Los Angeles as a digital place: The geographies of user-generated content

Andrea Ballatore and Stefano De Sabbata

Transactions in GIS, 2019
Author copy

Abstract Online representations of places are becoming pivotal in informing our understanding of urban life. Content production on online platforms is grounded in the geography of their users and their digital infrastructure. These constraints shape *place representation*, that is the amount, quality, and type of digital information available in a geographic area. In this article, we study the place representation of user-generated content (UGC) in Los Angeles County, relating the spatial distribution of the data to its geo-demographic context. Adopting a comparative and multiplatform approach, this quantitative analysis investigates the spatial relationship between four diverse UGC datasets and their context at the census tract level (about 685,000 geo-located tweets, 9,700 Wikipedia pages, 4M OSM objects, and 180,000 Foursquare venues). The context includes the ethnicity, age, income, education, and deprivation of residents, as well as public infrastructure. An exploratory spatial analysis and regression-based models indicate that the four UGC platforms possess distinct geographies of place representation. To a moderate extent, the presence of Twitter, OpenStreetMap, and Foursquare data is influenced by population density, ethnicity, education, and income. However, each platform responds to different socio-economic factors and clusters emerge in disparate hotspots. Unexpectedly, Twitter data tends to be located in more dense, deprived areas, and the geography of Wikipedia appears peculiar and harder to explain. These trends are compared with previous findings for the area of Greater London.

Key words: Information geography; crowdsourcing; volunteered geographic information; user-generated content; Los Angeles; geo-demography

Andrea Ballatore (✉)
Department of Geography, Birkbeck, University of London, London, UK
e-mail: a.ballatore@bbk.ac.uk

Stefano De Sabbata
School of Geography, Geology, and the Environment, University of Leicester, UK
e-mail: s.desabbata@le.ac.uk

1 Introduction

Online platforms enable users to produce an unprecedented variety of geo-referenced information, describing places and their locales and activities. Projects and services like Wikipedia, OpenStreetMap, and Twitter produce valuable data that is deployed in the social sciences to investigate geographic phenomena (e.g., Wang et al., 2018; Zagheni et al., 2014). The rapid expansion in information production occurred in the last decade includes phenomena that have been described as spatial crowdsourcing, volunteered geographic information (VGI), spatial social media, as well as big data (See et al., 2016). In this article, we use the term user-generated content (UGC) as it appropriately captures the diversity of spatial content produced by web users.

While such UGC sources are still popular, most research relies on isolated platforms, overlooking the geo-demographic context in which the production process takes place. Understanding how local geographical settings influence UGC production is critical to support its appropriate usage in research. Studies in Internet geography and data science have tended to focus at the global level, for example studying digital divides between countries (Graham et al., 2015a). However, a higher proportion of UGC is produced within cities, rather than in rural areas (Hecht and Stephens, 2014), and analyses at a finer granularity are needed to account for the intrinsic heterogeneity of this kind of data.

Among the many facets of informational geographies that can be operationalised, we consider *place representation* as the overall information available in a target geographic area for a given data source (Ballatore and De Sabbata, 2018). The representation of place is a central problem in GIScience (Purves et al., 2019), and UGC represents an important source of information *about* places. Data sources vary dramatically with respect to thematic, spatial, and temporal coverage, including, for example, geo-tagged photographs and articles about local monuments. Quantifying place representation is also useful to note areas of unexpected data scarcity, which is important when studying spatial phenomena (Robinson, 2019). To study place with UGC, it is crucial to know its intrinsic spatial structure and the factors that shape it. For this reason, comparing results across geographical locations, as we do in this article between Los Angeles and London, is important to establish to what extent factors tend to be global or place-specific.

This article contributes to our understanding of the informational dimension of place by comparing spatial UGC sources to their socio-economic context that constrain the production process. In a cross-platform comparison, we analyse four UGC sources, including geo-located tweets, Wikipedia articles, OpenStreetMap objects, and Foursquare venues. These data sources are extremely popular among researchers,[1] are easily accessible, and have specific geographies and biases that tend to be overlooked. As a study area, we consider 2,585 spatial units in the Los Angeles County, each including about 3,700 people on average. This area hosts about 10.1M

[1] As a proxy for research popularity of these platforms, we observed the search results on the Web of Science search engine, which provides more conservative results than Google Scholar. As of August 13th, 2019, Twitter has been mentioned in 25,120 articles, Wikipedia in 8,700, OSM in 1,203, and Foursquare in 1,019.

residents, and is a large, diverse, and complex area, and has been providing a rich context for the study of urban inequalities for generations of social scientists (Davis, 2006). The UGC sources are spatially related and quantitatively compared with authoritative sources of place information, such as the US census variables, deprivation index, and Federal Poverty statistics. The infrastructural context is included by considering the points of interest from the County's Location Management System. This study approaches the following research questions with quantitative methods:

RQ1 What are the similarities and differences between the geographies of the UGC sources (Twitter, Wikipedia, and Foursquare, and OpenStreetMap)?
RQ2 What is the relationship between place representation in UGC and their socio-economic and infrastructural context at the spatial level?
RQ3 Is there a relationship between socio-economic deprivation and place representation in UGC?
RQ4 What are the similarities and differences between the information geographies of this case study with other known areas (Los Angeles and London)?

Building on our prior research on Greater London (Ballatore and De Sabbata, 2018), this study contributes to the knowledge of the geographies of UGC, providing findings about what areas are over- and under-represented in these popular data sources. For the sake of replicability, in accordance with data source licenses, the data used in the study is made available online.[2] This analysis of digital divides and "informational ghettos" (Shaw and Graham, 2017, p. 4) can identify data-poor areas that can persist even in global cities like Los Angeles. It is important to note that this study does not aim at studying either social or natural phenomena using UGC as a data source, but analyses the data production itself. In the context of research about places, UGC can provide signals of other phenomena (e.g., geo-located pictures as a signal of public interest in a location), but its production can also be seen as distinct phenomenon that occur in a constrained context. The stakeholders of this place representation study are researchers and users of UGC, who can benefit from new knowledge of UGC's geographical structure and representativeness, evaluating its fitness-for-purpose. Mapping place representation can also inform producers and platform owners, who can embed more equal place representation in the platforms, particularly in critical applications (Schnebele and Cervone, 2013).

The remainder of this paper is organised as follows. Section 2 summarises related work in the areas of UGC and its socio-economic dimensions. The datasets used to characterise the informational geography of the study area are described in Section 3, including UGC and authoritative sources. Section 4 describes the analysis of Twitter, Wikipedia, Foursquare, and OpenStreetMap data located in Los Angeles County, starting with an exploratory analysis, a spatial auto-correlation analysis, and then continuing with a set of regression models. Section 5 summarises and discusses the study findings, while conclusions and future work directions are drawn in Section 6.

[2] `http://anonymous`

2 Place and user-generated content

Place is a central notion in human geography, and poses challenges to the traditional foundations of GIScience (Ballatore, 2016). To most geographers, place is mainly an experiential and phenomenological concept, making it hard to be representable in reference systems that were conceived to master euclidean or geodesic space. According to Agnew (2011), "places are specific time-space configurations made up of the intersection of many encounters between 'actants' (people and things)" (p. 325). The notion of *platial* information has been coined to refer to more qualitative descriptions of places, contrasted with geometrical and topological representations (Ballatore and Jokar Arsanjani, 2018). In this sense, as recently expressed by Purves et al. (2019), the representation of place needs not novel models, but appropriate combinations of existing ones, such as objects, networks, and fields. Once a precise epistemic cut is chosen, configurations of people and things can be represented and processed appropriately, capturing these dynamics in data sources.

2.1 Place and data

Disciplines like urban planning, human geography, and sociology aim at understanding how places function in terms of the human activities, from a variety of complementary perspectives. To study places, data can be collected by eliciting responses from people, for example in the case of traditional social scientific methods, or with automated, ambient approaches. The emergence of UGC in the mid 2000s has provided novel digital routes to reach place from space, ranging from photo sharing, micro-messaging, to collaborative cartography (Kitchin and McArdle, 2016).

After a decade of research, much is known about UGC. Some of these sources aim at a homogeneous spatial coverage, while others embed some locational information, but without a spatial objective (Antoniou et al., 2010). Contributors to Wikipedia and OpenStreetMap, for example, purposely aim at including all cities in the world in a systematic way, while geo-located tweets and Instagram photos are the by-product of a mediated communication process between users, which can then be used to study spatial collective behaviours and human dynamics.

Although these sources of digital information enable ways to understand places in their changing characteristics and interactions (Redi et al., 2018), their intrinsic limitations should not be overlooked. The geography of these datasets can be explored in terms of access (presence of digital infrastructure in a place), participation (who is contributing locally or non-locally), and representation (how much information is generated about a place) (Graham et al., 2015a). Much research focusses on the participation and production geography, studying contributor communities. Different UGC platforms attract different user groups in terms of gender, age, income, education, area of residence, and motivations, constraining the characteristics of the data (Acheson et al., 2017). Detailed market research by the Pew Research Center traces the demographic composition of online platforms over time (Smith and An-

derson, 2018). Exceptions can be found (Ballatore and Jokar Arsanjani, 2018), but contributor communities tend to be skewed towards wealthy, educated, white, and male users in the Global North (Crampton et al., 2013).

