Special Issue: Data Science for Developing Cities

Urban Analytics and City Science

EPB: Urban Analytics and City Science
2022, Vol. 0(0) 1–16

DOI: 10.1177/23998083221117830
journals.sagepub.com/home/epb
SAGE

Measuring the local complementarity of population, amenities and digital activities to identify and understand urban areas of interest

AQ1

Eduardo Graells-Garrido
Department of Computer Science, University of Chile, Chile

Rossano Schifanella
Department of Computer Science, University of Turin, Italy

Daniela Opitz
Data Science Institute, Universidad del Desarrollo, Chile

Francisco Rowe
Department of Geography and Planning, University of Liverpool, UK

Abstract

Identifying and understanding areas of interest are essential for urban planning. These areas are normally defined from static features of the resident population and urban amenities. Research has emphasised the importance of human mobility activity to capture the changing nature of these areas throughout the day, and the use of digital applications to reflect the increasing integration between material and online activities. Drawing on mobile phone data, this paper develops a novel approach to identify areas of interest based on the degree of complementarity of digital activities, available amenities and population levels. As a case study, we focus on the largest urban agglomeration of Chile, Santiago, where we identify three distinctive groups of areas: those concentrating (1) high *availability* of amenities; (2) high *diversity* of amenities and digital activities; and (3) areas lacking amenities, yet, presenting high usage of digital leisure and mobility applications. These findings identify areas where digital activities and local amenities play a complementary role in association with local population levels, and provide data-driven insights into the structure of material and digital activities in urban spaces that may characterise large Latin American cities.

Keywords

Mobile phone data, areas of interest, digital activities

Corresponding author:
Eduardo Graells-Garrido, Department of Computer Science, University of Chile, Av Plaza 680, Las Condes, Santiago 7610315, Chile.
Email: egraells@dcc.uchile.cl

Introduction

An Area of Interest (AoI) is often defined as an urban area where the activities of the citizens are linked to the built environment, such as educational areas and business districts (Crooks et al., 2015; Hu et al., 2015). The study of these areas is essential as their identification offers a variety of applications, such as identifying potential areas for new businesses, guidance on relevant amenities to locals and tourists, and urban planning interventions (Yuan et al., 2012). AoIs have been extensively studied using traditional data sources, such as surveys and population censuses (Herold et al., 2005). More recently, the development of new technologies and increased availability of non-conventional data sources, such as social media content, GPS traces and mobile phone records, have allowed the identification of AoI at finer granularities based on the activities occurring within them (Yuan et al., 2012; Wu et al., 2014).

Despite these developments, less is known about the ways digital activities relate to local urban amenities in the context of AoIs. Traditionally, digital activities have been conceptualised as distinctive recognisable elements associated with human behaviour (Fors and Wiberg, 2010). These activities, however, have become more integrated with our real-world behaviour via augmented reality and mobile phone applications, blurring the line between the online and physical world (Graham et al., 2013; Nagenborg et al., 2010; Zook et al., 2015). As a result, digital and material activities now tend to occur simultaneously complementing each other (Graham et al., 2013; Zook et al., 2015). Integrating the relationship between both types of activities to identify AoI is key to determine and understand local urban spaces today.

To address this gap, we develop a novel framework to identify urban AoIs based on the relationship between urban amenities and digital activities. Based on the Activity-Based Modelling framework proposed by Reichman (1976), we unified the classification of digital activities and urban amenities into a single taxonomy with digital and material aspects. We propose a two-stage approach. First, we apply Geographically Weighted Regression (GWR) modelling to measure spatial variations in the relationship between local levels of present population, urban amenities and digital activities. Second, we use the derived GWR coefficients as input into a HDBSCAN clustering (Campello et al., 2013) to identify AoIs.

We use mobile phone data from Santiago, Chile, to capture digital activities. Santiago is the capital and largest city of Chile, expanding into 641 Km^2 and accommodating 30% (5.614 m) of the national population. Santiago is considered to follow the Latin American city model characterised by Ford (1996). Large Latin American cities are often defined by a concentration of economic activity around a central spine running from the central business district to affluent neighbourhoods with large employment centres (Rodríguez-Vignoli and Rowe, 2017). Peripheral and radial areas are characterised by 'disamenities' or limited local amenities (Suazo-Vecino et al., 2019). Our findings surface unseen patterns in the complementarity of digital and material activities in Santiago; thus, they extend existing conceptual urban models of Latin American cities.

Related work in the analysis of areas of interest

There is an inherent relationship between AoIs in a city and the present population and available amenities in such areas (Chen et al., 2019; Hu et al., 2015). In this context, the data-driven identification of urban AoIs has mainly followed two approaches: the use of predefined spatial contours and the identification of customised boundaries. In the first case, the shape of the spatial units is known, such as zoning regulations or regular grids, which are then classified into functional areas (e.g. residential, industrial or commercial) using different clustering methods. Some of these techniques classify areas based on socio-economic traits, building shapes and available amenities (Xing and Meng, 2018), as well as clustering of mobility patterns and availability of amenities (Liu et al., 2021; Yuan et al., 2020).

The second approach involves spatial clustering, where the shape of the AoIs is inferred from cluster compositions. With point data, a common method is to consider the convex hull of a cluster as its boundary (Cranshaw et al., 2012), while other methods delimit areas based on the available urban infrastructure, such as the surrounding street network (Liu et al., 2021). A frequently used clustering method is Density-Based Spatial Clustering of Applications with Noise (DBSCAN, Ester et al., 1996) and its variations (Liu et al., 2021), as it automatically detects the number of clusters.

The adoption of mobile phone records has been studied in the context of urban mobility (Blondel et al., 2015). Similar to our work, Ren et al. (2019) proposed a methodology to classify Chinese administrative areas into socio-economic categories using amenities and fine-grained application traffic data. However, this work differs from our approach since it focuses on area classification, having specific amenities and application categories (such as Travel or Games) as input. In a different study, web browsing activities in indoor retail spaces have been successfully used to predict demographic characteristics of retail customers (Ren et al., 2018).

