Measuring spatio-textual affinities in twitter between two urban metropolises

Minda Hu1 · Mayank Kejriwal1

Received: 1 September 2020 / Accepted: 22 May 2021 / Published online: 2 June 2021 © The Author(s), under exclusive licence to Springer Nature Singapore Pte Ltd. 2021

Abstract
With increasing growth of both social media and urbanization, studying urban life through the empirical lens of social media has led to some interesting research opportunities and questions. It is well-recognized that as a ‘social animal’, most humans are deeply embedded both in their cultural milieu and in broader society that extends well beyond close family, including neighborhoods, communities and workplaces. In this article, we study this embeddedness by leveraging urban dwellers’ social media footprint. Specifically, we define and empirically study the issue of spatio-textual affinity by collecting many millions of geotagged tweets collected from two diverse metropolises within the United States: the Boroughs of New York City, and the County of Los Angeles. Spatio-textual affinity is the intuitive hypothesis that tweets coming from similar locations (spatial affinity) will tend to be topically similar (textual affinity). This simple definition of the problem belies the complexity of measuring it, since (re-tweets notwithstanding) two tweets are never truly identical either spatially or textually. Workable definitions of affinity along both dimensions are required, as are appropriate experimental designs, visualizations and measurements. In addition to providing such definitions and a viable framework for conducting spatio-textual affinity experiments on Twitter data, we provide detailed results illustrating how our framework can be used to compare and contrast two important metropolitan areas from multiple perspectives and granularities.

Keywords Urban informatics · Twitter · Social media analysis · Spatio-textual affinity · Embeddings

Mayank Kejriwal kejriwal@isi.edu

1 Information Sciences Institute, University of Southern California, Marina del Rey, Los Angeles, CA 90292, USA
Introduction

As summarized by Ritchie and Roser in 2018 [66] and revised in 2019, more than half of the global population currently live in urban areas, and increasingly so in cities that are highly dense. Such rapid and pervasive urbanization is a unique feature of modern life, never having been witnessed before to this extent in human history. Urbanization has profoundly influenced human activity and behavior, including the ways in which we live, work, commute and socialize [6, 7].

At the same time, the relative rise in technological fluency and adoption of novel technologies (especially among the younger populace), driven in no small part by the ubiquity and low cost of devices such as smartphones [70], as well as network externalities from growing social media platforms and services such as Twitter, Instagram, WhatsApp and WeChat [22, 59], has led to a new era for social science. The field of computational social science [3], which has emerged slowly but surely over the previous decade due to largely uncoordinated efforts in different fields, including urban studies, economics, sociology and network science, has generally been described as the ‘computational study of social phenomena’ and is primarily focused on emergent collective phenomena through analysis of interacting individual units (including both people and organizations) [12, 45]. With the availability of both public and proprietary datasets from governments, polling organizations (such as Gallup [21]) and enterprise (and non-profit) alike [41], computational social science has made great strides, especially in urban settings [4, 25, 71] (see also “Related work”). Foth et al. [26] describe the field of urban informatics as being ‘situated at the intersection of notions, trends and considerations for place, technology and people in urban environments’. The growth of social media, especially short, public messages from platforms such as Twitter, allows us to study all three of these features in painstaking detail using a range of visualization and machine learning technologies. While many advances have already been achieved in urban informatics, many questions also remain.

In this article, we study one such question, namely the tendency of people who are based in similar locations to discuss topically similar themes. We introduce computational spatio-textual affinity as an paradigm both for understanding urban differences in the way that people in specific urban areas (such as New York City and Los Angeles) communicate, as well as what they communicate. It is important to note that some differences can be explained due to inherent differences in the city’s geography, politics, national culture, language and even time zones. For example, it would not surprise us to know that tweets from Madrid are more likely to be in the Spanish language than tweets from the US Midwest, or that tweets from an electoral region that was known to favor a particular political candidate would express more favorable sentiment about that candidate than a region that was heavily opposed1. However, when studying two cities over an identical period of time, and that have similar traits with respect to language,

---

1 This is not to say that such questions are trivial or not important. The extent to which such social phenomena manifest on a social media platform, or among different demographics, is always interesting; furthermore, an unexpected finding can be fertile ground for new science, or a precursor to further phenomena that deviate from the norm.
politics and national affiliation, we ask the question: is it feasible to distinguish tweets in one city from another city based on their textual characteristics? Put another way, when we cluster tweets using (subsequently described) textual similarity measures, do clear spatial divisions emerge as a result?

For some tweets and topics, the answer would (presumably) be in the affirmative. Local events, including crimes, road closures and festivals, are all good examples: a local book festival in Los Angeles would not be expected to provoke conversation in New York. Even here, however, caution is called for. A shooting in Los Angeles would likely receive national press coverage and may become a topic of discussion everywhere. In general, due to homogenization of mass media and entertainment (such as Netflix) [57], and due to the special status of urban dwellers as ‘cosmopolitan’ consumers of such services [13], we would expect a lower degree of spatio-textual affinity between cities like New York and Los Angeles. While some of our experiments do bear out this conclusion, we provide key results that illustrate differences between the two cities, providing validation for the use of the framework when studying similar cities through the lens of similar data. Additionally, by studying distributions of textual and spatial affinity in aggregate, regional evidence of spatio-textual evidence is seen to emerge both at the levels of zipcodes and counties.

This article’s contribution to spatio-textual affinity is twofold. First, we present rigorous methodologies, designed for studies of multiple granularities, for measuring and studying both spatial and textual affinities. We do not rely on a single measure, but provide at least two ways to define either affinity. Second, we conduct a detailed empirical study in the context of two international metropolises (Los Angeles County and the Boroughs of New York City). Our study uses unbiased and geotagged data collected from the Twitter decahose (an approximately random 10% of tweets streaming at any given moment that has to be proprietarily obtained from Twitter) during a full month. By not relying on the public Twitter API for selective data acquisition or using automatic geolocation detection techniques, we mitigate a potential source of noise and bias that always raise questions of validity for such heavily empirical studies. We further maintain rigor by focusing on affinity as a correlational and non-causal phenomenon. We also use a judicious mix of qualitative methods, including visualizations and specific examples, to express some of the consequences of our findings and hypotheses.

