International Journal of Geo-Information

MDPI

Article

"Voting with Their Feet": Delineating the Sphere of Influence Using Social Media Data

David W. S. Wong [1,*] and Qunying Huang [2,*]

[1] Department of Geography and Geoinformation Science, George Mason University, Fairfax, VA 22030, USA
[2] Department of Geography, University of Wisconsin-Madison, Madison, WI 53706, USA
* Correspondence: dwong2@gmu.edu (D.W.S.W.); qhuang46@wisc.edu (Q.H.); Tel.: +1-608-890-4946 (Q.H.)

Received: 28 July 2017; Accepted: 25 October 2017; Published: 29 October 2017

Abstract: Delineating regional boundaries for places has a long tradition in geography, urban analysis and regional planning. Its theoretical basis may be traced back to the central place theory. The normative approach, using spatial interaction models, has been used, and the empirical approach, using commuting data, is also popular. While gathering commuting data using traditional methodologies (e.g., surveys) is costly, data capturing people's locations and their thoughts, are widely available through social media platforms. This article demonstrates that Twitter data can be used to delineate boundaries among competing places. A generic approach based on the density of place names mentioned in geo-tagged tweets was proposed to reflect the sphere of influence or dominance of places. Locations with the same levels of influence from competing places constitute the boundaries delineating the regions dominated by the respective places. The method was tested to determine the boundaries between two metropolitan regions, two local cities, and two neighborhoods or communities. Results from these simple case studies demonstrated the validity of the general approach for evaluating existing place boundaries and determining boundaries if they have not been delineated. The method is applicable to different levels of the place hierarchy and has practical values for planning of places of different sizes.

Keywords: community boundary; Twitter; cyberspace; place; region

1. Introduction

Defining regions and associated boundaries is one of geography's main research areas [1]. The geographic literature has long debated the concept of "regions" [2]. Nevertheless, regions may be classified into formal and functional regions [3], or uniform and nodal regions, respectively [4,5]. Formal regions are defined by areas sharing similar characteristics, whereas functional regions are defined by areas connected by certain activities. Haggett et al. [5] suggested "planning regions", formed in an ad hoc manner for planning purposes, are similar to functional regions. Regions are fundamental units for various map types, particularly choropleth maps, and discrete boundaries delineating regions are essential map elements.

Triggered by the need to deliver various federal programs more efficiently, the gravity-based model delineated national planning regions using the concept of sphere of influence introduced by Huff [6]. The concept is akin to the market boundary delineation process, which is a function of the distance from market centers and the merchandize available from the centers [7]. As a result, the entire United States (US) was delineated into 72 regions around the first-order urban places and 292 regions around the second-order urban places. Regions can be of different geographical scales. Physical regions in the US may form a nested hierarchical structure ([4], pp. 363–364). In the urban system context, the central place hierarchical structure is a well-established concept [8,9], and the sphere of influence concept can be used to delineate regions at different levels of the urban hierarchy [6].

Various quantitative methods have been used to determine how regions should be formed or how boundaries should be drawn [4]. Some of these regions are aggregates of smaller regions, but some are defined by boundaries delineated along the continuous geographical space, without following the pre-existing boundaries of smaller regions [10]. Some regions are defined functionally, including the planning regions, and some are as small as neighborhoods, which are defined in people's mental maps based upon their perceptions [11,12].

Traditionally, regions and their boundaries have been determined using data describing the characteristics of various locations or the relationships among locations. These data most often are acquired from surveys conducted at various scales, and are thus expensive, or have been gathered for administrative purposes [13]. Recently, large volumes of georeferenced data have been generated from social media, partly due to the pervasive adoption of information communication technology (ICT). To some extent, society has moved into the "big data" era and is flooded with data created constantly by individuals, corporations, and governments in various forms (e.g., number, text and image). The emergence of big data in general and georeferenced big data specifically, has revolutionized different aspects of society [14]. As some spatial scientists have focused on mapping and analyzing these data, particularly social media data, some scholars called for cautious use of these data and a mindfulness of their limitations [15,16]. Despite various concerns of using social media data, such as the biased representations of those Twitter users who share their locations [17,18], these data have been frequently used to delineate boundaries of different types of regions or places [19–22].

The major objective of this article is to demonstrate that social media data can be used to delineate boundaries of regions on various geographical scales using the sphere of influence concept. In this study, boundaries are defined as locations having the same level of influence from competing places. Although this definition of boundaries is conceptually the same as the boundaries defined in Huff's model, our method is entirely different from Huff's model. We argue that the influence of a place can be reflected by the intensity of the place name being mentioned, and thus the boundaries are determined empirically by comparing the density levels of competing place names being mentioned over the study region. In other words, we show an alternative way to define place boundaries, using the relatively inexpensive and accessible Twitter data rather than the expensive survey data. Different regional delineation methods, including the one proposed by Huff [6], will be discussed in the Section 2. Section 3 explains how we translate the typical boundary concept into a density-based concept and argues that boundaries can be determined by comparing the density levels of competing place names as they appear in geo-tagged tweets. Section 4 demonstrates how the proposed method can delineate the boundaries of places at different levels of the urban hierarchy, including neighborhoods at the lowest level of the hierarchy.

2. Delineating Boundaries of Places

Many previous studies of boundary delineation focused on functional regions around places, where places were regarded as central places of different orders in the Christaller's hierarchical landscape [8,9]. Places provide goods and services of different orders, which determine the orders of places. Higher order goods and services, offered in higher order centers, are consumed less frequently whereas lower order goods, offered in both higher and lower order centers, are consumed more regularly, according to the central place theory. Regions around central places are functional regions, as the boundaries are determined by the market areas served by the central places. The population within the market area of a central place obtains goods and services from that central place. In other words, the market area region is under the influence of the central place.

In delineating the functional region boundaries, Haggett et al. ([5], p. 453) suggested that the boundaries should be drawn such that

$$\frac{intra-region\ connections}{inter-region\ connections} = maximum. \quad (1)$$

The idea is simply to draw boundaries to maximize interaction within the areas and minimize interaction among areas. To implement these criteria, various quantitative techniques have been used to assign units into different groups to form regions. Effort spent over the decades has resulted in sophisticated regionalization tools for delineating formal regions considering multiple criteria [23–25]. In developing functional regions such as the metropolitan regions in the US, data describing commuting patterns between counties have been used [26–28]. These methods adopt an empirical approach in that the commuting patterns of the populations across regions determine how the boundaries are drawn and what data depicting population commuting activities (origins and destinations) are required. These methods also operate under the premise that the regions are formed by grouping together smaller or basic units with clear boundaries.

On the other hand, regions may be defined according to some axioms. The gravity or spatial interaction model adopts the axiom that locations closer to a city or place are more likely to be under the influence of that place or city than a place or city further away, and therefore the model uses the characteristics of places and distance between places to delineate regions [7,29]. These trade areas or regions, in general, represent the sphere of influence of their respective centers, which may be market centers, cities or central places. In a functional region, such as a metropolitan area, it is often named after the center, as locations within the region are dominated by, or under the influence of, the center(s). Thus, these regions are under "the spatial influence of cities" ([6], p. 323).

