JOURNAL OF SPATIAL INFORMATION SCIENCE
Number 10 (2015), pp. 47–66
doi:10.5311/JOSIS.2015.10.211

RESEARCH ARTICLE

Visualizing patterns in spatially ambiguous point data

Jonny J. Huck[1], J. Duncan Whyatt[2], and Paul Coulton[1]

[1]Imagination Lancaster, LICA, Lancaster University, UK
[2]Lancaster Environment Centre, Lancaster University, UK

Received: November 23, 2014; returned: December 30, 2014; revised: February 3, 2015; accepted: March 12, 2015.

Abstract: As technologies permitting both the creation and retrieval of data containing spatial information continue to develop, so do the number of visualizations using such data. This spatial information will often comprise a place name that may be "geocoded" into coordinates, and displayed on a map, frequently using a "heatmap-style" visualization to reveal patterns in the data. Across a dataset, however, there is often ambiguity in the geographic scale to which a place-name refers (country, county, town, street etc.), and attempts to simultaneously map data at a multitude of different scales will result in the formation of "false hotspots" within the map. These form at the centers of administrative areas (countries, counties, towns etc.) and introduce erroneous patterns into the dataset whilst obscuring real ones, resulting in misleading visualizations of the patterns in the dataset. This paper therefore proposes a new algorithm to intelligently redistribute data that would otherwise contribute to these "false hotspots," moving them to locations that likely reflect real-world patterns at a homogeneous scale, and so allow more representative visualizations to be created without the negative effects of "false hotspots" resulting from multi-scale data. This technique is demonstrated on a sample dataset taken from Twitter, and validated against the "geotagged" portion of the same dataset.

Keywords: spatial ambiguity, weighted redistribution, passively georeferenced data, false hotspots, geographic scale

1 Introduction

1.1 Passively georeferenced data

The volume of data containing spatial information is increasing rapidly, both in terms of the amount generated and its availability to the researcher. In many cases, however, the quality of the spatial information contained within this data is low, with only small proportions (circa 1% reported by this project, Craglia et al. [3] and Dredze et al. [7]) of such data typically containing direct georeferences (e.g., GPS-derived coordinate-pairs). The majority of spatial references associated with such data are more vague (e.g., place-names) and often exhibit ambiguity in either location or scale, making it difficult to define their exact locations. These ambiguities are collectively referred to as "spatial ambiguity" for the purposes of this paper. One set of data that typically suffers from such spatial ambiguity is that which may be referred to as "passively georeferenced" (PG) data.

PG describes data that contain an indirect spatial reference, but are distinguished because the spatial reference was not specifically intended as a means of locating that data on a map and as such can only be used to provide an approximate location. Examples of PG data include text that contains either a place-name or reference to features from which a place-name may be derived (such as the Eiffel Tower for Paris), or posts on a social networking website that are attached to a user profile including a "location" field that describes where they "are from." In the field of VGI (volunteered geographic information), such data may be referred to as "implicitly geographic" [3]. Conversely, "actively" georeferenced data would describe that which was intended to locate a person or entity in space at a known scale, such as where a user has identified their precise location on a map or with a GPS receiver, or provided a full postal address or postcode. In connection with VGI, such data may be referred to as "explicitly geographic" [3]. One common feature of PG data is that the location information will typically be in the form of a place-name. Place-names are described by Longley et al. [25] as the simplest form of georeferencing, which can be applied to any feature in the landscape (either physical or administrative), at any scale, and which may or may not be officially sanctioned. It is this variability in scale and meaning that causes many of the issues referred to in this paper.

An indirect spatial reference such as a place-name may be converted to a direct spatial reference by geocoding: the process of assigning a geographic identifier to a computer record that lacks it, thereby tying information to geographic space [15, 31, 34], normally in the form of a "representative point" [19]. Geocoding is "a process critical to nearly every academic, industrial, and government field that seeks to perform any type of spatial analysis or mapping" [14]; and is used ubiquitously in modern, spatially-aware web services [20]. Modern geocoding is typically performed "on the fly" by submitting an HTTP request to an online geocoding service such as the *Google Geocoding API* or *Yahoo! PlaceFinder* at little or no cost, and potentially with little appreciation of the uncertainties involved [30]. This can, in some instances, make geocoded data difficult to maintain in terms of quality, with a number of incorrect, non-official, or outdated address components being returned from some services [13].

Two main issues of spatial ambiguity arise from the use of place-names, the first of which relates to whether or not a word is spatial or aspatial (does "Reading" refer to a town in England, or a person who reads a book?); and to which spatial entity a place-name refers (in Great Britain, for example, there are nine places listed in the Ordnance

Survey 1:50,000 Gazetteer named "Whitchurch," and a further nine that include the word "Whitchurch" within their name, such as "Whitchurch-on-Thames"). Amitay et al. [1] refer to these ambiguities as "geo/non-geo," and "geo/geo" respectively. Both of these issues fall within the domain of geocoding, and solutions to both are far beyond the scope of this paper. The second spatial ambiguity arising from the use of place-names relates to the question of how the "place" itself may be spatially defined. For example, once a single occurrence of "Whitchurch" had been selected, there is no reliable way to define it in space. Some may choose to reduce it to its geometric centroid, providing a precise location, but one that is unlikely to represent the exact location to which the user was referring (an individual claiming to live in "Whitchurch" is highly unlikely to live at its geometric centroid). This is the default behavior of many of the online geocoding services, as they are intended primarily for zooming a map to a place, rather than defining that place in space. Alternatively, some may choose to generalize "Whitchurch" to a polygon representing its official administrative bounds. This approach does not, however, solve the problem of "where" within that boundary is being referred to, and may even have worsened the problem, as the official bounds may not even contain all of the areas that some individuals consider to be a part of that "place." The difficulties of defining "place" are well established in the literature [34], and are summarized effectively in [4]. A solution has, however, yet to be identified within the literature, and so it is this latter form of spatial ambiguity that this paper seeks to address.

