Charting the geographies of crowdsourced information in Greater London

Andrea Ballatore and Stefano De Sabbata

AGILE Conference 2018 - Lund, Sweden, 12-15 June
Author copy

Abstract Crowdsourcing platforms and social media produce distinctive geographies of informational content. The production process is enabled and influenced by a variety of socio-economic and demographic factors, shaping the place representation, i.e., the amount and type of information available in an area. In this study, we explore and explain the geographies of Twitter and Wikipedia in Greater London, highlighting the relationships between the crowdsourced data and the local geo-demographic characteristics of the areas where they are located. Through a set of robust regression models on a sample of 1.6M tweets and about 22,000 Wikipedia articles, we identify level of education, presence of people aged 30-44, and property prices as the most important explanatory factors for place representation at the urban scale. To some extent, this confirms the received knowledge of such data being created primarily by relatively wealthy, young, and educated users. However, about half of the variability is left unexplained, suggesting that a broader inclusion of potential factors is necessary.

Key words: Information geography; crowdsourcing; volunteered geographic information; geo-demographics; Twitter; Wikipedia

1 Introduction

Over the past decade, the diffusion of crowdsourcing platforms and GPS-enabled smartphones has enabled the large-scale production of spatial information. This phe-

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

nomenon has been variously characterised as spatial crowdsourcing, volunteered geographic information (VGI), spatial social media, and user-generated content (UGC) (Sui et al., 2012; See et al., 2016). Among others, Wikipedia articles, OpenStreetMap vector data, and geo-located tweets are popular data sources for countless studies in geography, demography, sociology, and even seismology (e.g., Earle et al., 2010; Zagheni et al., 2014).

Although the potential of these data sources is apparent, much research analysed very few platforms, such as OpenStreetMap (Mashhadi et al., 2015), often paying limited attention to the geo-demographic context in which the data was produced. Studies of information geographies have so far focused on large spatial units, such as countries (Graham et al., 2015a), with few works focusing on the urban or regional scale. The latter is however of particular importance, as VGI tends to be produced in urban areas (Hecht and Stephens, 2014). When using VGI, it is necessary to consider the socio-spatio-temporal processes that supported its generation, thinking about what data is missing, and not only about what is visibly present.

As part of our ongoing efforts to chart geographies of digital information (Ballatore et al., 2017; Graham et al., 2015a), this article investigates the spatial structure of two popular VGI sources at the urban scale. In particular, we consider geo-located Twitter posts and Wikipedia articles in Greater London, comparing and contrasting their spatial distribution. After providing descriptive statistics, the data from both sources is then studied in relation to a set of socio-demographic variables that characterise Greater London. Through a number of regression analyses, we explore the factors exhibit a similar presence or absence of information, including day-work population, ethnic composition, education level, and property prices, which might indicate a relationship to underlying geographies.

The term VGI encompasses diverse sources of spatial information that vary dramatically in terms of demographic, thematic, spatial, and temporal coverage. One of the objectives of this study is precisely to highlight the commonalities and differences of two very different data sources observed in the same geographic area. Moreover, this study analyses the *place representation*, i.e., the digital information available to characterise areas in heterogeneous data sources, regardless of the demographic characteristics of producers. For this reason, apart from a handful of prolific bots, we include all data available for each spatial unit of analysis, to see to what extent an area is either data-rich or data-poor.

This study contributes to the knowledge of VGI sources, providing findings about what areas are over- and under-represented in these two sources. These insights are relevant to VGI users, providing evidence about datasets' geographical structure, representativeness, and therefore fitness-for-purpose for studies and applications. Producers can also benefit by updating their platforms for a more equal place representation, along similar lines of studies of gender inequality in Wikipedia, which prompted a number of initiatives to increase participation of women.[1] From a more social-scientific perspective, this study can contribute to the study of digital divides and "informational ghettos" (Shaw and Graham, 2017, p. 4), data-poor ar-

[1] `https://en.wikipedia.org/wiki/Gender_bias_on_Wikipedia`

eas that persist even in wealthy, digitally over-represented global hubs like London. Our initial hypothesis–only partially confirmed–is that content in both Twitter and Wikipedia tends to be more representative of wealthier urban areas, inhabited by a younger, and more educated, and less ethnically diverse population than average.

The remainder of this paper is organised as follows. Section 2 summarises critically related work in this area. The socio-economic datasets used to characterise the geography of London are described in Section 3. Subsequently, Sections 4 describes our study of Twitter and Wikipedia located in Greater London, starting with a visual analysis and then continuing with a set of regression models. Section 5 summarises our findings. Finally, conclusions and future work directions are drawn in Section 6.

2 Related work

After a decade of research, much is known about VGI sources. Some sources are spatially-explicit, aiming at a spatial coverage, while others are spatially-implicit, embedding locational information as simply one of the attributes being expressed (Antoniou et al., 2010). For example, Wikipedia and GeoNames aim at comprehensive coverage of cities, while geo-located tweets and Instagram photos are the by-product of a mediated communication process between users located in cities. Unsurprisingly, urban areas tend to be better covered than rural ones (Hecht and Stephens, 2014), with the exclusion of sources limited to highly specific themes (e.g., hiking).

Different crowdsourcing platforms attract different demographic groups in terms of age, gender, income, education, area of residence, interests, and motivations, shaping the properties of the resulting dataset, each displaying its own idiosyncrasies (Acheson et al., 2017). Despite early claims of radical democratization and inclusion, user communities tend to be skewed towards Western, wealthy, educated, white, and male users (Crampton et al., 2013), although exceptions exist – countering the general trend, Wikimapia enjoys more uptake among Indian and Middle-Eastern users (Bittner, 2017). Beyond the VGI niche, the characteristics of social media users are widely studied, particularly in relation to their similarities and differences to the general population.

The 330M monthly active users of Twitter tend to be wealthier and more educated than the average population, and they generate 500M tweets a day. From a statistical perspective, Twitter users are actually not representative of *any* particular population (Blank and Lutz, 2017), but their sheer number and ease of access to large samples of the data attracts researchers and marketeers. As Sloan and Morgan (2015) point out, 0.85% of the Twitter feed output is geotagged with coordinates, which amounts to roughly 4M tweets a day, produced by a population only marginally different to the overall platform population. Geo-located tweets have been used for a variety of purposes, sensing for example urban activities (Lansley and Longley, 2016), emotions (Quercia et al., 2012), and beer-related behaviour (Zook and Poorthuis, 2014).