Formally, Huff [6] determined the boundary between two places, h and k, by the isoprobability lines, along which P_{im} and P_{in} are the same, where P_{ij} is the probability of an individual in location i traveling to place j, and j can be m and n. Specifically, the probability is defined as follows:

$$P_{ij} = \frac{\frac{V_j}{d_{ij}^{\alpha}}}{\sum_j^k \frac{V_j}{d_{ij}^{\alpha}}}, \tag{2}$$

where V_i is the size or a property of place i, d_{ij} is the distance between place i and place j with k total places (i.e., m and n) in the system, and α is the distance decay parameter. Using this method, and with some adjustments after considering the existing county boundaries, Huff [6] constructed planning regions for 72 first-order and 292 second-order urban places across the US. Subsequently, the method was applied to study the urban systems of Ireland [30] and Ghana [31]. In Huff's original formulation, the size of a city was derived from a composite measure of 14 dimensions from 97 variables describing the cities. To generalize Huff's idea, the boundary between two regions may be formulated as a set of locations, s_i, such that

$$I_A(s_i) = I_B(s_i) \quad s_i \in \mathcal{R}, \tag{3}$$

where $I_A(s_i)$ refers to the influence of center A on location s_i and $\mathcal{R}$ is a region. In other words, the boundary is composed of locations, where the influences from the two competing centers are at the same level.

Apparently, the general framework for delineating regions is not limited to determining the boundaries of trade areas of central places as demonstrated by Huff. It can be generalized to determine the spatial extent of the influence of places, and thus places do not have to be central places in the spatial economic system. Berry and Lamb [32] used newspaper circulation data to verify if the spatial interaction approach proposed by Huff was a valid method to delineate spheres of influence. In their study, N_{ij} was the number of newspapers published in city j and sold in county i. The probability a location i is under the influence of place or city j (Equation (2)) can be rewritten as

$$P_{ij} = \frac{N_{ij}}{\sum_j^k N_{ij}}, \tag{4}$$

assuming there are k places or cities in the study. Apparently, the probability in Equation (4) is also a density or intensity measure. With a large number of newspapers published from place j found in location i, it is more likely that location i is under the influence of place j.

An issue with the spatial interaction approach for delineating regional boundaries is that boundaries are part of the circle and subsequent adjustments of the arcs are often necessary, as boundaries are often linear. Another common, but more geometrically oriented approach for delineating regional boundaries is using the Voronoi diagram or Thiessen polygons. These partition space into regions surrounding points (or seeds), and the region boundaries have the same distances to the nearest points. To account for the differences between point characteristics, weighted Thiessen polygons have been suggested [33,34], and various versions of weighted Thiessen polygons have been used to study the regional structures of urban hierarchies by accounting for different socioeconomic variables pertaining to the cities [35].

Although settlements and places can be categorized into different types and sizes, a "place" in a layperson's mind may refer to regions of various sizes, including larger regions, such as states and provinces, and smaller regions, such as neighborhoods or communities [36]. This conceptualization of place is based upon the cognitive categorization of geographic objects proposed by Lloyd et al. [37]. They claimed that generic geographic regions, such as "country, region, state, city, neighborhood" in a US context, are basic-level categories that are specific instances of the superordinate category of place [38]. While some of these regions, such as the administrative regions of states and counties, have reasonably well-defined boundaries, some regions may not, including neighborhoods or communities and physiographical regions (e.g., Piedmont, Midlands, Lowcountry in the state of South Carolina) [10]. Even for those regions with well-defined boundaries, their boundaries may not align closely with people's perceptions of place boundaries.

The humanistic conceptualization of "place" provided by Cresswell concurs with the cognitive approach that "place" covers the geographical spectrum from a specific location to an extensive region [39]. The name of a place, not its site, in terms of latitude and longitude, carries meaning to people. Agnew's three-part definition of place, location, locale and sense of place also implies that people can identify with places [40]. While locale refers to the physical setting of a place, the sense of place refers to people's attachment to a place. In other words, when someone mentions a place name, the place means something, although we do not know if the place projects a positive or negative image toward that person. The feeling or attachment may be regarded as a form of influence. The study of the urban sphere of influence through the analysis of newspaper readership was based on the premise that people who read the newspaper of a city are concerned about, or under the influence of, that city and the spatial distribution of these readers reflects the sphere of influence of that city [32]. Our study's objective of identifying the spheres of influence for places is based on a similar premise. When a place name is mentioned by someone at a location, that person is concerned about or influenced by that place to a certain extent.

So far, few studies have used social media data to delineate market boundaries by Huff's model framework. A formal study used data extracted from Sina Weibo, a social media platform in China, to estimate users' trip patterns and derived statistics to calibrate the parameters of Huff's model to delineate the boundaries of five retail agglomerations in Beijing, China [41]. Our objective is not to calibrate Huff's model to determine trade areas, but to explore if the spheres of influence of competing places in general, can be empirically determined by Twitter data and thus to delineate boundaries of these regions or places.

Despite various spatial issues, such as positional uncertainty and vagueness in the boundaries of place names, there are promises in using crowd-sourced or social media data to determine the spatial extent of places [42]. Flickr data have been used to determine the spatial extent of city centers or cores [21]. On a different geographical scale, global-virtual Syrian communities were identified by extracting locations of tweets mentioning "Syria" [22]. Thus, mentioning "something" in cyberspace is linked to the presence of "something" in a geographical space. It is notable that these studies do not

spatially handle competing places or regions. Our study follows a similar approach, but the purpose is to delineate boundaries of places that compete for their influences over the region. Again, "places" may follow different levels of the central place hierarchy of the spatial economic system for goods and service provision, or the settlement-spatial administrative hierarchy, such as villages-communities, towns, and cities.

We would like to test the feasibility of using the proposed method in delineating local neighborhoods, although our notion of neighborhoods has much smaller geographical extents than the "territorial expressions of community life" in Huff's urban-city scale analysis. The boundaries of places derived from our method will unlikely match those boundaries defined officially, if these places have official boundaries. The differences between these boundaries likely reflect the discrepancies between people's perceptions of these places and their official definitions. For places without well-defined boundaries, our method offers a candidate that can be used to determine the boundaries of these places. To a large degree, we want to evaluate if people/ordinary citizens can tell us where the boundaries of places are drawn and thereby contribute to the creation of knowledge [13].

3. Methodology and Data

3.1. A Density-Based Method

When a place name is mentioned at a given location, it reflects that the mentioned place has a certain degree of influence in that location. Where a place name is mentioned may show the geographical extent of the influence of that place. How frequently the name is mentioned shows the strength or intensity of the influence. For instance, New York is mentioned numerously all over the world as its influence is trans-national. However, it is likely mentioned more frequently around the New York metropolitan region than in the central valley of California. Clearly, a place name can be mentioned for many reasons at various locations, not necessarily indicating the place has a strong influence in those locations. For instance, a basketball fan in city A may mention city B because the basketball team of city A was in a match with city B days ago. From a spatial sampling perspective, place names may be mentioned in selected locations non-systematically in low intensity and these mentions may be treated as random noises, but locations under the influence of a place should hit the place name consistently and with relatively high intensity levels. Thus, locations hitting a place name with relatively high frequencies, similar to those locations carrying newspapers published by a given city [32], should be included in the region's boundary, and locations with low frequencies can be ignored.