The rest of this article is structured as follows. We start with a review of some related work in “Related work”, followed by a description of data and methodology in “Data and methodology” and the experimental study itself in “Experiments”. “Discussion” provides a discussion of some core findings, as well as extended findings that expose the scope for future research. “Conclusion and future work” concludes the article.
Related work

Urban informatics and social media

Quantitative studies in an urban context have been on the rise with the release of data, with commonly studied domains including demographics, transportation, and economics [5, 8, 65]. Use of GPS and cellular data has also been the subject of various studies, especially in the last decade [56, 64, 65]. In many works, a city is compared in the context of world city networks, where pairs of cities are linked based on indicators such as whether they share headquarters of some companies [16, 18, 73], whether both are part of the same production chains [8], Internet [53], financial services [5] and existence of direct flight or ship routes [19, 20, 77]. Regional differences within cities have also been studied e.g., Grauwin [33] studied differences between several major cities in the world, along with regional studies based on patterns of mobile phone call and SMS. Yuan et al. [76] study land use and human mobility within the city using GPS trajectory of taxis. Our work complements, and adds to, this emerging body of work by studying a specific concept (spatio-textual affinity) for a linked pair of coastal metropolises (Los Angeles and New York City).

In fact, increasing use of social media (see, for example, [50]) has led to a growing body of work in urban informatics [71], and unprecedented opportunities to study urban differences based on big data with a global footprint and human-generated content. Based on social media data, previous research is often revisited and some authors have illustrated divergence from previous studies on topics such as human mobility [24, 46, 55], social media activities [2, 25], lifestyle [37], emotional ‘maps’ [4], language usage [49] and word interpretation [38] with variables including region, occupation, age and gender. Hu et al. [38] combine modern Natural Language Processing (NLP) and social media methodologies for such studies by leveraging word embeddings that are arguably better suited for capturing underlying patterns such as grammar, synonyms and morphological relations in text. Using such embeddings, differences in word interpretations on topics like fashion and sports have been found between genders and regions (NYC and the Bay Area on the US West Coast). Studies that measure some kind of ‘spatial affinity, such as the work in [75] that groups similar cities across the world based on a ‘Local-Based Social Network’ collected from Foursquare (showing spatial affinity of cities in same clusters) have also been on the rise. In a similar vein in [61], cities are compared based on spatial distribution of amenities. [40] further utilize textual data and graphical models to surface content that is characteristic for a region. However, all of these studies have independently studied spatial or textual affinity, without attempting to bring them together in a common empirical framework. In this article, we address this issue using both spatial affinity and NLP methods, including word embeddings and clustering.
Sociology

As urbanization has expanded, urban studies have also taken on a more prominent role in sociology, including the field of urban sociology [32]. Over the last few decades, there has also been an increased focus on data-driven, empirical studies. A work that is, in principle, similar to the studies in this paper is the study of demographic dynamism and metropolitan change for four different cities in the US, including New York City and Los Angeles [52]. However, since social media (as we know it today) did not yet exist, any such study could not take into account ‘grass-roots’ conversations coming from people in those cities through millions of digital conversations. Therefore, to our knowledge, studies of spatio-textual affinity at scale have been non-existent. In other relevant work, urban differences have been viewed and studied through the lens of globalization and world-system theory by Knox et al. [42]. Cities in the context of globalization have been an object of increasing academic and policy-oriented scrutiny for almost four decades now [17, 28, 35, 72].

Domain-specific applications

Attempts have also been made to apply findings from urban informatics and social media analytics to domain-specific cases like policymaking, city-planning [11, 29, 54], computational epidemiology [68] and automatic geolocation detection [10, 23, 62, 74]. For example, Chenge et al. [10] and Wing et al. [74] utilize word frequency to model spatial variations and predict geolocations of tweets. Eisenstein et al. [23], Hong et al. [36] and Emre et al. [9] use generative models to identify features of geographical areas. Frias et al. [27] determines land use in urban areas by clustering regions with similar tweeting activity patterns. Based on detected land use, Preotuc-Pietro and Cohn [60] cluster users and reveals communities with different lifestyle and preference signatures. Using similar methodologies, Sadilek et al. [68] models the associations between lifestyle and health conditions. While we do not focus on applications of spatio-textual affinity in this article, it could serve as a valuable feature in many of the applications above. For example, the positive evidence that we find in favor of spatio-textual affinity could help improve accuracy of geolocation detection systems using some of our findings in a weak supervision setting [48, 63].

Data and methodology

To explore our hypotheses on spatio-textual affinity empirically, we used social media data from the New York and Los Angeles metropolitan areas collected from the Twitter decahose (understood to be an approximately random 10% sample from globally streaming Twitter data that must be purchased directly from Twitter) over
all 31 days of the month of October, 2016. To remove any ambiguities concerning location, we filtered and used only the subset of the English tweets that had geotags embedded in the metadata (usually because the GPS is on and the tweeter has allowed the information to be embedded in the tweet metadata) and where the geotags could be triangulated to either lie in Los Angeles County or in the Boroughs of New York City (the Bronx, Brooklyn, Manhattan, Queens, and Staten Island). 2 Henceforth, we succinctly refer to these metropolitan areas as Los Angeles (LA) and New York City (NYC), respectively.

Besides their cultural significance, another good reason for choosing these two areas for exploring our hypotheses on spatio-textual affinity is that they are generally taken to be more similar than different. Politically, both are known to be liberal and lean heavily (much higher than the national average) towards the Democratic Party [51]. Both also have a healthy population of Twitter users; our data contains 247,068 tweets with geotags from LA and 197,854 tweets from NYC. Based on Fig. 1, we observe major ‘hotspots’ of twitter prevalence and density in areas such as Hollywood and downtown LA (Fig. 1a), and Manhattan and Long Island (Fig. 1b). However, there are also many smaller clusters in both plots.

Despite the similarities (primarily the fact that both cities are big and international metropolises), there are also many differences between them, not the least of which is the spatial dispersion and density of the city’s residents. In the recent era, the COVID-19 crisis showed how dense and interconnected NYC was, since the disease spread quickly through the city in the earliest days of the crisis [44]. In

2 Reverse geocoding of the geotags to the county/borough in this article is based on the Census County shapefile https://catalog.data.gov/dataset/tiger-line-shapefile-2017-nation-u-s-current-county-and-equivalent-national-shapefile.
popular culture, the two cities also exhibit differences [14]. A range of other studies have also paired New York City and Los Angeles [1, 15, 30, 34, 47]; hence, there is precedent for selecting these two cities for exploring the concept of spatio-textual affinity.

To study spatio-textual affinity, we have to precisely define both how to measure textual similarity and spatial similarity. For the latter, an established measure is the Haversine distance between the geotags (a latitude–longitude coordinate) of two tweets [69]. Specifically, given two latitude–longitude coordinates \((\text{lat}_1, \text{long}_1)\) and \((\text{lat}_2, \text{long}_2)\) in radians, we denote two variables \(d_1\) and \(d_2\) as \((\text{lat}_1 - \text{long}_1)\) and \((\text{lat}_2 - \text{long}_2)\), respectively. The Earth’s radius is given by the symbol \(r\). With these notations in place, the haversine distance \(HD\) between two latitude–longitude coordinates is defined as

\[
HD = 2 \times r \times \arcsin\left[\sqrt{\sin^2(d_1/2) + \cos(\text{lat}_1) \cos(\text{long}_1) \sin^2(d_2/2)}\right]
\]  

In some studies, a more city-specific version of spatial affinity is also considered, since the Haversine distance can sometimes be less meaningful due to the different densities of the two cities. Specifically, we consider zipcodes and counties: two tweets are deemed to have spatial affinity of 1 if they have the same zipcode or county (depending on the measure used) and 0 otherwise. In a similar vein, the coarsest-grained measure is city membership: two tweets are assumed to have spatial affinity of 1 if both are geolocated in the same city (and 0 otherwise). Despite their coarse granularity, we show later that these measures can be useful for macroscopic analysis utilizing many tweets.