Using geo-located tweets, Longley and Adnan (2016) conducted a geo-demographic analysis of tweets in London, identifying sub-groups in the user population and measuring the heterogeneity and the connectedness of places. Hahmann et al. (2014) investigated the spatial relationship between geo-located tweets and points of interest, showing correlations at the local scale for certain topics (e.g., "train station", "airport") and not for others ("pub", "bakery"). Using an approach similar to the one we adopt in our study, Li et al. (2013) explore tweets and Flickr photos spatio-spatial distribution in California, showing that photos are denser in natural parks and that tweets tend to originate from areas with educated, high-income people.

Our second source for the study of place representation is Wikipedia. Despite a prolonged decline in the number of contributors (Halfaker et al., 2013), the crowdsourced encyclopaedia is one of the top ten most visited websites worldwide, reaching more than 270M views per day,[2] and hosts 5.5M articles in English, edited by about 130,000 monthly active editors. Wikipedia shows an extreme gender imbalance, with about 84% of male editors, and less so in the readership, which is about 40% female (Hill and Shaw, 2013). From a geographical perspective, in 2013, about 730,000 articles in English were associated with a geo-location. The bulk of the editing of these articles occurs in the Global North, also for articles about places in the Global South, exhibiting a staggering bias towards Western European and North American contributors (Graham et al., 2015b). For this reason, the location of Wikipedia editing activity can be used as a proxy to knowledge capital of countries (Stephany and Braesemann, 2017). Editing of spatial features such as cities tends to be performed by local editors, as observed for OpenStreetMap (Johnson et al., 2016).

Much of this research aims at understanding the population of data producers, while the theme of place representation has been studied only marginally. The informational geographies of crowdsourced, VGI datasets have been charted at the global level (Graham et al., 2015a), drawing attention to the common bias towards relatively young, educated, wealthy users located in the Global North. To the best of our knowledge, no study has comparatively explored the properties of diverse VGI datasets at the urban scale, and their relationship with socio-economic texture of the places that the data describe. Moreover, the relationship between Twitter and Wikipedia from a spatial perspective has not been directly studied before.

3 Datasets

The area of our study is Greater London, which has a population of 8.87M, extended over 1,569 km^2. Three groups of datasets were collected and harmonised: socio-economic data from the UK Census, geo-located tweets, and Wikipedia articles. The summary statistics for the three datasets are shown in Table 1, showing minimum, median, maximum, mean, and standard deviation for all relevant vari-

[2] `https://www.alexa.com/siteinfo/wikipedia.org`

Table 1 Descriptive statistics for the socio-economic data for Greater London for 983 MSOAs, 1.6M geo-referenced tweets and 22,411 Wikipedia articles, both grouped in the MSOAs. (*) Black, Asian, and minority ethnic. (†) Post-high school qualifications.

Statistic (983 MSOAs)	Min	Median	Max	Mean	St. Dev.
Socio-economic variables					
Area (ha)	29.40	114.50	2,243.00	159.94	186.10
Population	5,184	8,156	14,719	8,315.30	1,448
Workday population	3,444	6,789	360,075	8,826	14,500
Age 0-15 (%)	6.05	19.77	35.90	19.83	4.14
Age 16-29 (%)	10.61	21.29	52.73	22.36	5.73
Age 30-44 (%)	13.44	24.86	38.26	25.25	4.36
Age 45-64 (%)	10.06	20.98	31.96	21.34	4.01
Age 65+ (%)	2.40	10.37	27.23	11.22	4.12
BAME* population (%)	3.80	37.30	93.90	39.42	19.31
Household one person (%)	12.60	30.50	56.40	30.82	7.22
Household couple (%)	8.70	19.00	30.30	18.89	4.41
Hh. couple with dep. child (%)	4.60	18.80	32.20	18.28	5.12
Qualification 4 or above† (%)	8.19	28.05	62.62	30.32	11.83
House price (2012, £)	130,000	284,000	2,930,000	333,675	186,641
Twitter					
Number of tweets	18	413	161,050	1,617	8,063
Number of Twitter users	9	128	58,286	678	3,03
Twitter entropy	0.70	3.93	10.14	4.13	1.51
Twitter Gini coefficient	0.08	0.54	0.90	0.55	0.14
Wikipedia					
Number of Wikipedia articles	0	9	1,857	22.80	87.50
Wikipedia cumulative length (bytes)	0	47,606	17,042,103	155,839	735,909
Wikipedia cumulative edits	0	450	68,739	1,088	3,512
Wikipedia cumulative minor edits	0	4	2,499	94.95	192.92

ables. All variables are captured at the level of spatial unit selected for the analyses, detailed in the remainder of this section.

3.1 London demographic data

The UK Census, the latest of which occurred in March 2011, provides detailed socio-economic information about London. Census data is structured in Output Areas (OA), each covering between 40 and 149 households, corresponding to an average of 300 people.[3] For small area statistics, OAs are grouped into Lower Layer Super Output Areas (LSOA), which contain from four to six OAs, with a mean population of 1,500. Given the spatially uneven distribution of Twitter and Wikipedia data, the OAs and even LSOAs are too granular, leaving many areas without data. For this reason, we use Middle Layer Super Output Areas (MSOA), which further

[3] `https://www.ons.gov.uk/methodology/geography/ukgeographies/censusgeography`

aggregates LSOAs into contiguous groups, with a minimum population of 5,000 and a national mean of 7,200.

Greater London contains 983 MSOAs, whose boundaries were collected from the UK Data Service,[4] while the Census 2011[5] aggregated variables used in our study were collected from the London Datastore[6] and Nomis.[7] The variables include the MSOA's area, total and workday population, age composition, household size, education, and house prices (see Table 1). While some variables tend to have parametric distributions (e.g., percentage of residents aged from 0 to 15), others are heavily skewed towards large outliers, which we will take into account in our analyses. Notably, some areas of London have extremely high workday population: notably, during the day, the City of London hosts about 360,000 workers, while having just 9,400 residents. Similarly, house prices are skewed by multi-million-pound properties that are common in Central London.