However, the region defining a place is rarely exclusively dominated by that place name. For instance, in Manhattan, the core of New York City, people mention other places occasionally (e.g., Philadelphia PA, Boston MA) but probably not as frequently as New York. Therefore, the area within the boundary of a place should be influenced by that place at a higher intensity level than the influence from other places. This intensity principle is similar to that discussed in [7] for delineating market area boundaries and in [32] for delineating planning regions. Despite differences in perspectives, these boundaries are the locations along which influence from the competing markets or places are equal.

The frequencies of place names being mentioned over space can be depicted as a density surface, using the concept of spatial kernel density estimation (KDE) [43,44]. Spatial KDE has been applied in many geographical studies, but most studies are related to the density of events or population [45,46]. In general, the surface is partitioned into grid cells. Let s_i be the location of the ith grid cell and $s_i \in \mathcal{R}$, a region in two-dimensional (2D) space. Points represent the locations of events or objects and are scattered over the study region. An estimated point density level is computed for each location s_i, based on the number of points surrounding s_i within distance γ. Thus, the point density for s_i is defined as

$$D(s_i) = K(s_i, \mathrm{h}) \text{ where } \gamma \leq \mathrm{h}, \text{ or } D(s_i) = 0 \text{ if } \gamma > \mathrm{h}, \quad (5)$$

where $K(s_i,\ \mathrm{h})$ is the kernel function applied to location s_i using bandwidth h. Many types of kernel function can be used (quartic is the most popular), but their general structure follows a distance decay structure such that locations farther away from the center have smaller weights and points farther away from the center of the kernel are counted less. The specific structure of the kernel function is less important than the bandwidth. Small bandwidths may create spikes on the density surface if points are not very dense, and large bandwidths will generate smooth surfaces.

In our study, each location where a specific place name (a) was mentioned in a tweet was treated as a point and the spatial KDE was used to compute the density of these points within the bandwidth of the kernel centered at location s_i to derive a density, $D(s_i) \mid a$ (i.e., density in location s_i mentioning place a). When the point density was computed for each point s_i in region $\mathcal{R}$ in reference to place a, a density surface could be generated. Dispersed locations with small numbers of points, representing the random mentioning of place a, should have low density levels, whereas clustered locations with large numbers of points, indicating the dominance of place a, should have high density levels. To determine the boundary between two places with names a and b, the boundary or boundaries should be formed by locations meeting the condition that

$$D(s_i)|a = D(s_i)|b. \tag{6}$$

These are points or locations with equal density levels referencing to the two places. Conceptually, the boundary consists of points where the influence from the two neighboring or competing places are the same, analogous to the iso-intensity described in Equation (3). This concept of boundaries is also applicable to different types of regions: formal and functional. In other words, the definition described in Equation (6) offers an operational and generic definition of regional boundaries.

An implicit assumption adopted by the definition described in Equation (6) is that the two places are compatible, meaning that the two places belong to the same category or order of place in the place hierarchy. For instance, if one place is a local community and the other is the surrounding state or county in the US, Equation (6) may not be applicable, although the community should have its own boundary. Even if the two places are compatible or of the same order, using Equation (6) may still encounter operational problems. An example is that a place is much larger than the other, in terms of population size. As our approach relies on "people's votes", larger places will likely receive more votes and thus their density surfaces will have higher levels. In this case, the boundary delineation will not only be determined by where "voters" are located, but will also be affected by the size of the voting population. To neutralize the population size effect, densities may be standardized by population sizes.

3.2. Data

In this study, messages (tweets) from Twitter were used to demonstrate the proposed methodology for deriving the conceptual boundaries among places. As one of the largest social networking sites, Twitter is widely used by the public for sharing information, reflecting personal witnesses and feelings through micro-blogging. Twitter Stream application programming interfaces (APIs) can be used to harvest 1% of total tweets. We followed the same method as described in [47] to create our database, in which each tweet was a document entry with metadata about that tweet message (e.g., user name, time stamp, location, tweet time, source generation, text content, and hashtags). Each tweet is limited to 140 characters and is highly unstructured. It may include a large number of abbreviations and a hashtag, which is a phrase without space, but prefixed with the sign "#", an identifier unique to Twitter. This identifier is often used to search for tweets sharing the same topic. Therefore, we selected all tweets with text content or a hashtag that included our target place names.

To explore whether our proposed method could be used to determine the boundary of places, we selected a few places of different orders. At the metropolitan level, we chose Washington, DC and Baltimore, MD (Washington and Baltimore thereafter) as a pair of neighboring metropolitan areas within a large urban region [48,49]. The official boundary between the two metropolitan areas follows

the Prince George's–Montgomery–Frederick and Anne Arundel–Howard–Carroll county divisions (Figure 1). Our question was whether people's perceived boundary between the two metropolitan areas followed the official division. At the city level of place hierarchy, we chose Rockville and Bethesda in Maryland as local cities. Between the two cities is North Bethesda, a place not well-recognized as a separate entity by the locals. Then, the question was where the boundary between Rockville and Bethesda should be? At the community level of the place hierarchy, we chose Ballston and Clarendon, two neighborhoods or communities in Arlington County, Virginia (VA). These are also station names along the Washington, DC Metro (subway) system, but these names are also associated with neighborhoods, or at least are used as place names. They do not have official boundaries, and there are many other place names of similar nature. They were paired mainly because of their compatibility and geographical proximity.

Since our study area focused on Maryland, Washington, DC, its surrounding states (Virginia, West Virginia, Delaware, New Jersey, and Pennsylvania) defined the study boundary (i.e., six-state region). Figure 1 shows the boundaries of the county-equivalent units in the region, and Table 1 reports the number of geo-tagged tweets between August 2013 and January 2015 mentioning the place names selected for the study. The percentages of tweets with geo-tags for these places were between 1.89% (Washington) to 7.31% (Ballston, VA) of all tweets, with higher percentages for smaller versus that for larger places (Table 1). It is reasonable to assume that the 'larger' or higher ordered places should be mentioned more frequently on Twitter.

Table 1. Number of different categories of tweets mentioned in the selected places.

Selected Places	Metropolitan Areas		Cities		Neighborhoods	
	Washington	Baltimore	Rockville	Bethesda	Ballston	Clarendon
Total number of tweets (nation-wide)	2,092,350	535,724	29,869	36,498	3334	9105
Number and percentage of geo-tagged tweets (nation-wide)	39,601 (1.89%)	15,406 (2.87%)	1761 (5.90%)	2393 (6.55%)	244 (7.31%)	380 (4.17%)
Number of geo-tagged tweets within the six-state boundary	21,772	12,471	1556	1137	238	169
Final number of geo-tagged tweets used for the analysis	6835	3016	1556	1137	238	169