Similarly, we exploit aggregated data at the level of mobile phone towers; we measure the relationship between several activity patterns and their role in explaining the size of the present population throughout the city. To account for local variations and influence from nearby towers, we use a GWR model (Brunsdon et al., 1996), and use its outputs to apply the HDBSCAN algorithm (Campello et al., 2013), which automatically determines the main parameter of DBSCAN. In addition to considering urban amenities, we include signals of mobile phone applications and Website usage, which allows us to discover relevant areas that do not necessarily have available urban amenities. To the extent of our knowledge, this approach linking digital activities to urban amenities and present population to identify AoIs in cities has not been proposed before. Such an approach can uncover structural patterns of urban inequalities in the availability and accessibility to local amenities and digital technology, and ultimately, help to inform urban policy interventions seeking to improve the supply of local services and digital infrastructure.

Data and methods

In this work, we seek to identify a new conceptualisation of AoIs based on material and digital activities. We propose a novel process of geographical analysis and machine learning, having as input several data sources, including mobile phone data, OpenStreetMap data and official data. The core unit of analysis when using mobile phone data are GSM towers, thus, we define an AoI as a set of towers within an enclosing polygon, with the common characteristic of presenting a high number of connected devices at a given time; here, the notion of high is local, that is, a high value in an area of the city may be a low value in another. In summary, our approach follows a two-step process after deriving features from the data (see Figure 1). First, we use regression to find the relationship between features derived from the data and the interest in the area associated with each tower. Second, we group towers into clusters based on how their features explain the interest in each area, proxied through the number of connections. In this section, we detail the data and the process proposed, including the analysis of intermediary results of the process.

Data

As a case study of our proposed method, we integrate mobile phone data, the OpenStreetMap data source and official data sources from Santiago, Chile. Here, we describe these sources.

Mobile phone data. We used two mobile phone datasets provided by the telecommunications operator Telefónica Movistar in Chile: aggregated traffic from Deep Packet Inspection (DPI) and trajectories from Extended Detail Records (XDR). Both datasets were generated from more than a

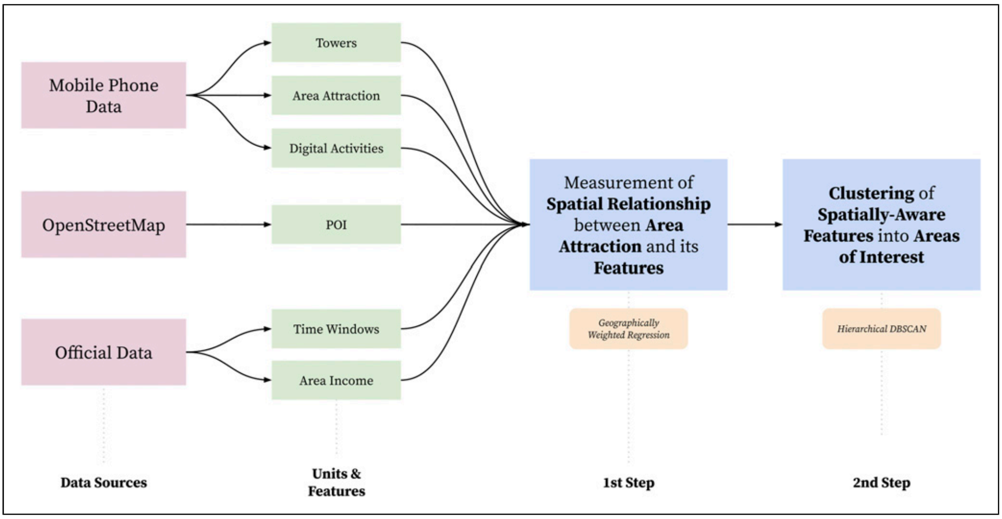
Mobile Phone Data
OpenStreetMap
Official Data
Towers
Area Attraction
Digital Activities
POI
Time Windows
Area Income
Measurement of Spatial Relationship between Area Attraction and its Features
Geographically Weighted Regression
Clustering of Spatially-Aware Features into Areas of Interest
Hierarchical DBSCAN
Data Sources
Units & Features
1st Step
2nd Step

Figure 1. A schema of the methods from this paper.

million phones between 27 July 2016, and 10 August 2016, with a temporal granularity of 15 minutes. Telefónica has nearly one-third of the Chilean market and in the urban area of Santiago operates 1374 towers distributed in the city, covering 34 municipal boundaries. Note that some towers have the same position (e.g. multiple level towers inside shopping centres). As such, we aggregated them, reducing the number of studied towers to 976. The spatial distribution of the reduced towers is shown in Figure S1 from Supplemental Material.

The DPI data contain the number of connections in each tower to the 5000 most accessed IP addresses, representing around 80% of the all accessed IPs in the country as is indicated by the operator (Graells-Garrido et al., 2018a). Through manual inspection of these addresses, it was possible to assign mobile phone applications and websites to them. We categorised these applications and websites into thematic categories, such as audio (e.g. Spotify), business (e.g. bank applications), games (e.g. Pokémon Go), social networks (e.g. Facebook) and transportation (e.g. Cabify and other ride-hailing applications). Table S1 in Supplemental Material shows the full list of categorised apps.

The XDR data correspond to sparse trajectories per device. These data are used to count the number of people who connect to cell phone towers, offering a way to estimate the local present population in an area at a given point in time. In Santiago, the distributions of home and work locations inferred from this data have a high correlation with those from household surveys, although the trajectories under-represent short trips in the city, as these trips cannot be seen due to the spatial resolution of towers (Graells-Garrido et al., 2018b).

OpenStreetMap (OSM). OpenStreetMap is a global, open geographical database that is freely available and fairly accurate for many cities (Haklay, 2010; Zhang and Pfoser, 2019). It contains several types of geographical features, including urban infrastructure and amenities. This dataset is particularly valuable for Santiago as there is no official dataset for amenities in the city. For this project, we used urban amenity data from OSM. OSM amenities can be classified into *education* (educational services at all age levels), *food* (restaurants, coffee shops, etc.), *professional* (amenities that offer professional services), *convenience* stores, *health* (including hospitals and clinics), *retail* (department stores and shopping malls), *government* facilities, *finance* (banks), *recreation* (such as parks), *entertainment* (which, unlike recreation amenities, require some form of payment, such as a theatre), *accommodation* (hotels), *nightlife* (including bars), *religion* (churches and similar places) and *money* (money exchange). These amenities are places that enable the population to perform activities within them; as such, we included them in our process.

AQ2

Official data: Santiago travel survey 2012. This survey collects data on travel behaviour, such as origin, destination, and mode of transportation of trips, and demographics of the resident population. It is administered by the Chilean Ministry of Transport and Telecommunications and is collected every 10 years. The survey defines a set of daily periods relevant for transportation, such as morning and afternoon commuting. In our analysis, we used the periods defined in the survey, as each period has a characteristic set of primary activities according to the daily routines in the city (see Table 1).

