the z-scores for the modularities $m \in M$ and the outliers $o \in O$ for each hyperapraeter combination i. The motivation for subtracting the z-score for the number of outliers is that we want to identify partitions with a low number of outliers. For each city we choose the partition that maximizes the metric s. Interestingly, we find that for each city an exponent $n = 1.5$ is optimal as a result of the cities having a similar distance decay pattern (see Fig. 4).

Outcome: The outcome is a set of functional neighbourhoods for each of the cities (See Table 1 for details and Fig. 5 for an illustration of the neighbourhoods for London and New York). Polygon boundaries are created by excluding the outliers and applying an Alpha Shape (Kittel, 2019) to the associated venue locations within each identified community. The algorithm proposed by Edelsbrunner, Kirkpatrick, and Seidel (1983) is used to automatically determine the alpha value that enables the tightest polygon that contains all points for each neighbourhood. Each functional neighbourhood corresponds to a community of nodes, representing a distinctive aggregate Foursquare user activity, and bringing together venues with a high degree of interdependence.

While we capture the outliers – individual venues not located in the spatial proximity of their community – we note the existence of clusters with few venues spatially separated from the community they belong to. By applying alpha shapes on each community we manage to visually identify these small clusters of venues, which highlight interesting underlying patterns. In London for instance the majority of venues within community 18 are located in the north east Holloway area. However, we observe that the community includes a sub-community of venues – that have not been classified as outliers – located at Heathrow Airport (see Sub-Fig. 5a). We find that 595 edges with venues belonging to the category "travel" connect this set of venues with the venues located around the Holloway area. The FCC data therefore captures the transportation links that exist between these two locations (e.g., via the Piccadilly line). Meanwhile, for New York we observe that a significant number of flows occur between community 14 and the East River Tunnels (see Sub-Fig. 5b).

3.2. Contextualizing venues locations

Venue linkage through user interaction and their geographic location drive the spatially differentiated functional neighbourhoods demonstrated in the previous section; however, at the local scale, these patterns of use are driven by the venue type (e.g. travel, food etc); and other contextual measures. Some literature within urban planning and architecture explores how the morphology of the built environment (e.g. street geometry) may influence activity within places and limit or enhance attraction between locations (Ratti, 2004). Previous studies have examined the impact of how space syntax variables relating to urban morphology influence the propensity for non-motorised transport modes (Rybarczyk & Wu, 2014) and walkable pedestrian spaces (Frank et al., 2009). In our case, the objective here is to capture a collection of contextual measures that characterise each

[4] Previous studies have often found n ~ 2.0 to be optimal across a range of data-sets (Chen et al., 2015).

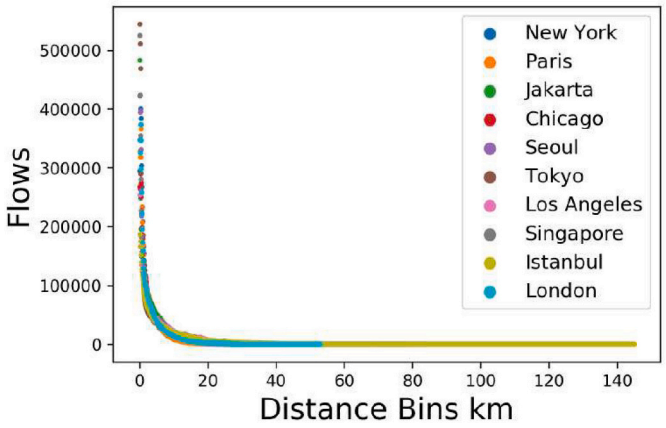


Fig. 4. The proportion of flows by distance using 100 m bins.

check-in by aspects of street topology, density, connectedness and behavioural patterns. Thus, we contextualise venues locally before aggregation within functional neighbourhoods to facilitate a comparison.

The first stage requires creation of a "catchment" area for each of the 330 k venues; which are defined as a ten minute walk around each Foursquare venue. A polygon delineating the bounding box of all venue locations within each city was used to create a NetworkX (Hagberg et al., 2008) graph from OpenStreetMap (OSM) data using the OSMnx library (Boeing, 2017). For each city level graph, a sub graph is induced outwards from the venue location. Those nodes and edges captured by each sub graph are selected according to a ten minute walking distance (see Fig. 6). Assuming the average person walks a speed of five kilometres per hour, an individual covers 800 m in a ten minute walk from each Foursquare venue. This is considered in the "Manual for Streets" by the Bradbury et al. (2007) as the timescale and distance of a walkable built environment, and so represents a sensible catchment reflective of the urban environment immediately accessible to the individual from the venue. From these walk radii, the convex hull of nodes within the sub graph that are accessible within a ten minute walk are then extracted; which are the catchment areas used in the subsequent analysis.

For each catchment area, we define a series of measures that are summarized in Table 2. These are derived from both Foursquare activity to highlight spatio-temporal dynamics of check-in behaviour, and a variety of street network measures that expose the morphological structure of the catchments. Measures of topology, density and connectedness have been shown to reveal insight into common mobility and design characteristics that differentiate pedestrian-orientated environments from those which are more auto-orientated. For example, 'average circuity' measures the extent that our catchments deviate from the spatial ordering logic of dense (orthogonal) grids to sparse networks of circuitous, curving streets that form loops and lollipops (Ewing & Cervero, 2010). Unlike street networks with high curvature, a gridiron geometry allows a longer line of sight that enables pedestrians to better visualize their surroundings and navigate across their environment (Hajrasouliha and Yin, 2015). Another measure, 'average street length', provides a linear approximation of block size (Boeing, 2017). Further topology-based characteristics such as 'street per node average' measures the mean number of physical streets that emanate from street intersections and dead-ends, which proxies the complexity of streets.

In addition to topology, we derive several density-based measures that describe characteristics teristics such as walkability of the urban form. 'Node density', the number of intersections divided by the area covering the network, relays information about the extent of street connectivity. More intersections are generally suggestive of an environment more amenable to pedestrian walkability (Frank et al., 2009), and have been used previously to derive scores that are accessible to the general public through online tools like Walkscore. Additional variables we include to describe street network density are the

Table 1
The frequency of functional neighbourhoods identified for each city, mean (μ) venues per community, alongside the mean (μ) and standard deviation (σ) of the edge distance in KM, as well as the Louvain algorithm's resolution parameter and the number of outliers.

City	Neighbourhoods (Count)	μ Venue	μ Distance	σ Distance	Venues (Total)	Edges (Total)	Resolution	Outliers
Istanbul	211	1387.2	7.397	9.251	113,752	6,450,218	1.5	6179
Paris	44	165.7	3.183	4.422	13,588	4,198,732	6.0	92
Seoul	66	189.5	4.074	4.584	15,545	4,248,317	5.5	205
Singapore	94	284.4	4.950	5.727	23,324	4,671,817	5.0	351
Tokyo	183	705.0	4.544	6.420	57,810	5,435,836	4.0	354
London	82	276.6	3.721	4.958	22,689	4,834,661	1.5	437
Los Angeles	75	193.5	6.833	7.763	15,868	5,026,393	1.5	157
Jakarta	72	266.0	4.206	4.357	21,813	4,251,002	2.0	1223
New York	79	402.0	3.409	4.593	32,971	5,965,441	2.5	642
Chicago	58	169.5	4.610	5.821	13,904	4,921,583	3.5	146

(a) London, England. (b) New York, USA.

Fig. 5. Functional neighbourhoods within London and New York. Both images contain partitions with interesting outliers that are not located in spatially adjacent neighbourhoods (blue). For London we see Heathrow Airport belongs to the same cluster situated around community 18 in the north east. For New York meanwhile the East River Tunnels belong to community 14 (Note; venues data for Brooklyn was not provided). (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

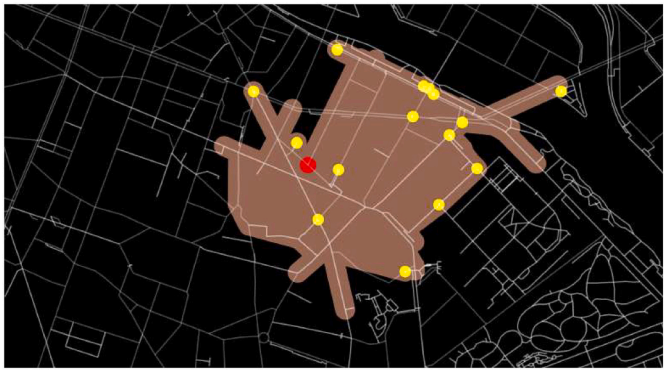


Fig. 6. A Catchment showing a ten minute walking radius along a street network in Paris built from a venue (red), with other venues within the catchment highlighted in yellow. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

absolute count of nodes (intersections) *n* and the mean number of inbound and outbound streets incident to the nodes, *k avg.* Alongside density, we also considered the 'eigenvector centrality ratio' which is a measure of street network connectedness; which has been correlated to footfall volume (Hajrasouliha & Yin, 2015). While OSMnx provide over forty measures describing the morphological and topological conditions of road networks, the majority of these exhibit significant positive correlation, which is undesirable for subsequent analyses like clustering due to implicit redundancy inherit in the data (Liu & Yu, 2005). To mitigate potential effects introduced by collinearity, we adopt a heuristic employed in the construction of Work-Place Zone classifications for London created by (Singleton & Longley, 2019) on behalf of the Greater London Authority, which we take as an example of best practice. Variable pairs with a correlation ± 0.8 were investigated, and for each, the sum of their correlation with every other variable taken. Thus, for each pair, we create a sum of correlations to every other variable in the data. Within the pair, the variable that has the highest sum of correlations to every other variable is removed from selection.

Alongside morphology and topology, we explore behavioural practices revealed though Foursquare check-ins within each catchment. Every spatial choice (or check-in) reflects a conscious decision-making process that can be used to infer types of urban activity. These insights are not just revealed from the mobility itself, but from meta-data that contextualises these experiences within the urban environment. Our variable 'Percent Same Type' describes the percentage of additional check-ins at venues located within the catchment that share the same Foursquare category tag to the check-in at the venue used to generate the catchment. This measure communicates the degree of heterogeneity in the kinds of check-ins within our catchments, with catchments intersected by a large percentage of similar activities ties offering more homogeneous consumption, leisure and service spaces. Our next variable, 'venue count', reflects a raw summation of the number of additional check-ins at venues that intersect the catchment of the original

Table 2
Measures used to compare venues catchment areas using the OSM network.