2.2 UGC platforms

This study includes four UGC platforms, ranging from privately-owned social media (Twitter and Foursquare) to commons-based collective editing projects (OSM and Wikipedia). Considering the heterogeneity and dynamic nature of UGC, it is not possible to obtain a representative probability sample of all existing UGC sources. For this reason, this selection is unavoidably a non-probability sample, which combines convenience sampling (i.e. sources that are accessible) and purposive sampling (diverse sources that are of interest to broad communities of researchers).

These four platforms cover different ontological dimensions of place representation. OSM and Wikipedia aim at systematically mapping stable geographical objects. Foursquare venues also aim at completeness, although limited to a narrower set of objects, e.g. cafes and restaurants, while check-ins represent user presence at the venues (and not included in the study). Finally, the geography of geo-located tweets differs radically from that of the other sources, as they capture the presence and activity of users in the target area. The content of tweets may refer to objects in the area or not, and we consider this aspect outside of the scope of the study—for this aspect of Twitter data, see Hahmann et al. (2014).

At the time of writing, these data sources are freely accessible either through data dumps or through proprietary APIs, with different constraints. Wikipedia and OSM data can be accessed in its entirety, while Twitter allows to access a sample from the stream of current data (access to historical data is possible, but not free). The Foursquare Places API has a free option with a limit on the number of queries per second. The terms of use of each source vary, and is generally forbidden either to re-publish the data or derive new datasets from them. It is important to note that this constraint limits the replicability of the study, but not its reproducibility (Ostermann and Granell, 2017). Data from other popular platforms, such as Instagram or Facebook, were not included in the study as they are not currently accessible to researchers.

As of 2018, Twitter was used by 21% of US adults, particularly in the 18-29 age group (36% of online adults), and among college educated users (29% of online adults) (Greenwood et al., 2018). As Blank and Lutz (2017) forcefully argued, Twitter users are actually not statistically representative of any particular population, but their large number (about 330M users) and the low access cost to samples of the data attracts researchers. From a spatial perspective, 0.85% of tweets are tagged with geo-coordinates, which amounts to roughly 4M tweets a day, produced by a population only marginally different to the overall platform population (Sloan and Morgan, 2015). It must be noted that the use of geo-coordinates is in decline, replaced by a

vaguer *place* field (e.g. "San Francisco" instead of a precise geo-location), mainly in relation to user privacy.

Geo-located tweets have been widely used in spatial research, tracing urban functional areas (Lansley and Longley, 2016), geo-demographic groups and connectness (Longley and Adnan, 2016), local activities and interactions (Steiger et al., 2015), urban mobility patterns (Wang et al., 2018), and potential areas of urban regeneration (Martí et al., 2019). Using an approach similar to our study, Li et al. (2013) explore tweets and Flickr photos spatio-spatial distribution in California at the county level, showing that photos are denser near natural attractions and that tweets tend to originate from areas with residents who are more educated and more affluent than average. This study adopts a problematic methodology, modelling the relationships between variables at a very large scale in an unclear way. By contrast, our study adopts a much finer spatial resolution (census tracts), and includes more variables and platforms.

Wikipedia is the second UGC source in this study. Despite a prolonged decline in the number of contributors, it is one of the top ten most visited websites worldwide, and hosts 5.8M articles in English, edited by about 130,000 monthly active editors, of which 84% are male.[3] From a geographical perspective, in 2013, about 730,000 articles in English were associated with a geo-location. As Graham et al. (2015b) pointed out, most editing occurs in the Global North, also for articles about places in the Global South, with a deep bias towards contributions from Western European and North American editors.

The third UGC source is OpenStreetMap (OSM), one of the most successful collaborative mapping platforms. 1m unique users contributed to its global vector dataset, which comprises 5.1b points and 570m polygons and polylines.[4] OSM data found its way as a spatial data source for many non-profit and commercial services, providing street networks and points of interest to countless research projects (e.g. Bright et al., 2018).[5] Although OSM purports to map the world systematically, its coverage is highly uneven. In their study in Greater London, Mashhadi et al. (2013) observed contextual factors that correlate with coverage, finding a relationship with population density and distance from the centre, with heteroscedasticity between Inner and Outer London. In a case study in Germany, the vitality of OSM mapping correlates positively with relatively high population density, high income, high rate of overnight stays, and high number of foreigners (Arsanjani and Bakillah, 2015).

Finally, we include Foursquare data. Launched in 2009 as a location-based service, Foursquare allows users to "check-in" into venues, such as restaurants, cafes, and bars, enabling a novel mode of collective location-sharing. As of 2018, the platform claims a global reach of 3 billion visits per month, 105 million mapped venues,

---

[3] https://web.archive.org/web/20180505163037/https://www.wired.com/2016/01/at-15-wikipedia-is-finally-finding-its-way-to-the-truth

[4] https://web.archive.org/web/20190221113106/https://blog.openstreetmap.org/2018/03/18/1-million-map-contributors

[5] https://web.archive.org/web/20171206130620/http://wiki.openstreetmap.org/wiki/Research

and 25 million active users.[6] From 2014, the company transitioned from a consumer app to an enterprise platform for location intelligence and analytics. Foursquare behaviour has been studied as source of human mobility patterns (Noulas et al., 2011). A user study showed the diverse social and cultural motivations of users who share their location by checking into venues (Lindqvist et al., 2011).

Much of the aforementioned UGC research aims at understanding the characteristics, motivation, and behaviour of data producers, while place representation has been addressed only marginally. To the best of our knowledge, no study has comparatively explored the properties of diverse UGC datasets at this spatial resolution, and their relationship with the socio-economic and infrastructural context in which the data is produced. In addition, the relationship between Twitter, Wikipedia, Foursquare, and OSM from a spatial perspective has not been directly studied before.

2.3 Studying Los Angeles

This study focuses on the Los Angeles County as a case study. The LA agglomeration provide a suitable context to explore the relationship between UGC and geographic context at the urban scale for several reasons. LA has fascinated generations of human geographers and urban planners, who explored it as a unique, global, complex, fragmented, and extraordinarily heterogeneous city, with some even selecting it as the "paradigmatic American metropolis" (Curry and Kenney, 1999). The Central City area is the historical core of LA, from which urbanisation extends from 100 km in all directions. Auto-driven urbanisation has resulted in a compact inner city with minority population, surrounded by a low-density suburban sprawl dominated by more affluent middle-classes. Central Los Angeles has the largest concentration of offices and jobs in Southern California. In recent decades, LA developed the unusual urban form of *city region* or *regional city*, in which it is hard to delineate urban and suburban areas on a very large spatial extension (Soja, 2014).

Since the 1980s, Soja (2014) argues, once marginal LA has become "the focal point for the development of new and widely recognised urban theories" (p. 13). According to him, the development of LA anticipated trends seen globally, including suburbanisation, deindustrialisation, and new sprawling peripheries becoming more dominant than decaying inner cities. *City of Quartz* (Davis, 2006) proved to be one of the most influential texts in urban studies, discussing LA's persistent ethnic tensions, surveillance practices, and socio-economic inequalities. Our quantitative study of the digital geography of LA can be linked to this tradition of urban studies. Furthermore, LA provides a useful case study for comparison to our prior research on UGC in Greater London, similar in terms of population and area (Ballatore and De Sabbata, 2018). The next section summarises the geographic area and datasets included in the study.

[6] https://web.archive.org/web/20180220183737/https://techcrunch.com/2018/01/19/foursquare-is-finally-proving-its-dollar-value

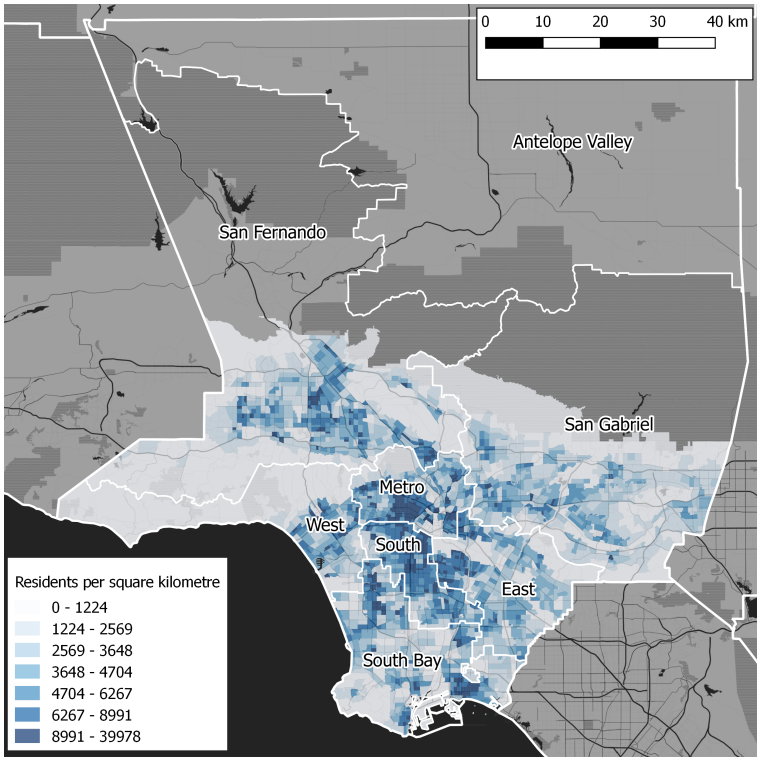


Fig. 1 The study area includes the Southern part of Los Angeles County, excluding the islands, grouped into 2,585 spatial units. The white boundaries are the Service Planning Areas (SPA). The population density is shown only for the study area and is grouped using 7 quantiles. Sources: 2016 population estimates by the American Community Survey and Tableau.