The survey provides the most recent household income measurement available with geographical coordinates. From it, we computed the base-10 logarithm of mean home income at every traffic analysis zone, and assigned that income to the towers within each zone. Note that there are 706 zones in the area under study, of which only 473 have towers within, although these areas cover all municipalities in the city (see Figure S1 in Supplemental Material for more details).

Activity definition

We study how people perform physical and digital activities through a unified lens. We built this lens from the Activity-Based Model (ABM) developed by Reichman (1976) and extended by Pas (1982), which characterises citizens as agents performing activities in places according to their lifestyle and moving from one place to another via transportation. ABM defines four types of activities:

- *Maintenance*: shopping and personal business, non-income activities required to maintain a household.
- *Subsistence*: work and school.
- *Discretionary*: leisure, recreation, optional activities engaged in for enjoyment.
- *Mobility*: moving between places to perform another activity.

Although digital technologies have been considered in the activity decision process (Ren and Kwan, 2009), the framework draws divisions between concepts that seem to be blurred today. For example, it assumes that travelling excludes other activities; however, nowadays mobile phones can enrich travel time with additional activities if given the right conditions (Jain and Lyons, 2008). Furthermore, the digital context supports multi-tasking natively, as many activities can be performed simultaneously.

Table 1. Names and extent of the hourly periods defined in the Santiago Travel Survey from 2012.

Period	Start	End	Comment
Morning peak	06:00	08:59	Most people are commuting
Morning valley	09:00	11:59	Work, study and other activities outside home are predominant
Lunch	12:00	13:59	Lunch break
Afternoon valley	14:00	17:29	(Same as morning valley)
Afternoon peaks	17:30	20:29	Most people are commuting back home
Night valley	20:30	22:59	Most people are at home
Night	23:00	05:59	Most people are sleeping at home

To extend the ABM framework, we define *digital activities* as the usage of mobile phones to perform tasks that fall in one of the activities previously introduced; that is, for a given application or Web site usage from DPI data, an activity can be assigned. We classified the categories of apps and websites in our DPI data into the four activities defined by ABM. Some categories were assigned directly (e.g. games → *discretionary*, business → *subsistence*, transportation → *mobility*), while others required discussion and context analysis (see Supplemental Material Section 3 for details on the assignment process). For instance, a messaging service may be used either as any of these activities; however, since the primary usage of messaging services is to maintain communication with others, not only coworkers, we assign it to *maintenance*. Conversely, Email tends to be used for formal communication, and as such it is categorised as *subsistence*. Reading and news applications were assigned *discretionary*, among others. The complete categorisation and its corresponding activities are shown in Table S1 from Supplemental Material.

In total, 74.35% of DPI traffic is caused by digital *discretionary* activities. The second greatest source of traffic are digital *mobility* activities (11.37%), which can be explained by the popularity of ride-hailing applications in Santiago. The third source of traffic are *maintenance* activities (11.2%). Finally, digital *subsistence* comprised only 3.05% of the traffic. This small value is arguably expected as their traffic could be generated at desktop computers or workplaces with non-mobile network connections.

Urban amenities enable material activities in a given spatial unit, and, as such, they have been used in activity modelling (Jiang et al., 2015). Thus, we classified amenity categories into one of the main activities according to its primary usage. The definition of primary usage allows assigning a single category to an amenity. For instance, a tea shop is primarily a *discretionary* amenity, though we recognise it could be a *subsistence* amenity for its workers. The final categorisation results into the following groupings:

- *Maintenance* (45.68% of amenities): *accommodation*, *convenience*, *finance*, *food*, *government*, *health*, *money*, *retail*, *sustainability.*
- *Subsistence* (45.43% of amenities): *professional*, *education.*
- *Discretionary* (8.89% of amenities): *entertainment*, *nightlife*, *religion*, *recreation.*

At this point we have described the datasets that serve as input to our process, and how we mapped part of the data into activities. Having a unified lens for amenities and mobile phone application usage enables us to move forward into defining how to identify AoIs in the city.

Units of analysis and features

The spatial unit of analysis is defined as a point centred on each mobile phone tower. For each tower t, we define a vector of features $\vec{u_t}$ that include the number of amenities per activity, a rate of digital activities performed by phones connected to the tower, a base-10 logarithm of mean income imputed to the tower based on the travel survey, and the diversity of activities (both, digital and material) held there. Next, we detail these attributes.

We counted the number of amenities in each activity category within a radius of 500 m (according to the mobile phone company, more than 80% of the devices connected to a tower are within this radius). Note that we discarded the use of a Voronoi grid as this would split the urban space in a binary way with respect to each tower. Tower coverage overlaps, and thus, it is not binary. For instance, two towers may be close enough to cover the same radius (and thus, have a similar group of amenities associated with them), whereas a Voronoi grid would impose that an amenity can be inside one cell only.

Regarding digital activities, we aim to compare the relative usage of applications rather than the absolute value, as the absolute value is correlated with the present population. Thus, we define a *Digital Activity Rate* (DAR) in a tower within a period of time according to the following expression

$$DAR_t(c, w) = \frac{\#\ of\ \ Accesses\ \ to\ \ c\ \ in\ \ tower\ \ t\ \ at\ \ period\ \ w}{Present\ \ Population(t, w)},$$

where: c is a digital activity, w is a period of the day as defined in the travel survey, t is a mobile phone tower, and the *Present Population* corresponds to the number of active devices connected to a tower t during a period w measured from XDR.

We also define the diversity of digital/physical activities by computing the Shannon Entropy formula

$$H_t = -\sum p(activity, t) log\ p(activity, t),$$

where H_t is diversity at tower t, $p(activity, t)$ is the probability of performing a request of an activity application at tower t, or the probability of finding an amenity for the corresponding activity at tower t.

We estimated these features for all towers in the dataset (see Figure 2). In spatial terms, digital *maintenance* traffic tends to be higher in the central parts of the city and near metro stations, whereas *subsistence* and *discretionary* activities follow inverse spatial distributions. With respect to amenities, *maintenance* and *discretionary* locations tend to be concentrated in the historical centre and in the emergent centre at wealthy areas, particularly near the north-east area of the city, as it concentrates most of the high income homes. Conversely, *subsistence* amenities have greater coverage of the city, which can be explained due to the availability of work and study places in the whole city. Areas with high diversity of digital activities also have high economic development. Furthermore, there are areas with high income but low diversity of amenities, and areas with low income but high diversity of amenities. All these differences motivate a spatially aware characterisation of them, with the ultimate aim of identifying AoIs.