1.2 Social media as a source of passively georeferenced data

In recent years there has been a dramatic rise in the use of social media services such as Facebook and Twitter [24] that has created a vast new set of PG data. This is because such data will generally be attached to a publicly accessible "user profile," that contains basic information about the user, such as name, birthday, likes, dislikes, and so on. Often, these profiles also include user-defined information relating to their "location" that may be accessed along with their published content through an Application Programming Interface (API). It is important to note that such "location" information is intended as a permanent human-readable description of "where the user is from," and not a repository for the ever-changing location of the user as they move in geographic space [10]. This data may, however, be used as a proxy for the location of their generated content, and the ease with which such data may be collected and located in this manner has led to a significant amount of mapping activity by researchers and the media alike.

One of the most prolific targets for such mapping has been Twitter, a social networking service that allows users to share information in the form of short text "tweets" that are limited to a length of 140 characters [27]. As a result of their wide user-base, Twitter is generating a vast amount of data, which has developed beyond "conversational" social interaction, to the publication of "terabytes of real-time 'sensor' data" [2], whereby the "sensors" are the numerous and geographically diverse users themselves. Goodchild [16] described this kind of mapping with the term "Humans as Sensors," though it is distinct from the "actively" georeferenced "volunteered geographic information" (VGI) to which he primarily refers. Nevertheless, this data provides the researcher with the opportunity not only to use the social media posts themselves, but also demographic, temporal, and geospatial data in order to derive information about human feelings or behavior in both time and space. There are many examples of data from Twitter being mapped for a vari-

ety of different purposes including: public health [7], weather reporting [10], earthquake detection [8], the assessment of Twitter itself as a platform for crowd sourcing and collaboration [6, 23], and the assessment of the "geographies of Twitter" [11, 22, 33]. These examples only represent a small portion of the available literature, but all follow a very similar methodology to map their respective data: extracting Tweets using the Twitter API, passing the contents of the "location field" of the user profile to a geocoder to be located, and then placing them on a map.

With an increase in the use of GPS-enabled Internet-connected devices such as smartphones, it is also increasingly possible for social network users to "geotag" their content with the location at which it was published. Mapping such "geotagged" data is a trivial yet powerful exercise, using the location of the Tweet as a proxy for the location of the individual [10]. There is a second large body of literature arising from mapping Tweets using this alternative methodology (e.g., [21]). Geotagged Tweets, however, locate the "Tweeter" at the time they created the data, and are therefore not necessarily comparable to the places described in their public user profile [3], although they are treated as such in some studies (e.g., [7,11]). Furthermore, such data represents a very small proportion of that made available by the social networks (circa 1% reported by this project, Craglia et al. [3] and Dredze et al. [7]), causing questions to arise about the representivity of such a small sample for the identification of patterns in the data. Other studies have attempted to extract location from the content of social media (e.g., [2]), or even to investigate other options such as the inference of location based upon the location of social connections (e.g., [5]).

1.3 Issues of spatial ambiguity and indeterminate scale

In the process of geocoding, a place-name is typically generalized to a single point in space, rather than the area that it actually represents, or indeed, the nature of the "place" that was intended by the creator of the data [19, 28]. This is helpful when placing a label on a map or zooming to a given location, but less so for the purpose of visualizing patterns in the data, as much of the information (spatial or otherwise) implied by the use of the place-name is lost through the use of a simple coordinate pair for its representation. Goldberg et al. [15] describe this as a "fundamental question" of geocoding: discussing whether or not the returned point should be the centroid of the geographic object in question, should be weighted by population distribution, or even whether a boundary should be returned instead. Either way, it is the resulting loss of spatial context that causes the spatial ambiguity to which this paper refers; whereby the approximate location is known (the coordinate pair returned from the geocoder), but the specifics or nature of the boundaries to which the data point originally referred are not. One result of this ambiguity is that it leaves no way to determine whether two locations are comparable based upon their place-names alone (e.g., is "Lancaster" a town, county, or country?). This is important, because only locations at the same scale (town, county, etc.) may be considered comparable [22], and the comparison of locations at multiple scales leads to spatial imprecision [35], and the formation of "false hotspots."

The author of a PG data point referring to Lancaster in north-west England, for example, is unlikely to be referring to the administrative area as a whole, and nor are they likely referring to its geometric centroid. In reality, the author is referring to an unknown point that likely (but not definitely) falls within the corresponding boundary. This unknown location is not likely adequately represented by the centroid of the enclosing administrative

area, which represents a generalization of the data that can lead to the formation of a visually misleading "false hotspot," the impact of which is dependent upon the geographic scale at which it is formed, and that at which it is visualized. The concept and impact of generalization is well established in the literature, and it is an accepted and necessary cartographic technique for dealing with the technical limitations of the drawing media being employed (either paper or digital) [17]. It is equally recognized, however, that generalization represents a loss of information and introduces uncertainty into the dataset. As such, cartographic decisions relating to generalization must be made carefully, and applied to objects based upon their size and importance relative to the scale at which they are being mapped [17]. In the case of geocoded PG data, however, the place-names are all generalized to a single point irrespective of scale (it may be assumed that all points are of equal importance for most applications). As a result, the level of generalization varies significantly within the dataset, with some data points generalized to the centroid of their country (a higher level of generalization), but others generalized to the centroid of their town or district (a lower level of generalization). It is the comparison of this multi-scale data with differing levels of generalization that results in the formation of "false hotspots" at the location to which the geocoder reduces each administrative area. These "false hotspots" will appear as a dense cluster of data-points on the map, but in reality represent nothing more than artifacts caused by the inability of the geocoder to locate those data-points at a larger scale.