Measure	Description	Label
Topological		
Average Circuity	Total edge length divided by sum of great circle distance between nodes incident to each edge.	circuity_avg
Self-loop Proportion	Proportion of edges in catchment that have a single incident node (or loop).	self_loop_proportion
Street Length Average	Mean edge length in undirected representation of catchment (meters).	street_length_avg
Street Length Total	Sum of edge lengths in undirected representation of catchment.	street_length_total
Street Per Node Average	Average number of streets per node.	street_per_node_avg
Density and connectedness		
k Average	Average degree of catchment.	k_avg
n	Number of nodes in catchment	n
Node Density Km	n divided by area in square kilometers.	node_density_km
Eigenvector Centrality Ratio	Ratio of eigenvector centrality between venues and other venues within catchment.	ratio_eig
Behavioural		
Percent Same Type	Percentage of check-ins to venues of same category inside catchment.	percent_same_type
Check-ins Count	Summation of total check-ins to all venues inside catchment.	n_pois
Average Morning	Average Foursquare check-ins of all venues in catchment between 06:00 and 09:59 (%).	avg_morning
Average Midday	Average Foursquare check-ins of all venues in catchment between 10:00 and 14:59 (%).	avg_midday
Average Afternoon	Average Foursquare check-ins of all venues in catchment between 15:00 and 18:59 (%).	avg_afternoon
Average Night	Average Foursquare check-ins of all venues in catchment between 19:00 and 23:59 (%).	avg_night
Average Overnight	Average Foursquare check-ins of all venues in catchment between 00:00 and 05:59 (%).	avg_overnight

venue. This allows us to proxy the popularity of the urban environment within the catchment. Finally, we expose temporal dynamics of check-in activity by examining distributional splits between morning, midday, afternoon, night and overnight.

Together, our many venue catchments nest into different functional neighbourhood definitions, with the tenant mix of these places (alongside their built environment characteristics) expressive of the underlying scene a neighbourhood projects to residents and passers-by. Venues as amenities are windows into the scenes of neighbourhoods which reveal plausible behavioural patterns that encode expressions of local traditions, taste and preferences (Silver & Clark, 2016). Ultimately, our approach contextualises these functional neighbourhoods by the multi-dimensional properties of different activities and experiences offered by each neighbourhood's range of venues and built environment characteristics. Across all venues inside each neighbourhood, we find that averaging the values for the variables identified in Table 2 best reflects the aggregate character and profile of the functional neighbourhoods.

3.3. Understanding and comparing human dynamics in functional neighbourhoods

As Shaw and Sui (2018) put it, "human dynamic research is not just about human". On the contrary, it requires a comprehensive characterization of the environmental factors that determine certain behavioural patterns (Shaw & Sui, 2018). Therefore, to gain an understanding of human dynamics in cities, we develop a multidimensional description of the functional neighbourhoods, implementing k-means clustering, that accounts for human and non-human elements, such as entities in the physical spaces. We posit three drivers of functional neighbourhood differentiation that include: diversity, context and mobility. While we outline each driver in detail below, we provide a short description for each in Table 3.

Diversity: The variety of venue categories within a functional neighbourhood provides a measure of diversity insofar as it allows people to perform different activities within these areas (i.e. dining, shopping, working). A widely used measure summarising diversity is the Shannon Index, which was first developed in the context of information theory to capture the predictability of certain content to appear in a message (Shannon, 1948). However, as implemented here, this provided us with an entropy measure characterizing the different mix of venue categories within each functional neighbourhood.

Entropy is therefore calculated according to the Shannon Index as follows:

$$H = -\sum_{i=1}^{s} p_i ln p_i, \tag{2}$$

where s is the number of categories, p_i is the proportion of venues of each category and ln is the natural log. A high value of H corresponds to highly entropic functional neighbourhoods, where venues tend to be classified by a wider variety of categories (richness), and it is more difficult to predict which category a new venues will fall within (evenness).

Context: Check-ins present descriptive measures for contextualizing human activity and behavioural patterns. The proportion of check-ins by time of day (morning, midday, …) represent spatio-temporal aspects of

Table 3
Functional neighbourhood differentiation.

Measure	Description
Diversity	Diversity is measured in the categories of venues by computing an entropy score (Shannon Index)
Context	The aggregation (mean) of venues information resulting from the catchment area morphological profiles (topological, density and connectedness metrics) and type and average check-ins by time of day (behavioural metrics) (see Table 2)
Mobility	Functional neighbourhood distance decay function

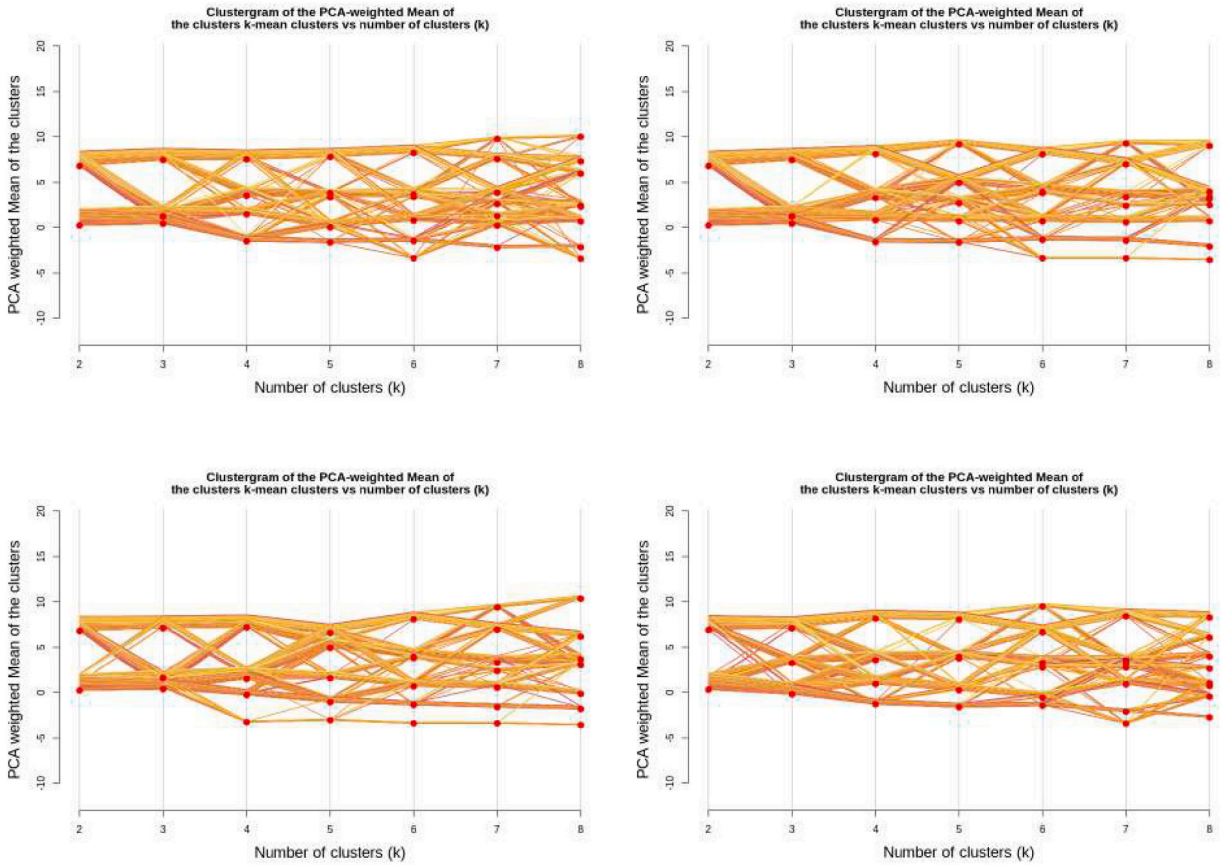


Fig. 7. Clustergrams showing how different iterations of k impacted the cluster PCA weighted mean separation.

human presence within each area, and are averaged over all venues within each neighbourhood. The means of the variables (n pois, percent same type, ratio eig, streets per node avg., street length total, street length avg., self loop proportion, circuity avg., k avg., n, node density km) characterizing catchment morphology and people preferences are reduced through the application of Principal Component Analysis. PCA is a statistical tool commonly used to reduce a large number of possibly correlated variables into a smaller set of uncorrelated components. A mathematical transformation is applied to maintain most of the information in the data accounting for as much of the variability as possible. We identify four principal components that explain 77% of the variance derived.

Mobility: The richness of the Foursquare data allows us to further elucidate different aspects of people's behaviour: notably spatial interaction. For these measures we examine trip length between check-ins to gain information on about human mobility patterns within the functional neighbourhoods. We define the aggregate neighbourhood mobility matemathically through a parameterized exponential decay function $f(x_{1\ldots n})$ in the Equation below, which establishes two parameters d and k to approximate the proportion of flows by distance (k moderates the decay rate d). Eq. (3) takes a vector $x_{1\ldots n}$ of distance bin values, and returns the approximated number of flows for each value in x_i. We use TensorFlow to obtain the best fit for each of the city's empirical observations $y_{1\ldots n}$, optimizing parameters d and k by minimizing the L_2 loss between $y_{1\ldots n}$ and $f(x_{1\ldots n})$. This provides a set of exponential decay parameters that estimated likely relative impedance between venues locations for given neighbourhoods.

$$f(x_{1\ldots n}) = max_i\, y_i \times (1.0 - exp.\times(-k \times d^{x1\ldots n})) \quad (3)$$

Comparing Multidimensional Characteristics: The methods presented above result in a set of analytics that describe the functional neighbourhoods through features of: venues diversity, context and human mobility. To obtain a multidimensional description and compare neighbourhoods across all cities we turn to clustering, bringing together those areas with similar features. A common method to group multidimensional data is k-means algorithm. It clusters data points to minimize data variance within a k number of partitions. While this method facilitates comparisons between all features across clusters of functional neighbourhoods, a challenge when implementing k-means clustering is to select an appropriate number of clusters. Here we utilize a clustergram, which along the x-axis plots a range of potential k values; and along the y-axis a weighted mean PCA score. Each line relates to a functional neighbourhood and relative score for each iteration of k; and dots represent the cluster average PCA scores. An indication of the best fit for k relates to a model where these centroids are well separated (See Fig. 7). From this figure, we select four clusters as an appropriate k value. Given that k-means is stochastic, we run the algorithm 10,000 times on the standardised and scaled data with k = 4, extracting the result that had the lowest total within cluster sum of squares; i.e., the most compact clusters. After clustering, we append the original data back onto the clusters and examine the resulting distributions relative to the global averages.

4. Results and discussion

Before presenting results from the cluster analysis in Section 4.1, we first discuss our findings along each of the three analytical dimensions of diversity, context and mobility. Correlation coefficients are calculated to complement interpretation of the results [5]. Diversity: The Shannon indices reveal that 66% of the functional neighbourhoods have entropy scores ranging from moderately low (1.7) to moderately high (2.4), while 204 of the functional neighbourhoods are highly entropic (> 2.4) and 81 have very low diversity [6].