Spatial variations in relationships between population and activities

Once the $\vec{u_t}$ vectors were built with these attributes, we fitted Negative Binomial (NB) Geographically Weighted Regression (GWR) (Fotheringham et al., 2002) to explain spatial variations in the present population as a function of the digital activities and the available amenities. We used a NB distribution as the present population can be a highly skewed count variable. The NB distribution accounts for over-dispersion in the data which often violates the equidispersion assumptions in Poisson regression models (Cameron and Trivedi, 1986). GWR can capture geographic variability in this relationship over space by estimating a set of parameters for each unit of analysis in the dataset. It does so by creating a NB regression for each unit of analysis, where observations in the data are inversely weighted according to their proximity to the focal location. The weight is determined by a kernel function. Our GWR model is expressed as follows

$$E[T_{tw}] = e\left[\beta_{0tw} + \sum\nolimits_k \beta_{(k)}\left(\vec{u_t}\right)x_{ktw}\right],$$

where T_{tw} is the present population, $\beta_{(k)}(\vec{u_t})$ is a vector of geographically varying parameters at a given position of T_{tw} and x_{ktw} is the kth-feature of tower t at period w.

Since our GWR model is not time-aware, we needed to select a specific time window to work with. We fitted multiple models, one per time period from the travel survey (see Table 1). In each, the independent variables were quantile-transformed before fitting, with normal distribution output.

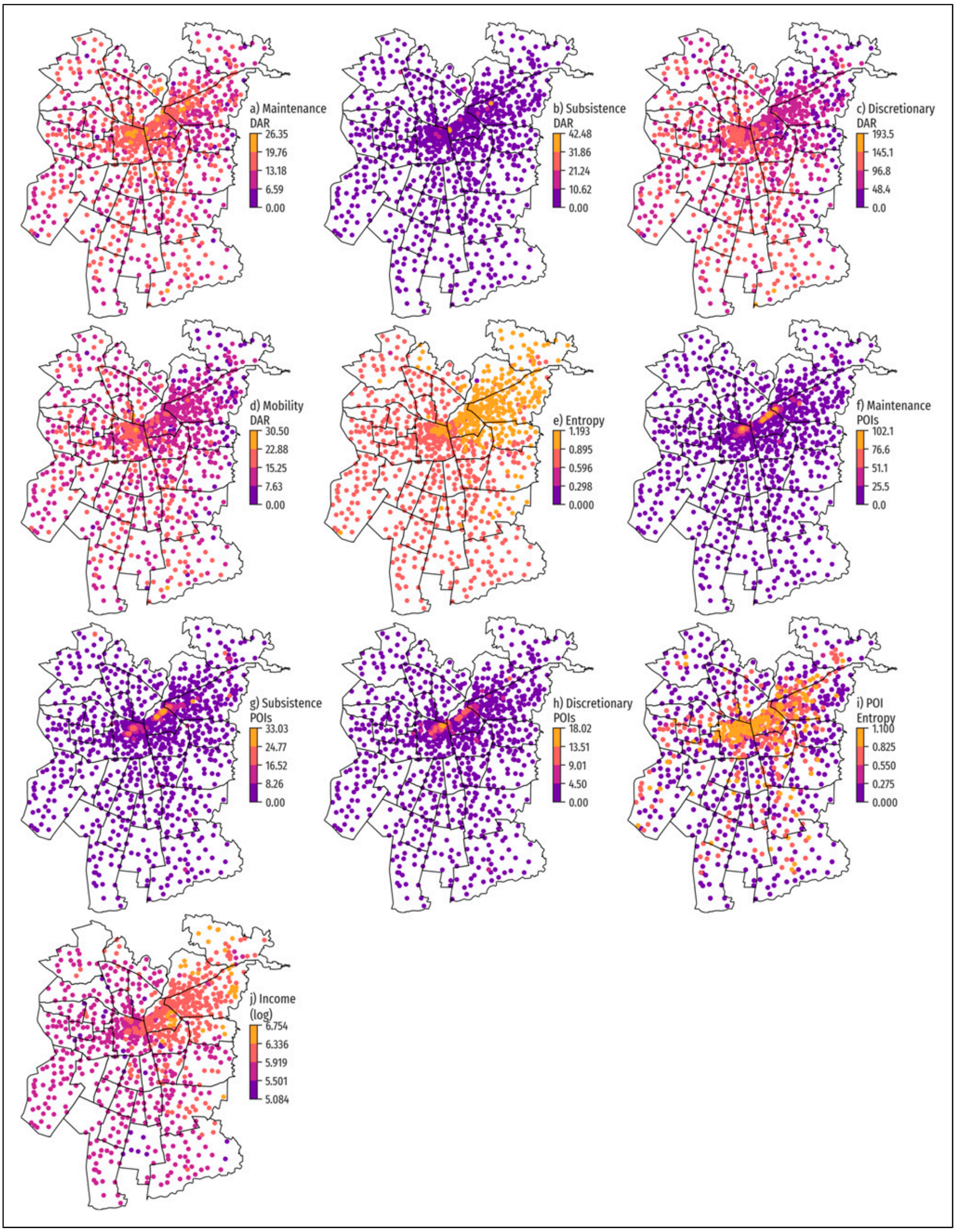

Figure 2. Spatial distribution of features.

Since each time window covers several hours, we averaged the size of the present population per tower during the corresponding time ranges.

To select a time window to focus our analysis on, we compared the corrected Akaike Information Criterion (AICc) of each model, a metric of information loss that accounts for multiple model comparisons and their sample size (Cavanaugh, 1997). Given that AICc is an information loss metric, lower values are preferred; in our case, the *morning valley* (from 9a.m. to 12p.m., a time characterised by subsistence activities) was the best model (AICc = 645.29). We fitted regular NB regression to compare with, and the GWR model consistently exhibited better AICc in all periods, suggesting that local models may be preferred at any time of the day. Additionally, we tested GWR and regular models without considering the digital activities, and found that those models present much higher information loss according to AICc (see Figure S7 in Supplemental Material). The result is consistent with our expected dynamics of the city, in the sense of the *morning valley* period being arguably the most stable in terms of activities in daily routines.