Because of the high dimensionality and complexity of the Census data, geodemographic classifications have been produced as a way to summarise the population into a set of discrete classes. Notably, the London Output Area Classification (LOAC) categorises each OA in Greater London into eight super-groups, such as "Urban Elites" and "Settled Asians", further classified into groups (Singleton and Longley, 2015). This classification is useful to detect the demographic structure of the urban space, and can be related to the place representation observed in the informational geographies.

3.2 Twitter data

All geo-referenced tweets produced in Greater London were collected from the Twitter API from October 2015 to May 2016, for a total of 2,076,588 tweets, produced by 222,719 users, excluding re-tweets. As we are interested in place representation and not in specific user behaviours, we retain low-activity users. The only category of users that we exclude from the analysis is high-activity bots, whose tweets do not capture the manual information production we intend to observe.

To identify bots, we combine two heuristics, measuring for each user (1) the number of tweets per day, and (2) the percentage of repeated tweets, assuming that bots, for advertising purposes, generate a high number of tweets, and tend to repeat the same content more than human users. Hence, we selected users that generated more than 10 tweets per day, and whose 10% of tweets were repeated at least once. These thresholds were identified by trial and error, and then observing a sample of excluded users to make sure they were all bots (e.g., *trendinaliaGB*). This process filtered out 1.4% of users, corresponding to 22.7% of tweets.

---

[4] `https://borders.ukdataservice.ac.uk/`

[5] `https://www.ons.gov.uk/census/2011census`

[6] `https://data.london.gov.uk/`

[7] `https://www.nomisweb.co.uk/`

The remaining dataset of 1,589,819 tweets was generated by 219,604 users. As expected, most users produced very few tweets: The number of tweets per user range from 0 to 1,950, with a median of just 2. From a linguistic viewpoint, 91.6% of tweets are in English, with the other larger groups being undefined (3.4%), Spanish (1.6%), and French (0.8%). The tweets were then grouped in the MSOAs of Greater London. The number of tweets per spatial unit ranges from 9 to about 160,000 in Westminster, with a median of about 400 tweets (see Table 1). The number of users active in each unit follows a similarly skewed distribution. We also calculated Shannon entropy and the Gini coefficient as measures respectively of contribution diversity and inequality.

3.3 Wikipedia data

The second VGI source in this study consists of geo-referenced Wikipedia articles located in Greater London. At the time of writing, Wikipedia only allows for geo-tags in the form of points, and even large geographical entities are geo-tagged to a point. For example, the article about the Palace of Westminster is associated with a latitude/longitude point.[8] The decision about where to locate entities is a combination of the platform guidelines and the editors' arbitrary choices. As a result, the same entity can be pin-pointed in different locations in different language editions. Such inconsistencies are common in collaborative editing (Ballatore and Mooney, 2015). For instance, at the time of writing, "England" in the English Wikipedia[9] is geo-tagged on the River Thames near the Palace of Westminster, whereas the Italian version geo-tags "Inghilterra" somewhere in the Borough of Bromley, near Biggin Hill. By contrast, the German Wikipedia selects a geometric centroid located east of Birmingham.

Using the databases available on Wikimedia Toolforge,[10] we extracted 22,411 articles, including features such as monuments, notable buildings, parks, and headquarters of organisations. About 41% of articles are in English, followed by 6% in French, 6% in German, and the remainder 47% in other languages. After grouping them in the MSOAs, it is possible to note that the articles are sparser than the tweets, with 32 spatial units (3%) without any article (see Table 1). Furthermore, 34% of units contain fewer than 6 articles. The densest parts of the distribution are found in an MSOA in Westminster (1,857 articles), and in the City of London (1,466).

[8] `https://en.wikipedia.org/wiki/Palace_of_Westminster`

[9] `https://en.wikipedia.org/wiki/England`

[10] `https://tools.wmflabs.org`

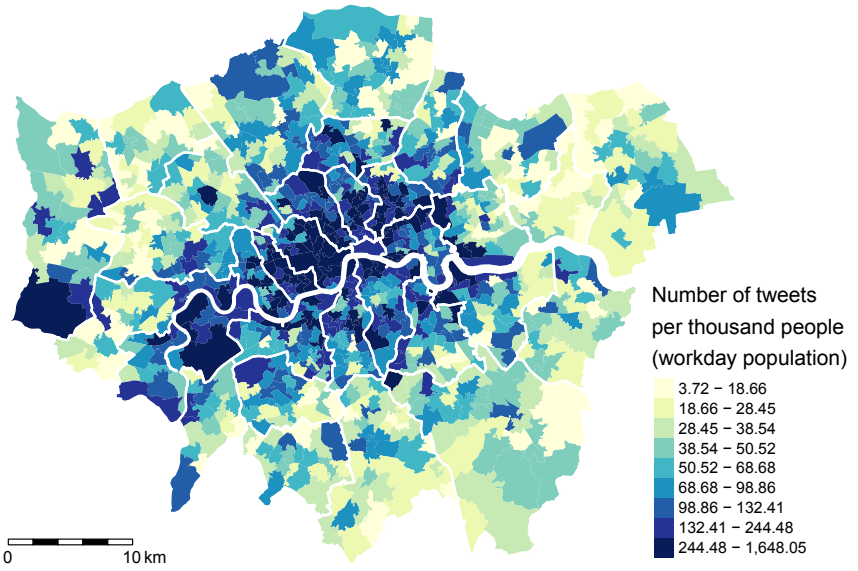


Fig. 1 Distribution of 1.6M geo-located tweets in Greater London, scaled by workday population, for 983 MSOAs. The data is grouped into 9 quantiles. The boundaries of the boroughs are outlined in white.