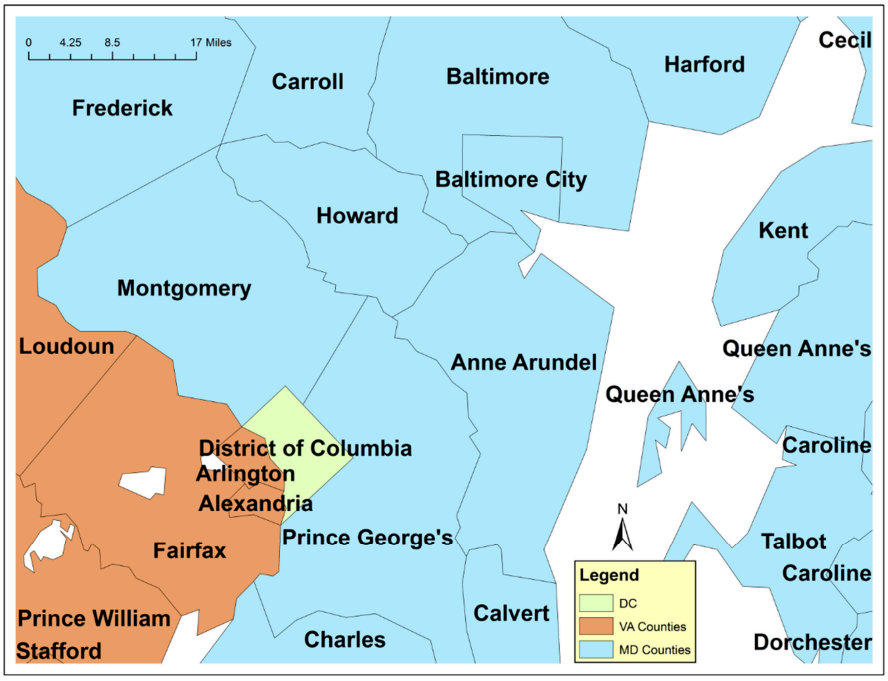

Figure 1. Administrative boundaries of county-equivalent units in the Washington, DC and Baltimore, MD region.

Tweets can be created by different sources, including iPhones, Android phones, websites, and other social media sites. Knowing the sources of tweets provides clues about the potential topics of the tweet content because different websites are used for different functions, including checking-in places (e.g., Foursquare), sharing photos and videos (e.g., Instagram and Hipstamatic), advertising (e.g., TweetMyJOBS), and reporting local news and alerts (Goldstar, Baltimore 311, dlvr.it, and screamradius). For example, tweets mentioning Washington and Baltimore were generated from 158 and 104 sources, respectively. The top 15 sources mentioning Washington and Baltimore are shown in Table 2. Messages (tweets) from certain sources are for information dissemination or propaganda purposes only and are created by corporations, organizations, or government agencies. Although tweets from these sources reflect the importance and influence of the respective cities at the given locations, their frequencies or intensities are related to their operational objectives and do not reflect people's perceptions and feelings about a specific place. In addition, these sites produce repeated tweets for the same locations (e.g., people check-in the same place using Foursquare). Therefore, including these messages would produce high intensities for selected locations (spikes), inflating the importance levels of the respective cities on these locations. To avoid upward biases introduced by these sources of tweets, they were removed in our analyses. Therefore, only tweets generated by iPhones, Android phones, Blackberry phones, iPads, Windows phones, Android tablets, and iPhone Operating System (IOS) were included, as tweets from these sources are more likely generated by individuals. As a result, we only had 6835, and 3016 geo-tagged tweets mentioning Washington and Baltimore, respectively, after the data processing procedure. For the other selected places, the numbers of geo-tagged tweets that mentioned these places were smaller, because they are less recognized and less famous places. Therefore, fewer websites are used to advertise these smaller places. Thus, tweets generated from all sources were used in the analysis.

Table 2. Numbers of tweets mentioning Washington, DC and Baltimore, MD from various sources.

Washington, DC		**Baltimore, MD**	
Source	**Count**	**Source**	**Count**
Twitter for iPhone	16,795	Twitter for iPhone	4608
Foursquare	8373	Foursquare	3265
Twitter for Android	8081	Instagram	2173
Instagram	6275	Twitter for Android	1692
Safetweet By Tweetmyjobs	804	Safetweet By Tweetmyjobs	1002
Twitter for Blackberry®	667	Baltimore 311	803
Twitter for iPad	604	dlvr.it	772
Hipstamatic	207	Twitzip	167
Goldstar	178	Tweetmyjobs	138
IOS	159	Twitter for iPad	96
dlvr.it	158	Goldstar	48
Twitter for Android Tablets	154	Untappd	40
Twitter for Windows Phone	153	Screamradius	40
Twitterfeed	145	IOS	40
Path	103	Twitter for Blackberry®	37

4. Exploring Geographical Boundaries in Cyberspace

4.1. Washington, DC versus Baltimore, MD

When using KDE analysis, the critical parameter is the bandwidth conceived as the search distance within which points are counted toward the evaluation of the density. Figure 2 shows the kernel surfaces of Baltimore and Washington, based on tweet locations mentioning the two places. These density surfaces use 10 km as the bandwidth and 1 km as the output cell size. The size of the bandwidth is the most influential factor in affecting the results in KDE. While we did not experiment exhaustively with different sizes of bandwidth, results using 10 km are presented partly because the results are

sufficient to demonstrate the concept and the proposed method. In addition, 10 km approximates two-thirds the length of one side of Washington (the shape of Washington can be thought of a rotated square with each side being about 10 miles or 16 km). In general, the highest densities were found around the centers of the respective metropolitan cities, as shown in Figure 2, but there was insufficient information to show the boundary between the two cities. To determine the boundary, the densities of the two places needed to be compared.

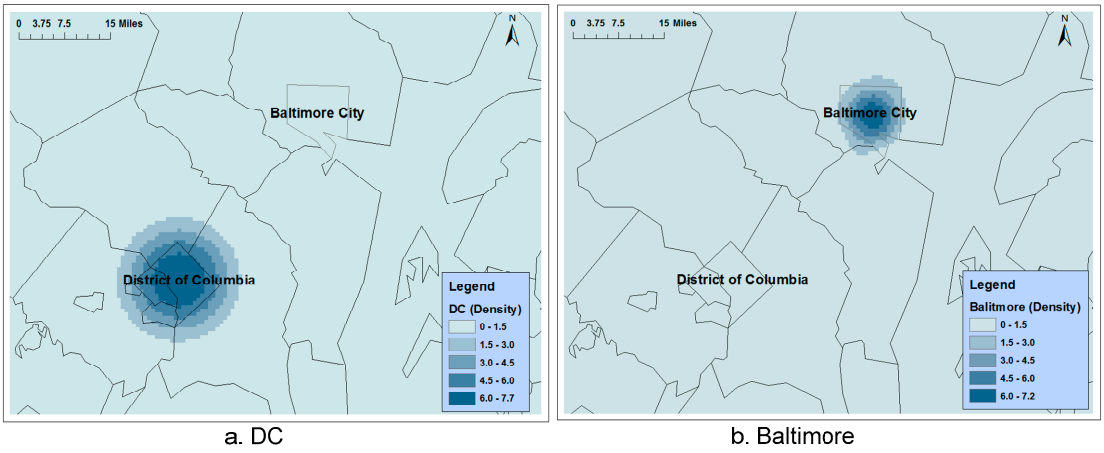

Figure 2. Density of tweets for Washington, DC (**a**) and Baltimore, MD (**b**) using 10 km as the bandwidth and 1 km as the cell size.