The selected model used an adaptive bi-square kernel with a bandwidth of 362 m, a value found using bisection search with AICc as criterion (Oshan et al., 2019). The residuals of each tend to be symmetric, and the local condition number for each spatial unit for all models was below than the common threshold of 30 in the literature (see Section 4 in Supplemental Material for the full assessment). Thus, we proceeded to analyse the regression coefficients, summarised in Table 2.

The local variations of each regression coefficient present a different picture than the raw features values (see Figure 3, c.f. Figure 2). For instance, the geographical variation of *subsistence* activities reflects the idiosyncrasy of the two main poles of the city: the historical centre is a hotspot of *subsistence* amenity and *amenity diversity*, whereas the new attraction pole in the wealthy area of the city is a hotspot of digital *subsistence*. The latter is more specialised, particularly with recent office skyscrapers. The digital *maintenance* values are positive in almost the entire city, in coherence with the coefficient from the regular model (which is statistically significant); however, the digital *discretionary*, while positive in more than half of the city's surface, exhibit an opposite trend: when *maintenance* has a high/positive explanatory effect, *discretionary* has a low/negative effect, and vice versa. Similarly, the *discretionary* amenities present negative or neutral effects in the whole city.

Clustering of spatially aware features

The coefficient distributions exhibit geographical variations when explaining the present population in each unit of analysis. These variations show concentration of feature values related to the present population, and thus, they may be the basis to identify AoIs if grouped.

AQ3

Table 2. Summary table of regression results in the global model (Negative Binomial Regression) and the local model (GWR with GLM Negative Binomial Regression).

	Regular model (AICc = 874.38)				GWR model (AICc = 645.29)				
Variable	β	SE	t (Est/SE)	*p*	β mean	β STD	β Min	β Median	β Max
Intercept	**10.126**	0.046	221.729	0.000	10.121	0.284	9.711	10.118	10.677
Maintenance DAR	**0.337**	0.086	3.935	0.000	0.459	0.231	−0.015	0.503	0.968
Subsistence DAR	0.097	0.075	1.287	0.198	0.115	0.172	−0.209	0.062	0.519
Discretionary DAR	**0.446**	0.108	4.114	0.000	0.234	0.476	−0.720	0.219	1.181
Mobility DAR	−**0.284**	0.083	−3.423	0.001	−0.146	0.180	−0.776	−0.105	0.294
Digital diversity	−0.035	0.066	−0.526	0.599	−0.188	0.263	−0.964	−0.160	0.406
Maintenance POIs	−0.023	0.023	−1.007	0.314	−0.037	0.057	−0.129	−0.043	0.099
Subsistence POIs	−0.020	0.017	−1.189	0.235	−0.008	0.027	−0.070	−0.005	0.047
Discretionary POIs	−**0.070**	0.015	−4.672	0.000	−0.045	0.025	−0.092	−0.045	−0.000
Amenity diversity	0.045	0.026	1.711	0.087	0.053	0.048	−0.080	0.065	0.132
Mean log income	−0.030	0.043	−0.709	0.478	0.074	0.132	−0.152	33	0.399
Model stats (*N* = 976)	DF = 965.00, Pearson chi2 = 514.58				Adaptive bi-square kernel, bandwidth = 362m, DF = 898.44				

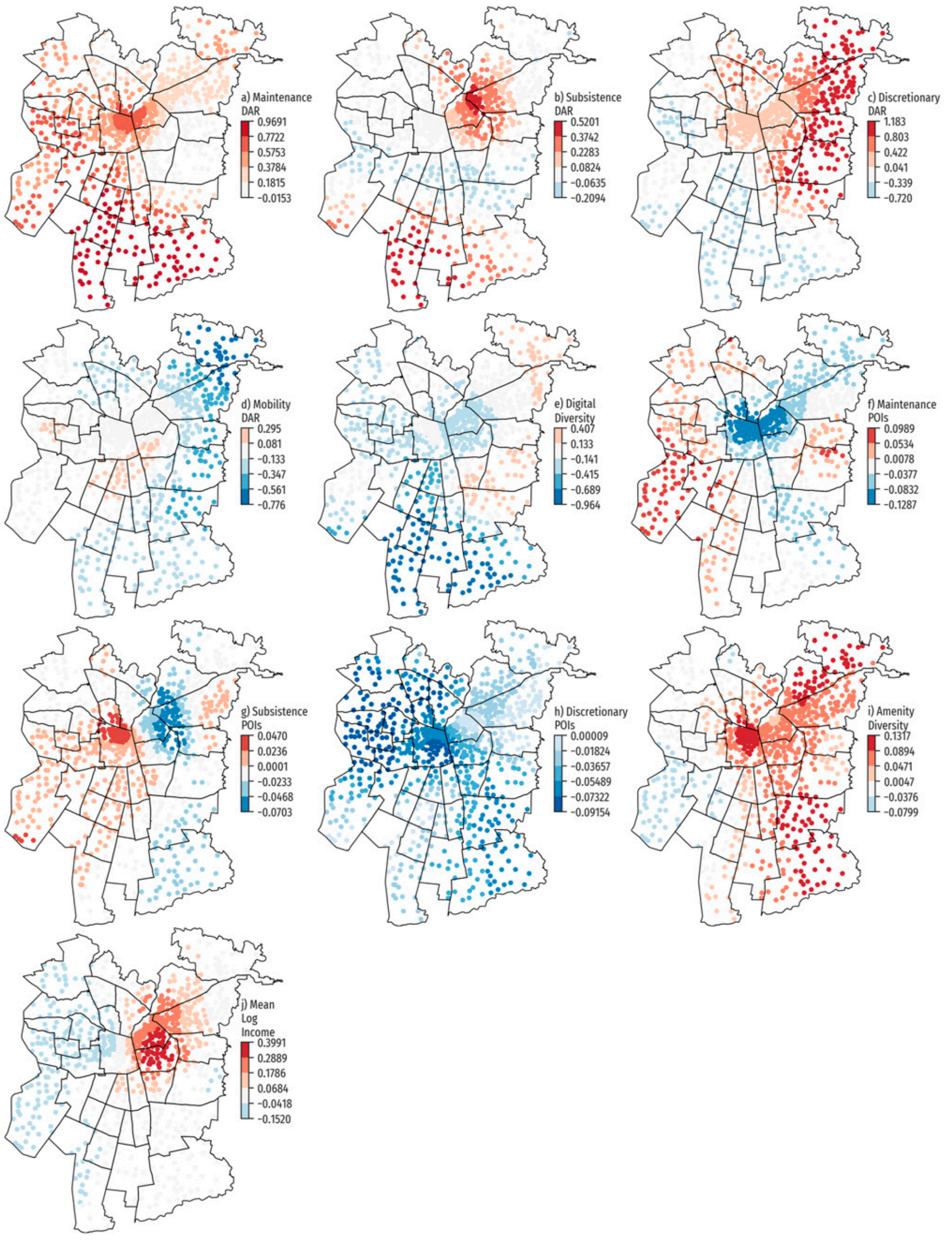

Figure 3. Geographically Weighted Regression coefficients. Towers are coloured according to the corresponding β coefficient value. A diverging colour scale was used in each coefficient to identify positive (red), neutral (grey) and negative values (blue).