3 Mapping UGC in Los Angeles

3.1 Study area and spatial units

The study area is located in the Los Angeles County. As of 2017,[7] this area had a population of 10.16m people over a territory of 10,510 km^2. Figure 1 shows the extension and population density of the study area.[8] The population density is derived

[7] https://web.archive.org/web/20181229113621/https://www.census.gov/quickfacts/fact/table/losangelescountycalifornia,ca/PST045217

[8] Note that all maps in this article are projected with UTM zone 11N in metres (NAD83 datum, EPSG:26911).

from 2016 estimates by the American Community Survey (ACS).[9] For reference, all maps include the 2012 County's Service Planning Areas (SPA).[10]

In 2016 estimates, the County's population includes Hispanic (49%), non-Hispanic White (28%), African-American (8%), and Asian (14%), and other (1%) residents. Most of the population is clustered in Southern part of the County, in Los Angeles (3.8m people), followed by Long Beach, Santa Clarita, and Glendale (each of which has a population larger than 200,000 people). The population density remains relatively high East of Los Angeles, while it is extremely low North of the city, with the exception of sparse urban areas. The population density is also low in the coastal area West of Los Angeles that includes Malibu. The County comprises two large islands: Santa Catalina Island (4,050 people) and San Clemente Island (a naval base officially uninhabited).

As this study focuses on the relationship between UGC and the geo-demographic and infrastructural context, we excluded areas in the Northern part of the County (all tracts North of latitude 34.36, including Santa Clarita, Lancaster, and Palmdale), as well as the two islands, which were predominantly low-density both in terms of population and data (see Figure 1). As spatial units, we adopted those from Population and Poverty Estimates dataset (2016), which contained core demographic variables. The units are 2010 Census tracts, split into sub-units when they overlap multiple Countywide Statistical Areas (CSA). For this reason, the study area includes 2,585 tract/CSA units, corresponding to 2,199 tracts. For simplicity, we will henceforth refer to the study spatial units as "tracts", even though some of them are smaller.

The study area can be partitioned into 6,111 census block groups, sub-units of tracts. The tracts include a median of 3,796 residents (between 0 and 9,492 people), while census block groups include between 600 and 3,000 people.[11] On average, each tract in the County contains 2.3 block groups. The study area corresponds to 92% of the administrative units and 93% of the County population (9.5m people). As detailed below, several datasets were collected from authoritative sources and UGC. Table 1 provides summary statistics of all variables included in the study, using tracts as unit of analysis. As is possible to notice, the distributions of variables vary considerably. Almost all variables are right-skewed, which is expected as these geographical phenomena tend to have large outliers in the right part of the distribution that boost the mean (Westerholt et al., 2016). The kurtosis is positive for all variables except the residents count, indicating a dominance of leptokurtic distributions (i.e. clustered around the mean). The distinct distribution of the resident count can be explained with the fact that the spatial units (tracts) are designed to obtain a distribution as uniform as possible.

The four UGC sources are diverse in terms of their thematic scope (micro-messages, notable spatial objects, or all spatial objects) and have different spatial distributions. To compare these data sources, we start by observing their densities

[9] https://www.census.gov/acs/www/data/data-tables-and-tools/data-profiles/2016

[10] https://egis3.lacounty.gov/dataportal/2012/03/01/service-planning-areas-spa-2012

[11] https://www.census.gov/geo/reference/gtc/gtc_bg.html

Table 1 Descriptive statistics for the study area in the Southern part of Los Angeles County, over 2,585 tracts, including skewness and kurtosis. Values above 1,000 are rounded. Abbreviations: *sqkm* per square km; *1kres* per 1,000 residents; *lmspoi* points of interests from Los Angeles County. Sources: Population and Poverty Estimates (2016), LMS points of interest, Twitter, Wikipedia, OpenStreetMap, and Foursquare.

Variable (2,585 tracts)	min	median	max	mean	stdev	skewness	kurtosis
Population							
residents	0	3,832	9,492	3,692	1,901	−0.23	−0.41
residents sqkm	0	4,105	39,979	5,163	4,555	2.37	9.68
age 18 34 perc	0	7.18	33.33	7.17	3.18	0.86	5.11
age 35 54 perc	0	13.49	40	13.29	4.89	−0.17	2.56
Infrastructure							
lmspoi	0	13	1,248	22.38	40.98	13.47	332.42
lmspoi sqkm	0	13.97	636.94	23.91	39.41	7.32	79.67
lmspoi 1kres	0	3.51	10,947	13.33	226.18	46.86	2,257
User-generated content							
tweets	0	157	3,590	198.47	192.07	4.60	53.28
tweets sqkm	0	179.41	3,853	263.94	307.73	3.88	25.24
tweets 1kres	0	43.66	1,306	57.61	68.87	8.22	102.73
wikip	0	0	139	2.79	7.83	7.70	88.29
wikip sqkm	0	0	248.80	3.01	10.70	11.32	182.27
wikip 1kres	0	0	719.30	1.40	15.74	40.12	1,794
osm	2	1,164	5,890	1,267	842.26	0.72	0.68
osm sqkm	5	1,350	5,600	1,351	649.64	0.16	0.27
osm 1kres	23	334.35	10,891	390.25	522.91	14.76	263.80
foursq	0	32	848	47.67	53.90	3.51	27.10
foursq sqkm	0	36.26	500	44.27	37.62	3.42	21.38
foursq 1kres	0	8.88	1,830	17.49	67.41	19.41	445.98

per tract. Figure 2 shows the distribution of density of objects for the four datasets on a logarithmic scale. Twitter, Foursquare, and OSM exhibit some convergence at different spatial scales. These three datasets have approximately log-normal distributions, with clear peaks around to the median value. Twitter and Foursquare have some areas with 0 values, while OSM has a tail of sparse areas. By contrast, most areas do not have Wikipedia pages, and this results in an extreme skew towards 0 values. Excluding 0 values, Wikipedia also exhibits an approximate log-normal distribution, with a median of 2.3 pages per km^2.

When calculating the object densities per km^2, data exhibits some tails of extreme values, observable in Figure 2. These values are the result of extremely low-density units, where people per km^2 are lower than 22 (corresponding the bottom 5% percentile). For the purpose of visualisation in Figures 3 and 4, we deem these areas to be outliers and we removed them.

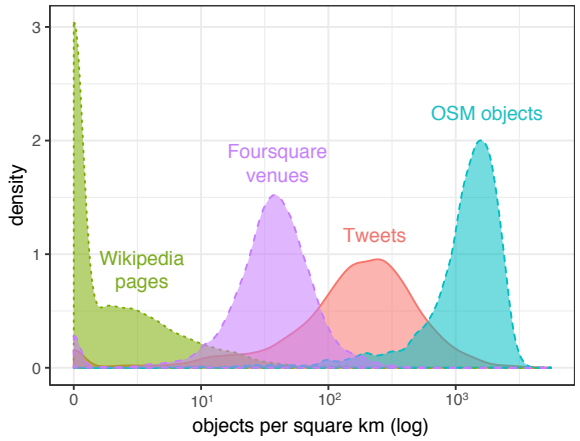


Fig. 2 Objects per km^2 over the 2,585 spatial units for the four UGC datasets (Wikipedia, OpenStreetMap, Twitter, and Foursquare). The density value is transformed with $\log_{10}(x+1)$.

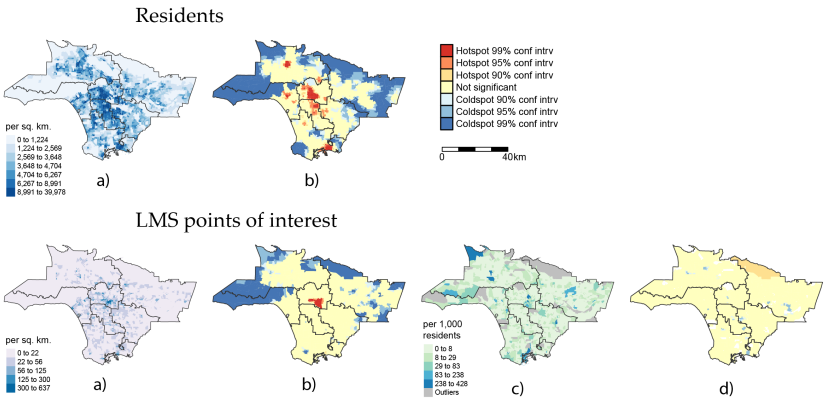


Fig. 3 Density of population and LMS points of interests (POIs), and LMS POIs per 1,000 residents over 2,585 tracts. Residents per km^2 (a) are grouped as quantiles, LMS POIs per km^2 (a) are grouped with Jenks, and LMS POIs per 1,000 residents (c) with kmeans. Hotspot analysis (b, c) performed with Getis-Ord local GI*. Sources: Population and Poverty Estimates (2016), LA County Location Management System (2016).