In order to illustrate this effect, Figure 1a shows a density map of Tweets relating to the 2011 wedding of the Duke and Duchess of Cambridge, which exhibits three distinct "hotspots" and many smaller ones, all of which may be considered as "false." Of the three most distinct hotspots: two are located in the major cities of Manchester and London (labeled A and B respectively), and represent all of the Tweets that were geocoded to these cities. The data comprising these hotspots, however, is all located at the geometric centroids of these cities rather than being distributed across them at the locations from which the Tweets originated, meaning that they must be considered "false." Of greater concern is the third hotspot, which is located in the West Midlands (labeled C), away from any significant centers of population. The addition of bounding-boxes to each country in Figure 1b (shown in red) helps to identify the cause of this hotspot, clearly demonstrating that this point represents nothing more than the centroid of the bounding box of England (labeled in blue), and as such is a "false hotspot" comprised of all the data that could not be geocoded to a greater level of detail than "England." The bounding boxes also make it clear that there are similar (though less significant) hotspots visible at the centroids of Scotland and Wales (also labeled in blue).

As described by Silvan et al. [32], the extent to which a "false hotspot" is visually misleading may be considered to be the result of the relationship between the scale at which the object was observed (an infinitely precise point in space, generalized to the centroid of the enclosing administrative area), and that at which it is represented (the scale at which it is drawn on the map). As the size of the administrative area increases, more generalization is required to reduce the location to its administrative centroid, and so the more misleading the point may be. Similarly, the larger the geographic scale of the map, the more significant the apparent impact of the generalization, and so the more visually misleading a "false hotspot" will be. A county-scale "false hotspot," for example, would be extremely misleading on a large-scale map (e.g., of a city), but would be much less misleading on a small-scale map (e.g., of a continent). It has previously been suggested that most current

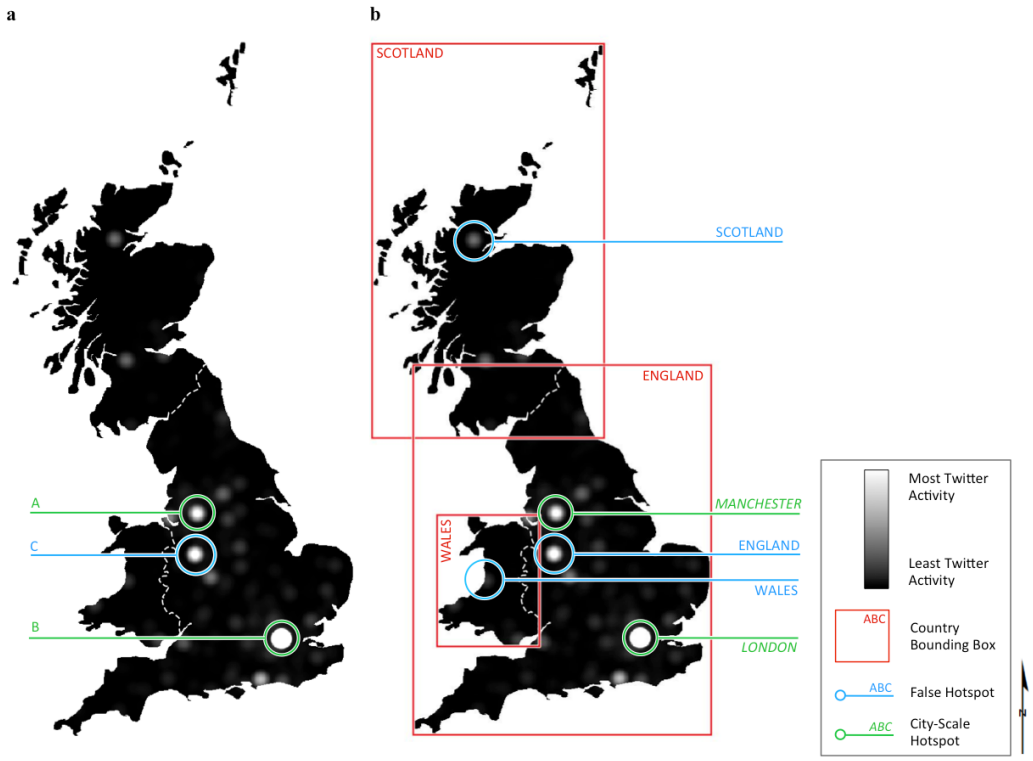

Figure 1: (a) A density map of geocoded Tweet locations in the UK. (b) Annotated with bounding boxes for each component country (in red), the associated "country level" false hotspots (in blue), and examples city-scale hotspots (in green). *The dataset is displayed using a greyscale 2%–99% histogram stretch, and was produced using a* 18 *km* × 18 *km density kernel at 1 km resolution.*

systems hide uncertainty arising from the comparison of multi-scale data "under the carpet" [35] and, following a thorough review of the causes and nature of "false hotspots," there is no evidence of any alternative algorithms or approaches to dealing with this issue within the literature. The only possible exception to this would be [18], which represents an earlier iteration of this work, and highlighted the issue of "false hotspots" but did not propose a formal solution.