After the density surfaces were generated using KDE, the densities of the two surfaces for each grid cell were compared (density of Washington minus density of Baltimore for each location). In general, locations with positive grid cell values were dominated by Washington whereas locations with negative grid cell values were dominated by Baltimore. Figure 3 shows grid cell values in three classes: largely positive, largely negative and approaching zero (−0.05 to 0.05). The last category includes locations in which the influences from the two metropolitan cities were about the same and may be conceived as the boundaries dividing the two regions. A large proportion of the area falls into this category, but most of this area is farther away from one or both cities (Figure 3). This observation is not surprising as locations farther away from the cities have relatively low densities, and their differences tend to approach zero. Instead, the area between the two cities should be our focus.

Figure 3. Differences in tweet density (density of Washington minus density of Baltimore) between Washington, DC and Baltimore, MD for the six-state region (**a**) and a zoomed-in view (**b**).

Theoretically, locations with a zero-density difference (Figure 3b) depict the boundary between the two regions (Equation (6)). Instead of using a crisp line, we used a region with a density difference of ± 0.05 to depict the boundary. We cannot be certain that for any location within this boundary region the influence of one city is more than the other. This elongated boundary region cuts across Howard and Anne Arundel counties (Figure 3b), both officially within the Baltimore metropolitan region. In fact, locations very likely dominated by Baltimore (with density differences less than -0.05), occupied less than half of Howard and only a small portion of Anne Arundel counties. In other words, the influence of Washington intrudes into the official boundary of Baltimore to a large degree. Conversely, the influence of Baltimore was well recognized in the Western part of Maryland, including the cities of Hagerstown and Cumberland–Frostburg (Figure 3a). Areas under the influence of Baltimore extended from Morgantown West Virginia, to Western and Southern Pennsylvania, including the city of Hanover and several cities in York County. However, the dominance of the Washington influence extended farther than Baltimore to the north in the larger cities, including Philadelphia, PA and New York City (Figure 3a). It is noted that this analysis included only Baltimore and Washington. The influence of Washington to locations north of Baltimore merely reflects the relatively influences of the two cities, excluding the influence of Philadelphia, PA and New York City.

The densities of the two metropolitan cities were based on the total number of tweets. Being the capital of the nation, Washington was more popular and mentioned more often than Baltimore. The number of geo-tagged tweets mentioning Washington was more than twice (2.266 times) the number mentioning Baltimore (Table 1). To neutralize this size factor, the density of Baltimore tweets was raised by a factor of 2.266 before the Baltimore density surface was compared with that of Washington (Figure 4). Although the overall geographical pattern did not change substantially, the details were noticeably different. Not only did the areas dominated by Baltimore in Howard and Anne Arundel counties become larger than those in the unscaled situation, the dominance of Baltimore in other Maryland counties surfaced. Frederick County (Maryland) is in the Washington metropolitan region but was clearly dominated by Baltimore. More counties to the east of Baltimore and in Southern and Western Pennsylvania showed the influence of Baltimore. Even in New York City, Philadelphia, and part of New Jersey, the dominance of Baltimore was recognizable, whereas the influence of Washington in those same places diminished.

a. Six-state region b. Zoom-in view

Figure 4. Difference in tweet density between Washington, DC and Baltimore, MD (density of Washington minus density of Baltimore after scaling up Baltimore by a factor of 2.266) in the six-state region (**a**) and in the zoomed-in view (**b**).

To compare the derived with the existing boundaries of two competing places, P_1 and P_2, an accuracy ratio (AR defined as n_0/n) was introduced where n is the number of cells (pixels) of the existing boundaries and n_0 is the number of existing boundary cells aligned with the derived

boundaries or boundary region. If AR approaches one, the derived boundaries spatially align with the existing boundaries well. Letting n_1 and n_2 be the numbers of existing boundary cells found inside the derived regions of P_1 and P_2, respectively, a bias ratio (BR defined as $(n_1 - n_2)/n$) indicates the positional bias of the derived boundaries. If the bias ratio is positive, more cells of the existing boundaries are located in the derived region of P_1 than those found in the derived region of P_2, meaning that the derived boundaries of P_1 have intruded into the existing boundaries of P_2, and vice versa. The magnitude of the ratio reflects merely the bias, in terms of cell numbers of derived boundaries not aligning with the existing boundaries, but not the spatial extent of the misalignments. The two indices, AR and BR, need to be considered together. If AR is larger (close to one indicating a strong alignment of the existing with the derived boundaries or boundary region), the absolute value of BR should be small. Regardless of the value of AR, the sign of BR indicates which derived region has intruded into the other more.

Assuming P_1 is Washington and P_2 is Baltimore, AR and BR (without scaling Baltimore) were 0.37 and 0.58, respectively (Figure 3) (instead of using the difference of density between -0.05 and 0.05 to define the boundary region in the resultant maps, we used a more stringent range of -0.01 to 0.01 to depict the boundaries more precisely.) The AR was not high, indicative of a weak spatial correspondence between the derived boundaries and the official boundaries delineating the two metropolitan areas. Conversely, the BR was high, indicating that more cells of the official boundaries landed inside the derived region of Washington than that of Baltimore. This implies that the derived boundaries of Washington intrude into the official territory of Baltimore. It is clear (Figure 3) that the derived Washington region encroaches upon Howard and Anne Arundel counties, which officially belong to the Baltimore region.

After scaling the Baltimore intensity level (Figure 4), AR and BR were 0.38 and 0.16, respectively. While the accuracy did not change, the positional bias of the derived boundaries diminished substantially. The derived boundaries were better aligned with the official boundaries (Figure 4) than the result illustrated in Figure 3. Nevertheless, the derived boundaries of Washington still intrude into the official territory of Baltimore.

4.2. Rockville versus Bethesda, MD

Rockville, MD is an incorporated city with a government and an administrative boundary, whereas Bethesda is a census designated place (CDP) with no government, but with a boundary. While these two places have different administrative and statistical statuses in the US census geography, they have a similar, if not an identical order, in the settlement-administrative hierarchy, as they are both "local cities" with well recognized names in the region. Conversely, between Rockville and Bethesda is North Bethesda (CDP), whose existence is often ignored by the local communities in the metropolitan region. As these are not "large" or high order places, our experiment was to explore if Twitter data could help define the sphere of influence of these places and thus the perceived boundary between these two places.

Clearly the boundary between Rockville and Bethesda cuts through North Bethesda (Figure 5). Bethesda was more influential toward the south. The dominance of Rockville within its vicinity was solid. It is important to note that the points (locations) of tweets mentioning the two places were relatively sparse spatially, as compared to the case of Baltimore–Washington, and the circular shape of the dominance areas is an artifact of using a circular kernel. Thus, some of the dominance areas could be the results of just a few "votes" of the places on Twitter. Expanding the temporal window to harvest more tweets is likely to produce more robust results with less noises. As Rockville and Bethesda are not adjacent (with North Bethesda in between), the quantitative evaluation framework introduced above is not applicable.