3.2 The demographic context

Ethnicity, age, income, and education

To obtain a detailed picture of the population characteristics of LA County, we relied on the Population and Poverty Estimates dataset (2016).[12] Figure 3 shows the spatial

[12] https://egis3.lacounty.gov/dataportal/2014/09/09/population-and-poverty-estimates

distribution of population, including an auto-correlation analysis based on Getis-Ord local Gi*.[13] This dataset includes socio-economic variables from the 2010 census, compounded with more recent estimates from the ACS. For each tract, counts are broken down by gender, age, and ethnicity. Federal poverty levels (at 100%, 130%, 133%, and 200%) and median household income are also available. As an indicator of education relevant to our study, we calculated the percentage of residents with higher education (any college degree).[14] The population density exhibits hotspots in Downtown LA, Long Beach, Brentwood, Westwood, and Panorama City.

Infrastructure

To describe the infrastructural context that supports UGC production, we retrieved the points of interest (POIs) from the County's Location Management System.[15] This dataset contains 64,982 POIs that are part of the infrastructure managed by the County as of 2016. The POIs are classified using 270 unique types in a category tree of 3 levels. For the purpose of this study, we consider all POIs using the highest level of the tree, which groups them in broad categories, such as *Communications*, *Education*, *Environment*, *Transportation*, and *Emergency*. The spatial distribution of these POIs is displayed in Figure 3. Note that each group of maps is binned using a different strategy (quantile, Jenks, and k-means) to improve their readability. The only hotspot is located in Downtown LA.

Deprivation index

Deprivation indices are aggregate indicators of social, economic, and cultural exclusion (Smith et al., 2015). To the best of our knowledge, no official deprivation index is available for the US, but public health researchers devised the Area Deprivation Index (ADI) that provides a comparable indicator for 2013 (Kind et al., 2014). The ADI is calculated at the census block group level (6,425 units for the County), and it ranks small geographic areas from 1 (least disadvantaged) to 10 (most disadvantaged). The index is calculated at the national and at the state level. As our area of study is located in a single state, we consider the ADI that ranks census block groups within California.

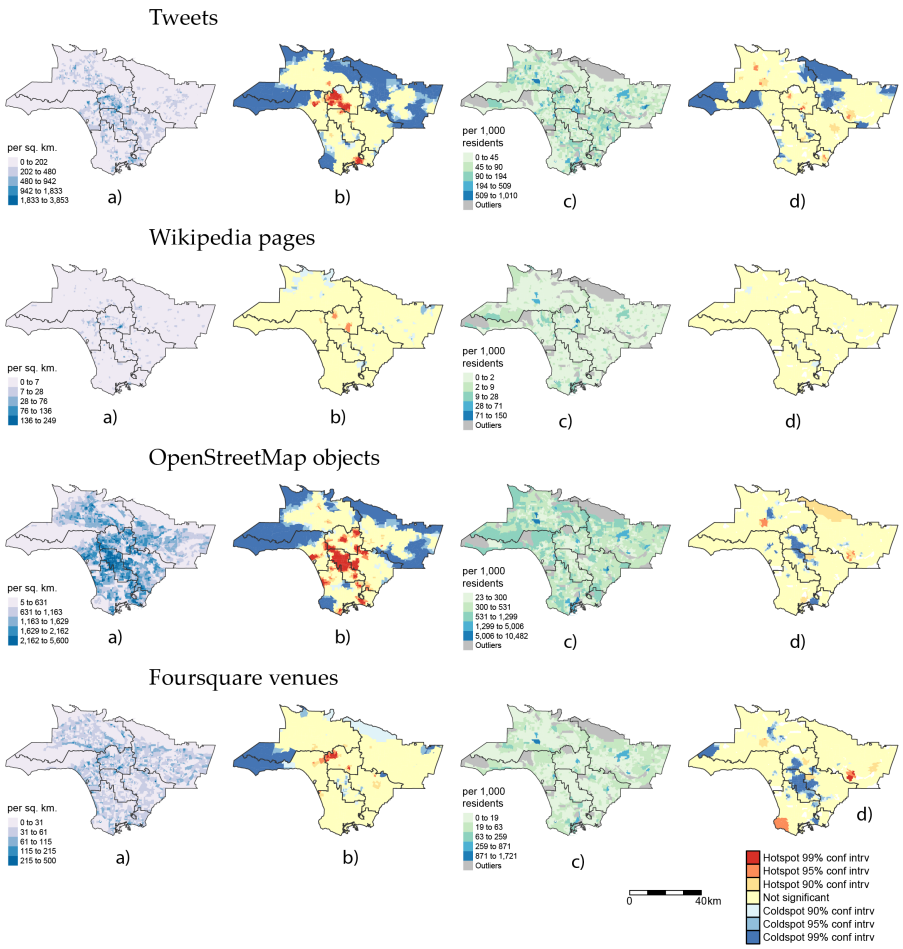


Fig. 4 Density of UGC datasets, displaying density per km^2 (a) and density per 1,000 residents (c), and corresponding hot and cold spots (b, d), over 2,585 tracts. To ensure readability, variables per km^2 (a) are grouped with quantiles and variables per 1,000 residents (c) with k-means. The hotspot analysis (b, d) is performed with Getis-Ord local GI*. Sources: Twitter, Wikipedia, OpenStreetMap, Foursquare, Population and Poverty Estimates dataset (2016).

3.3 Exploration of UGC datasets

Geo-located tweets

To answer *RQ1*, we explored the geographies of the four UGC platforms in the study area. This study considers a dataset of 684,793 geo-located tweets, collected through

[13] Implementation from R package `https://cran.r-project.org/package=spdep`

[14] Sum of census variables from B15003_019 to B15003_025.

[15] `https://egis3.lacounty.gov/dataportal/2016/01/14/locationspoints-of-interest-lms-data`

the public API from October 2014 to May 2015, produced by 85,200 unique users. Geo-located tweets are by-products of the communication process between users, unlike the other sources that aim at building coherent representations of target areas (e.g. Wikipedia and OpenStreetMap). It is worth noting that, overall, geo-located tweets are declining, plunged from just over 3.5% in 2012 to 1.5% in 2019, as users increasingly prefer to share vaguer place names to geo-coordinates.[16] After excluding the 102 bots, the final dataset contained 672,538 tweets by 85,098 unique users (98% of the initial dataset).

Figure 4 shows the distribution of tweets. In the study area, the density of tweets ranges from 0 to 3,852 per km^2 (179 median). Based on visual inspection, corroborated by hotspot analysis, the three largest contiguous blocks of high-density areas are Downtown Los Angeles, Long Beach, and Hollywood. The top decile by density (563 tweets per km^2 or more) tend to be located in down towns, universities, and shopping areas, including areas in Northridge, Downtown Burbank, Downtown Santa Monica, Playa del Rey, Loyola Marymount University, West Hollywood, University of Southern California, and Boyle Heights. When observing the number of tweets compared to the resident population, a small hotspot is evident in Downtown Los Angeles. Affluent areas of the city in the Rolling Hills gated community, Malibu, and around Pasadena are identified as coldspots when looking at tweet density per km^2. The two latter areas are also marked as coldspots when comparing tweets with the resident population.

Wikipedia pages

The second UGC source in this study consists of geo-referenced Wikipedia articles. Currently, Wikipedia only allows for geo-tags in the form of points, and even large geographical entities are geo-tagged to a point. For example, Union Station in Los Angeles is associated with a point that represent its centroid.[17] Using the resources available on Wikimedia Toolforge,[18] we retrieved 9,704 articles in November 2017, including notable buildings, parks, monuments, and organisation headquarters (such as United Artists). The dataset contains pages in multiple languages, most of which are in English (41%), but including French (7%), Spanish (7%), and other 75 languages. The relatively low prominence of content in Spanish appears surprising, considering that in 2013 29% of Californians spoke Spanish at home.[19] As shown in Figure 4, the Wikipedia dataset is very sparse, with a median of 0 pages, a 3rd quartile of 2.2 pages per km^2, and a maximum of 249 pages per km^2. Visual in-

[16] https://web.archive.org/web/20190306220513/https://www.forbes.com/sites/kalevleetaru/2019/03/06/the-era-of-precision-mapping-of-social-media-is-coming-to-an-end

[17] https://en.wikipedia.org/wiki/Union_Station_(Los_Angeles)

[18] https://tools.wmflabs.org

[19] https://web.archive.org/web/20190404073709/https://www.census.gov/data/tables/2013/demo/2009-2013-lang-tables.html

spection and hotspot analysis show that pages are clustered in Downtown LA and Hollywood, with other non-significant areas of high density.