The purpose of this paper, therefore, is to propose a new algorithm that allows the mapping of patterns in spatially ambiguous PG data, whilst avoiding issues associated with the lack of any explicit scale information. This algorithm will permit the removal of the "false hotspots" that would otherwise obscure patterns in the data, and thus allow PG data to be utilized more effectively in the visualization of these patterns. The algorithm "Weighted Redistribution" is therefore described in order to redistribute data from the centroid of their respective administrative areas in a manner that will reflect likely patterns in the dataset as a whole. It is not suggested that redistributed data will be either more or less accurate or precise than the "raw" PG data on a point-by-point basis, but rather that the resulting pat-

tern across a large dataset will likely be more representative of the unknown "real" patterns in phenomena. In this way, the algorithm may be considered akin to other cartographic techniques such as dasymetric mapping, whereby data is depicted using boundaries of relative homogeneity with respect to the underlying statistical surface [9]. In the case of dasymetric mapping, the cartographer may be required to make both objective and subjective decisions with regard to the division of data into boundaries, leading to the inevitable potential for human error [9, 29]. It is recognized amongst cartographers, however, that any such error may be offset by the additional information offered by the technique [29], and this is also true of Weighted Redistribution, which will not return a result that is quantifiably "correct" at the level of the individual data point, but which may reveal valuable information about the overall distribution of those data points within the context of their dataset.

Furthermore, each of the redistributed data-points will be represented by a distribution of values indicating "likelihood" of location as opposed to discrete point objects. This approach serves to better represent the spatially ambiguous nature of the true location of the point, as opposed to merely "guessing" a precise point in space, or generalizing to an entire area. The effect of this algorithm is demonstrated by case study using a sample dataset collected from Twitter relating to the highly publicized wedding of the Duke and Duchess of Cambridge, which took place in 2011. Whilst the examples used within this paper will therefore relate specifically to PG data located within Great Britain and collected from Twitter, the proposed algorithm is equally applicable to PG data from any source, or global location.

2 Methodology

Data returned from a geocoder will typically be divided into a number of distinct scales, ranging from the country or even continent level, down to street level. In accordance with the terminology dictated by Gibson et al. [12], these distinct scales will be referred to as "levels" hereafter. The number of levels that are exhibited within a given geocoded data point is simply a function of the scale to which it could be georeferenced. Data can easily be generalized to a higher level in the hierarchy (e.g., from town to country), but cannot be artificially enhanced to a lower level in the hierarchy (e.g., from country to town), as the required information (i.e., which town the data point should be attributed to) is unknown. The comparison of data at different scales is the fundamental cause of the "false hotspot" issue to which this paper refers.

The proposed algorithm iterates through every level within the dataset, every administrative area within each level, and every data point within each administrative area in order to intelligently redistribute them within their enclosing administrative area (the polygon at the given level to which they belong) based upon the parameters and weighting surface provided by the user. Rather than claiming to redistribute each individual point to its "correct" location (which would be impossible and untestable), this process moves each point to a "likely" location which, across a large dataset, will contribute to a pattern that is more representative of the "real" patterns in phenomena than the "raw" PG dataset (which suffers from "false hotspots"). Once relocated in this way, each data point is then represented as a distribution of values on a raster output surface based around the new "seed" location, with each cell in the distribution denoting the "likelihood" of a data point being located at

any point within it. The effect of this approach is to better reflect likely "true" patterns in the data, and so prevent the formation of "false hotspots" due to the comparison of multi-scale geocoded point data.

Weighted Redistribution requires two user variables and three input datasets. The user variables are a positive integer w, which determines the influence of the weighting surface over the results and a value s of between 0 and 1, which represents the level of spatial ambiguity in the dataset. The input datasets are a set of points at each level for redistribution, a set of administrative boundaries for each level, and a weighting surface. The algorithm is given as pseudo-code in Figure 2, and for the purposes of this project has been built into a Java application using the GeoTools library (http://geotools.org).

Algorithm: Weighted Redistribution (w, s, *pointData*, *weightingSurface*, *administrativeAreas*):

```
where:
    w = [user defined] desired influence of weighting surface
    s = [user defined] desired level of spatial ambiguity
    pointData = the input point dataset to be redistributed
    weightingSurface = [user defined] raster data
    administrativeAreas = [user defined] polygons for relevant administrative areas at each "level"

outputSurface = new RasterDataset(...)
for each level in administrativeAreas as level:
    for each administrativeArea in level as admin:
        centroid = [get the geometric centroid of admin]                    (Standard Geographical Operation)
        br = [get the bounding radius of admin]                             (Standard Geographical Operation)
        points[] = [get all data points from pointData within admin]        (Standard Geographical Operation)
        for each point in points[] as p:
            for i = 0 to i < n:
                do until p[i] is within admin:                              (Standard Geographical Operation)
                    p[i] = [random point within br]
                loop
                value = [value from weighting surface at p[i]]
                if value > max then:
                    max = i
                end if
            next i
            point = p[max]
            r = [based upon the geometric area of admin and variable s]     (Equation 1)
            for each cell in distribution [radius r around point p]:
                cell = [calculate cell value in distribution]               (Equation 2)
            next cell
            [add distribution to outputSurface]
        next p
    next admin
next level
return outputSurface
```

Figure 2: A pseudo-code representation of the proposed algorithm for Weighted Redistribution.