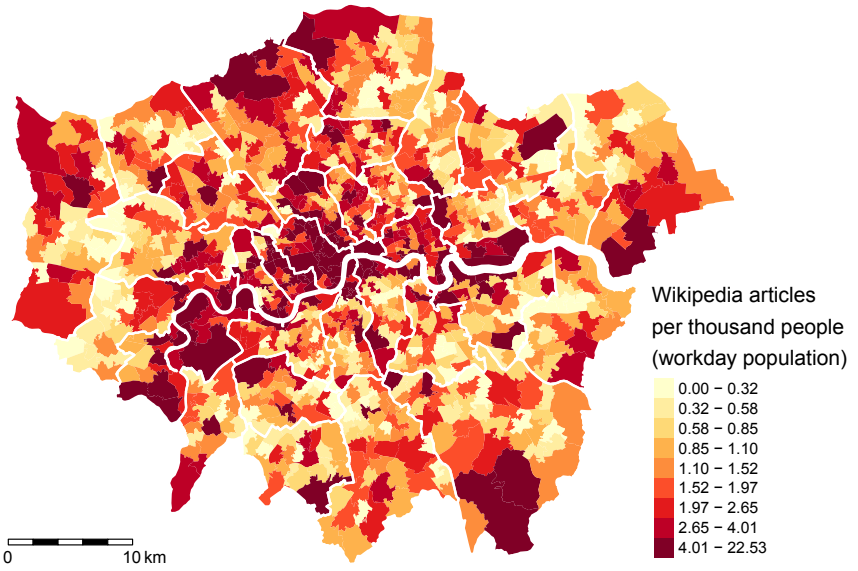


Fig. 2 Distribution of 22,411 geo-located Wikipedia articles in Greater London, scaled by workday population, for 983 MSOAs. The data is grouped into 9 quantiles. The boundaries of the boroughs are outlined in white.

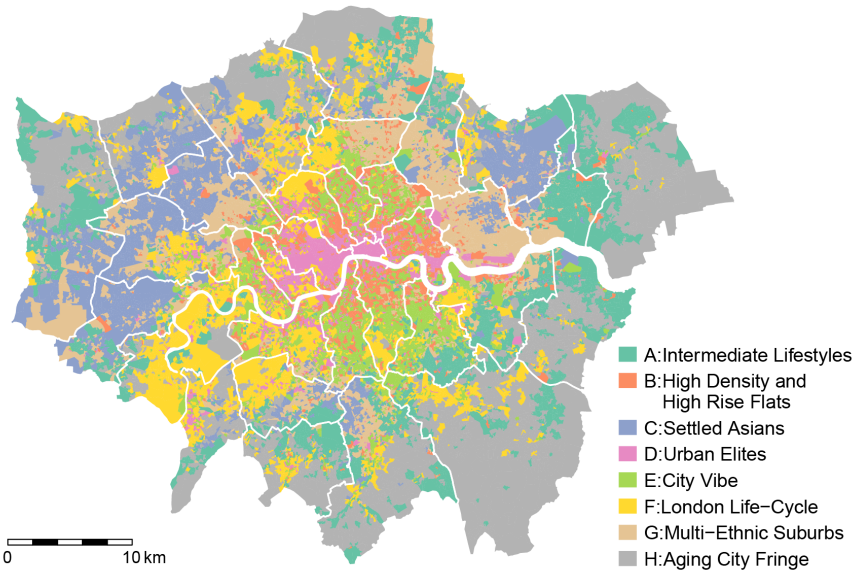


Fig. 3 London Output Area Classification (LOAC), showing super-groups. Source: Singleton and Longley (2015).

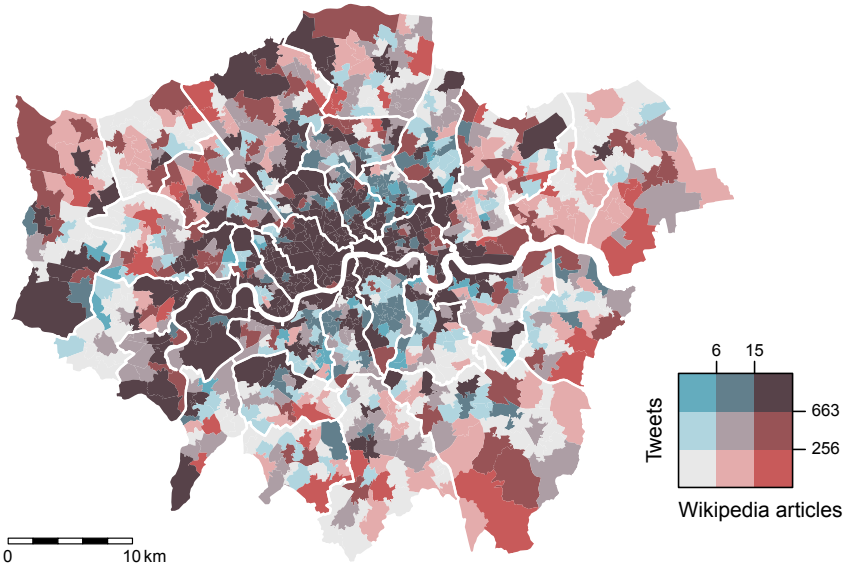


Fig. 4 Comparison of the distribution of 1.6M tweets and 22,411 Wikipedia articles in Greater London in 983 MSOAs. The boundaries of the boroughs are outlined in white.