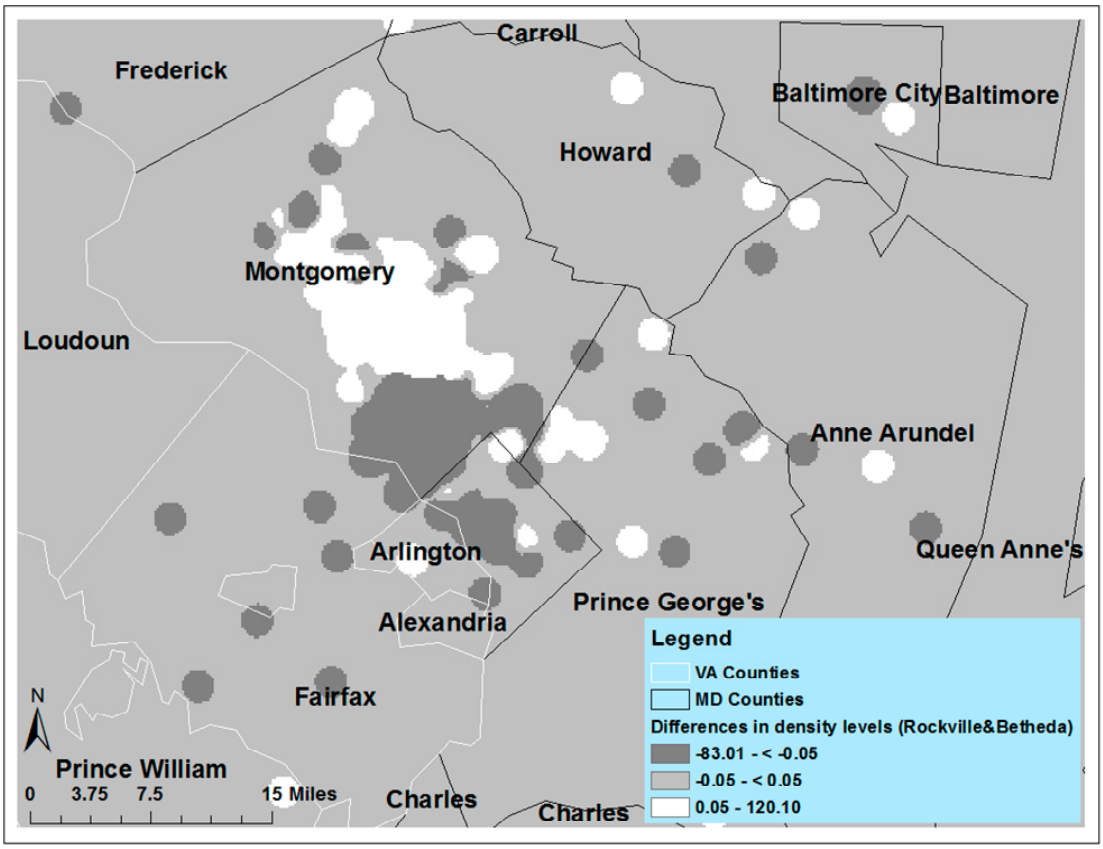

Figure 5. Difference in tweet density between Rockville and Bethesda (density of Rockville minus density of Bethesda; the density surfaces of Rockville and Bethesda used 2.5 km as the bandwidth).

4.3. Ballston versus Clarendon, VA

In this last example, Ballston and Clarendon are neighborhoods or communities in Arlington County, Virginia. These two communities have no clear boundaries, and this attribute differs from the previous examples. As these are local communities, they are not that popular in the cyber world and therefore received much fewer votes than the higher order places in the other examples (Table 1). Thus, delineating boundaries between these places with sparse points could be challenging.

The results shown in Figure 6 are quite promising. Ballston and Clarendon are names in Arlington County used in two stations along the Washington, DC Metro system. In between the two stations is a third station—"Virginia Square". While both Ballston and Clarendon are characterized as "populated places" in the US Geographic Names Information System (GNIS), "Virginia Square" is not similarly classified [50]. The analysis showed a clear boundary between the two populated places, splitting the region near the "Virginia Square" station, between the two. Due to the small numbers of votes and the circular shape of the kernel, circles of dominance are presented irregularly, indicating noise in the data and possibly the effect of small sample sizes.

The place boundaries defined by the proposed method reflect the spheres of influence of competing places based on people's feelings. Boundaries of this sentimental nature are different from the administrative or political boundaries of places in two of our case studies (Washington versus Baltimore; Rockville versus Bethesda). The third case in Virginia does not have existence boundaries with which to compare.

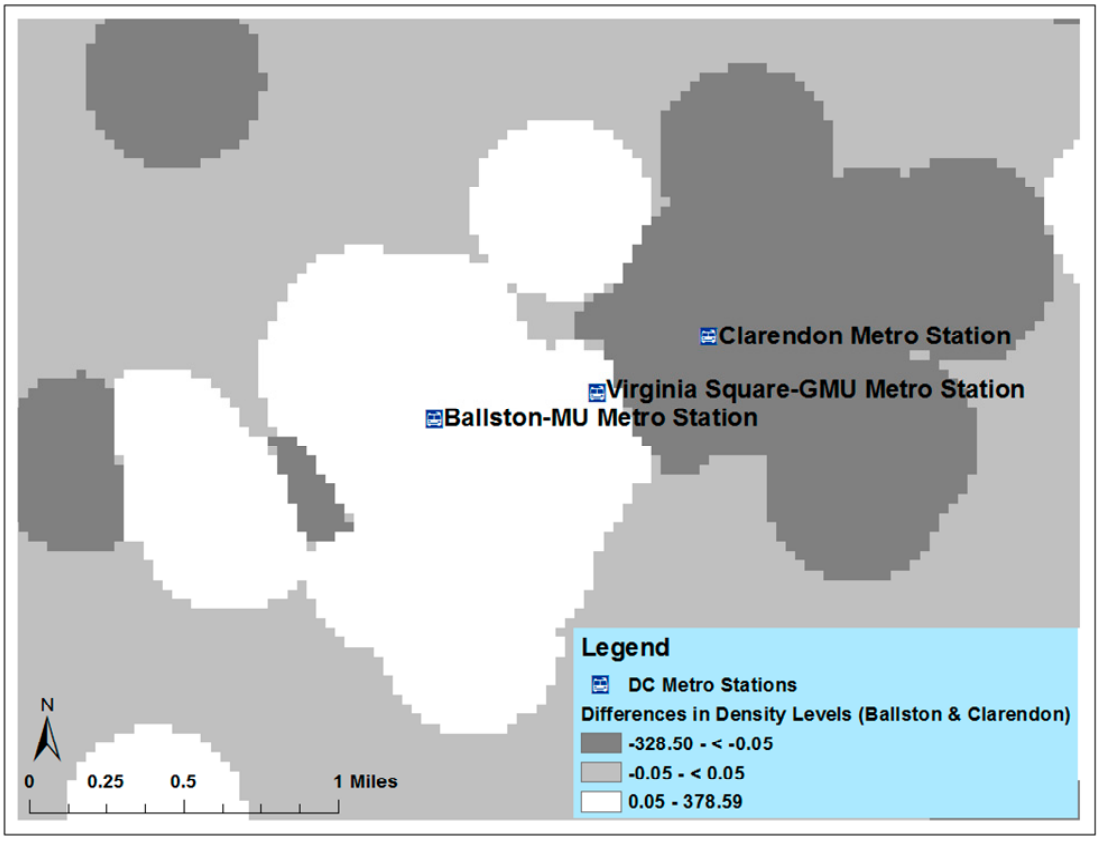

Figure 6. Differences in tweet density between Ballston and Clarendon with OpenStreetMap as the basemap (density of Ballston minus density of Clarendon; the density surfaces of Ballston and Clarendon use 1 km as the bandwidth).