For each administrative area, all of the points at the corresponding level that are geometrically "within" it must be extracted. This type of spatial query is standard within geographical software and software libraries (including GeoTools), and so does not warrant specific discussion within this methodology. The geometric centroid and bounding

radius for that administrative area are then calculated as illustrated in Figure 3a, and are used to select w random locations within the administrative area to be evaluated against the weighting surface, as illustrated in Figure 3b, which are determined by selecting a random distance d between 0 and the bounding radius, and a random azimuth a. The "random" numbers used in this implementation are returned from the standard Java Random() class, which returns pseudo-random numbers at an approximately uniform distribution within the specified range. The coordinates are then calculated for a point at distance d and azimuth a from the centroid using simple trigonometry, and assessed as to whether or not they are located within the parent polygon. If so, then the coordinates become one of the w locations; if not, then the point is discarded. Finally, each of the w locations is evaluated against the weighting surface, as illustrated in Figure 3c, and the one with the greatest weighting value becomes the new "seed" location for that point, as illustrated in Figure 3d.

A distribution is constructed around the selected seed location, based upon the premise that there is no way of knowing what specific location within a given administrative area the PG data point was intended to refer to, and so a discrete point may not be the most appropriate representation for it. Each cell in the distribution represents a value of between 1 (at the seed location itself), and 0 (at distance r from the seed location), which serves to represent the "likelihood" of that cell containing the correct location of that data point. The radius r is computed from the radius of a circle equal in area to the administrative area and the user variable s:

$$r = \sqrt{\frac{As}{\pi}} \tag{1}$$

where A is the area of the parent polygon of the seed. The value v of each cell in the distribution is therefore computed as a linear radial distance from the seed location, scaled between 1 and 0:

$$v = 1 - \frac{\sqrt{(seed_x - cell_x)^2 + (seed_y - cell_y)^2}}{r} \tag{2}$$

where r is the radius determined from the input variable s. The assignment of values to the distribution is illustrated in Figure 4 for clarity. The distribution is then added to an output surface that, in the case of the Java application created for this project, is written to a GeoTiff file.

As already discussed, the intention of Weighted Redistribution is not to place each individual data point in its "correct" location, since any attempt to do so would be untestable because the "correct" location of each PG data point is unknown. Rather, the intention of this process is to identify an area within which the data point may likely be located. Whilst this is not useful for a single or even a few PG data points; when applied across an entire PG dataset containing thousands or millions of data points, likely spatial patterns in the data may be inferred to a given degree of "weighting" and "spatial ambiguity." The effect of this is the reproduction of patterns at a larger scale than that at which the data could originally be geocoded, providing a pattern that may be considered overall to be more representative of the (unknown) true patterns in the phenomena in question and thus permitting a greater level of understanding of those patterns. The representivity of the output pattern will, however, depend upon the suitability of the weighting surface, and the appropriate selection of variables w (weighting) and s (spatial ambiguity in the dataset) by the researcher. Given the known low proportion of datasets that will typically exhibit direct georeferences (e.g., a GPS-derived coordinate pair), it may be considered in some circumstances, and for some

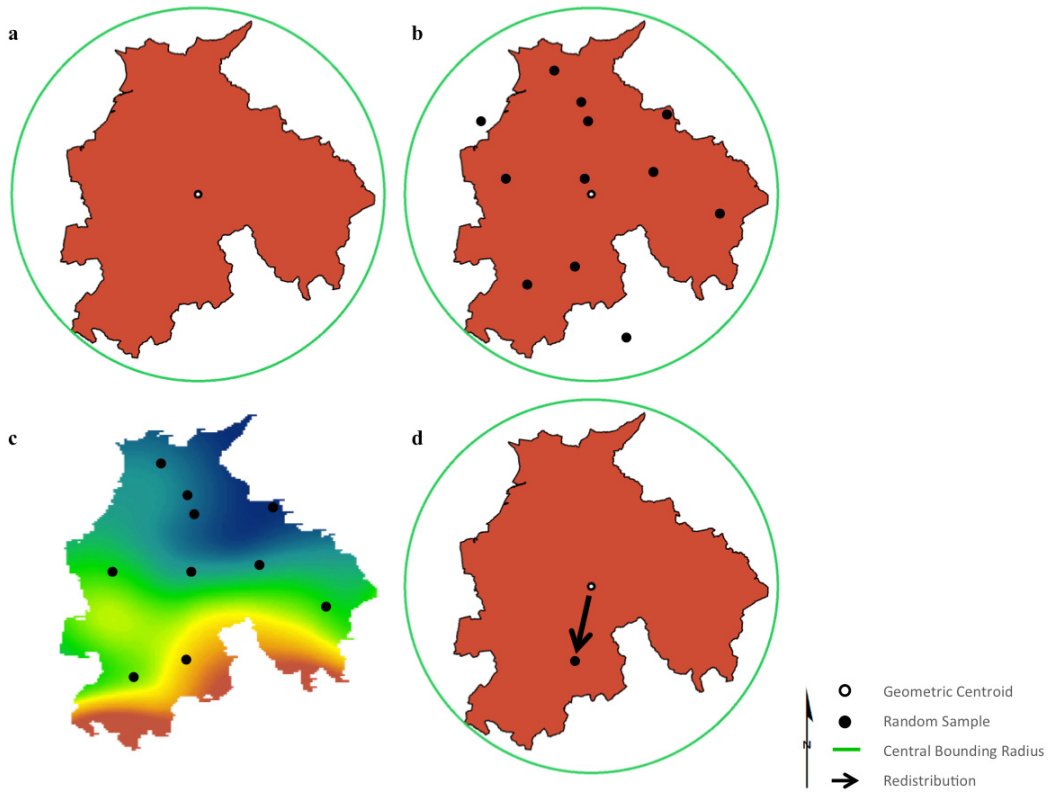

Figure 3: Illustration of the identification of seed locations for Weighted Redistribution: (a) Identification of the geometric centroid (in white) and central bounding radius (in green) around the county of Lancashire. (b) Identification of w random locations (in black) within the bounds of Lancashire using the central bounding radius. Locations falling outside of the bounds are discarded. (c) Use of the weighting surface to identify which of the w locations has the greatest value taken from the provided weighting surface. (d) Movement of the point from its original location at the geometric centroid (in white), to the new seed location. A distribution will be constructed about this seed location.