To identify AoIs, we created a vector space using the $\beta_{(k)}(\vec{u_t})$ GWR coefficients from each spatial unit. We clustered these vectors using HDBSCAN (Campello et al., 2013; McInnes et al., 2017). The HDBSCAN algorithm extends DBSCAN (Ester et al., 1996), a density-based method that detects the number of clusters using two parameters: the search radius and the minimum number of points reachable within the search radius. HDBSCAN automatically determines local search radiuses. It needs one parameter only: the minimum number of points. A common heuristic to define this parameter is to use the natural logarithm of the total number of observations (Birant and Kut, 2007), which is approx. six in this context. By applying this method, we found 20 AoIs. We analyse them in the next section.

Results

Identifying complementarity of digital and material activities

We identified 19 distinctive Areas of Interest (displayed and labelled in Figure 4(a)), all of them with well defined, non-overlapping boundaries. Area 0 extends a large geographical area, comprising some of the poorest areas of Santiago and displaying deficient availability of most types of amenities and high availability of digital *maintenance*, *discretionary* and *mobility* activities. This suggests that digital activities in urban socio-economically deprived urban spaces tend to complement material activities. Mobile phone applications may offer a way for deprived households to access basic services, such as access to education and to coordinate social activities. In more affluent areas, the level of complementarity of these activities appears to be different, as described later in this section.

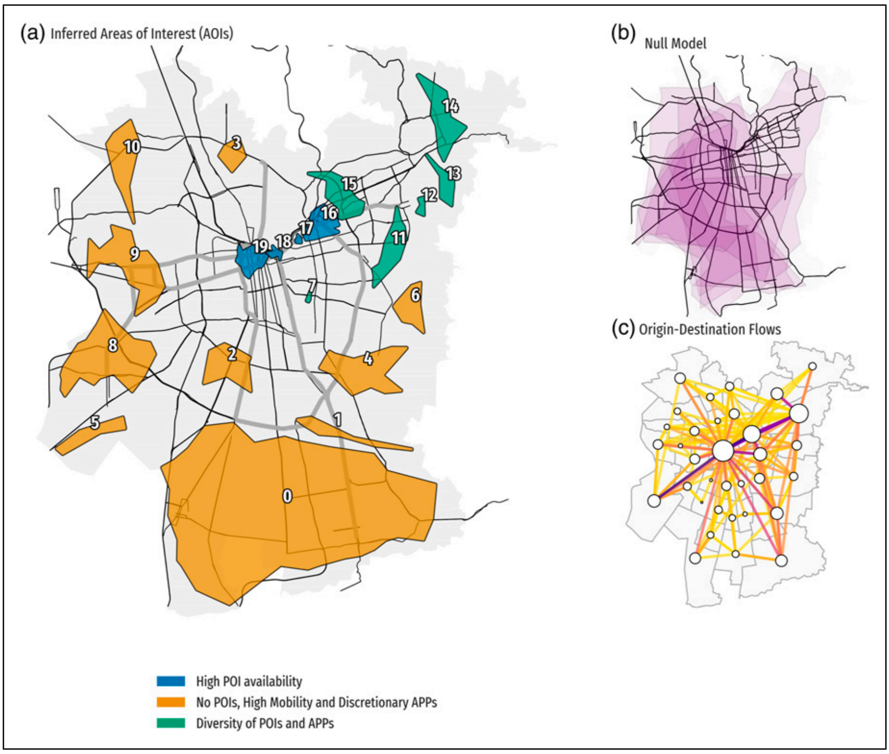
(a) Inferred Areas of Interest (AOIs)
0
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
(b) Null Model
(c) Origin-Destination Flows
🟦 High POI availability
🟧 No POIs, High Mobility and Discretionary APPs
🟩 Diversity of POIs and APPs

Figure 4. Areas of Interest inferred using HDBSCAN of towers during the morning valley period. a) AoIs identified by clustering the Geographically Weighted Regression coefficients. Each cluster is labelled. b) Clustering performed on tower features, as a null model. c) Origin–Destination movements in the preceding period to analysis (morning peak). Node size represents in-degree and darker edges have a greater number of trips.

Similar patterns of deficient availability of urban amenities and high availability of digital *maintenance*, *discretionary* and *mobility* activities tend to characterise Areas 8 and 9. They also show high population levels of middle income households and are densely populated. They currently serve as transition areas to commute to employment centres in Santiago. There, a lack of urban amenity infrastructure seems to be complemented by digital activities. This has been exemplified in recent years by the Pokémon Go videogame, which increased the use of the physical space in many areas without amenities (Graells-Garrido et al., 2017).

Areas 16 and 19 are the historical city centre and regenerated central neighbourhoods. Areas surrounding them (i.e. areas 15, 17, 18) may be merged into the known business centres in the future according to their development, as they are situated along a structural axis of the city (the Alameda-Providencia-Apoquindo avenues).

The rest of the areas have smaller present populations than average, and are characterised mainly by a lack of amenities. Areas close to the wealthy sectors of the city have less digital activities than average, but higher digital *diversity*, whereas clusters in the middle- and lower-income sectors have the characteristics discussed above. This suggests that more affluent populations consume a more diverse diet of digital information, signalling a potential digital divide in the city (Warf, 2001).

To further understand the AoIs, we performed two additional analyses. First, to validate these results, we compared our areas with a baseline null model using HDBSCAN based on raw tower features and present population. The resulting clusters from the null model are considered to have no meaningful geographical interpretation, as they display overlapping and seemingly random shapes (Figure 4(b)). Second, we estimated origin–destination movements between municipalities from XDR in the *morning peak* period of time, when people tend to commute to work (see Figure 4(c)). The stronger edges from the network are aligned among the structural axis of Alameda-Providencia-Apoquindo, reflecting a relatively centralised urban structure despite having multiple centres of employment (Limtanakool et al., 2007).