The functional neighbourhoods belonging to the latter group are

[5] Pearson's correlation coefficients can be found in Figure A. 15 in Appendix A

[6] Classes are based on Jenks breaks.

found across the ten cities and tend to be smaller and located in peripheral areas; in contrast, 50% of the most entropic functional neighbourhoods are all placed within the city of Istanbul. To understand which characteristics of the functional neighbourhoods may drive diversity we examine the correlation with the other variables, and with the venues count within each functional neighbourhood (see Appendix A.15). We found a high (0.4) and statistically significant correlation between venues count and diversity which reflects the entropy characterizing Istanbul's functional neighbourhoods – which is the sample city with the largest number of venues.

The time and context of human activities: Check-ins are mostly made between midday and night within in each city (see Appendix B.16). By analyzing check-in temporal distribution we observe a strong and significant negative correlation (-0.77) of the midday check-ins with those recorded at night. Bearing in mind that the majority of check-ins are almost always made at midday, during these two time intervals the share of check-ins have a very dissimilar geographic distribution. This suggests a tendency towards specialization within certain neighbourhoods of the city: some areas clearly function to perform midday activities with a very low proportion of night check-ins; at the same time, areas with higher proportion of night check-ins show a lower than average share in midday check-ins (see Fig. 8). We also note that there is no direct correlation between check-ins in those time frames and diversity in venue categories (see Fig. A.15). This suggests that the temporal specialization of functional neighbourhoods is likely to be independent from specific category types. This result is in line with previous studies, investigating the relation of temporal signatures with venues categories, and identifying both types with strong regional variability and similar patterns across cities (McKenzie, Janowicz, Gao, & Gong, 2015).

Next, we consider findings from our Principal Component Analysis which aims to capture the context of venues within each functional neighbourhood. These findings are summarized by biplots in Fig. 9, where each black circle identifies the score of functional neighbourhoods on the first four principal components, while the arrows identify the loading (or influence) of each variable in Table 2 on the components. To aid interpretation, the vector angles in the figure can be used to identify correlations between groups of variables. We characterise these components as follows:

- PC1 value is driven up by self-loops and circuity. Self loop designs like cul-de-sacs might be more common in suburban residential areas that have less predictable, grid-like street geometries. On the contrary, high number of intersections (*n*) and strong network connectivity (*k avg*) – generally characterizing pedestrian-oriented areas – have a negative load on PC1 (see Fig. 9a).
- PC2 positive value corresponds to functional neighbourhoods of considerable size, therefore with longer total street length. On the other hand, the average number of streets per node loads negatively onto the component. These two variables are negatively correlated (as suggested by the diverging arrows) and complemented by contrasting user behaviours. While variegated user preferences (*n pois*) play a role in driving up the value of this component, the check-in behaviour of individuals patronising neighbourhoods with negative PC2 value reflect less diverse, and more homogeneous consumption spaces (percent same type).
- PC3 and PC4 only explain 23% of the variability in the data and are mostly influenced by four variables moving in opposite directions. PC3 captures neighbourhoods described by a higher than average self loop proportion, which is indicative of street networks that consist of many loops and lollipop roads. In the other direction, the loading for the streets per node avg. variable indicates a negative correlation to the component, which suggests that intersections in these neighbourhoods are typically connected by a higher number of streets. Following a similar trend, neighbourhoods with a high value in PC4 are more characterized by users with preference towards a narrow sub-set of venues categories while low values are significantly driven by high connectivity (ratio eig).

Mobility: As outlined in Section 3.3, to capture mobility patterns we extrapolate two parameters describing the best fit distance decay function for each neighbourhood on distance bins with a size of 100 m per bin. Parameter *d* has the most direct impact on the curve's slope, while *k* acts as a moderator. Both parameters determine to what extent long distances are travelled between venues within a neighbourhood. A steep slope (low *d* and *k*) is indicative of neighbourhoods where there is a sharp decay in the number flows as the distance between venues increases, i.e., we observe a larger number of flows between more proximal venue pairs. In contrast, a gradual slope can be observed within neighbourhoods where there is a slower decay in the number of flows per bin as the distance between venue pairs increases, i.e., people more frequently travel longer distances between venues. As expected parameter *d* has a negative correlation with the number of venues, allowing users to remain local, and thereby travel shorter distances between venues. The majority of the neighbourhoods are characterized by a rather steep function – 98% of the neighbourhoods have a $d < 0.4$ – while the few neighbourhoods with a more gradual decay are mostly located in Los Angeles and Chicago. Fig. 10 shows two examples of how the modelled estimates capture the mobility structure of a neighbourhood where we see more flows between close venues – Fig. 10a – or where entries with longer distances between venues are more frequent - Fig. 10b. However, we observe that a few very small neighbourhoods located on the cities' outskirts exist, that are not properly described by our function (e.g., see Fig. 11).

4.1. Clustering

The four clusters resulting from the k-means algorithm combine the analytics reported above into a multidimensional description of neighbourhood types. We describe the characteristics of each neighbourhood cluster below [7].)

- Cluster 1 (size 169) classifies neighbourhoods mostly located in Singapore and Istanbul. This cluster is characterized by having a very high entropy score. Furthermore, the proportions of night and overnight check-ins - the latter represents the least frequent check-in periods within the FCC dataset [8] – are above average. The mobility structure of the neighbourhood shows a less than average steepness in the distance decay function. These neighbourhoods are characterized by cul − de − sac design (high PC1 and PC3).
- Cluster 2 (size 257) classifies neighbourhoods which have a diversity score almost on average (the standardize value approaches zero), and are popular locations for morning and midday activities. They are more likely to be residential areas (high PC1) and favour rather fixed behaviours (low PC3 mostly influenced by check-ins in venues of the same type). Short distance movements are preferred as emerging from a mobility structure well approximated by a significantly steep distance decay function (both k and d lower than average).
- Cluster 3 (size 129) comprises neighbourhoods with moderately high entropy. We also observe a high number of check-ins during commuting peak periods: morning and afternoon. PC2 is generally positive, suggesting that such neighbourhoods tend to be of a considerable size (with longer total street length). Accordingly, this cluster's mobility structure has the most gradual distance decay slope, therefore users typically travel longer distances between venues compared to the neighbourhoods within the other clusters.

[7] See Fig. 12 for details regarding the standardised averages for each cluster-variable pair.

[8] See month-period check-ins visualization in Appendix B.

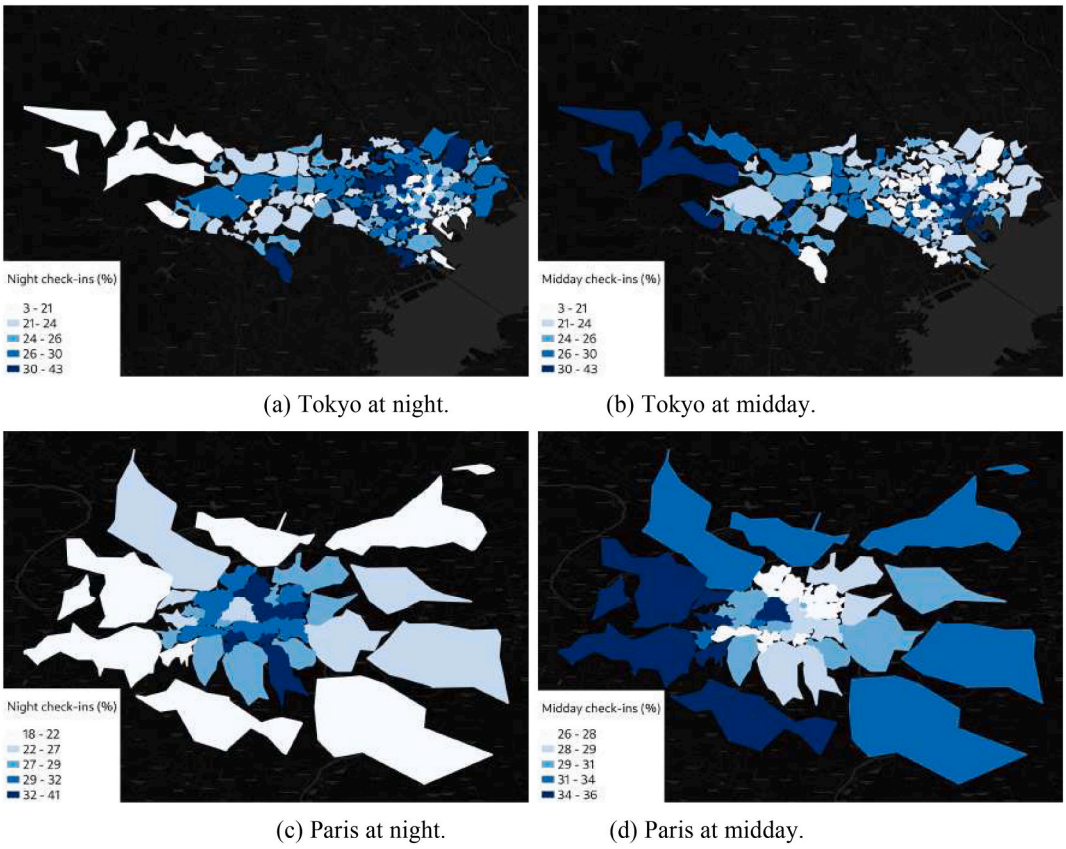


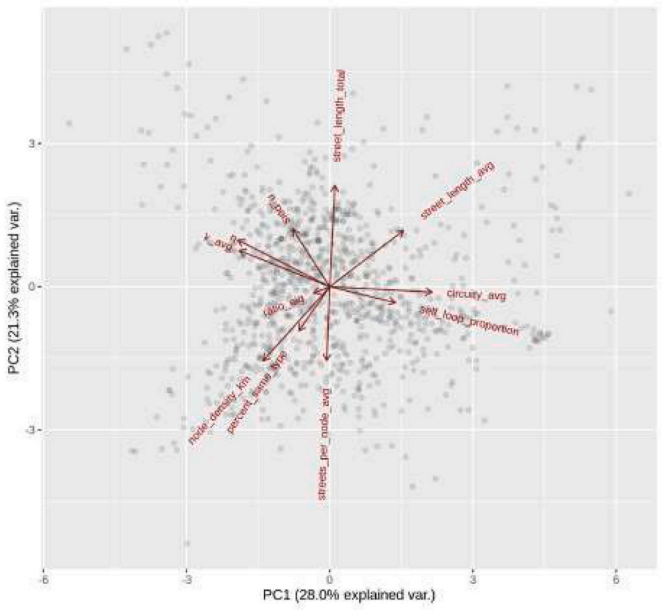
Fig. 8. Two exemplary cities sharing similar temporal geography: check-ins are not evenly distributed by day time in the functional neighbourhoods with the most striking distinction in check-ins between the classes of midday and night.

- Cluster 4 (size 312) is characterized by a very low diversity combined with high number of afternoon and night check-ins. In these neighbourhoods short distance trips between venues are the most common, resulting in the most steep distance decay function slope, and the area tends to be the most walkable (PC1 and PC2 lower than average, demonstrating the high connectivity of the road network).