Textual similarity is a more complicated and subjective matter, especially if automated methods from natural language processing (such as word embeddings [39]) are considered. We consider two separate methodologies for measuring textual similarity. The ‘conservative’ methodology is to only consider topics that are explicitly and unambiguously declared by the tweeter through hashtags. To avoid topical overlap and ambiguity, we limit our corpus in hashtag-based experiments to tweets that have exactly one hashtag. Tweets that share a common hashtag are judged to be similar, otherwise dissimilar. Applying this single-hashtag constraint yields a filtered set of 34,047 tweets and a total of 17,243 unique hashtags. Note that this approach is not only conservative, but also binary (similar to the coarse-grained spatial affinity measures): either a pair of tweets is judged to be similar or not similar. In “Experiments”, we refer to this methodology as the hashtag-based methodology for computing textual affinity.

This conservative methodology, while likely to lead to results that are more precise and unambiguous, also discards much of the data, since many tweets may have more than one hashtag or even no hashtags. Its binary nature also precludes more interesting and nuanced studies, especially since it does not give any weight to the actual content of the tweets (the words) besides the lone hashtag. Therefore, we consider a second methodology, namely subword-based embeddings embodied by the fastText package released by Facebook Research [39]. These
embeddings, which can be used to convert words and entire sentences (and hence, tweets) into dense, continuous, real-valued vectors are particularly apt for handling the kinds of spelling variations common on social media. We used the publicly available embedding described in [31] that had been previously trained on 400 million Twitter English tweets and found to yield excellent performance on multiple NLP tasks.

Using the embedding methodology, the textual affinity can be computed by computing the cosine similarity between the vector representations of the two tweets. Given two vectors $v_1$ and $v_2$, the cosine similarity between them is

$$\frac{v_1 \cdot v_2}{\|v_1\| \times \|v_2\|}.$$  

It is an established way in the NLP community for measuring textual affinity in a more nuanced way. In “Experiments”, we refer to this methodology as the embedding-based methodology for computing textual affinity.

**Experiments**

**Measuring spatio-textual affinity: hashtag-based methodology**

As a first step, we study spatio-textual affinity using the hashtag-based methodology described in the previous section. Specifically, we group our tweet collection by hashtag, and then calculate, for each such cluster, the ratio $\text{ratio}_{LA}$ of tweets that fall within the LA metropolitan area.

In total, there are 17,239 clusters, since there are 17,239 unique hashtags. In Fig. 2, we plot the cluster–frequency ratio $\text{ratio}_{LA}$. We find a significant topical separation between LA and NYC, with about 16,089 clusters being relatively ‘local’ ($0.8 \leq \text{ratio}_{LA}$ or $\text{ratio}_{LA} \leq 0.2$), and with only about 1150 clusters having a

---

3 Recall from the previous section that only tweets with exactly one hashtag were considered minimize potential topical ambiguity.
'moderate’ mix of participants from both cities (\(0.8 \geq \text{ratio}_{\text{LA}} \geq 0.2\)). This is perhaps the most direct evidence of clear spatio-textual affinity.

Further analysis showed that most of these hashtags have a long-tailed frequency distribution themselves, with most having been used only once in our considered time period. To better understand the spatial relationships of these hashtags, and the common trends across the two metropolitan areas, we plot the top 20 most frequently occurring hashtags in LA and NYC in Figs. 3 and 4, respectively. From these figures, we find that hashtags such as ‘job’, ‘hiring’ and ‘CareerArc’ have similarly high ranks in both cities, which is understandable, since the economy and the workplace are predominant areas of conversation in any large-scale social setting.
However, some evidence of location-specificity starts to emerge even in these top 20 hashtags, since some are related to locations (e.g., ‘NYC’, ‘Torrance’ and ‘Longbeach’). Furthermore, other hashtags may not directly pertain to location, but seem to have high ranking in one metropolis but not the other. For example, ‘diabetes’ is present in NYC but not Los Angeles (although ‘Healthcare’ is) and ‘VPDebate’ is present in both top 20 lists but at relatively different ranks. To obtain an independent trend for these topics, we also plot hashtag usage (for some hashtags) using Google Trends (Fig. 5). Similarities and differences between LA and NYC are clearly manifested, e.g., ‘clinton’ has wide disparity between both cities, while ‘healthcare’ is nearly coincidental and ‘job’ is highly synchronous. Therefore, the google trends
support the Twitter results, permitting us to verify (at least in part) the empirical validity of our experiments.

Even though the hashtag clustering is conservative, it can shed light on conversations on social media that are national in scope, spanning cities and regions. To obtain these insights, we conduct the following experiment that utilizes both

Fig. 5 Daily Google Trends index for selected keywords over October, 2016
hashtags and sentence embedding. Recall that there were 1150 hashtag-based clusters with a ‘moderate’ mix of participants. For each such cluster within the set of clusters \( C = \{t_1, \ldots, t_{1150}\} \), respectively, with \(|C_k|\) tweets within the \(k\)th cluster, we partition the pairwise set\(^4\) derived from that cluster into three sub-clusters: \(C_{\text{LA}}\), \(C_{\text{NY}}\) and \(C_{\text{hetero}}\). \(C_{\text{LA}}\) and \(C_{\text{NY}}\) each contain those pairs, where both tweets in the pair are (respectively) from LA and New York. \(C_{\text{hetero}}\) contains the remainder, i.e., (by definition) all pairs, where one tweet each is from LA and New York region. To control for the fact that tweets may be coming from the same user, we remove all pair of tweets from the same users in these sub-clusters for this experiment.

Next, for each sub-cluster, we compute a textual affinity distribution in embedding space by computing the cosine distance between the embeddings of the tweets in the pair.\(^5\) Since there are always three non-empty\(^6\) sub-clusters for each cluster (i.e., hashtag), we get three distributions that we can plot. Figure 6 provides illustrative

---

\(^4\) There are \(C_k^2\) pairs in the pairwise set, since we do not distinguish the order in the pairs.

\(^5\) In terms of the spatio-textual affinity framework, textual affinity is a combination of hashtag clustering and embedding-based cosine similarity (as just described), while spatial affinity is simpler and binary (same-city membership).