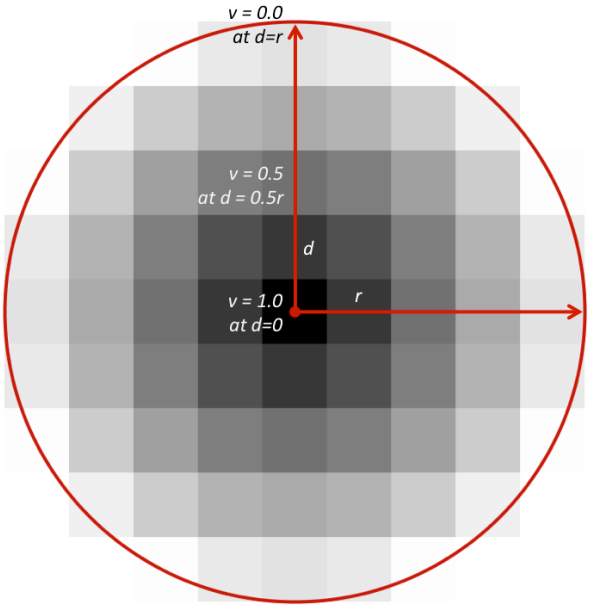

Figure 4: Figure illustrating the calculation of cell values (v) for the output distribution, where r is the distribution radius determined in Equation 1, and d is the distance from the seed location. The calculation for the calculation of these values is given in Equation 2.

applications, that a redistributed 99% sample of the data could be more representative of overall patterns than a directly georeferenced 1% sample.

3 Illustrative example

In order to illustrate the effect of this algorithm upon a PG dataset, it is applied here to a database of Tweets in connection with the wedding of the Duke and Duchess of Cambridge, which took place on Friday 29th April 2011 (the "Royal Wedding"), and received significant media coverage and discussion on the social networks. For the purpose of this exercise, only Tweets that were successfully geocoded to locations within Great Britain were used, which amounted to 550,171 Tweets. The dataset was collected using the *Twitter API* (http://dev.twitter.com) and only the textual location data from within the "location" field of the user profile were used to locate the Tweets (any GPS coordinates were ignored). The dataset was geocoded by simply feeding the "location" text directly to the *Yahoo! PlaceFinder* online geocoding service (http://developer.yahoo.com), and the results were loaded into a relational database. The sample data exhibited 4 levels: country (38,701 Tweets), county (40,905 Tweets), town/city (409,911 Tweets), and "better" than town/city (60,654 Tweets). Administrative boundaries for each of these levels were used as the "boundary" data; with a simple surface of population density constructed from 2001 census data to act as the weighting surface. This population surface has not been filtered or weighted to reflect the demographic distribution of Twitter users, though this may be

an appropriate approach for some analyses. A detailed discussion of the Tweet dataset, the manner in which it was collected, and the construction of the weighting surface have not been included here, as they merely represent a sample data for the illustration of the proposed algorithm, and as such are not of material importance to this work.

The Tweet data was processed using the above algorithm, with s values of 0.001, 0.01, and 1 for country, county, and "town and better" respectively; and a w value of 20 was used for all levels. These values were chosen based upon the perceived spatial ambiguity inherent within the dataset, the variation in size of the various administrative areas, and the desired influence of the weighting surface. These decisions were made subjectively by the authors, based upon a detailed exploration of the impact of adjustments to each variable with a subset of this dataset (referred to later in this paper, and illustrated in Figure 6). As with many similar GIS algorithms (such as kernel density estimation, or clustering for example), the specific input variables can rarely be fully justified, which is discussed further later in this paper. The dataset is shown prior to processing in Figure 5a and post processing in Figure 5b. Visual comparison between the two figures illustrates the significant effect that Weighted Redistribution has upon a dataset, with the data in Figure 5b representing a smooth surface, as opposed to the collection of multi-scale hotspots that are evident in Figure 5a. All of the "false hotspots" evident in Figure 1b and Figure 5a have been redistributed to areas of appropriate population within their respective administrative areas. This is evident, for example, in the case of Manchester (labeled in Figure 1b), which has shifted from a single "false hotspot" at the centroid of the city to a reflection of the urban extents of the city and surrounds; and in the case of England, where the misleading hotspot at its center (labeled in Figure 1b) has been completely removed to other more likely locations within the country. The removal of the "false hotspots" has also significantly lowered the range of values contained within the surface, allowing a much more effective use of the color ramp, which reveals patterns of activity in areas that were once hidden by the magnitude of values contained within the "false hotspots."

The use of distributions as opposed to discrete points has allowed the formation of a surface of values indicating the likelihood of Tweet activity relating to the "Royal Wedding" originating from any given point within the study area. As the weighting surface for this illustrative example was population density, the probability is effectively derived from the spatial coincidence of Tweet activity as recorded in the database (erroneously located at the centroid of the parent polygon), and areas of high population density within the parent polygon. The patterns given in Figure 5b are therefore far more useful as an approximation of true patterns of activity than those patterns evident in the "raw" data (Figure 5a; Figure 1) as, although neither could be described as "correct," they reflect real patterns of related phenomena (population density in this case), as opposed to non-relevant geometric approximation. In the case of data such as that from Twitter, this algorithm could therefore productively be applied to any data intended for the identification of patterns of activity. Examples could include attempts to map spatial patterns in Tweets relating to a given topic (such as the "Royal Wedding," "Olympics," or "World Cup" for example); or in the semantic properties of the text contained within those Tweets (such as evidence of feelings such as happiness or sympathy, or linguistic devices such as spatial variation in the use of certain colloquialisms for example).