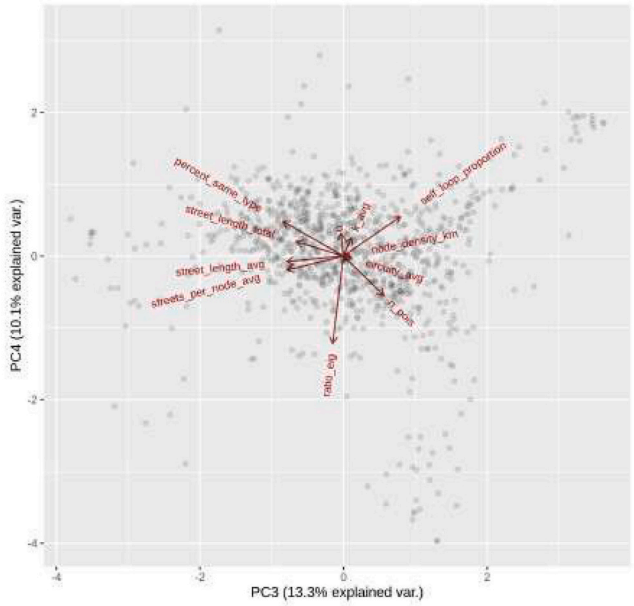
Human Dynamics within Global Cities: From the clustering analysis we can finally uncover human dynamics emerging from the data across the ten cities. The frequency of venues within each of the functional neighbourhood clusters is shown in Fig. 13. Cluster 1 classifies almost exclusively neighbourhoods in Singapore and Istanbul, where we observe a high entropy score, an active nightlife and a street network with long street segments and an high number of self-loops. Cluster 2 and 4 neighbourhoods are the most common across the ten cities. These clusters split neighbourhoods based on typical activity time, until midday the former and from the afternoon onward the latter. Despite temporal differences, short distance movements are prevalent in both cases, as well as a moderately high connectivity of the street network. Half of Los Angeles venues are located in Cluster 3 neighbourhoods, followed by Istanbul and Chicago. A remarkable characteristic of these areas is a mobility structure that favours long distance movements. Unsurprisingly, a car-centric city as Los Angeles sees the higher concentration of venues in such Cluster.

The utility of these distributions can be illustrated by plotting archetypal functional neighbourhoods across cities. We do this by examining the input scores for each neighbourhood, and selecting those that are closest to the average for a particular cluster (see Fig. 14). We would expect these functional neighbourhoods to have similar characteristics in terms of venues diversity, urban morphology, user check-in behaviour and mobility structure. Our approach results in global typologies that bring together very different cities. In line with Robinson (2006)'s idea of ordinary cities, we aim at highlighting the diversity and similarity of human dynamics in a variety of contexts worldwide. At the same time, such comparative endeavour does not exclude the possibility to investigate differences within each city.

Previous studies turned to clustering, using different kind of data sources, to profile urban areas by the prevalent activity types (Assem et al., 2016; Calabrese, Reades, & Ratti, 2009; Gao et al., 2017; Lenormand et al., 2015) or to identify land use patterns (D'Andrea, Ducange, Loffreno, Marcelloni, & Zaccone, 2018; Grauwin, Sobolevsky, Moritz, Godor, & Ratti, 2015). While these works result in typologies describing the urban environment from a single perspective - the use of space - our profiling approach is designed to combine elements of the built environment, such as the street morphology and the diversity in venues type, with information about human behaviours. A direct application of our framework is urban planning. Outcomes of our framework would help urban planners to identify meaningful relations between aspects of the built environment – such as its diversity and morphological structure – and certain behaviours. While the study of the interaction between human activity and the built environment is at the core of urban planning practice (Gehl & Svarre, 2013), it is mostly based on observation (Gehl & Svarre, 2013), rendering evidence-based studies time consuming and focused only on limited areas (Ertio, 2015). In line with data intensive approaches to urban planning (Batty, 2013; Singleton et al., 2017), our framework capitalizes on the availability of data at a global scale to investigate human dynamics, describing the way behavioural patterns are combined with characteristics of the built environment. While relations across variables are not linear (see Fig. A.15 in Appendix A) – with exception of a significant temporal divergence when neighbourhoods are more popular – through a clustering approach we maintain a multi dimensional neighbourhood characterization. This enables to identify various declination of similar patterns across cities. For example cities such as Singapore or Istanbul have



(a) PC1 & PC2.



(b) PC3 & PC4.

Fig. 9. Variables contributing to PC1 & PC2 (a) and PC3 & PC4 (b).

several neighbourhoods where a high number of night check-ins is linked with a walkable environment with high diversity (see Cluster 1); in Tokyo or Seoul neighbourhoods characterized by the same temporal signature are instead combined with a low diversity and prevalence of short-distance movements (see Cluster 4).

5. Conclusion

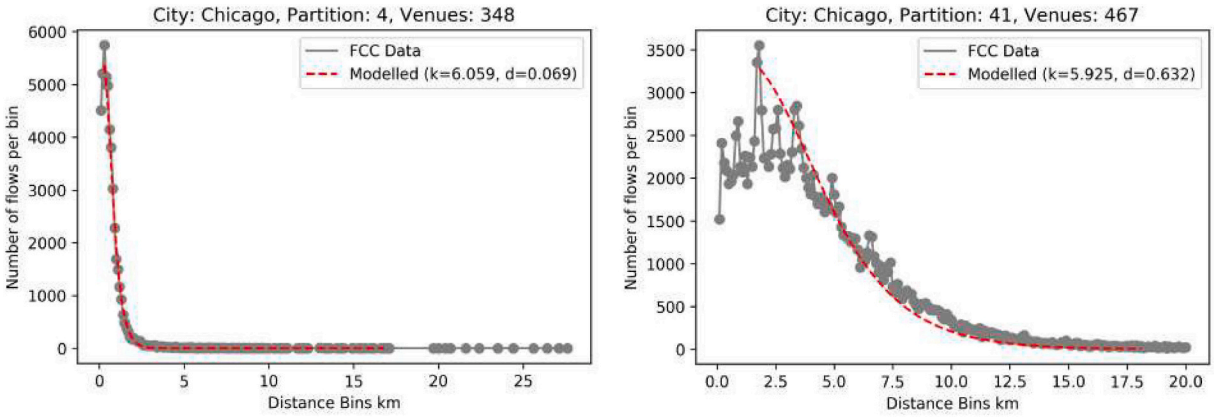
A well rehearsed problem of working with crowdsourced information relates to its representativeness (Blank & Lutz, 2017; Mart'ı, Serrano-Estrada, & Nolasco-Cirugeda, 2019). Some studies (Ballatore & De Sabbata, 2018, 2019) explore the geo-demographic context where the production process of user generated information takes place and – comparing Greater London and Los Angeles – have shown that the way information and socio-demographic geographies overlap can vary considerably among cities. In light of this, we acknowledge that FCC dataset will also be biased towards those venues added by the Foursquare user-base, and as such, differences may exist between reality and the geographical coverage of the dataset used in this work. Such issues might be explored in future work examining the correspondence between Foursquare check-in locations and the universe of other potential sites recorded in ancillary points of interest data.

It is also important to state that the crowdsourced nature of the FCC data may impact upon the shape, scale and extent of the derived representations and their characterization. For example, the entropy scores are based on categories of venues that are added by the Foursquare user-base. This indicates that venues might not reflect comprehensively the true variety of businesses within cities. Some categories of venues – probably those perceived as more interesting by the users – or when the user is a business owner looking to advertise their activity, e.g., food or drinking establishments – are over-represented (as evident from the check-ins per category illustrated in Fig. 2). For example, the low diversity score found in Paris city center – where a large proportion of venues can be categorised as being touristic services such as accommodation and food – are likely driven by the users' interests within that area. By contrast, Los Angeles follows a different and perhaps more expected pattern where downtown is the most diverse area. The limited universe of venues within Foursquare's database also affects the check-in counts, which are therefore only possible in their mapped venues. In addition, we hypothesize that the number check-ins during each period will be influenced by the users' daily routines, i.e., certain venue types will be less likely to experience check-ins during what one would consider to be typical working hours. Therefore, results on the check-in percentage by time of day should be interpreted by taking such considerations into account. Furthermore, while the FCC dataset does not allow us to differentiate check-ins that occur during the typical working week (Monday – Friday) from the weekend, we consider that obtaining this additional level of granularity could open up interesting avenues for future research.

Catchment areas – used to characterise the venue's context – are computed based on the Open Street Map (OSM) road network which is also populated by user generated data. OSM data quality has been assessed in the UK (Haklay, 2010) and France (Girres & Touya, 2010) through a comparison framework to evaluate such data against national reference geographic databases. Results show that while the positional accuracy of the road network is fairly adequate, disadvantaged and rural areas have weaker coverage.

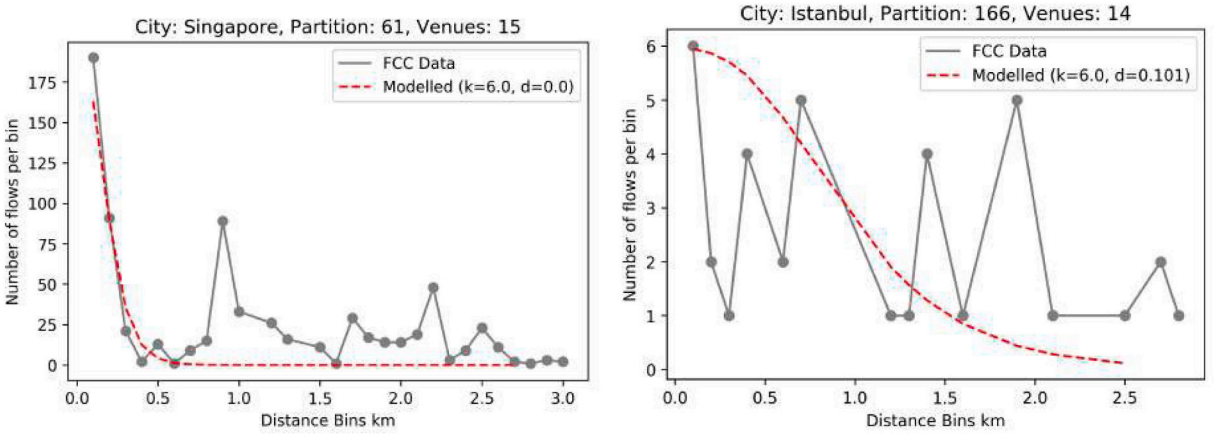
Although we recognize that detecting biases in crowdsourced data is still a major challenge to the application of geographic data science methods (Mart'ı et al., 2019), it must be acknowledged that the unprecedented volume, velocity and variety (De Mauro, Greco, & Grimaldi, 2015) of information afforded by big data open up novel opportunities for obtaining insights about urban contexts. Evaluations of crowdsourced geographic information, along with mobile phone data, have provided knowledge on a variety of urban facets, e.g., highlighting issues of intra-neighbourhood segregation, mobility and inequality (Shelton, Poorthuis, & Zook, 2015), automatically suggesting routes that are both short and emotionally pleasant (Quercia, Schifanella, & Aiello, 2014), and providing insights into the physical aspects of cities and the spatial distribution of urban functions (Arribas-Bel, Kourtit, Nijkamp, & Steenbruggen, 2015).