OpenStreetMap objects

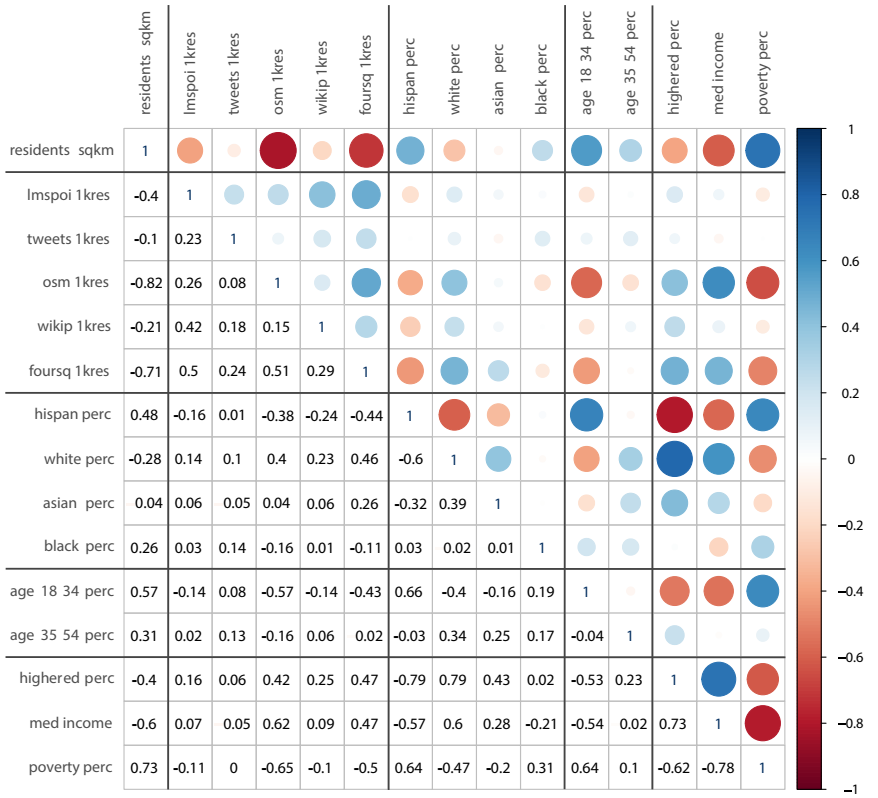
OpenStreetMap (OSM) is the most popular source of open vector data at the global level. For this study, we retrieved the full history OSM dataset for Southern California up to February 2019, describing 3,999,548 spatial objects.[20] Based on visual inspection of Figure 4, OSM data is denser and less clustered around fewer hubs than Twitter and Wikipedia, and enjoys an overall less skewed distribution, with a median of 1552.5 objects per km^2. This can be explained with the fact that OSM is a spatially explicit resource that aims at providing even coverage of the planet (Antoniou et al., 2010).

The top decile of the tracts in terms of OSM density (more than 2,200 objects per km^2) is located mostly in South Los Angeles, with hotspots also identified in Boyle Heights, East Compton, East Hollywood, Marina del Rey, and in Long Beach. Coldspots include again areas in Rolling Hills and Malibu. However, when the number of objects in OSM is weighted against the number of residents, a large coldspot is clearly visible spreading from Downtown Los Angeles to West Hollywood, along with smaller ones in Long Beach, North Hills East, and Glendale.

Foursquare venues

As our study does not include an explicit temporal dimension, we consider 178,814 Foursquare venues collected in 2013. This dataset includes venues classified in 9 top-level categories, which contain 422 sub-categories. The categories are residence (23%), shop & service (21%), professional & other places (18%), food (17%), outdoors & recreation (7%), travel & transport (4%), arts & entertainment (4%), nightlife spot (3%), and college & university (3%). The top decile by density (areas with more than 79 Foursquare venues per km^2) includes areas in Downtown Los Angeles, Long Beach, and Pasadena. Unlike other UGC, the dataset shows high density areas along major streets, such as Santa Monica Boulevard and Ventura Boulevard. As visible in Figure 4, a notable hotspot of venues stretches from Hollywood to Westwood, with a smaller hotspot in Downtown LA. However, when comparing the number of venues to the resident population, coldspots are visible in South Los Angeles, Long Beach, East Compton, the areas surrounding Koreatown, and North Hills East. The next section will proceed to study the relationships between these UGC datasets and their geo-demographic context.

[20] Data retrieved from `http://download.geofabrik.de/north-america/us/california/socal.html`. Aggregate statistics generated with PyOsmium `https://github.com/osmcode/pyosmium`



	residents sqkm	lmspoi 1kres	tweets 1kres	osm 1kres	wikip 1kres	foursq 1kres	hispan perc	white perc	asian perc	black perc	age 18 34 perc	age 35 54 perc	highered perc	med income	poverty perc
residents sqkm	1														
lmspoi 1kres	-0.4	1													
tweets 1kres	-0.1	0.23	1												
osm 1kres	-0.82	0.26	0.08	1											
wikip 1kres	-0.21	0.42	0.18	0.15	1										
foursq 1kres	-0.71	0.5	0.24	0.51	0.29	1									
hispan perc	0.48	-0.16	0.01	-0.38	-0.24	-0.44	1								
white perc	-0.28	0.14	0.1	0.4	0.23	0.46	-0.6	1							
asian perc	0.04	0.06	0.05	0.04	0.06	0.26	-0.32	0.39	1						
black perc	0.26	0.03	0.14	-0.16	0.01	-0.11	0.03	0.02	0.01	1					
age 18 34 perc	0.57	-0.14	0.08	-0.57	-0.14	-0.43	0.66	-0.4	-0.16	0.19	1				
age 35 54 perc	0.31	0.02	0.13	-0.16	0.06	0.02	-0.03	0.34	0.25	0.17	-0.04	1			
highered perc	-0.4	0.16	0.06	0.42	0.25	0.47	-0.79	0.79	0.43	0.02	-0.53	0.23	1		
med income	-0.6	0.07	0.05	0.62	0.09	0.47	-0.57	0.6	0.28	-0.21	-0.54	0.02	0.73	1	
poverty perc	0.73	-0.11	0	-0.65	-0.1	-0.5	0.64	-0.47	-0.2	0.31	0.64	0.1	-0.62	-0.78	1

Fig. 5 Spearman's correlation coefficients between demographic and UGC variables (2,585 tracts). Abbreviations: *sqkm* per square km; *1kres* per 1,000 residents. All coefficients greater than $\pm.03$ are significant at $p < .001$. Sources: Twitter, OpenStreetMap, Wikipedia, Foursquare, LA County's LMS, Population and Poverty Estimates (2016).

4 Modelling UGC geographies

4.1 Correlation analysis

As a first step towards modelling the relationship between UGC and their geo-demographic context, we calculated correlation coefficients between all relevant variables at the tract level. Considering the non-parametric distributions of the data, we calculated the ties-adjusted rank correlation coefficient (Spearman's ρ).

Correlations between ethnicity, age, education, and income

To interpret the distribution of UGC datasets, it is beneficial to observe the correlations between the core variables that structure LA's geo-demography, summarised in Figure 5. Population density shows a positive correlation with poverty (meaning the FPL 100%) (.73, all correlations at $p < .001$) and Hispanic residents (.48), and inverse correlations with higher education ($-.40$), median household income ($-.60$), and percentage of white residents ($-.28$). Median household income is correlated with the percentage of White residents (.6), inversely correlated with the percentage of African-Americans ($-.21$) and poverty ($-.78$). White and Asian residents tend to live in the same tracts (.39), while the presence of Black and Hispanic people is not mutually correlated. Higher education exhibits strong correlations with White (.79) and Hispanic ($-.79$) residents, weaker with Asians (.43), and no correlation with African-Americans.

These correlations are rooted in patterns of socio-economic inequality common in the US, particularly between White and Asian, who tend to have relatively high income and educational attainment, with the Hispanic and Black population having generally lower income and education (Fontenot et al., 2018). This is reflected in the tendency of dense inner city areas to be more Black and Hispanic, while less dense suburban areas are more White and Asian—as a result, denser tracts are generally more deprived. From an infrastructural viewpoint, the density of LMS POIs per 1,000 residents is inversely correlated with population density ($-.4$), confirming the intuition that more deprived areas tend to have relatively less infrastructure.

Correlations between UGC and geo-demographics

To investigate *RQ2*, Figure 5 shows the main UGC variables, including the number of tweets, Wikipedia pages, Foursquare venues, and OSM objects per 1,000 residents ("p.r." henceforth) in each tract. Twitter p.r. has overall weak relationships with demographic variables, and a slightly higher correlation with LMS POIs (.23). The lack of correlation with demographic variables indicates that the place representation is decoupled from the residence of participants. OSM p.r. exhibits a very strong correlation with population density ($-.82$), which indicates that more objects per residents appears high in low density areas. OSM p.r. is also related to income (.62), higher education (.42), percentage of White residents (.4), and poverty ($-.65$).

Wikipedia pages p.r. correlates positively with LMS POIs p.r. (.42) and negatively with Hispanic residents ($-.24$) and population density ($-.21$). Finally, Foursquare venues p.r. is negatively correlated with population density ($-.71$), with Hispanic residents ($-.44$), and age group 18-34 ($-.43$). Positive correlations are visible with LMS POIs (.5), White residents (.46), as well as income and education (both at .47).

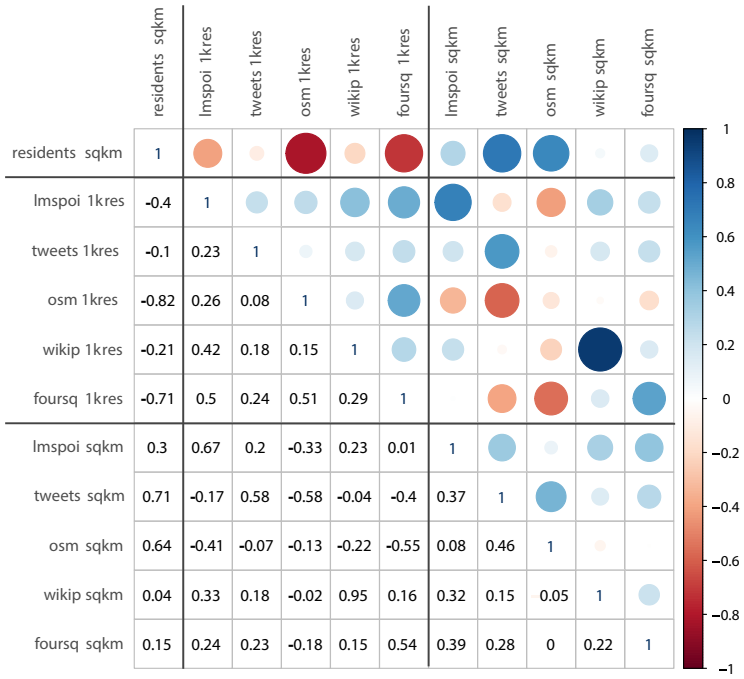


Fig. 6 Spearman's correlation coefficients between UGC variables, comparing density per km^2 and per 1,000 residents (2,585 tracts). Abbreviations: *sqkm* per square km; *1kres* per 1,000 residents. All coefficients greater than $\pm.03$ are significant at $p < .001$. Sources: Twitter, OpenStreetMap, Wikipedia, Foursquare, LA County's LMS, Population and Poverty Estimates (2016).