\(^6\) Guaranteed, since \(\text{ratio}_{\text{LA}}\) is between 0.2 and 0.8 in these ‘mixed’ clusters, by definition.
results for four such hashtags. Intriguingly, we find inter-city differences in these distributions even though they are well-mixed. #job and #TheWalkingDead, for example, have distributions with non-trivial non-overlap, while #debate and #Trump (normally considered polarizing topics) have almost identical overlap, which may (hypothetically) be due to the similar political affiliations in both cities. Furthermore, from the trend of daily hashtag usage shown in Fig. 7, we obtain a similar conclusion as in Fig. 6 and (earlier) in Fig. 5: some hashtags have more distinctive
spatial signatures (i.e., are more predictive of, and tied to, their locations) than others, despite seeming like national or pop-culture topics on the surface.

Measuring spatio-textual affinity: embedding-based methodology

To better address the brittleness that can be caused by hashtag-only clustering, we conduct a deeper study of spatio-textual affinity of tweets using k-means clustering on the tweet embeddings. Since k-means assumes $k$ (the number of clusters) as given, we used the established elbow method to obtain an optimal $k$ value of 1000 [43]. Once we obtain the clusters, we compute the distribution of clusters based on $\text{ratio}_{\text{LA}}$, analogous to the distribution in Fig. 2 for hashtag-based clustering. As shown in Fig. 8, the resulting distribution, while still having localized peaks at the extremes, now has a more robust mixture distribution compared with Fig. 2, with many more clusters (even as a ratio of the total number of clusters having moderate $\text{ratio}_{\text{LA}}$ values).

Considering the significant change that manifested only when taking a finer-grained view of textual similarity (continuous-space embeddings versus discrete
hashtags), the question arises as to whether a similar change could be observed by taking a finer-grained view of spatial similarity inside LA and NYC regions rather than just classifying tweets as being from LA or NYC. To this end, we use the Haversine distance presented in Eq. 1 for the next set of experiments.

First, for the 1000 embedding-based clusters of which the $\text{ratio}_{\text{LA}}$ distribution is plot in Fig. 8, we took the 70 clusters with $\text{ratio}_{\text{LA}} \geq 0.8$. The union $C = \bigcup_{i=1}^{70} C_i$ of these clusters contains 15,320 (=$|C|$) tweets in total. We start by computing the Haversine distance $\text{dist}(t_i, t_j)$ (using the lat-long coordinates in $t_i$ and $t_j$) between the two tweets in every possible pair of tweets from LA. Next, we partition the distances into two distributions $\text{in}$ and $\text{between}$, where $\text{dist}(\{t_i, t_j\})$ is placed into $\text{in}$ if the tweets $t_i$ and $t_j$ fall within the same cluster, and placed into $\text{between}$ otherwise. We plot these two sets of distances as probability distributions in Fig. 9 using a Kernel Density Estimation (KDE) algorithm [58, 67]. Although visually apparent in the figure, we also used the Chi-squared test (specifically, by computing the D-statistic) to reject the null hypothesis at the 99.99% confidence level that the two distributions are the same (statistically). This difference between the plots indicates a high degree of spatio-textual affinity: simply controlling for the city (i.e., high $\text{ratio}_{\text{LA}}$) is not enough to yield strong affinity, since whether two tweets have the same cluster membership (in the fine-grained text embedding space) clearly matters.

More evidence of spatio-textual affinity is indicated when we repeat the experiment for the ‘mixed’ clusters, i.e., those clusters with $\text{ratio}_{\text{LA}}$ between 0.2 and 0.8, by again plotting the $\text{in}$ and $\text{between}$ distributions in Fig. 10. Despite the visual suggestion in the figure, we find that even in this case we can reject the null hypothesis that the two distributions are the same, but the D-statistic ($D = 0.0119$) is very different (almost 15× lower) from that of the experiment above ($D = 0.1754$). Hence, spatio-textual affinity declines considerably as we start ‘loosening’ spatial constraints (as we do here by only considering mixed-location clusters). In other words,

---

7 We believe that this is also an artifact of the relatively large numbers of points in our distribution, since even small differences can become significant in large datasets.
both spatial and textual dimensions make strong positive contributions in measurements of spatio-textual affinity.

While not shown here, the conclusion for NYC was found to be the same: for the set of clusters with low ratio$_{LA}$ (and hence, high concentration of tweets from NY), a similar finding as in Fig. 9 was observed.

We further explore affinity observed above on a more intuitive level of geographical region rather than at the fine-grained level of latitude–longitude coordinates or at the highly coarse-grained level of the entire city. For the next set of experiments, therefore, we choose the zipcode, city and county as our spatial affinity variables, not dissimilar to using the hashtag earlier as the textual affinity variable. Namely, if two tweets have the same value for either the city, zipcode or county (depending on the experiment) they have a spatial affinity of 1, and 0 otherwise. Formally, we can use the Kronecker Delta Function (KDF) to express this notion. Given a function $\mathcal{F}$:

$$
\delta^{\mathcal{F}}_{ij} = \begin{cases} 
0 & \text{if } \mathcal{F}(i) \neq \mathcal{F}(j), \\
1 & \text{if } \mathcal{F}(i) = \mathcal{F}(j).
\end{cases}
$$

(2)

Here, $\delta^{\mathcal{F}}_{ij}$ is the KDF. We compute three types of KDFs: $\delta^{\text{zipcode}}_{ij}$, $\delta^{\text{city}}_{ij}$ and $\delta^{\text{county}}_{ij}$. For example, $\delta^{\text{zipcode}}_{ij} = 1$ for precisely those (unordered) pairs of tweets $\{t_i, t_j\}$ that have the same zipcode.

Furthermore, since these are nominal variables, $\chi^2$ test and Cramer’s V value are used to determine the significance of the difference in distributions of tweet pairs falling within ‘in’ or ‘between’ clusters in the same vein as in Figs. 9 and 10. The Cramer’s V value is defined by

$$
V = \sqrt{\frac{\chi^2/n}{\min(c - 1, r - 1)}}
$$

(3)

Here, $\chi^2$ is derived from Chi-squared test, $n$ is the total of samples, and $c$ and $r$ are (respectively) the number of columns and rows (the respective value set observed for the two nominal variables). Cramer’s V ranges from 0 to 1, with 0 corresponding to no association between the variables and 1 corresponding to complete association (Cramer’s V can reach 1 only when each variable is completely determined by the other).