As Miller and Goodchild (2015) put it, the widening use of big data in geography should be interpreted as evolutionary, rather than revolutionary, to geo-spatial research, complementing and augmenting existing data sources and methods. In particular, crowdsourced geospatial data offers the researcher a glimpse into the intangible aspects of urban life, such as people's spatial behaviours and preferences (Sui & Goodchild, 2011). These domains are well beyond the scope of official geographic data provided by institutional agencies, and have been traditionally captured through observational methods (Gehl & Svarre, 2013) at higher cost, slower speed and with much more limited coverage (Quercia, Aiello, Schifanella, & Davies, 2015).



(a) Partition 4. (b) Partition 41.

Fig. 10. Two modelled line examples for Chicago.



(a) Singapore – Partition 61. (b) Istanbul – Partition 166.

Fig. 11. Two examples of neighbourhoods with a small number of venues and flows in Singapore and Istanbul.

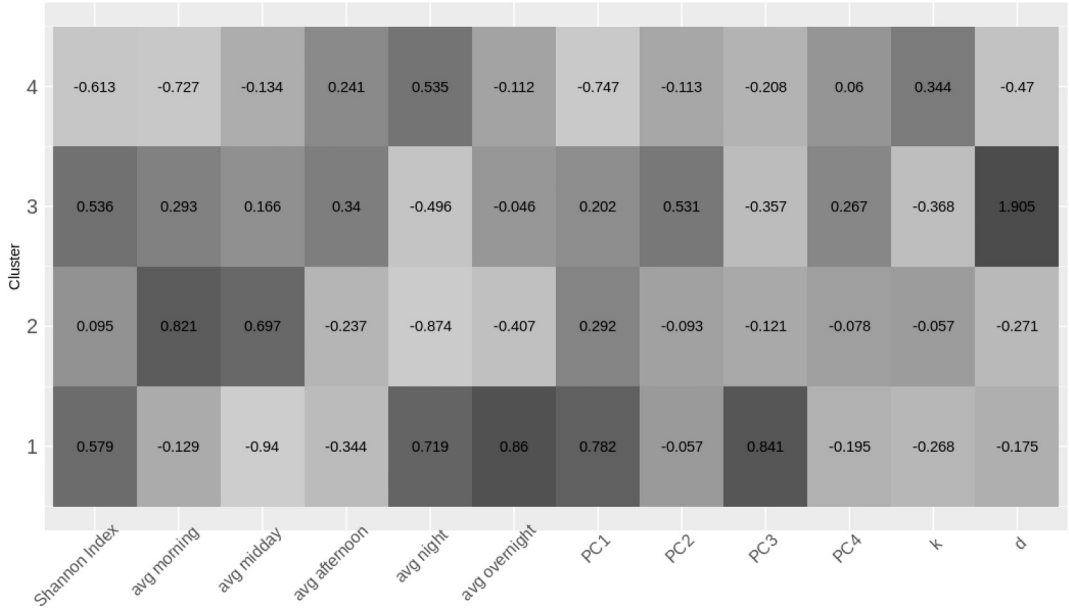


Fig. 12. Standardised averages for each cluster-variable pair.

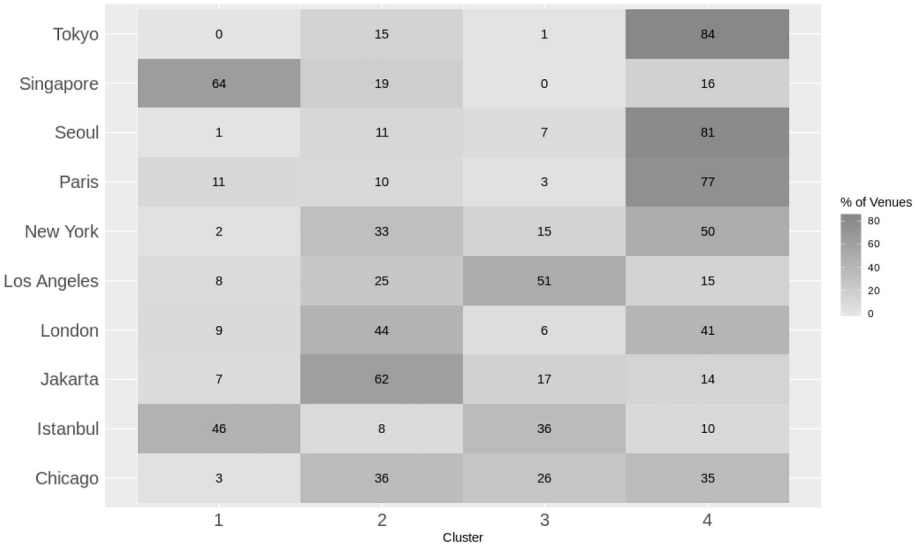


Fig. 13. The % of venues within functional neighbourhood clusters by city.

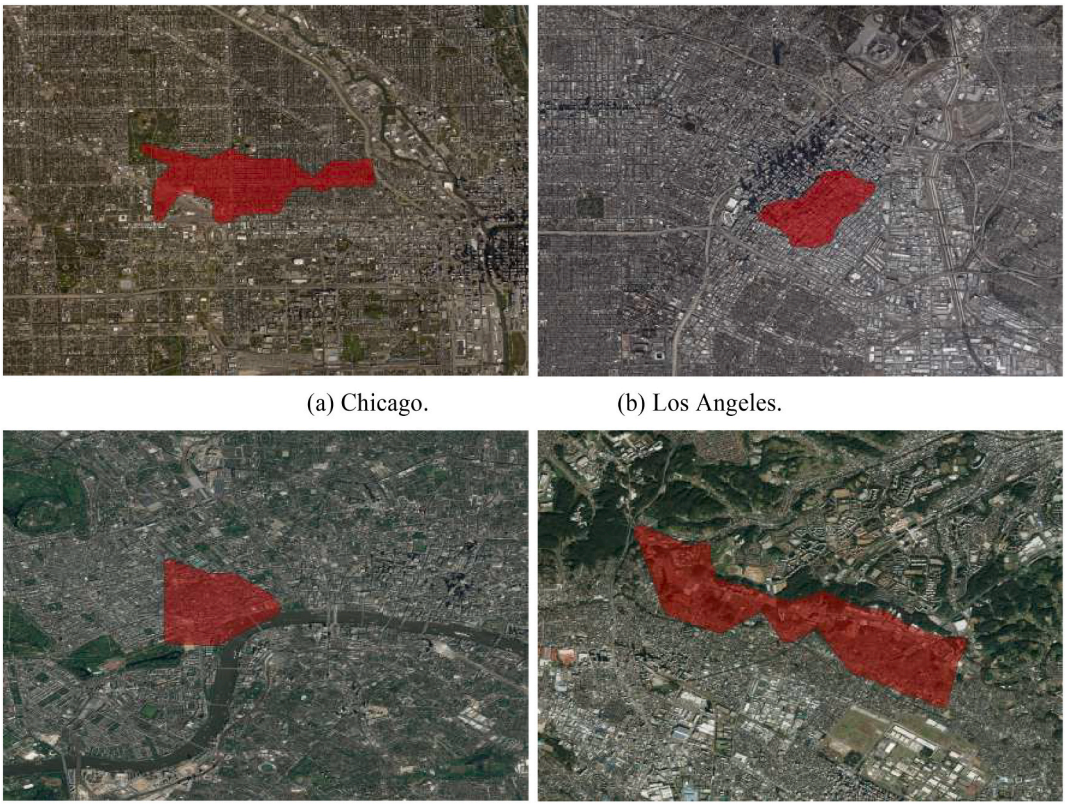


Fig. 14. Archetypal functional neighbourhoods from Cluster 4. The neighbourhoods are characterized by a low entropy score, are very dense areas with a highly connected street network within which people predominately move short distances between venues, and are typically frequented at night.

Furthermore, platform users contribute to the data collection voluntarily which, according to the Hawthorne effect, might be an advantage compared to traditional methods since the studied population is unaware of being observed (Mart'ı et al., 2019).

In addition, the increasing use of such media on a global scale allows scholars and analysts to employ internationally integrated data. This facilitates comparisons across cities around the world, overcoming existing gaps in the availability and interoperability of institutional open geographic data.

Results of our framework can be input for the development of urban theories and, consequently, planning strategies in a variety of contexts. We capitalize on the human content provided by this data, such as people behaviours in terms of temporal movements and venues check-ins, bringing about a human-centric approach to geographic data science methods.

First, we generate geometries - functional neighbourhoods - entailing venues with strong interdependence over geographical space. Differently from administrative units, these are areas emerging from human interaction with the built environment in their activity space and as such they provide scholars and planners with the spatial forms underlying human dynamics.

Second, a set of urban analytics are developed to describe the main features of functional neighbourhoods. We select variables to mix behavioural patterns and characteristics of the built environment into a multidimensional characterization.

Third, functional neighbourhoods are clustered and profiled. This final output is well suited towards identifying and describing similar neighbourhoods between and within cities.

Outputs of our framework can significantly address further investigation into areas of specific interest. In particular, by centering the unique mix of methods that we propose on the interaction between humans and the built environment, our framework can be a valuable tool to support urban planning strategies in different contexts.

There are numerous avenues for future research. As we mention in Section 2, the data have been aggregated on a month-year and period basis, determined by the date and time when users arrive at the destination venue. We are therefore unable to establish if a direct path was taken by users between venue1 and venue2, thereby adding noise. However, more granular time stamps would enable the exploration of potential modes of transit linking venue sequences.

In our current evaluation we discover functional neighbourhoods for each city using all the provided flows. However, as mentioned in Section 2, the FCC dataset includes movement data for venue pairs aggregated for month-year and period of the day. This information opens up interesting opportunities for future research in this area. For instance, applying our framework to individual periods of the day could enable a comparison of how functional neighbourhoods change over time. Utilizing the month-year column meanwhile may enable a longitudinal study. We leave such considerations to future work.

Additional variables to characterise and profile the neighbourhoods can be included, i.e. pollution and congestion data. At the same time new metrics to estimate to what extent crowdsourced data such as Foursquare is representative of the reality can be developed to ensure more robust results.

Furthermore, in its current state our distance decay function is best suited towards approximating monotonically decaying distance bin flow counts. However, as we observe in Fig. 11, the number of entries per bin is not always monotonically decaying, raising the question whether a more suitable parameterized approach can be found.