Cross-platform correlations

As shown in Figure 6, the UGC sources p.r. show positive correlations to one another, showing that, despite their differences, there is a general trend for which the same areas tend to have more data across platforms (*RQ1*). Foursquare and OSM is the highest (.51), followed by Foursquare and Wikipedia (.29), and Foursquare and Twitter (.24). The other relationships are all positive but weaker, in interval $[.08, .18]$. Figure 6 also summarises the inter-UGC correlations, comparing object densities across the four UGC sources and LMS POIs that indicate the presence of public infrastructure. Interestingly, the four datasets show variability in the correlation between their density with respect to area (objects per km^2) and to population (objects per 1,000 residents). Wikipedia has the strongest correlation (.95), followed by Twitter (.58) and Foursquare (.54). By contrast, OSM shows a weak inverse correlation ($-.13$), because its objects are more evenly spread, and thus less influenced by population density.

Correlations between UGC and deprivation

Subsequently, to investigate *RQ4*, we calculated correlations between 2013 deprivation index (ADI) and UGC variables over the 6,111 census block groups (CBGs), which are sub-units of census tracts. Because of the high number of ties, Kendall's τ was adopted as the correlation coefficient. The state deprivation rank (from 1, least disadvantaged, to 10, most disadvantaged) shows some positive correlation with the LMS POIs (.22, $p < .001$). The tweet density also correlates with deprivation (.28, $p < .001$), indicating that, to a moderate extent, more tweets tend to be generated in deprived areas. This is a counter-intuitive result that holds at the tract level too ($\rho = .56$ between tweet density and poverty), and stems from the fact that the geography of participation to the platform is different to its place representation: While Twitter users are more educated and affluent than average, tweets can be generated in either work or leisure locations, for example at offices, bars, restaurants, and public spaces, which are denser in inner city areas, such as Downtown LA and the Fashion District. By contrast, no correlation was found with Wikipedia density ($-.06$, $p < .001$), nor for Foursquare and OSM densities ($\tau \in [-.05, .05]$). For this reason, we did not pursue deprivation as an explanatory factor for the other UGC sources.

4.2 Linear regression models

Based on a correlation analysis outlined above, we constructed several explanatory linear regression models. It is important to note that these models do not offer causal explanations, but only correlations. Due to the skewed distribution of most of the variables, we normalised them with the inverse hyperbolic sine, abbreviated as $[ihs]$ (Burbidge et al., 1988). Interestingly, none of these models could be defined as robust, even when using normalised variables and filtering extreme values. The skeweness of the UGC variables and the effects of functional segregation in the city are such that all models were affected by non-normally distributed and heteroskedastic residuals. Nonetheless, in this section we briefly report on the most promising models that we developed, which are summarised in Table 2, as some insight can be extrapolated from them.

As suggested by the correlation coefficients, the number of residents, the area size, higher percentages of White and Asian residents, as well as higher levels of income and education seem to explain a relevant amount of the variance in the amount of content available in the platforms. Among the demographic variables, we excluded poverty and deprivation, due to their strong correlation with the combined effect of area size and number of residents (i.e. population density). The remainder of the variables used in the models also display some significant levels of correlation, but not to the point of being of concern, as all the average VIF scores are in the range $[1, 2.5]$.

Table 2 Linear regression models to explain variability of UGC variables in the study area (tract level). All F-statistic at $p < .001$. [ihs]: variable transformed with inverse hyperbolic sine.

No.	Dependent	Independent	Adj. R^2	F-statistic
1	tweets [ihs]	residents [ihs], area [ihs], high ed. (perc), income, asian (perc)	0.577	$F(5,2579) = 705.32$
2	osm [ihs]	residents [ihs], area [ihs], high ed. (perc), asian (perc)	0.774	$F(5,2580) = 2210.02$
3	foursq [ihs]	residents [ihs], area [ihs], high ed. (perc), income, asian (perc), white (perc)	0.474	$F(6,2578) = 389.71$
4	wikip [ihs]	residents [ihs], area [ihs], high ed. (perc), income, white (perc)	0.178	$F(5,2579) = 113.31$
5	tweets [ihs]	osm [ihs]	0.514	$F(1,2583) = 2739.03$
6	tweets [ihs]	foursq [ihs]	0.432	$F(1,2583) = 1966.84$
7	tweets [ihs]	wikip [ihs]	0.058	$F(1,2583) = 161.41$
8	osm [ihs]	foursq [ihs]	0.494	$F(1,2583) = 2526.96$
9	osm [ihs]	wikip [ihs]	0.061	$F(1,2583) = 168.22$
10	foursq [ihs]	wikip [ihs]	0.124	$F(1,2583) = 367.09$

The explanatory power is particularly strong for OSM (model 2), where much of the variance would be explained by those factor (Table 2). Similarly, the same factors can also explain about half of the variance in Twitter (model 1) and Foursquare (model 3). However, the same does not hold in case of Wikipedia (model 4), which obtained a much lower R^2. This finding is also corroborated by the models exploring the relationship between the UGC variables. Twitter, OSM, and Foursquare could reciprocally explain about half of the variability of the other (models 5, 6, 8), whereas their relationship with Wikipedia appears negligible (models 7, 9, 10).

It is particularly interesting to note how the beta coefficients presented in Table 3 suggest that income has a negative effect on the amount of UGC content in an area. This corroborates observations detailed above, as wealthy areas such as Rolling Hills and Malibu have been identified as coldspots for tweets, as well as for content in Wikipedia and, partially, for OpenStreetMap. These findings contradict observations made by Li et al. (2013) at the county level. The number of residents seems to be the strongest factor driving both the number of tweets and OSM content, whereas the size of the area appears as the main factor driving content in Fourquare and Wikipedia, and it partially impacts on OSM too.

The maps and Q-Q plots in Figure 7 display the distribution of the residuals of the first four models and again illustrate the clear similarities between Twitter, OSM and Foursquare, compared to Wikipedia. Further analysis of the distribution of the residuals revealed that all models exhibited spatially auto-correlated residuals. This clearly indicates that these linear models, although providing some explanatory power, are not currently robust, especially in dealing with their spatial component. As such, further analysis will be necessary to establish reliable models, for example using approaches such as geographically weighted regression (GWR).

Table 3 Beta coefficients for the linear regression models presented in Table 2, indicating the weights of each independent variable on the final estimate.

No.	Dependent	Independent					
		residents [ihs]	area [ihs]	high ed. (perc)	income	asian (perc)	white (perc)
1	tweets [ihs]	0.689	0.179	0.109	-0.299	-0.098	
2	osm [ihs]	0.708	0.446	-0.045		-0.053	
3	foursq [ihs]	0.363	0.498	0.123	-0.179	0.067	0.075
4	wikip [ihs]	0.053	0.358	0.193	-0.278		0.130

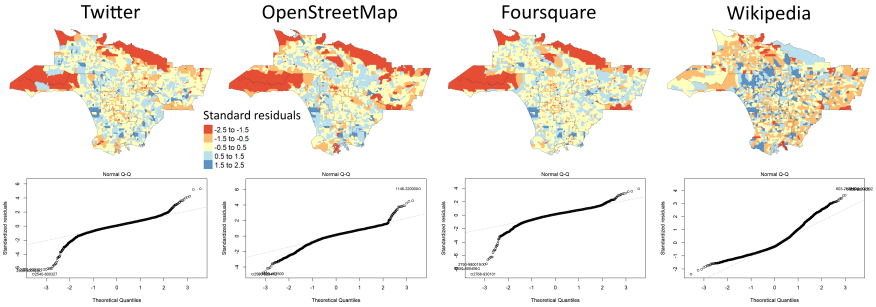


Fig. 7 Spatial distribution and Q-Q plots of the standard residuals of the regression models 1 to 4 presented in Table 2.

5 Discussion

This analysis of UGC datasets provides insights about the place representation at the urban scale across our four research questions, considering US census tract as spatial units. In the study area, a particularly important role is played by population density, which is a co-variate of many other socio-demographic variables, such as presence of Hispanic, young, and deprived residents (see Figure 5). For this reason, we consider more important the density per resident population, rather than by area. As the correlation and regression analyses revealed, UGC platforms show a remarkable diversity in their place representation, both in terms of density of objects in the geographic space and in factors correlating with their density. In other words, each UGC platform generates a unique geography, influenced by contextual factors in different ways ($RQ1$, $RQ2$).