We compute the three KDFs ($\delta^{\text{zipcode}}_{ij}$, $\delta^{\text{city}}_{ij}$ and $\delta^{\text{county}}_{ij}$) for all possible pairs of tweets from LA and NYC in both local and mixed clusters, where zipcode($i$), city($i$) and county($i$) correspond, respectively, to the zipcode, city and county that tweet $t_i$ belongs to. For each type of function and cluster, we partition the KDFs into two distributions in and between, depending again on the condition imposed on ratio$_{LA}$ (analogous to the in and between distributions in Figs. 9 and 10). The counts, specifically for all LA tweets, are tabulated in Table 1. Note that computing all pairs of tweets from mixed clusters is not computationally viable due to the large number of tweets that fall in mixed clusters compared to local clusters; hence, for each mixed cluster, we randomly subsample 1000 tweets (if there are more than 1000 tweets in
the cluster), to obtain approximately $C_{2}^{1000}$ pairs (that could be further sub-divided into ‘in’ and ‘between’ pairs).

The counts show that, in local clusters, there is higher count of tweet pairs in the condition $\delta_{ij}^{\text{zipcode}} = 0$ (compared to condition $\delta_{ij}^{\text{zipcode}} = 1$), but the situation reverses in mixed clusters, showing that spatio-textual affinity strongly asserts itself at the more macroscopic level rather than at the microscopic level. However, at the county level, spatio-textual affinity is completely dominant, i.e., both for local and mixed clusters, there is much higher count of tweet pairs for the condition $\delta_{ij}^{\text{zipcode}} = 1$.

Concerning significance (as well as analysis of the results in the context of both NY and LA clusters), we tabulate the Cramer’s V values in Table 2. One nominal variable is the KDF, while the other variable simply records whether two tweets (an observation for the purposes of this experiment is always a pair of tweets) are in an ‘in’ condition (belong to the same cluster) or the ‘between’ condition (belong to two different clusters). The meanings of ‘local’ and ‘mixed’ clusters are the same as earlier (depending on the $\text{ratio}_{\text{LA}}$), but we break out local clusters further based on whether the clusters are NYC or LA clusters. While all values are low in the table, Cramer’s V of local clusters are about 10x larger than that of mixed clusters, indicating stronger spatio-textual affinity in local clusters compared to mixed clusters. As

Table 1 Counts of pairs of tweets on each Kronecker Delta Function (KDF) setting. ‘In’ (‘between’) refers to the case when both tweets belong to the same (different) embedding-based cluster. ‘Local’ (‘mixed’) refers to clusters that have $\text{ratio}_{\text{LA}} \geq 0.8$ ($0.2 \leq \text{ratio}_{\text{LA}} \leq 0.8$). Tweets are never paired across ‘local’ and ‘mixed’ clusters

| Spatio-textual affinity on zipcode level | $\delta_{ij}^{\text{zipcode}} = 1$ | $\delta_{ij}^{\text{zipcode}} = 0$ |
|---|---|---|
| In local LA clusters | 346,968 | 3,971,972 |
| Between local LA clusters | 2,651,037 | 97,683,069 |
| In mixed clusters | 13,783,819 | 46,047,860 |
| Between mixed clusters | 37,151,247 | 127,897,592 |

| Spatio-textual affinity on city level | $\delta_{ij}^{\text{city}} = 1$ | $\delta_{ij}^{\text{city}} = 0$ |
|---|---|---|
| In local LA clusters | 1,237,258 | 3,081,682 |
| Between local LA clusters | 11,543,638 | 88,790,468 |
| In mixed clusters | 14,587,640 | 45,244,039 |
| Between mixed clusters | 39,000,613 | 126,048,226 |

| Spatio-textual affinity on county level | $\delta_{ij}^{\text{county}} = 1$ | $\delta_{ij}^{\text{county}} = 0$ |
|---|---|---|
| In local LA clusters | 4,297,336 | 21,604 |
| Between LA clusters | 99,675,204 | 658,902 |
| In mixed clusters | 58,605,826 | 1,225,853 |
| Between mixed clusters | 161,650,749 | 3,398,090 |
Table 2 Two variables involved in the Cramer’s V calculations are ‘in/between’ and the KDF in each row. For an explanation of ‘In/Between’, see caption of Table 1

Cramer’s V result for NY tweets

| Experiment                                      | Cramer’s V |
|-------------------------------------------------|------------|
| In/Between NY clusters vs. $\delta_{ij}^{\text{zipcode}}$ | 0.1125***  |
| In/Between mixed clusters vs. $\delta_{ij}^{\text{zipcode}}$ | 0.0175***  |
| In/Between NY clusters vs. $\delta_{ij}^{\text{city}}$   | 0.0627***  |
| In/Between mixed clusters vs. $\delta_{ij}^{\text{city}}$ | 0.0077***  |
| In/Between NY clusters vs. $\delta_{ij}^{\text{county}}$ | 0.0626***  |
| In/Between mixed clusters vs. $\delta_{ij}^{\text{county}}$ | 0.0079***  |

Cramer’s V result for LA tweets

| Experiment                                      | Cramer’s V |
|-------------------------------------------------|------------|
| In/Between LA clusters vs. $\delta_{ij}^{\text{zipcode}}$ | 0.0056**   |
| In/Between mixed clusters vs. $\delta_{ij}^{\text{zipcode}}$ | 0.0642***  |
| In/Between LA clusters vs. $\delta_{ij}^{\text{city}}$   | 0.1041***  |
| In/Between mixed clusters vs. $\delta_{ij}^{\text{city}}$ | 0.0078***  |
| In/Between LA clusters vs. $\delta_{ij}^{\text{county}}$ | 0.0038***  |
| In/Between mixed clusters vs. $\delta_{ij}^{\text{county}}$ | 0.0003***  |

NY and LA clusters, respectively, refer to clusters with \(\text{ratio}_{LA} \leq 0.2\) and \(\text{ratio}_{LA} \geq 0.8\), respectively.  

***p value is less than 0.001

Fig. 11 Spatial distribution of geotagged tweets from LA in local and mixed clusters (different color represents different clusters and each point represents a single tweet)
Discussion

The statistical results and visualizations in the previous section show that while different levels of affinity exists in local and mixed cluster, far greater affinity exists in local clusters. Interestingly we can see that affinity is the strongest for LA clusters when measured on the city level, while affinity is strongest for NY clusters when measured on the zipcode level.

To further understand the reason why statistically significant differences in affinity exist between NY/LA-dominant clusters and mixed clusters, we looked into the content of selected tweets from these clusters. In mixed clusters (with ratio\textsubscript{LA} between 0.2 and 0.8), many clusters tend to have general topics shared by both LA and NYC e.g. political tweets such as "@realDonaldTrump @HillaryClinton HRC talk about important issues like jobs I security health care Got nothing on Trump to move the needle", but also national sports e.g.,"@chedsy22 Cool. I will root for the Giants big time if they play the Dodgers or Nationals.". In contrast, in local clusters (with ratio\textsubscript{LA} ≤ 0.2 or ratio\textsubscript{LA} ≥ 0.8) many topics are (understandably) discussing local events, e.g., "Awesome! RickAndMorty NYCC @Javits Center" and "Halloween Horror Nights 2016 🎃 😈 #GetScaredEasily #AlmostShittedMyself # BarelyMadeItOutAlive?". Local clusters are also more likely to include notifications for local incidents and information, e.g. local crimes: "Incident on #QM2Bus EB from 6th Avenue: 34th Street to 6th Avenue: 59th Street https://t.co/RmU4nfTP4E" and job opportunities: We’re #hiring!
These qualitative results lend further support to the quantitative data discussed earlier.