We also note that the euclidean distance metric used to identify partitions offers a simplification of true physical distance, as individuals rarely travel from point A to point B in a straight line. Describing distance in this way likely underestimates true physical proximity, potentially decreasing the level of confidence that can be ascribed to the detected communities which are contingent on spatially-weighted network edges. We leave the computing of these distances using routing algorithms for future work.

This work has outlined how social media data such as Foursquare can be utilized to provide insight into the structure and function of cities. Through our Geographic Data Science workflow we create a framework for identification and description of functional neighbourhoods across global contexts, linking a range of measures ascribed down to the local level. The framework is portable to other geographic contexts where interaction data are available to bind different localities into functional agglomerations, and provide insight into their contextual and human dynamics.

Appendix A. Correlation coefficients

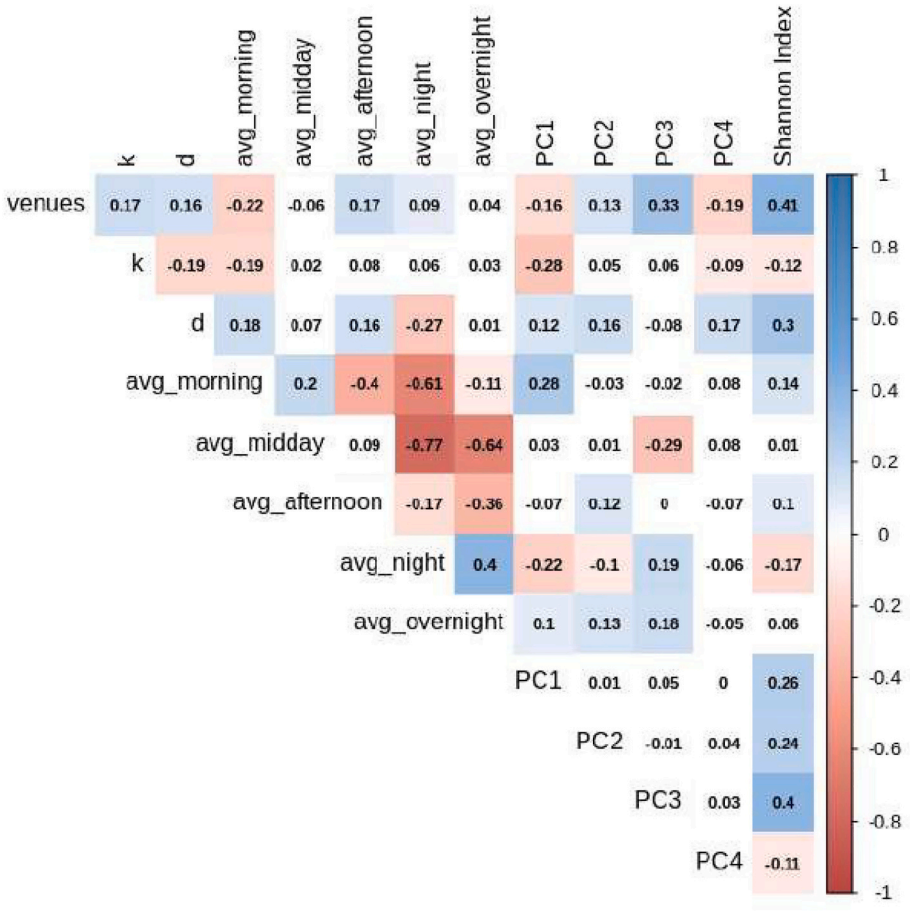


Fig. A.15. Pearson's correlation coefficients among variables characterizing the functional neighbourhoods. All coefficients which are not significant at $p < .01$ are left blank.

Appendix B. Month-period check-ins visualization

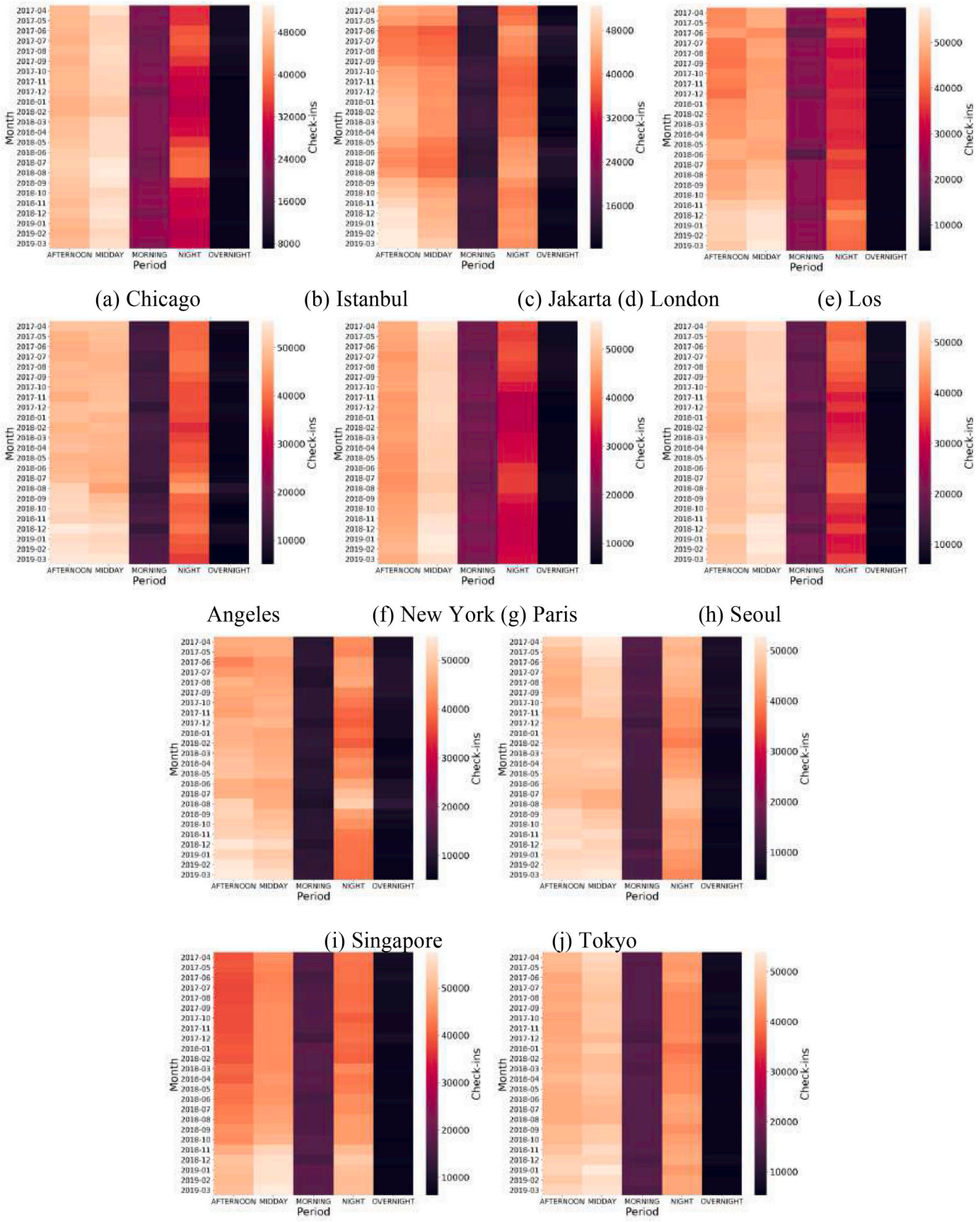


Fig. B.16. Heatmaps illustrating the number of check-ins for each month-period combination.

Appendix C. Modularity visualization

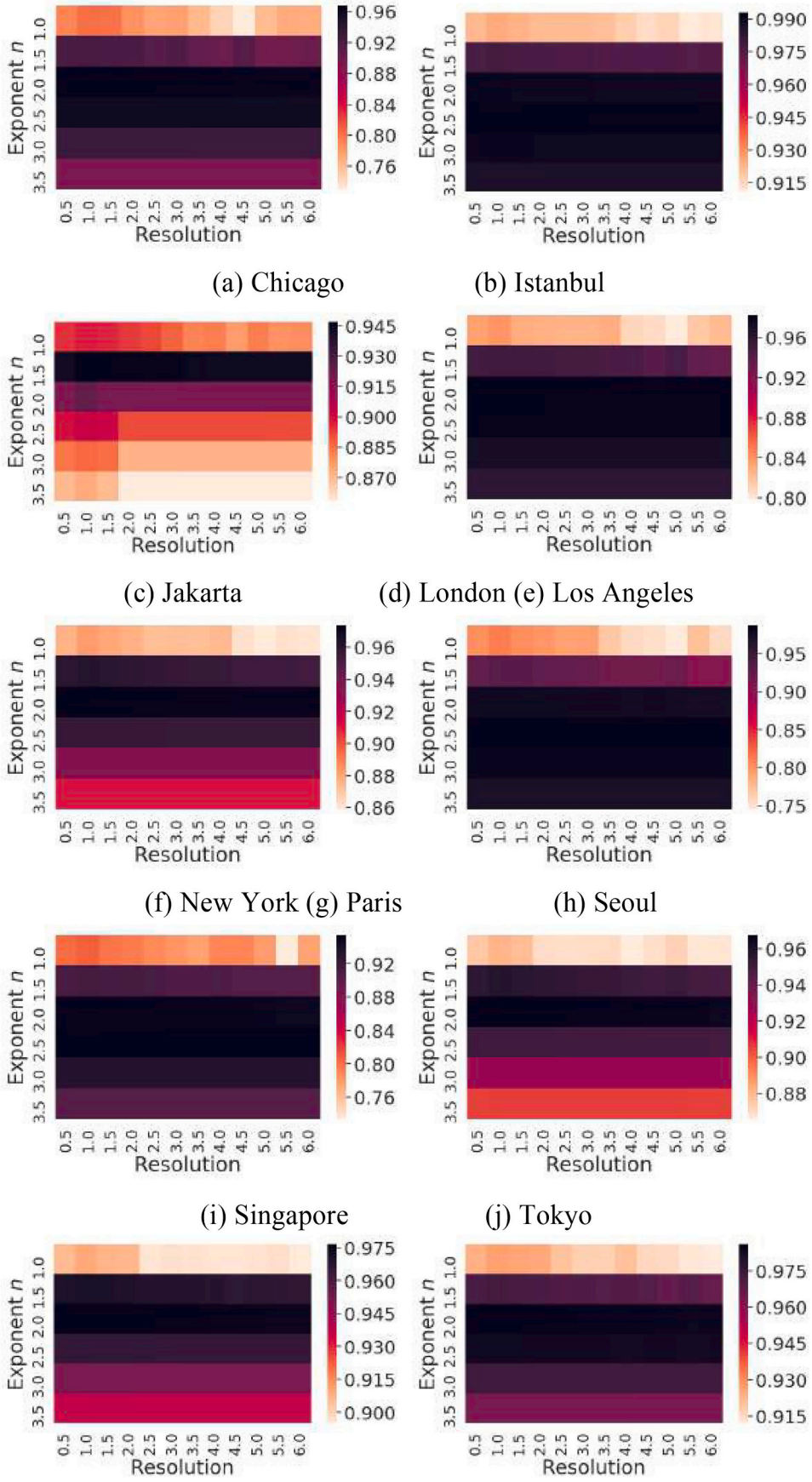


Fig. C.17. Modularity scores for each exponent decay n and resolution value combination.