Tweets tend to appear in high-density, urban, deprived areas, showing a substantial divergence between the geography of participation (i.e. where users live) and where they produce tweets, contradicting previous research (Li et al., 2013). OSM objects tend to be located in White and non-Hispanic areas, with older and educated residents. A similar pattern is followed by Foursquare venues and, with much weaker correlations, Wikipedia pages. This corroborates the view that these information geographies are unique and, statistically speaking, only representative of themselves (Ballatore and De Sabbata, 2018).

A core finding is that, while participation studies show that UGC producers are relatively white, young, urban, and affluent, the same generalisation does not necessarily hold for place representation. When considering UGC per 1,000 residents, data generated in Los Angeles County exhibits a significant bias towards areas characterised by a younger and higher-qualified population for spatially-explicit platforms (OSM, Wikipedia, and Foursquare), and much less for Twitter, which obtains substantially lower coefficients. Surprisingly, wealth does not have a noticeable impact.

In a city as segregated as LA, ethnicity plays a role in shaping the geographies of place representation. With the exception of Twitter, areas dominated by White residents are associated with more content, while Hispanic are associated with less. Areas with Black and Asian residents show weaker associations. Income, higher education, and poverty are strong co-variates of the presence of White and Asian residents. The presence of infrastructure, estimated with the LMS POIs per 1,000 residents, is a positive factor for all four UGC, in a weak form for Twitter and OSM, and more strongly for Wikipedia and Foursquare. This suggests that Wikipedia and Foursquare tend to be denser in areas with more public services per inhabitant.

When considering the regression models in Table 2, the variability of all UGC objects appear explainable to some by number of residents, area, level of higher education, and median income, ranging from a very high R^2 for OSM (.78) to a much lower one for Wikipedia (.18). While this partially confirms some common assumptions (e.g. higher population density drives higher amount of content), it also contradicts some received wisdom. Digital content, although probably produced by relatively affluent and educated users, does not concentrate in affluent areas, which in LA tend to be functionally residential. It also highlights the strong variability and differences between platforms. This also shows a clear relationship between density of population and density of UGC objects for three platforms (OSM, Twitter, and Foursquare), while Wikipedia, possibly for its extremely sparse distribution, appears harder to explain.

5.1 Comparing UGC in LA and London

To answer *RQ4*, these findings can be contrasted with those we obtained for Twitter and Wikipedia in Greater London, where level of education of residents, share of population aged group 30-44, and property prices were independent variables (Ballatore and De Sabbata, 2018). This comparison is beneficial to probe to what extent our findings are idiographic (specific to a locale) or nomothetic (general). The two target areas share several characteristics, being both multi-core cities in the Global North with high-income, ethnically diverse, similarly sized, and largely unequal populations. However, they diverge in their human geographies along several dimensions, such as ethnic segregation (lower in London) and functional segregation (also lower in London) (Johnston et al., 2007).

In both case studies, population size, higher education, and affluence carry some explanatory power, adopting property prices in London and household income in LA as estimates, but the relationship bears opposite effects (positive in London, negative in LA). In this context, a striking difference between Greater London and LA County is also visible in the different relationship that Twitter and Wikipedia engender in the two cities. In Greater London, the variability in the number of Wikipedia articles explains 49% of the variation in the number of tweets, showing a convergence of the two geographies (Ballatore and De Sabbata, 2018). In LA, by contrast, it could be said, at best, that the variability in the former would account just for about 6% of variability in the latter.

Behind the complex patterns of a multi-core metropolis, in our prior study, the place representation of Greater London showed unequivocally a data-rich central core surrounded by sparse peripheries, with considerably less data. We expected this picture to be inverted in post-industrial LA, which famously exhibits a hollow core and affluent suburbs, but the spatial structure is more complex. Areas dominated by educated, affluent White and Asians are over-represented, but much data is produced in deprived inner city areas, where more economic activities are located. This strongly indicates that, as much as possible, local geographies must be taken into account when considering any UGC platform.

5.2 Limitations

The limitations of the data and methods should be borne in mind when considering these findings. The study focusses on LA County and, as discussed above, the specific human geography of the area constrains the UGC geographies, in ways that are place-specific, as shown in the comparison to London. In general, spatially-explicit studies over census units can suffer from the ecological fallacy, also known as aggregation bias (Voss, 2007). However, our analysis of place representation does not aim at drawing inferences about individuals who live and produce data in the target areas, but only about the areas themselves. Furthermore, this study is agnostic regarding the relationship between the geography of informational content (i.e. where digital content is located) and the residential geography of content producers (i.e. where users live). These two geographies are likely to have some degree of overlap, but we consider this research question outside of the scope of this article. From a temporal perspective, the datasets were collected in slightly different time frames, all in period 2013–2018. We deem this to be an acceptable time gap for the study, but some rapid changes in the urban fabric might be missed. The regressions, while informative, are not robust, exhibiting non-normal residuals, indicating that the relationships between UGC and its context are not trivial to model. Different approaches are needed to move from this kind of correlational analysis to causal explanations.

6 Conclusions

In this article, we conducted a quantitative investigation of place representation for four UGC platforms (Twitter, Wikipedia, OSM, and Foursquare), using several geo-demographic variables to capture the platial context of the data. As a case study, we considered the Southern part of Los Angeles County, adopting the methodology we devised for Greater London (Ballatore and De Sabbata, 2018). To trace these information geographies, we harvested about 685,000 geo-located tweets, 9,700 Wikipedia pages, 4m OSM objects, and 180,000 Foursquare venues, with 65,000 LMS POIs to represent public infrastructure. Guided by four research questions, an exploratory spatial analysis, followed by a correlation analysis and regression modelling, produced the following main findings about the place representation of UGC:

Population density. In LA County, higher population density is correlated with higher poverty, more Hispanic residents, younger population, fewer White residents, lower income, and lower education. Higher density corresponds to more tweets and more OSM objects.

Twitter. Geo-located tweets tend to appear in denser areas. Almost 60% of the variability in number of tweets can be explained. When considering tweets per resident, no clear demographic factor emerges. Counter-intuitively, tweet density correlates with the deprivation index (more deprived areas obtain more tweets), because LA business and leisure hotspots are often located in inner city areas.

Wikipedia. Pages are very sparsely distributed. Pages per resident correlate with LMS POIs, and weakly with White residents and, negatively, with Hispanic residents. Only 18% of variability can be explained through demographic factors.

OSM. When considering OSM objects per resident, these objects tend to occur in White, educated, affluent, older, non-dense areas. The distribution of OSM objects is the most predictable, with 77% of variability related to demographic factors.

Foursquare. Venues per resident exhibit positive correlations with LMS POIs per resident, and with White, educated, affluent, older, non-dense areas. 47% of variability is venues can be predicted through demographic variables.

Cross-platform comparison. Each of the four UGC sources has a distinct place representation, with different spatial distributions. In the regression models, Twitter, OSM, and Foursquare show a broadly similar structure, while Wikipedia appears more unique. Even sources that show relatively similar correlations (e.g. Foursquare, Twitter, OSM) have statistically different hot and coldspots.

Los Angeles and London. When comparing these findings with an analogous study of Greater London (Ballatore and De Sabbata, 2018), similarities and differences emerge. In both cities, population size and the share of residents with higher education have explanatory power. Indicators of affluence also have some explanatory power, but with seemingly opposite effects (positively for property prices in London, negatively for household income in LA). While the ethnic composition of areas bears no explanatory power in London, the presence of White and, to a

lesser extent, Asian residents is associated with more data in LA. The age group 30-44 has a clear contribution to data variability in London, but it is not a factor in LA. Half of the variability of Wikipedia and Twitter appears explainable in London, geo-demographic factors in LA explain almost 60% of Twitter and only 20% of Wikipedia. In London, the variability in Wikipedia is linked to up to 49% of that in Twitter, while only up to 6% in LA.

Future directions for this research include evaluating other modelling techniques to quantify these spatial relationships between UGC and geo-demographic variables. Poisson regressions might be provide a suitable tool for count data (e.g., number of tweets). As our linear regressions possess spatially auto-correlated residuals, geographically-weighted regressions (GWR) could provide a better fit. Moreover, this study focussed on coarse variables, such as the total count of Foursquare venues or OSM objects in a census tract. These UGC datasets contain semantic details that could be fruitfully harnessed in further analysis, for example considering individual sub-categories (e.g. bars, restaurants or schools). In addition, to account for the functional segregation of California, variables about landuse and zoning categories should be included in our models.

The study in this article is single scale, considering all the data at the tract level (about 3,700 people per unit). In order to further investigate the relationships across scales, it will be critical to aggregate the data in coarser units and perform sensitivity analyses. While the details of this article are idiographic, being focussed on a specific city at the census tract level, such case studies are necessary to build a corpus of empirical evidence to ground nomothetic thinking about place and information. Conversely, exploring small area estimation techniques would provide insights into the extent to which it is possible to model UGC variables in units smaller than tracts. Finally, a composite index of data availability (in analogy with deprivation indices) would provide a conceptually simple and yet powerful tool to quantify socio-spatial inequalities at the urban scale. As the use of unrepresentative UGC is still widespread (Bowker, 2014), more research in place representation can help understand and reduce the biases embedded in these web-based geographies, taking into account the structural variations in the human geography in different locales.

Acknowledgements The authors wish to thank Grant McKenzie (McGill University) for sharing the Foursquare and Twitter datasets and for his input in the study design. The article relies on open datasets from US government agencies and Los Angeles County, including Population and Poverty Estimates (2016) and the LA County Location Management System (2016). The authors thank Twitter, OpenStreetMap, Wikipedia, and Foursquare for kindly making their data available to researchers.