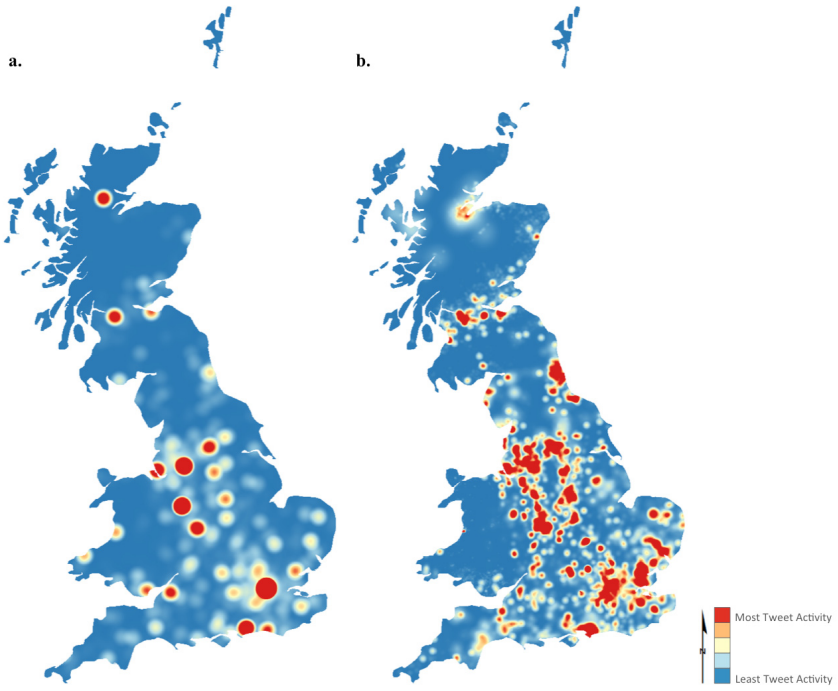

Figure 5: (a) A "raw" density map of geocoded Twitter activity relating to the wedding of the Duke and Duchess of Cambridge, exhibiting a pattern of multi-scale "false hotspots." (b) The result of performing Weighted Redistribution upon the same dataset, exhibiting a smooth surface representing a likely pattern in the data. Both datasets displayed using a 5-color, 2% clip histogram stretch, and (a) was produced using a 18 km $\times$ 18 km density kernel at 1 km resolution.

4 Discussion and conclusions

Passively geolocated (PG) data is a vast and important source of spatial data that continues to grow alongside the web, digital text resources and the social networks. Advances in natural language processing permit an increasing depth of understanding to be obtained from many forms of such text-based data, but suffer from the variety and inexplicit nature of the geographic scale at which the spatial component of each data point is produced. Weighted Redistribution represents a technique by which patterns in PG data may be reconstructed and visualized whilst avoiding the introduction of "false hotspots." In this way, these patterns may be explored using more representative visualizations than would be possible using only directly georeferenced data, and without the erroneous patterns inherent in PG datasets. Whilst the examples contained within this paper have been based around data collected from Twitter in the UK, the algorithm is equally applicable to any PG dataset, from any source, and relating to any other country or part of the world.

As with other raster-based GIS techniques, such as kernel density estimation for example, the output of Weighted Redistribution is heavily dependent upon the user. As such, an understanding of the algorithm, and particularly the effect of these variables, is required

before a meaningful visualization may be created using Weighted Redistribution. Figure 6 provides an illustration of the effect that adjustments in the variables s (spatial ambiguity in the dataset), and w (weighting) have upon Tweets collected in the county of Lancashire in the UK. A matrix is presented with a number of Weighted Redistribution layers created from the same dataset, but with different values for the two variables, with w increasing along the x-axis, and s increasing along the y-axis. The effects of these adjustments are very clear: with highly randomized datasets to the left of the matrix (where data is spread across the whole county), and highly population-weighted datasets to the right (where data is concentrated towards population centers). Similarly, the matrix exhibits relatively clear data boundaries towards the bottom of the matrix, and more ambiguous boundaries towards the top. Another clear pattern revealed by this figure is the inherent conflict between the two variables, with the effect of adjustments in w at the top (most spatially ambiguous) row of the matrix greatly diminished in effect when compared to the bottom (least spatially ambiguous) row. The selection of appropriate values for a given dataset is likely to depend upon factors such as confidence in the administrative boundary data and weighting surface, as well as the goals of the study, and in reality will likely be arrived at through a process of trial and error, rather than specific rules or calculations. Users of the Weighted Redistribution algorithm or software are encouraged to experiment widely with the input variables and weighting surface, and to vary them at each level, not just for each dataset.