In an attempt to gain further insight into how regional variables could affect spatial distribution of tweets, we looked into the relationship between crime rates and tweet prevalence rates (the number of tweets geotagged in a zipcode divided by the population of that zipcode). Fig. 13 illustrated a somewhat surprising result on a preliminary experiment conducted for NYC tweets: when zipcode populations are controlled, there is actually a positive correlation between the tweet prevalence rate and the crime rate, with weaker correlation observed for LA. Further study of this result is warranted and may involve going deeper into the tweets of high crime/
capital zipcodes to understand the nature of such tweets. Since there is high spatio-textual affinity at the zipcode level for NYC (Table 2), sampling a few hundred tweets (or hashtags) for the high crime/capita zipcodes may be all that is necessary to get the ‘pulse’ of the neighborhood (at least approximately). It is important to note that while a positive correlation would have been expected if the variables had not been per capita (since the population would have been the obvious explanation for an upward contemporaneous trend in both variables), the interesting aspect about the plot on the left in Fig. 13 is the upward trend even when controlling for population. The LA plot seems more normal and expected in this regard.

Most likely there are other variables in the NYC case that could serve as explanations. For example, if there are more nightclubs in the high-density neighborhoods in NYC, then that may partially explain both high crime/capita and high tweet prevalence. The same may not be true for LA, which is much less dense and far more spread-out than NYC, as even earlier plots (such as Fig. 1) have illustrated. Another possible explanation is the unemployment rate. Earlier, we showed that there were hashtags in the dataset related to jobs (including #job itself). Since high crime and unemployment are themselves correlated, it may be that there is a positive correlation between tweets per capita and between the unemployment rate, both measured at a per-zipcode level, just like in Fig. 13. We plot this data in Fig. 14 and find that there is some evidence for this. There is a positive correlation for both cities, and the differences between the two cities seem to have been reduced. Interestingly, therefore, the tweets per capita could be used as an informal, real-time barometer for measuring changes in phenomena such as the unemployment rate without always relying on surveys, especially at the fine-grained level of zipcodes in large cities such as NYC and LA.

As another qualitative study illustrating city-specific dependencies on tweeting behavior, we also plot the number of tweets sent out each hour in local time. The results in Fig. 15 illustrate the clear lags in the orange peak showing that NYC tweeters tend to reach their peak in late evening, even though LA sent out more tweets (at least within the full month of the data). We are planning to rigorously study such differences on a larger scale in the future. For example, using our textual affinity variables, we can study not only whether, after controlling for the time zone

Fig. 15 Number of geotagged tweets for each hour of a day in local time (Pacific Time for Los Angeles and Eastern Time for New York) over the entire period of the dataset (October, 2016)
differences, the two cities are tweeting at the same rates during the same local hours (which, according to Fig. 15, they are not) but also whether they are tweeting about the same things or topics.

**Conclusion and future work**

In this article, we developed spatio-textual affinity as a comparative paradigm for profiling geotagged citizen social media, and understanding the differences between dense and diverse urban metropolises through such comparative profiles. To illustrate the promise of this paradigm, especially at the scale of hundreds of thousands of tweets, we collected a dataset that contains an unbiased sample of tweets across the New York and Los Angeles metropolitan areas within a reasonably compressed time period in the United States. Our dataset has a high degree of control due to low variation between the cities on several important factors that are known to heavily influence social media content (including politics, demographics and nationalities). Within our dataset, we found clear evidence of spatio-textual affinity. Our findings are reasonably robust to the specific measures of affinity adopted for both the spatial and textual dimensions.

However, we re-iterate that it is important to consider these results in their context, and not as absolute or causal truths. Like many urban computing and computational social studies, the data was collected through observation rather than intervention. However, the volume of the data, consistency among the results both when using hashtag-based and embedding-based textual similarity, and the mechanisms of control that we described in “Data and methodology”, lend some measure of credibility to the use of spatio-textual affinity as a framework for understanding urban social media differences, even between two otherwise similar cities.

Our studies reveal that there are multiple promising avenues for future research. In particular, the more qualitative results in “Discussion” can be formalized further, not dissimilar to the studies in “Experiments”. Other such studies of a similar nature can also be designed and investigated. Investigating similar questions in the context of other cities both within the US, but also in other countries, is an issue that future research could investigate. In particular, the question remains open whether similar findings will be observed when we consider cities that are big but politically disparate (such as New York and Houston), cities that are dissimilar in size, and cities that fall within different countries. Multi-lingual studies could also be considered, but may be more challenging to calibrate without adequate translation of Twitter lingo. Replicating such studies in non-Twitter social media may also yield interesting empirical findings. Finally, replicating such a study in the context of COVID-19 may yield even more useful insights, since most people are currently working from home and not commuting, potentially leading to more robust estimates of spatial affinity.

**Acknowledgements** Not applicable.
Author contributions Hu primarily contributed to the design and execution of the experimental study, and statistical analysis, while Kejriwal was responsible for writing, editing and ideation.

Funding This research was done pro-bono and has not received independent funding.

Declarations

Conflict of interest The authors declare that they have no conflict of interest.

Availability of data and materials The underlying data available for these studies was acquired from the Twitter firehose during the month of October, 2016, and then pre-processed so that only geotagged tweets within the particular cities considered in this article were retained. The authors pledge to make the tweet-IDs available on a public portal.

Code availability All code and software that has been used in support of this paper are from established open-source packages. We have provided links and references in the article where applicable.