5. Conclusion and Discussions

The region concept is essential for the planning, administration, and delivering of goods and services, despite their varying natures. Although boundaries can be delineated based on the concept of the sphere of influence of places, historical, political, and administrative reasons often determine the resultant boundaries. As reviewed herein, many methods and data have been used to delineate regional or place boundaries. The proposed method, based on the locations of individuals while they "voted" for places in the cyberspace, seems appropriate to determine the geographical sphere of influence of a place. The method leverages the location information shared by individuals that indicates the place's influence. The fundamental concept is that locations along place boundaries receive the same levels of influence from competing places (Equation (6)). Thus, locations receiving the same intensity of votes from competing places can be treated as the boundaries between the two. The proposed method attempts to implement the intensity-based concept in delineating place boundaries in a spatial competition context using social media data.

Using selected metropolitan regions, local cities, and neighborhoods, this study shows that the proposed method can be used to depict the boundaries delineating the spheres of influence of respective places. Apparently, boundaries defined by the proposed method have different utilities for different types of places. In the metropolitan examples (Washington and Baltimore), the boundaries determined by the proposed method (Figures 3 and 4) indicated that the influence of Washington penetrates into the official territory of Baltimore. The two proposed indices, *AR* and *BR*, provide a quantitative assessment of the misalignments between the derived and official boundaries. The positional bias of the derived boundaries reflects the relative economic and political influences of the two cities in the areas surrounding them. Leaders of these cities may refer to the boundary when they

need to know the "spatial reach" of their cities. However, in the two-neighborhood case in Arlington County, Virginia, the boundary may be used to guide local community development programs or activities, as the boundary reflects the spatial extent of the "sense of place." In the example using the two Maryland cities, the boundary has little relationship to the jurisdictional extents of the two cities, but indicates, to some extent, the competitiveness or popularity of the two places over the region. Therefore, competing places can use the derived boundaries to assess their sphere influence on the people in the surrounding region, offering an important piece of information that the official boundaries may not reflect accurately.

The current study is not exhaustive, as only one pair of places for each level of the place hierarchy was examined. More comprehensive studies are needed to include more places across different levels of the place hierarchy. When more than two competing places are involved in the vicinity, the situation will be more complicated, although the general concept is still applicable. Each location (a cell in a grid data format) will have multiple densities, each corresponding to the density surface of one of the competing places. A location is under the dominant influence of a place (or within the boundary of that place) if the density of that place is the highest in that location. Extending the two-place situation (Equation (6)) to situations with multiple competing places, a location is part of the boundaries or boundary region of competing places if its densities corresponding to the competing places are the same or very similar. The required data for determining the boundaries of these multi-place situations will be enormous. The current study collected tweets posted within a one-year period. Using data for a longer period is possible and desirable, putting aside the likely challenge of dealing with the massive number of tweets, as the longer-term data may be more "representative" of the places, thereby reducing the effect of noise (e.g., messages related to special events). However, using longer-term data implicitly assumes that the spatial relationships between places are stable over time, potentially failing to capture the place dynamics that may be manifested over relatively short time frames.

As one place grows in popularity, the place may receive more votes, raising the intensity of influence but not necessarily expanding its geographical sphere of influence. In the case of Baltimore-Washington, adjusting the densities of Baltimore to the levels comparable to Washington isolates the influence of intensity differences, focusing only on the geographical extent of the influence. More experiments are needed to evaluate how the proposed method performs in other places, especially in comparing places of different orders, such as cities versus local communities.

The suggested method and concepts are rather simple but reasonably effective in revealing the sphere of influence of places and thus helping to delineate the boundaries of regions. The current study did not apply stringent criteria to "clean" the tweets to remove noise in the data. Place names mentioned, for whatever reasons, were included in most of our analyses, but the proposed simple method is robust enough to handle noise. It is not clear if the results would be dramatically different or improved if noise is removed through the use of more sophisticated text-mining and text-analysis techniques. Therefore, the proposed method shows promises of using crowd-sourced data to map place boundaries. Although our objective is similar to that in Berry and Lamb [32] in determining the spheres of influence of competing places, our method is density—rather than gravity—based and can employ contemporary data with a relatively low acquisition cost. In addition, the proposed method can be used to monitor the dynamics of the urban systems, in terms of relative competitiveness, among places as place characteristics vary over time.

Author Contributions: Both David Wong and Qunying Huang designed the methodology, Qunying Huang gathered and processed the data, both David Wong and Qunying Huang analyzed the results and wrote the paper.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Hartshorne, R. The nature of geography: A critical survey of current thought in the light of the past. *Ann. Assoc. Am. Geogr.* **1939**, *29*, 173–412. [CrossRef]