In Santiago, no other studies have systematically identified AoIs. Previous work has focused on measuring and understanding spatial residential segregation (Cox and Hurtubia, 2021; Fuentes et al., 2022) and urban and rural regions (Sotomayor-Gomez and Samaniego, 2020). We highlight that the AoIs obtained from our method are novel as they improve the understanding of the unequal distribution of spatial material and digital activities. Existing work focuses on defining urban AoIs based on material or population movement activities only.

Extent of complementarity

We computed standardised coefficient values for individual AoIs (see Figure S8 in Supplemental Material), and performed agglomerative clustering over these coefficients to identify different types of AoIs (Figure 4(a)). There are three main groups of AoIs: those characterised by high availability of urban amenities (blue AoIs in Figure 4(a)), those characterised by a high diversity of amenities and digital activities (green AoIs in Figure 4(a)), and those characterised by lack of amenities, but a high amount of digital *mobility* and *discretionary* activities. These groups exhibit geographical patterns: ‘high availability of amenities’ is placed in the Alameda-Providencia-Apoquindo axis, and ‘high diversity of amenities and digital activities’ is located at the end of the spine projected from that axis toward the wealthiest sector of the city. Conversely, the group without amenities contains AoIs located in the periphery of the city, in a radial distribution around the spine. This description matches the Latin American city model proposed by Ford (1996), which defines a central spine starting from a central business district and ending in areas with shopping centres; surrounding this spine is the wealthy residential sector. In the model, the city has prominent rings of periphery and radial areas of ‘disamenities’.

Our results are consistent with the literature and provide new insights regarding CBDs and areas that could be further developed according to the saliency of digital activities.

Conclusion

We proposed a methodological framework to identify urban Areas of Interest (AoI) by capturing the extent of complementarity between material and digital activities. The proposed framework is novel as it uses amenities and metadata of digital signals from mobile phone records to capture digital activities. Applying our methodology to data from Santiago, Chile, we identified distinctive AoIs characterised by the local availability and diversity of amenities and digital activities and how they interact with local present population levels. The resulting evidence reveals inequalities in the availability and use of material infrastructure and digital technology which could serve to identify potential areas of urban policy interventions, targeting deficient local urban amenity infrastructure that cannot be currently complemented by the use of digital technology available locally.

As we observed, there are AoIs within the city that may be undetected when digital activities are not considered, as they have a high present population but do not have enough amenities within them. These areas could be developed further and provide new work and service destinations, informing future plans in moving toward a polycentric city with evenly-distributed locations regarding socio-economic status and new working patterns, such as working from home.

To conclude, mobile phone data is a sensible source with enormous potential for urban planning, however, it poses several risks regarding privacy if not treated correctly. We have shown that aggregating the data at spatial (GSM towers) and conceptual (activities) levels generates meaningful insights with respect to urban behaviour, without disclosing individual information. We highlight that towers are the next level of analysis after individual data, and thus, we propose them as a unit of analysis for other studies. An alternative would be the assignment of towers to their corresponding administrative units; however, since tower distribution is not uniform throughout the city, this risks potential distortions in the results.

As future work, we propose to control for bias in the non-official data sources. Assessing this bias for mobile phone data is challenging because the data is aggregated and anonymised, and there are no formal methodologies to handle this particular situation. We also propose to study the stability of data. We used information-based criteria to select a period to study; however, it could be expected to also use a discipline-focused approach, which would require the data to be equally good in information terms for all periods of the day. In addition, a methodological line of work is to evaluate other approaches to local models, as well as to analyse the temporal coherence and dynamics of the Areas of Interest. Finally, these results should be validated with domain experts. All in all, cities are growing and changing faster than expected, and data-driven methods may help to improve quality of life in cities.

Acknowledgements

We thank GPA Movistar for facilitating the data for this study, and Loreto Bravo and Patricio Reyes for insightful discussion.

Declaration of conflicting interests

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

The author(s) disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: Daniela Opitz was partially funded by ANID #PAI77190057. Parts of the data are ©OpenStreetMap contributors.

Availability of data

The Telefónica Movistar mobile phone records have been obtained directly from the mobile phone operator through an agreement between the Data Science Institute from Universidad del Desarrollo and Telefónica

R&D. This mobile phone operator retains ownership of these data and imposes standard provisions to their sharing and access which guarantee privacy. Anonymised datasets are available from Telefónica R&D Chile for researchers who meet the criteria for access to confidential data. The other datasets used in this study are publicly available.

ORCID iD

Eduardo Graells-Garrido https://orcid.org/0000-0003-0722-5969

Supplemental Material

Supplemental material for this article is available online.

References

Birant D and Kut A (2007) ST-DBSCAN: an algorithm for clustering spatial-temporal data. *Data & Knowledge Engineering* 60(1): 208–221.

Blondel VD, Decuyper A and Krings G (2015) A survey of results on mobile phone datasets analysis. *EPJ Data Science* 4(1): 1.

Brunsdon C, Fotheringham AS and Charlton ME (1996) Geographically weighted regression: a method for exploring spatial nonstationarity. *Geographical Analysis* 28(4): 281–298.

Cameron AC and Trivedi PK (1986) Econometric models based on count data. comparisons and applications of some estimators and tests. *Journal of Applied Econometrics* 1(1): 29–53.

Campello RJ, Moulavi D and Sander J (2013) Density-based clustering based on hierarchical density estimates. In: *Pacific-Asia conference on Knowledge Discovery and Data Mining, Gold Coast, Australia*. Berlin, Germany: Springer, pp. 160–172.

Cavanaugh JE (1997) Unifying the derivations for the Akaike and corrected Akaike information criteria. *Statistics & Probability Letters* 33(2): 201–208.

Chen M, Arribas-Bel D and Singleton A (2019) Understanding the dynamics of urban areas of interest through volunteered geographic information. *Journal of Geographical Systems* 21(1): 89–109.

Cox T and Hurtubia R (2021) Latent segmentation of urban space through residential location choice. *Networks and Spatial Economics* 21(1): 199–228.

Cranshaw J, Schwartz R, Hong J, et al. (2012) The Livehoods project: Utilizing social media to understand the dynamics of a city. In: Sixth International AAAI Conference on Weblogs and Social Media, Dublin, Ireland, June 4–7 2012. Menlo Park, CA: AAAI Press.