4 Explaining crowdsourced geographies

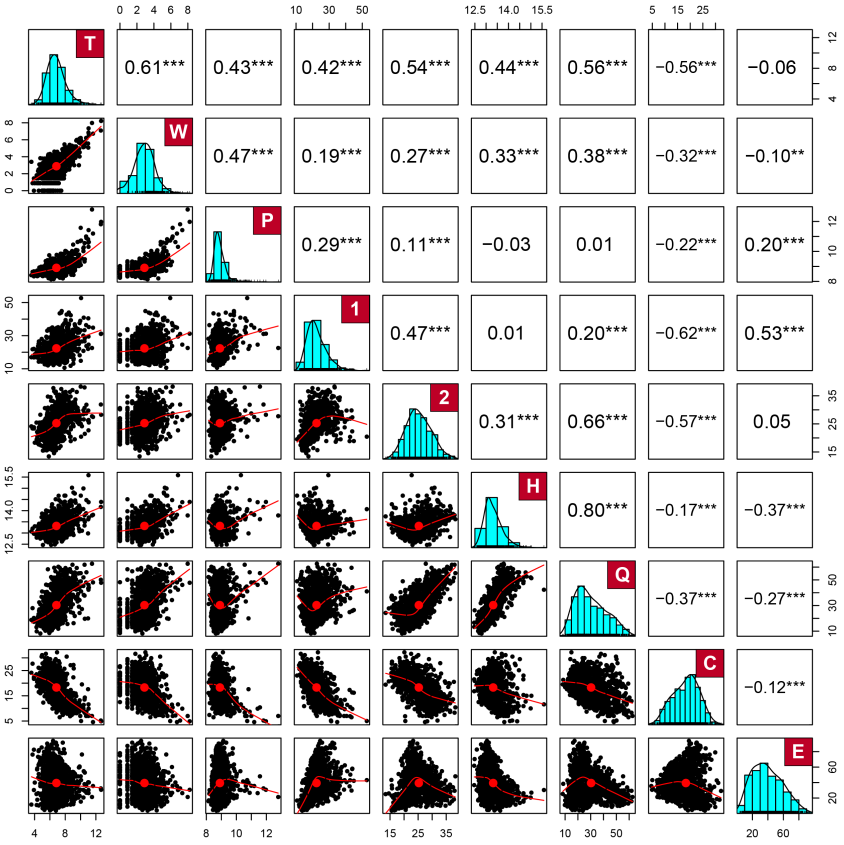
To understand the factors that shape the geography of the Twitter and Wikipedia data, we start by observing the properties of their spatial distribution. Figure 1 and Figure 2 show the number of tweets and Wikipedia articles in Greater London, scaled by workday population. While both distributions show, as largely expected, high density in Central London, the maps also suggest differences, for example in Southern parts of the city, which deserve more investigation.

To relate Twitter and Wikipedia data to the demographic geography of London, we display in Figure 3 the London Output Area Classification (LOAC) by Singleton and Longley (2015), as a summary of the demographic characteristics of each area. In order to allow for a visual comparison of tweets and Wikipedia articles, Figure 4 displays a bi-variate choropleth map generated from the intersections of the three quantiles of each of the two variables. The darkest areas in this map represent the highest quantiles of both Twitter and Wikipedia content, indicating the data-richest areas. These areas tend to correspond with OAs classified as "Urban Elites" (i.e., young professionals in science, technology, and finance) and "London Life-Cycle" (i.e., relatively low numbers of students and households with dependent children, highly qualified professionals, predominantly white) in the City and Westminster, as well as Richmond, Merton, and the south part of Newham.

Heathrow Airport, located at the Western edge of Greater London, shows extremely high content density, as already noted by Longley and Adnan (2016). Many areas in Southwark and Hackney, classified as "City Vibe", that is, single professionals and students in communal establishments, display a high density of Twitter content, but relatively low Wikipedia content. This might be due to not only to socio-demographic characteristics, but to the relatively low density of notable urban features in those districts. More generally, low-tweet areas seem to align with OAs classified as "Intermediate Lifestyles" and "Aging City Fringe", both associated with households in later stages in life-cycle.

This visual examination indicates that within Greater London there are substantial differences in the amount of content representing the different areas of the metropolis, broadly following the spatial distribution of the demographic characteristics that have been linked to VGI content production in the literature (e.g., Crampton et al., 2013). Our initial hypothesis is that content in both Twitter and Wikipedia tends to be representative of wealthier areas, inhabited by younger sections of the population, with access to higher levels of qualifications. In the remainder of this section, we perform a series of regression analyses to explain the relationships between the socio-demographic characteristics discussed above, as independent variables, and the number of tweets and Wikipedia articles as dependent variables, aggregated at the level of MSOA.

Fig. 5 Distribution and correlations between tweets *[ihs]* (T), Wikipedia articles *[ihs]* (W), workday population *[log]* (P), percentage of population aged between 16 and 29 (1), and between 30 and 44 (2), level four qualifications or above (Q), couples with dependent children (C), minority groups (E), and house prices *[ihs]* (H), in 983 MSOAs in Greater London. Significance level: $^{*}p<0.1$; $^{**}p<0.05$; $^{***}p<0.01$.



4.1 Variable selection and normalisation

The selection of the independent variables was based on a correlation analysis: Figure 5 illustrates the distribution of and correlations between the variables used in the regression models. The normalised values have been created using the natural logarithm (when $[log]$ is added to the variable name) and the inverse hyperbolic sine (when $[ihs]$ is added to the variable name) (Burbidge et al., 1988; Pence, 2006).

The percentage of population between 16 and 29 and between 30 and 44 were initially considered, but only the latter was included in the models, as it shows higher

Table 2 Linear regression models to explain the spatial variation in the 1.6M geo-located tweets over 983 MSOAs in Greater London. Four models were devised (standard errors between parentheses, significance levels: *p<0.1; **p<0.05; ***p<0.01)

	Dependent variable:			
	Number of tweets [ihs]			
	(Tw1)	(Tw2)	(Tw3)	(Tw4)
Workday pop. [log]	1.451***	1.296***	1.287***	1.225***
	(0.057)	(0.078)	(0.056)	(0.074)
Age 30-44 (%)	0.092***	0.089***		
	(0.006)	(0.006)		
House price [ihs]	0.977***	0.902***		
	(0.066)	(0.067)		
Qual. 4+ (%)			0.040***	0.040***
			(0.002)	(0.002)
Couple w/child. (%)			−0.071***	−0.064***
			(0.005)	(0.005)
Constant	−21.397***	−19.039***	−4.530***	−4.144***
	(0.927)	(1.160)	(0.547)	(0.689)
Observations	983	922	983	922
R^2	0.612	0.483	0.660	0.548
Adjusted R^2	0.611	0.481	0.659	0.546
Res. Std. Error	0.783 (df = 979)	0.712 (df = 918)	0.733 (df = 979)	0.666 (df = 918)
F Statistic	515.135***	285.995***	632.371***	370.803***
	(df = 3; 979)	(df = 3; 918)	(df = 3; 979)	(df = 3; 918)

correlation with both dependent variables. This was then combined with the house prices in the first models (Tw1, Tw2, and Wk1 below), which is the variable that show lower correlation among the other independent variables here considered. The percentage of households with dependent children and the percentage of population with level 4 qualifications or above are then combined in subsequent models (Tw3 and Wk2 below), as they show a lower correlation between each other.