Appendix D. Outliers

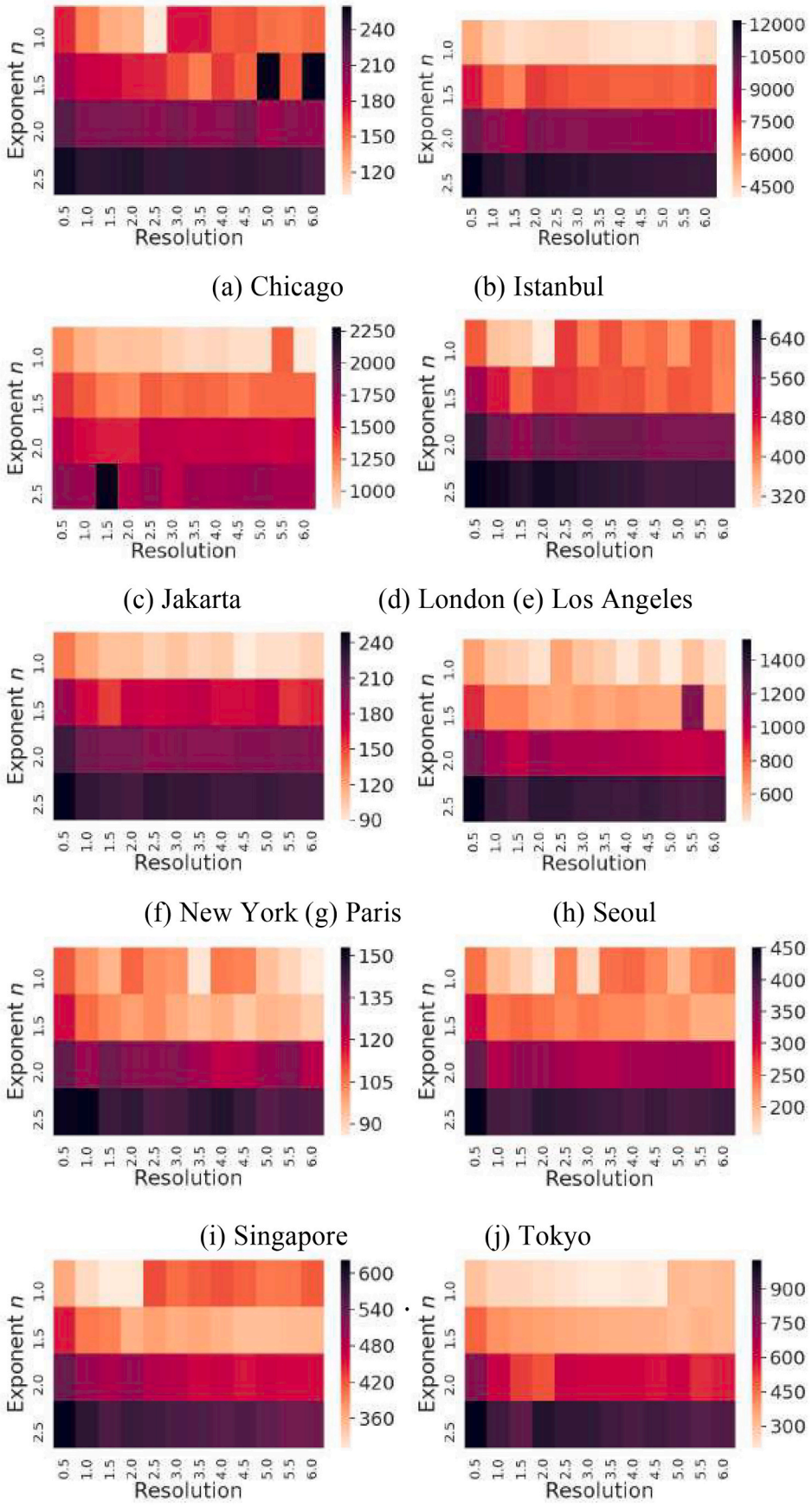


Fig. D.18. Outliers for each exponent decay n and resolution value combination.

Appendix E. Modularity-outlier metric visualization

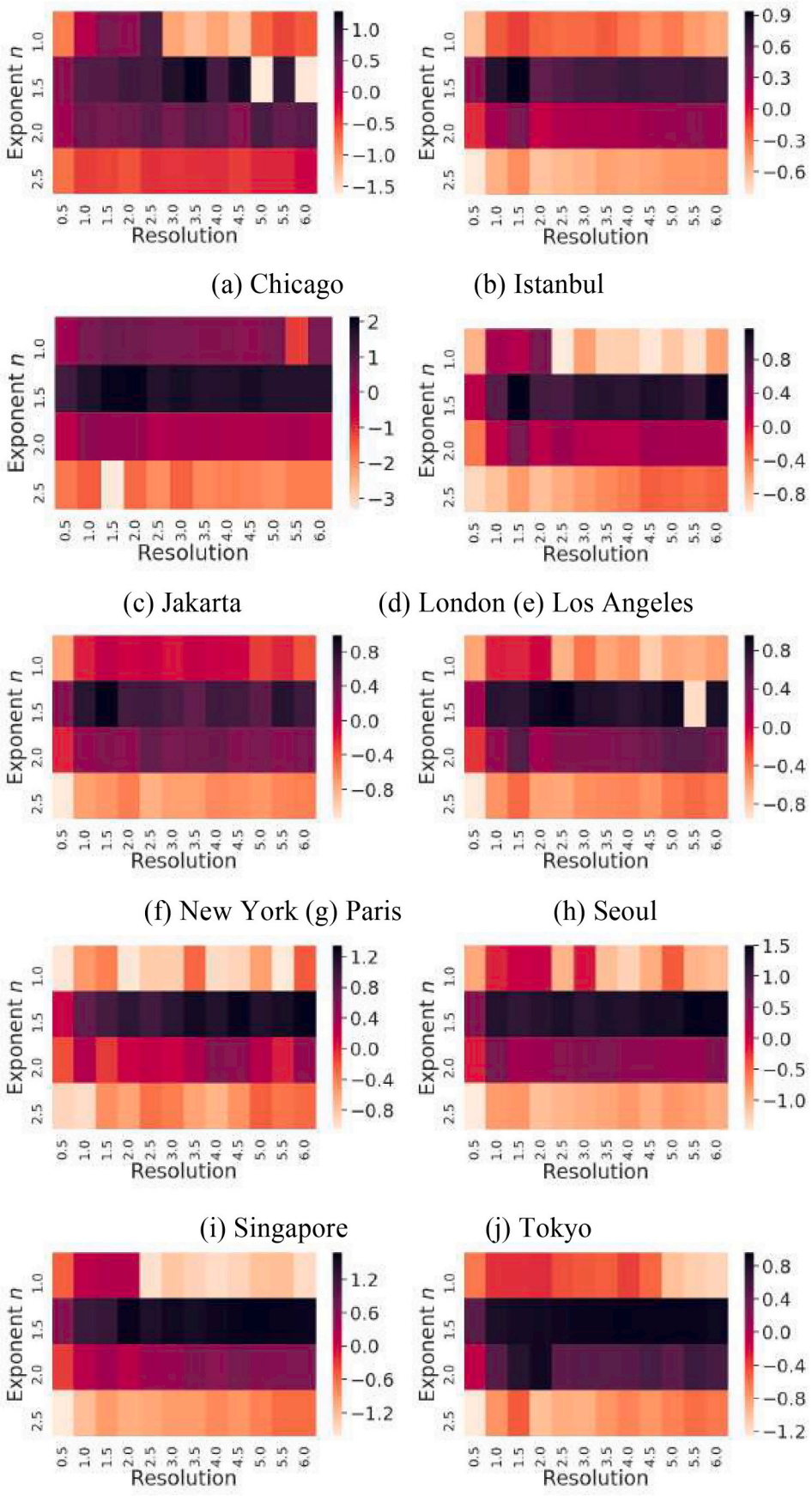


Fig. E.19. Heatmaps illustrating the Modularity-Outlier Metric for each city.

References

Arribas-Bel, D. (2014). Accidental, open and everywhere: Emerging data sources for the understanding of cities. *Applied Geography, 49*, 45–53.

Arribas-Bel, D., & Bakens, J. (2019). Use and validation of location-based services in urban research: An example with dutch restaurants. *Urban Studies, 56*(5), 868–884.

Arribas-Bel, D., & Reades, J. (2018). Geography and computers: Past, present, and future. *Geography Compass, 12*(10), e12403.

Arribas-Bel, D., Kourtit, K., Nijkamp, P., & Steenbruggen, J. (2015). Cyber cities: Social media as a tool for understanding cities. *Applied Spatial Analysis and Policy, 8*(3), 231–247.

Assem, H., Xu, L., Buda, T. S., & O'Sullivan, D. (2016). Spatio-temporal clustering approach for detecting functional regions in cities. *2016 IEEE 28th international conference on tools with artificial intelligence (ICTAI)* (pp. 370–377). IEEE.

Ballatore, A., & De Sabbata, S. (2018). Charting the geographies of crowdsourced information in greater london. *The annual international conference on geographic information science* (pp. 149–168). Springer.

Ballatore, A., & De Sabbata, S. (2019). Los Angeles as a digital place: The geographies of user-generated content. *Transactions in GIS, 24*(4).

Batty, M. (2013). *The new science of cities.* The MIT Press.

Batty, M. (2019). Urban analytics defined. *Environment and Planning B: Urban Analytics and City Science, 46*(3), 403–405.

Bettencourt, L. M., Lobo, J., Helbing, D., Kuhnert, C., & West, G. B. (2007). Growth, innovation, scaling, and the pace of life in cities. *Proceedings of the National Academy of Sciences, 104*(17), 7301–7306.

Blank, G., & Lutz, C. (2017). Representativeness of social media in great britain: Investigating facebook, linkedin, twitter, pinterest, google+, and instagram. *American Behavioral Scientist, 61*(7), 741–756.

Boeing, G. (2017). Osmnx: New methods for acquiring, constructing, analyzing, and visualizing complex street networks. *Computers, Environment and Urban Systems, 65*, 126–139.

Bradbury, A., Cameron, A., Castell, B., Jones, P., Pharoah, T., Reid, S., & Young, A. (2007). Manual for streets. *Technical report.* Department for Transport.

Calabrese, F., Reades, J., & Ratti, C. (2009). Eigenplaces: Segmenting space through digital signatures. *IEEE Pervasive Computing, 9*(1), 78–84.