Crooks A, Pfoser D, Jenkins A, et al. (2015) Crowdsourcing urban form and function. *International Journal of Geographical Information Science* 29(5): 720–741.

Ester M, Kriegel HP, Sander J, et al. (1996) A density-based algorithm for discovering clusters in large spatial databases with noise. In: Second International Conference on Knowledge Discovery and Data Mining, Portland, USA, August 2–4 1996, pp. 226–231. Menlo Park, CA: AAAI Press.

Fotheringham S, Brunsdon C and Charlton M (2002) *Geographically Weighted Regression*. Hoboken, NJ: John Wiley & Sons.

Fors AC and Wiberg M (2010) Digital materiality as imprints and landmarks: the case of northern lights. *The International Review of Information Ethics* 12: 5–10.

Ford LR (1996) A new and improved model of Latin American city structure. *Geographical Review* 86(3): 437–440.

Fuentes L, Truffello R and Flores M (2022) Impact of land use diversity on daytime social segregation patterns in Santiago de Chile. *Buildings* 12(2): 149.

Graells-Garrido E, Caro D, Miranda O, et al. (2018) The WWW (and an H) of mobile application usage in the city: the what, where, when, and how. In: Companion Proceedings of The Web Conference 2018, Lyon, France, April 23–27 2018, pp. 1221–1229. New York, NY: ACM.

Graells-Garrido E, Caro D and Parra D (2018) Inferring modes of transportation using mobile phone data. *EPJ Data Science* 7(1): 49.

Graells-Garrido E, Ferres L, Caro D, et al. (2017) The effect of Pokémon Go on the pulse of the city: a natural experiment. *EPJ Data Science* 6(1): 23.

Graham M, Zook M and Boulton A (2013) Augmented reality in urban places: contested content and the duplicity of code. *Transactions of the Instit*ute of British Geographers 38(3): 464–479.

Haklay M (2010) How good is volunteered geographical information? A comparative study of OpenStreetMap and Ordnance Survey datasets. *Environment and Planning B: Planning and Design* 37(4): 682–703.

Herold M, Couclelis H and Clarke KC (2005) The role of spatial metrics in the analysis and modeling of urban land use change. *Computers, Environment and Urban Systems* 29(4): 369–399.

Hu Y, Gao S, Janowicz K, et al. (2015) Extracting and understanding urban areas of interest using geotagged photos. *Computers, Environment and Urban Systems* 54: 240–254.

Jain J and Lyons G (2008) The gift of travel time. *Journal of Transport Geography* 16(2): 81–89.

Jiang S, Alves A, Rodrigues F, et al (2015) Mining point-of-interest data from social networks for urban land use classification and disaggregation. *Computers, Environment and Urban Systems* 53: 36–46.

Limtanakool N, Dijst M and Schwanen T (2007) A theoretical framework and methodology for characterising national urban systems on the basis of flows of people: empirical evidence for France and Germany. *Urban Studies* 44(11): 2123–2145.

Liu Y, Singleton A, Arribas-bel D, et al. (2021) Identifying and understanding road-constrained areas of interest (AoIs) through spatiotemporal taxi GPS data: a case study in New York City. *Computers, Environment and Urban Systems* 86: 101592.

McInnes L, Healy J and Astels S (2017) hdbscan: Hierarchical density based clustering. *Journal of Open Source Software* 2(11): 205.

Nagenborg M, Anders A, Klamt M, et al. (2010) On ICT & the City. *The International Review of Information Ethics* 12: 2–4.

Oshan TM, Li Z, Kang W, et al. (2019) mgwr: A Python implementation of multiscale geographically weighted regression for investigating process spatial heterogeneity and scale. *ISPRS International Journal of Geo-Information* 8(6): 269.

Pas EI (1982) Analytically derived classifications of daily travel-activity behavior: Description, evaluation, and interpretation. *Transportation Research Record*: 879.

Reichman S (1976) Travel adjustments and life styles: a behavioral approach. *Behavioral Travel-Demand Models* 1976: 143–152.

Ren F and Kwan MP (2009) The impact of the Internet on human activity-travel patterns: analysis of gender differences using multi-group structural equation models. *Journal of Transport Geography* 17(6): 440–450.

Ren Y, Tomko M, Salim FD, et al. (2018) Understanding the predictability of user demographics from cyber-physical-social behaviours in indoor retail spaces. *EPJ Data Science* 7(1): 1.

Ren Y, Xia T, Li Y, et al. (2019) Predicting socio-economic levels of urban regions via offine and online indicators. *Plos One* 14(7): e0219058.

Rodríguez-Vignoli JR and Rowe F (2017) ¿Contribuye la migración interna a reducir la segregación residencial? El caso de Santiago de Chile 1977-2002. *Revista Latinoamericana de Población* 11(21): 7–45.

Sotomayor-Gomez B and Samaniego H (2020) City limits in the age of smartphones and urban scaling. *Computers, Environment and Urban Systems* 79: 101423.

Suazo-Vecino G, Muñoz JC and Fuentes Arce L (2019) The displacement of Santiago de Chile's downtown during 1990-2015: travel time effects on eradicated population. *Sustainability* 12(1): 289.

Warf B (2001) Segueways into cyberspace: multiple geographies of the digital divide. *Environment and Planning B: Planning and Design* 28(1): 3–19.

Wu L, Zhi Y, Sui Z, et al. (2014) Intra-urban human mobility and activity transition: evidence from social media check-in data. *Plos One* 9(5): e97010.

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

Yuan G, Chen Y, Sun L, et al. (2020) Recognition of functional areas based on call detail records and point of interest data. *Journal of Advanced Transportation* 2020: 1–16.

Yuan J, Zheng Y and Xie X (2012) Discovering regions of different functions in a city using human mobility and POIs. In: Proceedings of the 18th ACM SIGKDD international conference on Knowledge discovery and data mining, Beijing, China, August 12–16 2012. pp. 186–194. New York, NY: ACM.

Zhang L and Pfoser D (2019) Using OpenStreetMap point-of-interest data to model urban change—A feasibility study. *Plos One* 14(2): e0212606.

Zook M, Graham M and Boulton A (2015) Crowd-sourced augmented realities: Social media and the power of digital representation. In: *Mediated Geographies and Geographies of Media*. Dordrecht, the Netherlands: Springer, 223–240.