All models also include workday population independent variable, to account for the varying presence of people in each MSOA. Workday population was preferred to resident population due to its higher correlation with the dependent variable. Finally, we compare the amount of content present in the two VGI platforms, which is a novel approach to studying these information geographies (model TwWk below). It is important to note that these models are understood as explanatory of relationships to common underlying geographies, without claims to causality.

4.2 Twitter models

The first model in Table 2 (Tw1) includes the percentage of population between 30 and 44 and house prices as independent variables. However, the residuals of the

model are not normally distributed (Shapiro-Wilk test, $W = 0.99$, $p < 0.001$), due to a handful of MSOA overrepresented in the Twitter dataset. To overcome this issue, we devised Model Tw2, which replicates Model Tw1, but excluding all MSOAs having an average of 10 or more tweets per day. In this second model, the independent variables account for 48% of the variation in the number of tweets. Model Tw2 is fit and robust, as the residuals are normally distributed ($W = 1$, $p = 0.808$), and satisfy the homoscedasticity assumption (Breusch-Pagan test, $BP = 1.56$, $p = 0.668$). The errors are slightly positively correlated, but this does not raise concerns (Durbin-Watson test, $DW = 1.65$, $p < 0.001$), and no multicollinearity has been identified in this model (average VIF is 1.07).

Model Tw3 includes the percentage of households with dependent children and the percentage of population with level 4 qualifications or above as independent variables (see Table 2). As above, the residuals are not normally distributed ($W = 0.99$, $p < 0.001$), thus we create Model Tw4 by excluding all MSOAs having an average of 10 or more tweets per day. In this model, the independent variables account for 55% of the variation in the number of tweets. Model Tw4 is fit and robust, as the residuals are normally distributed ($W = 1$, $p = 0.471$), and satisfy the homoscedasticity assumption ($BP = 9.27$, $p = 0.026$). The errors are slightly positively correlated, but this is not a cause for concern ($DW = 1.74$, $p < 0.001$), and no multicollinearity has been identified (average VIF is 1.12). By observing the spatial distribution of residuals of models Tw2 and Tw4, we did not find evidence of spatial clustering through local Moran's I.

Based on Model Tw2, a one percent increase in the population aged 30 to 44 is linked to an increase the number of tweets produced in the MSOA of about 9%. Similarly, a ten percent increase in house prices is linked to an increase the number of tweets in the MSOA of about 9%. Based on Model Tw4, other things being equal, a one percent increase in the population with level 4 qualifications or above is linked to an increase the number of tweets of about 4%, and a one percent increase in households composed by couples with dependent children is linked to a decrease the number of tweets of about 6%.

4.3 Wikipedia models

After having assessed the distribution of tweets, we proceed to observe possible explanatory factors of the presence of Wikipedia articles in a given MSOA in London. For the statistical modelling, we excluded the 3% of MSOAs that do not contain any Wikipedia article. We generated two regression models, Model Wk1 and Model Wk2, which are robust and fit (see Table 3). The residuals of Model Wk1 are normally distributed (Shapiro-Wilk test, $W = 1$, $p = 0.219$), and satisfy the homoscedasticity assumption (Breusch-Pagan test, $BP = 5.05$, $p = 0.168$). The errors are independent (Durbin-Watson test, $DW = 1.89$, $p = 0.088$), and no multicollinearity has been identified (average VIF is 1.08). The independent variables

Table 3 Linear regression models to explain the spatial variation in the 22,411 Wikipedia articles over 983 MSOAs in Greater London. Two models were devised (standard errors between parentheses, significance levels: *p<0.1; **p<0.05; ***p<0.01).

	Dependent variable:	
	Number of Wikipedia articles [ihs]	
	(Wk1)	(Wk2)
Workday pop. [log]	1.344***	1.312***
	(0.062)	(0.065)
Age 30-44 (%)	0.023***	
	(0.007)	
House price [ihs]	0.822***	
	(0.071)	
Qual. 4+ (%)		0.029***
		(0.003)
Couple w/child. (%)		−0.018***
		(0.006)
Constant	−20.522***	−9.265***
	(1.012)	(0.633)
Observations	951	951
R^2	0.445	0.452
Adjusted R^2	0.443	0.450
Res. Std. Error (df = 947)	0.844	0.838
F Statistic	252.669***	260.448***
	(df = 3; 947)	(df = 3; 947)

account for 44% of the variation in the number of Wikipedia articles, when aggregated by MSOA.

Similarly, the residuals of Model Wk2 are normally-distributed residuals ($W = 1$, $p = 0.158$), and satisfy the homoscedasticity assumption ($BP = 3.52$, $p = 0.318$), In this model too, the errors appear to be independent ($DW = 1.86$, $p = 0.04$), and no multicollinearity has been identified (average VIF is 1.21). The independent variables account for 45% of the variation in the number of tweets, when aggregated by MSOA. As for the models presented in the previous section, the spatial distribution of residuals shows no sign of spatial clustering.

Based on Model Wk1, other things being equal, a one percent increase in the population aged 30 to 44 is linked to an increase the number of Wikipedia articles in the MSOA of about 2%. Similarly, a ten percent increase in house prices is linked to an increase the number of Wikipedia articles in the MSOA of about 8%. Based on Model Wk2, other things being equal, a one percent increase in the population with level 4 qualifications or above is linked to an increase the number of Wikipedia articles in the MSOA of about 3%, and a one percent increase in households composed by couples with dependent child is linked to a decrease the number of Wikipedia articles in the MSOA of about 2%.