2. James, P.E. Toward a further understanding of the regional concept. *Ann. Assoc. Am. Geogr.* **1952**, *42*, 195–222. [CrossRef]
3. Norton, W. *Human Geography*, 8th ed.; Oxford University Press: Oxford, UK, 2014.
4. Haggett, P. *Geography: A Global Synthesis*; Pearson Education: London, UK, 2001.
5. Haggett, P. *Locational Methods*, 2nd ed.; John Wiley & Sons: New York, NY, USA, 1977.
6. Huff, D.L. The delineation of a national system of planning regions on the basis of urban spheres of influence. *Reg. Stud.* **1973**, *7*, 323–329. [CrossRef]
7. Huff, D.L. Defining and estimating a trade area. *J. Mark.* **1964**, *28*, 34–38. [CrossRef]
8. Berry, B.J. *Geography of Market Centers and Retail Distribution*; Prentice-Hall Englewood Cliffs: Upper Saddle River, NJ, USA, 1967.
9. Preston, R.E. The dynamic component of christaller's central place theory and the theme of change in his research. *Can. Geogr. Géogr. Can.* **1983**, *27*, 4–16. [CrossRef]
10. Lloyd, R.; Steinke, T. The identification of regional boundaries on cognitive maps. *Prof. Geogr.* **1986**, *38*, 149–159. [CrossRef]
11. Coulton, C.J.; Jennings, M.Z.; Chan, T. How big is my neighborhood? Individual and contextual effects on perceptions of neighborhood scale. *Am. J. Commun. Psychol.* **2013**, *51*, 140–150. [CrossRef] [PubMed]
12. Lee, B.A.; Campbell, K.E. Common ground? Urban neighborhoods as survey respondents see them. *Soc. Sci. Q.* **1997**, *78*, 922–936.
13. Elwood, S.; Goodchild, M.F.; Sui, D.Z. Researching volunteered geographic information: Spatial data, geographic research, and new social practice. *Ann. Assoc. Am. Geogr.* **2012**, *102*, 571–590. [CrossRef]
14. Kitchin, R. *The Data Revolution: Big Data, Open Data, Data Infrastructures and Their Consequences*; Sage: Beverly Hills, CA, USA, 2014.
15. Crampton, J.W.; Graham, M.; Poorthuis, A.; Shelton, T.; Stephens, M.; Wilson, M.W.; Zook, M. Beyond the geotag: Situating 'big data'and leveraging the potential of the geoweb. *Cartogr. Geogr. Inf. Sci.* **2013**, *40*, 130–139. [CrossRef]
16. Rzeszewski, M. Geosocial capta in geographical research—A critical analysis. *Cartogr. Geogr. Inf. Sci.* **2016**, 1–13. [CrossRef]
17. Rzeszewski, M.; Beluch, L. Spatial characteristics of twitter users—Toward the understanding of geosocial media production. *ISPRS Int. J. Geo-Inf.* **2017**, *6*, 236. [CrossRef]
18. Sloan, L.; Morgan, J. Who tweets with their location? Understanding the relationship between demographic characteristics and the use of geoservices and geotagging on twitter. *PLoS ONE* **2015**, *10*, e0142209. [CrossRef] [PubMed]
19. Ratti, C.; Sobolevsky, S.; Calabrese, F.; Andris, C.; Reades, J.; Martino, M.; Claxton, R.; Strogatz, S.H. Redrawing the map of great britain from a network of human interactions. *PLoS ONE* **2010**, *5*, e14248. [CrossRef] [PubMed]
20. Sobolevsky, S.; Szell, M.; Campari, R.; Couronné, T.; Smoreda, Z.; Ratti, C. Delineating geographical regions with networks of human interactions in an extensive set of countries. *PLoS ONE* **2013**, *8*, e81707. [CrossRef] [PubMed]
21. Hollenstein, L.; Purves, R. Exploring place through user-generated content: Using flickr tags to describe city cores. *J. Spat. Inf. Sci.* **2010**, *2010*, 21–48.
22. Stefanidis, A.; Cotnoir, A.; Croitoru, A.; Crooks, A.; Rice, M.; Radzikowski, J. Demarcating new boundaries: Mapping virtual polycentric communities through social media content. *Cartogr. Geogr. Inf. Sci.* **2013**, *40*, 116–129. [CrossRef]
23. Cockings, S.; Martin, D. Zone design for environment and health studies using pre-aggregated data. *Soc. Sci. Med.* **2005**, *60*, 2729–2742. [CrossRef] [PubMed]
24. Folch, D.C.; Spielman, S.E. Identifying regions based on flexible user-defined constraints. *Int. J. Geogr. Inf. Sci.* **2014**, *28*, 164–184. [CrossRef] [PubMed]
25. Openshaw, S. A geographical solution to scale and aggregation problems in region-building, partitioning and spatial modelling. *Trans. Inst. Br. Geogr.* **1977**, *2*, 459–472. [CrossRef]
26. Adams, J.S.; VanDrasek, B.J.; Phillips, E.G. Metropolitan area definition in the united states. *Urban Geogr.* **1999**, *20*, 695–726. [CrossRef]
27. Dahmann, D.C. New approaches to delineating metropolitan and nonmetropolitan settlement: Geographers drawing the line. *Urban Geogr.* **1999**, *20*, 683–694. [CrossRef]

28. Rain, D.R. Commuting directionality, a functional measure for metropolitan and nonmetropolitan area standards. *Urban Geogr.* **1999**, *20*, 749–767. [CrossRef]
29. Haynes, K.E.; Fotheringham, A.S. *Gravity and Spatial Interaction Models*; Sage: Beverly Hills, CA, USA, 1984; Volume 2.
30. Huff, D.L.; Lutz, J.M. Ireland's urban system. *Econ. Geogr.* **1979**, *55*, 196–212. [CrossRef]
31. Huff, D.L.; Lutz, J.M. Urban spheres of influence in ghana. *J. Dev. Areas* **1989**, *23*, 201–220.
32. Berry, B.J.; Lamb, R.F. The delineation of urban spheres of influence: Evaluation of an interaction model. *Reg. Stud.* **1974**, *8*, 185–190. [CrossRef]
33. Boots, B.N. Weighting thiessen polygons. *Econ. Geogr.* **1980**, *56*, 248–259. [CrossRef]
34. Aurenhammer, F.; Edelsbrunner, H. An optimal algorithm for constructing the weighted voronoi diagram in the plane. *Pattern Recognit.* **1984**, *17*, 251–257. [CrossRef]
35. Mu, L.; Wang, X. Population landscape: A geometric approach to studying spatial patterns of the us urban hierarchy. *Int. J. Geogr. Inf. Sci.* **2006**, *20*, 649–667. [CrossRef]
36. Agarwal, P. Operationalising 'sense of place'as a cognitive operator for semantics in place-based ontologies. In *Spatial Information Theory*; Springer-Verlag: Berlin Heidelberg, Germany, 2005; pp. 96–114.
37. Lloyd, R.; Patton, D.; Cammack, R. Basic-level geographic categories. *Prof. Geogr.* **1996**, *48*, 181–194. [CrossRef]
38. Mark, D.M.; Smith, B.; Tversky, B. Ontology and geographic objects: An empirical study of cognitive categorization. In *Spatial Information Theory. Cognitive and Computational Foundations of Geographic Information Science*; Springer-Verlag: Berlin Heidelberg, Germany, 1999; pp. 283–298.
39. Cresswell, T. *Place: An Introduction*; John Wiley & Sons: New York, NY, USA, 2014.
40. Agnew, J. *Place and Politics: The Geographical Mediation of State and Society*; Routledge: Abingdon, UK, 1987; Volume 3.
41. Wang, Y.; Jiang, W.; Liu, S.; Ye, X.; Wang, T. Evaluating trade areas using social media data with a calibrated huff model. *ISPRS Int. J. Geo-Inf.* **2016**, *5*, 112. [CrossRef]
42. Vasardani, M.; Winter, S.; Richter, K.-F. Locating place names from place descriptions. *Int. J. Geogr. Inf. Sci.* **2013**, *27*, 2509–2532. [CrossRef]
43. Bailey, T.C.; Gatrell, A.C. *Interactive Spatial Data Analysis*; Longman Scientific & Technical Essex: Harlow, UK, 1995; Volume 413.
44. Brunsdon, C. Estimating probability surfaces for geographical point data: An adaptive kernel algorithm. *Comput. Geosci.* **1995**, *21*, 877–894. [CrossRef]
45. O'Sullivan, D.; Wong, D.W. A surface-based approach to measuring spatial segregation. *Geogr. Anal.* **2007**, *39*, 147–168. [CrossRef]
46. Thurstain-Goodwin, M.; Unwin, D. Defining and delineating the central areas of towns for statistical monitoring using continuous surface representations. *Trans. GIS* **2000**, *4*, 305–317. [CrossRef]
47. Huang, Q.; Xu, C. A data-driven framework for archiving and exploring social media data. *Ann. GIS* **2014**, *20*, 265–277. [CrossRef]
48. Census. District Of Columbia—Core Based Statistical Areas (CBSAs) and Counties. 2013. Available online: https://www2.census.gov/geo/maps/metroarea/stcbsa_pg/Feb2013/cbsa2013_DC.pdf (accessed on 27 October 2017).
49. Census. Maryland—Core Based Statistical Areas (CBSAs) and Counties. 2013. Available online: https://www2.census.gov/geo/maps/metroarea/stcbsa_pg/Feb2013/cbsa2013_MD.pdf (accessed on 27 October 2017).
50. GNIS. US geographic names information system. Available online: http://geonames.usgs.gov/ (accessed on 27 October 2017).