Campbell, A. T., Eisenman, S. B., Lane, N. D., Miluzzo, E., Peterson, R. A., Lu, H., Zheng, X., Musolesi, M., Ahn, G.-S., et al. (2008). The rise of people-centric sensing. *IEEE Internet Computing, 12*(4), 12–21.

Chen, Y., Xu, J., & Xu, M. (2015). Finding community structure in spatially constrained complex networks. *International Journal of Geographical Information Science, 29*(6), 889–911.

Clauset, A., Newman, M. E., & Moore, C. (2004). Finding community structure in very large networks. *Physical Review E, 70*(6), Article 066111.

Crooks, A., Pfoser, D., Jenkins, A., Croitoru, A., Stefanidis, A., Smith, D., ... Lamprianidis, G. (2015). Crowdsourcing urban form and function. *International Journal of Geographical Information Science, 29*(5), 720–741.

D'Andrea, E., Ducange, P., Loffreno, D., Marcelloni, F., & Zaccone, T. (2018). Smart profiling of city areas based on web data. *2018 IEEE International Conference on Smart Computing (SMARTCOMP)* (pp. 226–233). IEEE.

De Mauro, A., Greco, M., & Grimaldi, M. (2015). What is big data? A consensual definition and a review of key research topics. *AIP conference proceedings. vol. 1644. AIP conference proceedings* (pp. 97–104). American Institute of Physics.

DESA, U. (2018). Revision of world urbanization prospects. *UN Department of Economic and Social Affairs, 16*.

Donoho, D. (2017). 50 years of data science. *Journal of Computational and Graphical Statistics, 26*(4), 745–766.

Dunford, M. (2009). Regional development models. International encyclopedia of geography: People, the earth, environment and technology: People, the earth, environment. *Technology*, 1–15.

Edelsbrunner, H., Kirkpatrick, D., & Seidel, R. (1983). On the shape of a set of points in the plane. *IEEE Transactions on Information Theory, 29*(4), 551–559.

Ertio, T.-P. (2015). Participatory apps for urban planning—Space for improvement. *Planning Practice and Research, 30*(3), 303–321.

Ewing, R., & Cervero, R. (2010). Travel and the built environment. *Journal of the American Planning Association, 76*(3), 265–294.

Frank, L., Sallis, J., Saelens, B., Leary, L., Cain, K., Conway, T., & Hess, P. (2009). The development of a walkability index: Application to the neighborhood quality of life study. *British Journal of Sports Medicine, 44*, 924–933.

Gao, S., Liu, Y., Wang, Y., & Ma, X. (2013). Discovering spatial interaction communities from mobile phone data. *Transactions in GIS, 17*(3), 463–481.

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*(3), 446–467.

Gehl, J., & Svarre, B. (2013). *Jan Gehl & Birgitte Svarre.* Island press.

Girres, J.-F., & Touya, G. (2010). Quality assessment of the french openstreetmap dataset. *Transactions in GIS, 14*(4), 435–459.

Glaeser, E. (2011). *Triumph of the City.* (Pan.).

Grauwin, S., Sobolevsky, S., Moritz, S., Godor, I., & Ratti, C. (2015). Towards a comparative science of cities: Using mobile traffic records in new york, london, and hong kong. *Computational approaches for urban environments* (pp. 363–387). Springer.

Guo, D., Jin, H., Gao, P., & Zhu, X. (2018). Detecting spatial community structure in movements. *International Journal of Geographical Information Science, 32*(7), 1326–1347.

Hagberg, A., Swart, P., & S Chult, D. (2008). *Exploring network structure, dynamics, and function using networkx. Technical report.* Los Alamos, NM (United States): Los Alamos National Lab.(LANL).

Hajrasouliha, A., & Yin, L. (2015). The impact of street network connectivity on pedestrian volume. *Urban Studies, 52*(13), 2483–2497.

Haklay, M. (2010). How good is volunteered geographical information? A comparative study of open-streetmap and ordnance survey datasets. *Environment and Planning. B, Planning & Design, 37*(4), 682–703.

He, M., Glasser, J., Pritchard, N., Bhamidi, S., & Kaza, N. (2019). *Demarcating geographic regions using community detection in commuting networks.* (arXiv preprint arXiv:1903.06029).

Hecht, B., & Stephens, M. (2014). A tale of cities: Urban biases in volunteered geographic information. *Eighth international AAAI conference on weblogs and social media*.

Horton, F. E., & Reynolds, D. R. (1971). Effects of urban spatial structure on individual behavior. *Economic Geography, 47*(1), 36–48.

Jiang, B., & Miao, Y. (2015). The evolution of natural cities from the perspective of location-based social media. *The Professional Geographer, 67*(2), 295–306.

Kallus, R., & Law-Yone, H. (2000). What is a neighbourhood? The structure and function of an idea. *Environment and Planning. B, Planning & Design, 27*(6), 815–826.

Kitchin, R. (2014). *The data revolution: Big data, open data, data infrastructures and their consequences.* Sage.

Kittel, T.. Alpha-Shapes. (2019). https://github.com/timkittel/alpha-shapes (Accessed: 2019-11-15).

Knaap, E., Wolf, L., Rey, S., Kang, W., & Han, S. (2019). *The dynamics of urban neighborhoods: A survey of approaches for modeling socio-spatial structure.* (SocArXiv).

Lambiotte, R., Delvenne, J.-C., & Barahona, M. (2008). *Laplacian dynamics and multiscale modular structure in networks.* (arXiv preprint arXiv:0812.1770).

Lazer, D., & Radford, J. (2017). Data ex machina: Introduction to big data. *Annual Review of Sociology, 43*, 19–39.

Lazer, D., Pentland, A., Adamic, L., Aral, S., Barab'asi, A.-L., Brewer, D., Christakis, N., Contractor, N., Fowler, J., Gutmann, M., et al. (2009). Computational social science. *Science, 323*(5915), 721–723.

Lenormand, M., Picornell, M., Cantu-Ros, O. G., Louail, T., Herranz, R., Barthelemy, M., ... Ramasco, J. J. (2015). Comparing and modelling land use organization in cities. *Royal Society Open Science, 2*(12), 150449.

Lindqvist, J., Cranshaw, J., Wiese, J., Hong, J., & Zimmerman, J. (2011). I'm the mayor of my house: examining why people use foursquare-a social-driven location sharing application. *Proceedings of the SIGCHI conference on human factors in computing systems* (pp. 2409–2418). ACM.

Liu, H., & Yu, L. (2005). Toward integrating feature selection algorithms for classification and clustering. *IEEE Transactions on Knowledge and Data Engineering, 17*(4), 491–502.

Mart'ı, P., Serrano-Estrada, L., & Nolasco-Cirugeda, A. (2019). Social media data: Challenges, opportunities and limitations in urban studies. *Computers, Environment and Urban Systems, 74*, 161–174.

McKenzie, G., Janowicz, K., Gao, S., & 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*, 336–346.

Miller, H. J., & Goodchild, M. F. (2015). Data-driven geography. *GeoJournal, 80*(4), 449–461.

Nelson, G. D. (2020). Communities, complexity, and the 'conchoration': Network analysis and the ontology of geographic units. *Tijdschrift Voor Economische en Sociale Geografie*.

Nelson, G. D., & Rae, A. (2016). An economic geography of the United States: From commutes to megaregions. *PLoS One, 11*(11), 1–23.

Noulas, A., Scellato, S., Mascolo, C., & Pontil, M. (2011). An empirical study of geographic user activity patterns in foursquare. *Fifth international AAAI conference on weblogs and social media*.

Noulas, A., Scellato, S., Lambiotte, R., Pontil, M., & Mascolo, C. (2012). A tale of many cities: Universal patterns in human urban mobility. *PLoS One, 7*(5).

Organizers, Kang, W., Oshan, T., Wolf, L. J., Discussants, Boeing, G., ... Xu, W. (2019). A roundtable discussion: Defining urban data science. *Environment and Planning B: Urban Analytics and City Science, 46*(9), 1756–1768.

Patterson, Z., & Farber, S. (2015). Potential path areas and activity spaces in application: A review. *Transport Reviews, 35*(6), 679–700.

PCAST (2016). PCAST report to the president on technology and the future of cities. *Technical report.* Office of Science and Technology Policy. President's Council of Advisors on Science and Technology (PCAST).

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

Quercia, D., Aiello, L. M., Schifanella, R., & Davies, A. (2015). The digital life of walkable streets. *Proceedings of the 24th international conference on world wide web* (pp. 875–884). International World Wide Web Conferences Steering Committee.

Ratti, C. (2004). Space syntax: Some inconsistencies. *Environment and Planning. B, Planning & Design, 31*(4), 487–499.

Ratti, C., Sobolevsky, S., Calabrese, F., Andris, C., Reades, J., Martino, M., ... Strogatz, S. H. (2010). Redrawing the map of great britain from a network of human interactions. *PLoS One, 5*(12).

Robinson, J. (2006). *Ordinary cities: Between modernity and development.* Psychology Press.

Rybarczyk, G., & Wu, C. (2014). Examining the impact of urban morphology on bicycle mode choice. *Environment and Planning. B, Planning & Design, 41*(2), 272–288.

Sampson, R. J. (2019). Neighbourhood effects and beyond: Explaining the paradoxes of inequality in the changing american metropolis. *Urban Studies, 56*(1), 3–32.

Sennett, R. (2018). *Building and dwelling: Ethics for the city.* Farrar, Straus and Giroux.

Shannon, C. E. (1948). A mathematical theory of communication. *Bell System Technical Journal, 27*(3), 379–423.

Shaw, S.-L., & Sui, D. (2018). Introduction: Human dynamics in perspective. *Human dynamics research in smart and connected communities* (pp. 1–11). Springer.

Shelton, T., & Poorthuis, A. (2019). The nature of neighborhoods: Using big data to

rethink the geographies of atlanta's neighborhood planning unit system. *Annals of the American Association of Geographers, 109*(5), 1341–1361.
Shelton, T., Poorthuis, A., & Zook, M. (2015). Social media and the city: Rethinking urban socio-spatial inequality using user-generated geographic information. *Landscape and Urban Planning, 142*, 198–211.
Silver, D., & Clark, T. (2016). *Scenescapes: How qualities of place shape social life.* University of Chicago Press.
Singleton, A., & Arribas-Bel, D. (2019). Geographic data science. *Geographical Analysis*.
Singleton, A. D., & Longley, P. A. (2019). Data infrastructure requirements for new geodemographic classifications: The example of london's workplace zones. *Applied Geography, 109*, Article 102038.
Singleton, A. D., Spielman, S., & Folch, D. (2017). *Urban analytics.* Sage.
Spielman, S. E., Folch, D., & Nagle, N. (2014). Patterns and causes of uncertainty in the american community survey. *Applied Geography, 46*, 147–157.
Sui, D., & Goodchild, M. (2011). The convergence of gis and social media: Challenges for giscience. *International Journal of Geographical Information Science, 25*(11), 1737–1748.