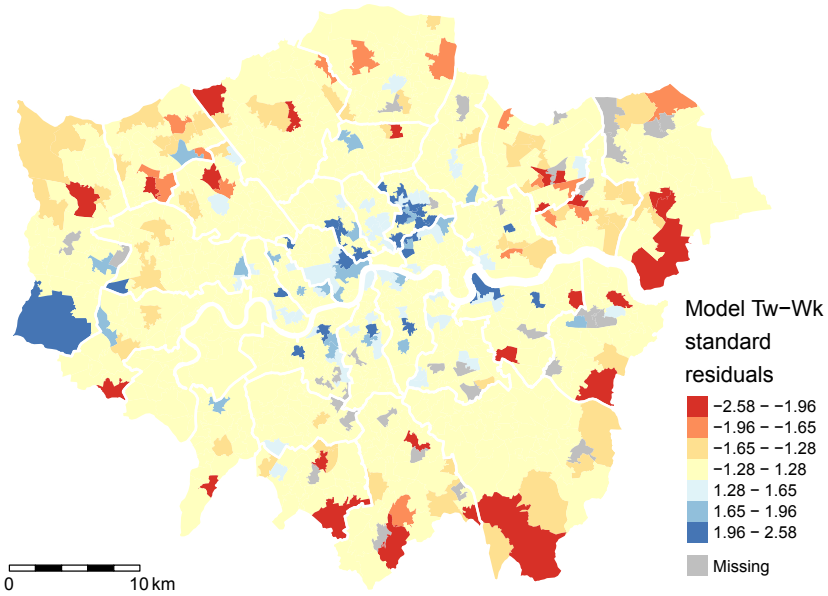


Fig. 6 Distribution of residuals for Model Tw-Wk, which relates Twitter and Wikipedia data. Blue areas have more tweets than expected based on the Wikipedia distribution, while red areas display lower Twitter density compared to Wikipedia.

4.4 Comparison between Twitter and Wikipedia

The geographies of place representation that we analysed above can be directly compared. For this purpose, we created a model using the number of Twitter posts as the dependent variable and the number of Wikipedia article, to observe to what extent they converge, and where they differ. As this is a simple regression, the choice of either variable as dependent or independent does not affect the outcome of the analysis, and thus it was made arbitrarily Model Tw-Wk is fit and mostly robust. The residuals normally distributed (Shapiro-Wilk test, $W = 1$, $p = 0.649$), and satisfy the homoscedasticity assumption (Breusch-Pagan test, $BP = 4.7$, $p = 0.03$). The errors appear to be slightly positively correlated, but this is not a cause for concern (Durbin-Watson test, $DW = 1.34$, $p < 0.001$). Overall, the variability in the number of Wikipedia articles accounts for 49% of the variation in the number of tweets, when aggregated at the MSOA level.

Based on Model Tw-Wk, other things being equal, a ten percent increase in the number of Wikipedia articles is linked to an increase in the number of tweets produced in the MSOA of about 8%. Indeed, these are not to be interpreted as causal

relationships. For this model, the map in Figure 6 shows the spatial distribution of residuals. Unlike the previous models, this map shows some clustering in Central London and Heathrow Airport, where Twitter activity is higher than expected based on Wikipedia content, whilst the outskirts exhibit lower tweet density.

5 Discussion

The exploratory and explanatory analyses discussed above highlight a number of aspects of the information geographies of Twitter and Wikipedia at the urban scale. Overall, the explanatory analyses in Section 4 confirm our general hypothesis based on the literature and the exploratory analysis. Crowdsourced information generated in Greater London exhibits a significant bias towards areas characterized by a wealthier, younger, and higher-qualified population (Crampton et al., 2013).

All models exhibit similar explanatory power, but the variability of content in both Twitter and Wikipedia seems to be more closely liked to level of education and average house prices, when aggregated at the MSOA level. At the same time, the standardized estimates presented in Table 5 suggest that the presence of population has a stronger influence on Wikipedia content than on Twitter, while both the percentage of people aged 30-44 and households composed by couples with dependent children have a stronger influence (positive and negative, respectively) on Twitter content. By contrast, house prices exert a strong influence on the presence of Wikipedia content. This might be linked to the tendency in Wikipedia content to document landmarks, heritage sites, and notable buildings, which tend to correspond to high property value.

The level of education seems to be a crucial factor for both datasets, suggesting that low-education level, indeed, constitute a barrier to accessing these platforms

Table 4 Linear regression model Tw-Wk, which relates Twitter and Wikipedia data, over MSOAs in Greater London (standard errors between parentheses, significance levels: *p<0.1; **p<0.05; ***p<0.01).

	Dependent variable:
	Number of tweets [ihs]
Number of Wikipedia articles [ihs]	0.778***
	(0.026)
Constant	4.562***
	(0.082)
Observations	951
R^2	0.490
Adjusted R^2	0.490
Res. Std. Error	0.896 (df = 949)
F Statistic	913.117***
	(df = 1; 949)

Table 5 Standardized beta estimates (β values) for the models Tw2, Tw4, Wk1, and Wk2, useful to evaluate the explanatory weight of independent variables.

	Models:			
Dependent variable	Tw2	Tw4	Wk1	Wk2
Workday pop. [log]	0.397	0.375	0.529	0.516
Age 30-44 (%)	0.391		0.087	
House price [ihs]	0.337		0.293	
Qual. 4+ (%)		0.455		0.301
Couple w/child.		-0.314		-0.083

and take part in the digital production. Yet, our models highlight that Twitter and Wikipedia do not share the same geography. The fact of being different platforms used for different purposes is reflected in their informational geographies and in the way they over- and under-represent places. In particular, our comparison of geo-located tweets and Wikipedia articles indicate how the former is shaped more by the presence of population with given characteristics, while the latter by notable urban features.

However, these findings must be qualified: The proposed regression models are robust, but account for about 44–55% of the variability, suggesting that much of the variation is not captured by socio-demographic variables alone. More explanatory factors, such as tourism-related activities and amenities, must be included to better capture the place representation outcomes. Interestingly, while most explanatory variables we considered bear some relationships with both crowdsourced datasets, the ethnic composition of each MSOA does not exhibit a link to neither. This could be related to the comparatively low level of residential segregation in the UK (Johnston et al., 2007).