References

Acheson, E., De Sabbata, S., and Purves, R. S. (2017). A quantitative analysis of global gazetteers: Patterns of coverage for common feature types. *Computers,*

Environment and Urban Systems, 64:309–320.

Agnew, J. (2011). Space and Place. In Agnew, J. and Livingstone, D., editors, *Handbook of Geographical Knowledge*, chapter 29, pages 316–330. Sage Publications, London.

Antoniou, V., Morley, J., and Haklay, M. (2010). Web 2.0 Geotagged Photos: Assessing the spatial dimension of the phenomenon. *Geomatica*, 64(1):99–110.

Arsanjani, J. J. and Bakillah, M. (2015). Understanding the potential relationship between the socio-economic variables and contributions to OpenStreetMap. *International Journal of Digital Earth*, 8(11):861–876.

Ballatore, A. (2016). Prolegomena for an ontology of place. In Onsrud, H. J. and Kuhn, W., editors, *Proceedings of the Vespucci Institutes: Advancing Geographic Information Science*, pages 91–103. GSDI Association Press, Needham, MA.

Ballatore, A. and De Sabbata, S. (2018). Charting the geographies of crowdsourced information in Greater London. In Mansourian, A., Pilesjö, P., Harrie, L., and van Lammeren, R., editors, *AGILE 2018, Lecture Notes in Geoinformation and Cartography*, pages 149–168, Berlin. Springer.

Ballatore, A. and Jokar Arsanjani, J. (2018). Placing Wikimapia: an exploratory analysis. *International Journal of Geographical Information Science*. In press.

Blank, G. and Lutz, C. (2017). Representativeness of Social Media in Great Britain: Investigating Facebook, LinkedIn, Twitter, Pinterest, Google+, and Instagram. *American Behavioral Scientist*, 61:741–756.

Bowker, G. C. (2014). Emerging configurations of knowledge expression. In Gillespie, T., Boczkowski, P. J., and Foot, K. A., editors, *Media Technologies: Essays on Communication, Materiality, and Society*, pages 99–118. MIT Press, Boston, MA.

Bright, J., De Sabbata, S., Lee, S., Ganesh, B., and Humphreys, D. K. (2018). OpenStreetMap data for alcohol research: Reliability assessment and quality indicators. *Health & place*, 50:130–136.

Burbidge, J. B., Magee, L., and Robb, A. L. (1988). Alternative transformations to handle extreme values of the dependent variable. *Journal of the American Statistical Association*, 83(401):123–127.

Crampton, J. W., Graham, M., Poorthuis, A., Shelton, T., Stephens, M., Wilson, M. W., and Zook, M. (2013). Beyond the geotag: Situating 'big data' and leveraging the potential of the GeoWeb. *Cartography and Geographic Information Science*, 40(2):130–139.

Curry, J. and Kenney, M. (1999). The paradigmatic city: Postindustrial illusion and the Los Angeles School. *Antipode*, 31(1):1–28.

Davis, M. (2006). *City of Quartz: Excavating the Future in Los Angeles*. Verso Books, London, 2 edition.

Fontenot, K., Semega, J., and Kollar, M. (2018). Income and Poverty in the United States: 2017. Current Population Reports, P60-263, United States Census Bureau.

Graham, M., De Sabbata, S., and Zook, M. A. (2015a). Towards a study of information geographies: (im)mutable augmentations and a mapping of the geographies of information. *Geo: Geography and Environment*, 2(1):88–105.

Graham, M., Straumann, R. K., and Hogan, B. (2015b). Digital divisions of labor and informational magnetism: Mapping participation in Wikipedia. *Annals of the Association of American Geographers*, 105(6):1158–1178.

Greenwood, S., Perrin, A., and Duggan, M. (2018). Social Media Update 2016. Technical report, Pew Research Center, Washington, D.C. `http://www.pewinternet.org/2016/11/11/social-media-update-2016`.

Hahmann, S., Purves, R. S., and Burghardt, D. (2014). Twitter location (sometimes) matters: Exploring the relationship between georeferenced tweet content and nearby feature classes. *Journal of Spatial Information Science*, 2014(9):1–36.

Hecht, B. and Stephens, M. (2014). A Tale of Cities: Urban Biases in Volunteered Geographic Information. In *Proceedings of the Eighth International AAAI Conference on Weblogs and Social Media*, pages 197–205.

Johnston, R., Poulsen, M., and Forrest, J. (2007). The Geography of Ethnic Residential Segregation: A Comparative Study of Five Countries. *Annals of the Association of American Geographers*, 97(4):713–738.

Kind, A. J., Jencks, S., Brock, J., Yu, M., Bartels, C., Ehlenbach, W., Greenberg, C., and Smith, M. (2014). Neighborhood socioeconomic disadvantage and 30-day rehospitalization: a retrospective cohort study. *Annals of Internal Medicine*, 161(11):765–774.

Kitchin, R. and McArdle, G. (2016). What makes Big Data, Big Data? Exploring the ontological characteristics of 26 datasets. *Big Data & Society*, 3(1):2053951716631130.

Lansley, G. and Longley, P. A. (2016). The geography of Twitter topics in London. *Computers, Environment and Urban Systems*, 58:85–96.

Li, L., Goodchild, M. F., and Xu, B. (2013). Spatial, temporal, and socioeconomic patterns in the use of Twitter and Flickr. *Cartography and Geographic Information Science*, 40(2):61–77.

Lindqvist, J., Cranshaw, J., Wiese, J., Hong, J., and Zimmerman, J. (2011). I'm the mayor of my house: Examining why people use foursquare-a social-driven location sharing application. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pages 2409–2418. ACM.

Longley, P. A. and Adnan, M. (2016). Geo-temporal Twitter demographics. *International Journal of Geographical Information Science*, 30(2):369–389.

Martí, P., García-Mayor, C., and Serrano-Estrada, L. (2019). Identifying opportunity places for urban regeneration through LBSNs. *Cities*, 90:191–206.

Mashhadi, A., Quattrone, G., and Capra, L. (2013). Putting ubiquitous crowdsourcing into context. In *Proceedings of the 2013 conference on Computer supported cooperative work - CSCW '13*, pages 611–621, New York, New York, USA. ACM Press.

Noulas, A., Scellato, S., Mascolo, C., and Pontil, M. (2011). An empirical study of geographic user activity patterns in Foursquare. In *Fifth International AAAI Conference on Weblogs and Social Media*.

Ostermann, F. O. and Granell, C. (2017). Advancing science with VGI: Reproducibility and replicability of recent studies using VGI. *Transactions in GIS*, 21(2):224–237.

Purves, R. S., Winter, S., and Kuhn, W. (2019). Places in information science. *Journal of the Association for Information Science and Technology*.

Redi, M., Aiello, L. M., Schifanella, R., and Quercia, D. (2018). The Spirit of the City: Using Social Media to Capture Neighborhood Ambiance. In *Proc. ACMHum.-Comput. Interact. 2, CSCW*, volume 2.

Robinson, A. C. (2019). Representing the presence of absence in cartography. *Annals of the American Association of Geographers*. in press.

Schnebele, E. and Cervone, G. (2013). Improving remote sensing flood assessment using volunteered geographical data. *Natural Hazards and Earth System Sciences*, 13:669–677.

See, L., Mooney, P., Foody, G., Bastin, L., Comber, A., Estima, J., Fritz, S., Kerle, N., Jiang, B., Laakso, M., et al. (2016). Crowdsourcing, citizen science or volunteered geographic information? The current state of crowdsourced geographic information. *ISPRS International Journal of Geo-Information*, 5(5):55.

Shaw, J. and Graham, M., editors (2017). *Our Digital Rights to the City*. Meatspace Press, Oxford, UK.

Sloan, L. and Morgan, J. (2015). Who tweets with their location? Understanding the relationship between demographic characteristics and the use of geoservices and geotagging on Twitter. *PloS one*, 10(11):e0142209.

Smith, A. and Anderson, M. (2018). Social Media Use in 2018. Technical report, Pew Research Center, Washington, D.C. `https://www.pewinternet.org/2018/03/01/social-media-use-in-2018`.

Smith, T., Noble, M., Noble, S., Wright, G., McLennan, D., and Plunkett, E. (2015). The English indices of deprivation 2015. Technical report, Department for Communities and Local Government, London.

Soja, E. W. (2014). *My Los Angeles: From urban restructuring to regional urbanization*. Univ of California Press, Berklely, CA.

Steiger, E., Westerholt, R., Resch, B., and Zipf, A. (2015). Twitter as an indicator for whereabouts of people? Correlating Twitter with UK census data. *Computers, Environment and Urban Systems*, 54:255–265.

Voss, P. R. (2007). Demography as a spatial social science. *Population research and policy review*, 26(5-6):457–476.

Wang, Q., Phillips, N. E., Small, M. L., and Sampson, R. J. (2018). Urban mobility and neighborhood isolation in America's 50 largest cities. *Proceedings of the National Academy of Sciences*, 115(30):7735–7740.

Westerholt, R., Steiger, E., Resch, B., and Zipf, A. (2016). Abundant topological outliers in social media data and their effect on spatial analysis. *PloS one*, 11(9):e0162360.

Zagheni, E., Garimella, V., and Weber, I. (2014). Inferring international and internal migration patterns from Twitter data. In *World Wide Web 2014 Companion*, pages 439–444, New York. ACM.