References

1. Abu-Lughod, J.L., et al.: Race, space, and riots in Chicago, New York, and Los Angeles. Oxford University Press, New York (2007).
2. Adnan, M., Leak, A., Longley, P.: A geocomputational analysis of twitter activity around different world cities. Geo-spatial Information Science 17(3), 145–152 (2014).
3. Alvarez, R. M. (2016). Computational social science. Cambridge: Cambridge University Press
4. Ashkezari-Toussi, S., Kamel, M., Sadoghi-Yazdi, H.: Emotional maps based on social networks data to analyze cities emotional structure and measure their emotional similarity. Cities 86, 113–124 (2019).
5. Bassens, D., Derudder, B., Witlox, F.: Searching for the mecca of finance: Islamic financial services and the world city network. Area 42(1), 35–46 (2010).
6. Berry, B.J.: Urbanization. In: Urban ecology, pp. 25–48. New YorK: Springer (2008)
7. Boustan, L. P., Bunten, D. M., & Hearey, O. (2013). Urbanization in the united states, 1800–2000. National Bureau of Economic Research, Technical report
8. Brown, E., Derudder, B., Parnreiter, C., Pelupessy, W., Taylor, P.J., Witlox, F.: World city networks and global commodity chains: towards a world-systems’ integration. Global Networks 10(1), 12–34 (2010).
9. Çelikten, E., Le Falher, G., Mathioudakis, M.: Modeling urban behavior by mining geotagged social data. IEEE Transactions on Big Data 3(2), 220–233 (2016).
10. Cheng, Z., Caverlee, J., Lee, K.: You are where you tweet: a content-based approach to geo-locating twitter users. In: Proceedings of the 19th ACM international conference on information and knowledge management, pp. 759–768 (2010)
11. Ciuccarelli, P., Lupi, G., Simeone, L.: Visualizing the data city: social media as a source of knowledge for urban planning and management. Springer Science & Business Media (2014)
12. Conte, R., Gilbert, N., Bonelli, G., Cioffi-Revilla, C., Deffuant, G., Kertesz, J., Loreto, V., Moat, S., Nadal, J.P., Sanchez, A., et al.: Manifesto of computational social science. The European Physical Journal Special Topics 214(1), 325–346 (2012).
13. Corpus Ong, J.: The cosmopolitan continuum: locating cosmopolitanism in media and cultural studies. Media, Culture & Society 31(3), 449–466 (2009).
14. Currid, E., Williams, S.: The geography of buzz: art, culture and the social milieu in Los Angeles and New York. Journal of Economic Geography 10(3), 423–451 (2010).
15. Currid, E., Williams, S.: Two cities, five industries: Similarities and differences within and between cultural industries in New York and Los Angeles. Journal of Planning Education and Research 29(3), 322–335 (2010).
16. Derudder, B., Witlox, F., Catalano, G.: Hierarchical tendencies and regional patterns in the world city network: a global urban analysis of 234 cities. Regional Studies 37(9), 875–886 (2003).

17. Derudder, B.: International handbook of globalization and world cities. Cheltenhem: Edward Elgar Publishing (2012).

18. Derudder, B., Witlox, F.: Assessing central places in a global age: on the networked localization strategies of advanced producer services. Journal of Retailing and Consumer Services 11(3), 171–180 (2004).

19. Derudder, B., Witlox, F.: On the use of inadequate airline data in mappings of a global urban system. Journal of Air Transport Management 11(4), 231–237 (2005).

20. Derudder, B., Witlox, F.: Mapping world city networks through airline flows: context, relevance, and problems. Journal of Transport Geography 16(5), 305–312 (2008).

21. Diener, E., Tay, L.: Subjective well-being and human welfare around the world as reflected in the Gallup world poll. International Journal of Psychology 50(2), 135–149 (2015).

22. Duggan, M., Ellison, N.B., Lampe, C., Lenhart, A., Madden, M.: Social media update 2014. Pew Research Center 19, 1–2 (2015).

23. Eisenstein, J., O’Connor, B., Smith, N.A., Xing, E.P.: A latent variable model for geographic lexical variation. In: Proceedings of the 2010 conference on empirical methods in natural language processing, pp. 1277–1287. Stroudsburg: Association for Computational Linguistics (2010).

24. Ferrara, E., Varol, O., Menczer, F., Flammini, A.: Traveling trends: social butterflies or frequent fliers? In: Proceedings of the first ACM conference on Online social networks, pp. 213–222 (2013).

25. Förster, T., Lamerz, L., Mainka, A., Peters, I.: The tweet and the city: Comparing twitter activities in informational world cities. In: Proceedings of the 3rd DGI Conference, pp. 101–118 (2014).

26. Foth, M., Choi, J.H., Satchell, C.: Urban informatics. In: Proceedings of the ACM 2011 conference on Computer supported cooperative work, pp. 1–8 (2011).

27. Frias-Martinez, V., Frias-Martinez, E.: Spectral clustering for sensing urban land use using twitter activity. Engineering Applications of Artificial Intelligence 35, 237–245 (2014).

28. Friedmann, J., Wolff, G.: World city formation: an agenda for research and action. International Journal of Urban and Regional Research 6(3), 309–344 (1982).

29. Giatsoglou, M., Chatzakou, D., Gkatziaki, V., Vakali, A., Anthopoulos, L.: Citypulse: A platform prototype for smart city social data mining. Journal of the Knowledge Economy 7(2), 344–372 (2016).

30. Gladstone, D.L., Fainstein, S.S.: Tourism in us global cities: a comparison of New york and Los Angeles. Journal of Urban Affairs 23(1), 23–40 (2001).

31. Godin, F.: Improving and interpreting neural networks for word-level prediction tasks in natural language processing. Ph.D. thesis, Ghent University, Belgium, 2019. 35 (2019).

32. Gottdiener, M.: New urban sociology. The Wiley Blackwell Encyclopedia of Urban and Regional Studies, pp. 1–5 (2019).

33. Grauwin, S., Sobolevsky, S., Moritz, S., Gődör, I., Ratti, C.: Towards a comparative science of cities: Using mobile traffic records in New York, London, and Hong Kong. In: Computational approaches for urban environments, pp. 363–387. New York: Springer (2015).

34. Halle, D., et al. (2003). New York and Los Angeles: politics, society, and culture—a comparative view. Chicago: University of Chicago Press.

35. Harrison, J., Hoyler, M.: Megaregions: globalization s new urban form? Cheltenham: Edward Elgar Publishing (2015).

36. Hong, L., Ahmed, A., Gurumurthy, S., Smola, A.J., Tsioutsiouliklis, K.: Discovering geographical topics in the twitter stream. In: Proceedings of the 21st international conference on World Wide Web, pp. 769–778 (2012).

37. Hu, T., Bigelow, E., Luo, J., Kautz, H.: Tales of two cities: Using social media to understand idiosyncratic lifestyles in distinctive metropolitan areas. IEEE Transactions on Big Data 3(1), 55–66 (2016).

38. Hu, T., Song, R., Abbarian, M., Ding, P., Xie, X., Luo, J.: A world of difference: Divergent word interpretations among people. In: Eleventh international AAAI conference on web and social media (2017).

39. Joulina, A., Grave, E., Bojanowski, P., Mikolov, T.: Bag of tricks for efficient text classification. arXiv preprint arXiv:1607.01759 (2016).

40. Kafsi, M., Cramer, H., Thomee, B., Shamma, D.A.: Describing and understanding neighborhood characteristics through online social media. In: Proceedings of the 24th international conference on world wide web, pp. 549–559 (2015).