This study contributes to the deeper and paramount issue of representativeness in crowdsourced, "big data" sources. As Blank (2016) argues, Twitter users are generally younger and wealthier than the rest of the population, and do not strictly represent any demographic group. Therefore, Twitter users cannot be used to corroborate claims about society at large. For this reason, knowing the platform biases is important to inform researchers about what they can and, most cogently, cannot expect from such data. In a similar spirit, we argue that the information geographies we investigate in this study are unique and – strictly speaking – only representative of themselves. Hence, studying their socio-demographic biases is key to support their effective usage in research.

6 Conclusions

In this article, we investigated the information geographies of two well-known crowdsourced data sources, Twitter and Wikipedia, observing their place representation in Greater London. A set of 1.6M geo-referenced tweets and about 22,000

Wikipedia articles located in Greater London was studied with respect to socio-economic variables. MSOAs were selected as the spatial unit of analysis, allowing for a granular analysis of the spatial variation in both tweets and articles. Linear regressions revealed that about half of the variability can be explained through variabilities in the level of education of residents, share of population aged group 30-44, and property prices. These factors explain half of the variability, while the other half remains unknown, calling for further work.

This study focussed only on one city. A comparative approach with other cities, in the UK and elsewhere, is necessary to observe the geographical variation in the relationships that between place and its information. Moreover, our models need to include land use and tourism data in order to capture the role of urban function and mass travel, prominent in many data-rich areas in central London. In the early stages of this analysis, we used an Output Area Classification (OAC) as a compact description of the demographic structure of the city (Singleton and Longley, 2015), and we plan to conduct further quantitative analysis, exploring patterns beyond what is visible. Furthermore, we intend to explore the spatial clustering of the residuals in Figure 6. Geographically weighted regressions (GWR) might provide more detailed explanations of the relationship between Twitter and Wikipedia. Characteristics of the built environment, such as building density and presence of other urban features, are likely to provide good independent variables for the analysis.

As future work, more comparative analyses between urban information geographies are needed to reveal the socio-spatial structure of these data sources. To refine these models beyond simple Census variables, more interaction between GIScience and quantitative human geography is needed. Charting the factors that shape these geographies and their place representation can produce insights about the real value, limits, and uncertainty when using such informational assets for research and knowledge extraction. In a context where the problematic use of unrepresentative data is widespread (Bowker, 2014), more research is needed to devise analytical techniques to reduce the spatial and social biases embedded in all emergent, online datasets.

Acknowledgements The demographic data used in this work have been provided by the Greater London Authority and Nomis under the Open Government Licence v2.0. The content analysed in this article was produced by Twitter users and Wikipedia contributors, and obtained through the web services by Twitter, Inc. and Wikimedia Foundation, Inc., under the respective licences. The maps contain data from CDRC LOAC Geodata Pack by the ESRC Consumer Data Research Centre; National Statistics data Crown copyright and database right 2015; Ordnance Survey data Crown copyright and database right 2015.

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.

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.

Ballatore, A., Graham, M., and Sen, S. (2017). Digital Hegemonies: The Localness of Search Engine Results. *Annals of the American Association of Geographers*, 107(5):1194–1215.

Ballatore, A. and Mooney, P. (2015). Conceptualising the geographic world: The dimensions of negotiation in crowdsourced cartography. *International Journal of Geographical Information Science*, 29(12):2310–2327.

Bittner, C. (2017). Diversity in volunteered geographic information: comparing OpenStreetMap and Wikimapia in Jerusalem. *GeoJournal*, 82(5):887–906.

Blank, G. (2016). The Digital Divide Among Twitter Users and Its Implications for Social Research. *Social Science Computer Review*, pages 679–697.

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.

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.

Earle, P., Guy, M., Buckmaster, R., Ostrum, C., Horvath, S., and Vaughan, A. (2010). OMG earthquake! Can Twitter improve earthquake response? *Seismological Research Letters*, 81(2):246–251.

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.

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.

Halfaker, A., Geiger, R. S., Morgan, J. T., and Riedl, J. (2013). The rise and decline of an open collaboration system: How Wikipedia's reaction to popularity is causing its decline. *American Behavioral Scientist*, 57(5):664–688.

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.

Hill, B. M. and Shaw, A. (2013). The Wikipedia gender gap revisited: characterizing survey response bias with propensity score estimation. *PloS one*, 8(6):e65782.

Johnson, I. L., Lin, Y., Li, T. J.-J., Hall, A., Halfaker, A., Schöning, J., and Hecht, B. (2016). Not at Home on the Range: Peer Production and the Urban/Rural Divide. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems - CHI '16*, pages 13–25.

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.

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.

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

Mashhadi, A., Quattrone, G., and Capra, L. (2015). The impact of society on volunteered geographic information: The case of OpenStreetMap. In Jokar Arsanjani, J., Zipf, A., Mooney, P., and Helbich, M., editors, *OpenStreetMap in GIScience*, pages 125–141. Springer, Berlin.

Pence, K. M. (2006). The role of wealth transformations: An application to estimating the effect of tax incentives on saving. *The BE Journal of Economic Analysis & Policy*, 5(1).

Quercia, D., Capra, L., and Crowcroft, J. (2012). The Social World of Twitter: Topics, Geography, and Emotions. In *International Conference on Web and Social Media, ICWSM*, pages 298–305, Palo Alto, CA. AAAI Press.

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.

Singleton, A. D. and Longley, P. (2015). The internal structure of Greater London: a comparison of national and regional geodemographic models. *Geo: Geography and Environment*, 2(1):69–87.

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.

Stephany, F. and Braesemann, F. (2017). An Exploration of Wikipedia Data as a Measure of Regional Knowledge Distribution. In *Social Informatics: 9th International Conference, SocInfo 2017, Oxford, UK*, pages 31–40, Berlin. Springer.

Sui, D. Z., Elwood, S., and Goodchild, M., editors (2012). *Crowdsourcing Geographic Knowledge: Volunteered Geographic Information (VGI) in Theory and Practice*. Springer, Berlin.

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.

Zook, M. and Poorthuis, A. (2014). Offline brews and online views: Exploring the geography of beer tweets. In Patterson, M. and Hoalst-Pullen, N., editors, *The Geography of Beer*, pages 201–209. Springer, Berlin.