41. Kitchin, R. (2014). The data revolution: Big data, open data, data infrastructures and their consequences. London: Sage.
42. Knox, P. L., Knox, P. L., Knox, P. L., & Taylor, P. J. (1995). *World cities in a world-system*. Cambridge University Press.

43. Kodinariya, T.M., Makwana, P.R.: Review on determining number of cluster in k-means clustering. International Journal 1(6), 90–95 (2013).

44. Kuchler, T., Russel, D., Stroebel, J.: The geographic spread of covid-19 correlates with structure of social networks as measured by facebook. Technical report, National Bureau of Economic Research (2020).

45. Lazer, D., Pentland, A., Adamic, L., Aral, S., Barabasi, A.L., Brewer, D., Christakis, N., Contractor, N., Fowler, J., Gutmann, M., et al.: Social science. computational social science. Science (New York, NY) 323(5915), 721–723 (2009).

46. Lenormand, M., Gonçalves, B., Tugores, A., Ramasco, J.J.: Human diffusion and city influence. Journal of The Royal Society Interface 12(109), 20150473 (2015).

47. Logan, J.R., Zhang, W., Alba, R.D.: Immigrant enclaves and ethnic communities in ew York and Los Angeles. American Sociological Review, 67, 299–322 (2002).

48. Meng, Y., Shen, J., Zhang, C., Han, J.: Weakly-supervised hierarchical text classification. In: Proceedings of the AAAI conference on artificial intelligence, vol. 33, pp. 6826–6833 (2019).

49. Mocanu, D., Baronchelli, A., Perra, N., Gonçalves, B., Zhang, Q., & Vespignani, A. (2013). The twitter of babel: mapping world languages through microblogging platforms. *PloS One*, 8(4), e61981

50. Mossberger, K., Wu, Y., Crawford, J.: Connecting citizens and local governments? social media and interactivity in major us cities. Government Information Quarterly 30(4), 351–358 (2013).

51. Motyl, M.: "Liberal and conservatives are (geographically) dividing. University of Illinois, Chicago(2016)

52. Myers, D.: Demographic dynamism and metropolitan change: comparing ew York and Los Angeles, Chicago, and Washington, DC. *Housing Policy Debate* 10(4), 919–954 (1999).

53. Neal, Z.: Differentiating centrality and power in the world city network. Urban Studies 48(13), 2733–2748 (2011).

54. Nikolaidou, A., Papaioannou, P.: Utilizing social media in transport planning and public transit quality: Survey of literature. Journal of Transportation Engineering, Part A: Systems 144(4), 04018007 (2018).

55. Noulas, A., Mascolo, C., Frias-Martinez, E.: Exploiting foursquare and cellular data to infer user activity in urban environments. In: 2013 IEEE 14th International Conference on Mobile Data Management. IEEE, vol. 1, pp. 167–176 (2013).

56. Noulas, A., Scellato, S., Lambiotte, R., Pontil, M., & Mascolo, C. (2012). A tale of many cities: universal patterns in human urban mobility. *PloS One*, 7(5), e37027.

57. Ott, B. L., & Mack, R. L. (2020). *Critical media studies: An introduction*. New York: John Wiley & Sons.

58. Parzen, E. (1962). On estimation of a probability density function and mode. *Ann. Math. Statist.*, 33(3), 1065–1076. https://doi.org/10.1214/aoms/1177704472.

59. Perrin, A.: Social media usage (pp. 52–68 ). Washington, DC: Pew Research Center (2015).

60. Preoţiuc-Pietro, D., Cohn, T.: Mining user behaviours: a study of check-in patterns in location based social networks. In: Proceedings of the 5th annual ACM web science conference, pp. 306–315 (2013).

61. Preoţiuc-Pietro, D., Cranshaw, J., Yano, T.: Exploring venue-based city-to-city similarity measures. In: Proceedings of the 2nd ACM SIGKDD international workshop on urban computing, pp. 1–4 (2013).

62. Rimmer, P.J.: Transport and telecommunications among world cities. In *Globalization and the world of large cities*, pp. 433–470 (1998).

63. Ras wasia, N., Vasconcelos, N.: Scene classification with low-dimensional semantic spaces and weak supervision. In: 2008 IEEE conference on computer vision and pattern recognition. IEEE, pp. 1–6 (2008).

64. Reades, J., Calabrese, F., Sevtsuk, A., Ratti, C.: Cellular census: Explorations in urban data collection. IEEE Pervasive Computing 6(3), 30–38 (2007).

65. Rimmer, P.J.: Transport and telecommunications among world cities. In *Globalization and the world of large cities*, pp. 433–470 (1998).

66. Ritchie, H., Roser, M.: Urbanization. Our World in Data (2018). https://ourworldindata.org/urbanization

67. Rosenblatt, M. (1956). Remarks on some nonparametric estimates of a density function. *Ann. Math. Statist.*, 27(3), 832–837. https://doi.org/10.1214/aoms/1177728190. doi: https://doi.org/10.1214/aoms/1177728190.
68. Sadilek, A., Kautz, H.: Modeling the impact of lifestyle on health at scale. In: Proceedings of the sixth ACM international conference on Web search and data mining. pp. 637–646 (2013)
69. Sinnott, R.: Virtues of the haversine. Sky and Telescope. 68, 158 (1984)
70. Smith, A.: Smartphone ownership-2013 update, vol. 12. Pew Research Center Washington, DC (2013).
71. Tasse, D., Hong, J.I.: Using social media data to understand cities. In: Proceedings of NSF workshop on big data and urban informatics, pp. 64–79. NSF Chicago, IL (2014)
72. Taylor, P., Derudder, B., Saey, P., & Witlox, F. (2006). Cities in Globalization: Practices, policies and theories. London: Routledge.
73. Taylor, P.J.: Specification of the world city network. Geographical Analysis 33(2), 181–194 (2001).
74. Wing, B.P., Baldridge, J.: Simple supervised document geolocation with geodesic grids. In: Proceedings of the 49th annual meeting of the association for computational linguistics: Human language technologies, vol. 1, pp. 955–964. Association for Computational Linguistics (2011)
75. Yang, D., Zhang, D., Qu, B.: Participatory cultural mapping based on collective behavior data in location-based social networks. ACM Transactions on Intelligent Systems and Technology (TIST) 7(3), 1–23 (2016).
76. Yuan, J., Zheng, Y., Xie, X.: Discovering regions of different functions in a city using human mobility and pois. In: Proceedings of the 18th ACM SIGKDD international conference on knowledge discovery and data mining, pp. 186–194 (2012)
77. Zook, M.A., Brunn, S.D.: Hierarchies, regions and legacies: European cities and global commercial passenger air travel. Journal of Contemporary European Studies 13(2), 203–220 (2005).

Publisher’s Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.