The construction of an appropriate weighting surface will also have a significant effect upon the quality of the output, with the magnitude of its effect proportional to the variable w. If an inappropriate weighting surface is used (for example, one based upon incomplete, unsuitable or irrelevant data), then this will lead to the formation of unrealistic patterns in the data, and as such will have a dramatic impact upon Weighted Redistribution process. It is vital, therefore, that an appropriate level of consideration is given to the creation of weighting surfaces that are representative of the issues relevant to the data: population distribution, demographic data, proximity to a given phenomenon, or habitat suitability for example. It is also important that the weighting surface data is at an appropriate resolution to allow sufficiently detailed redistribution of data points. A city-scale redistribution across a weighting surface with a resolution of 1 km, for example, would not yield useful results, as there would be insufficient variation in weighting values across the study area to form meaningful patterns. At the country scale, however, a weighting surface at 1 km resolution would provide an excellent level of detail and variation, and permit very detailed patterns to form.

Unsuitable boundary data may fall foul of issues similar to the Modifiable Aerial Unit Problem (MAUP), described by Openshaw [26]. This is because the redistribution is heavily influenced by the spatial properties of each data point's enclosing polygon. Given that administrative boundary data typically change over time, and do so with relative frequency, the researcher must be careful to select boundary data that not only represents the appropriate scale as closely as possible (town, county etc.), but also the temporal nature of that geography. The town of Beverley in the UK, for example was, prior to 1996, located within the now abolished county of "Humberside," but has since been reassigned to the county of the "East Riding of Yorkshire." As such, different boundary data would need to be used in order to perform Weighted Redistribution on a dataset collected prior to, and since, 1996. For datasets with a large temporal variation, therefore, it may be necessary to subdivide the data temporally in order to reflect changes in administrative boundaries, process each sub-

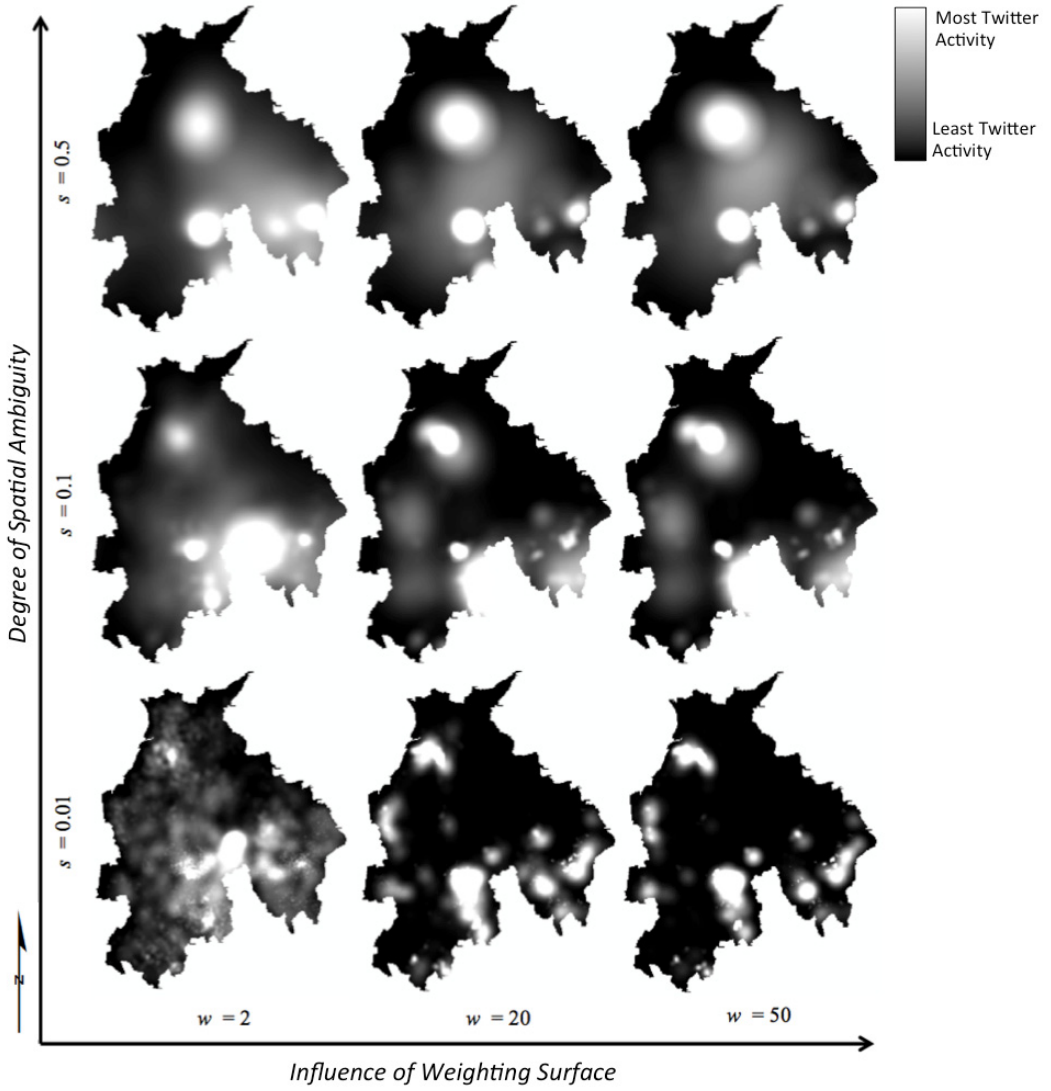

Figure 6: Weighted Redistribution surfaces for Tweets collected within Lancashire, with a range of values for w (weighting), and s (spatial ambiguity). The effect of an increased influence of the weighting surface from left to right (reducing randomness in the resulting patterns), and an increased level of spatial ambiguity from bottom to top (reducing the level of detail in the resulting patterns) are clearly visible.

division individually, and then aggregate the results in order to produce the final output data.

Weighted Redistribution could be described as an attempt to "get something for nothing," whereby it is attempting to introduce an increased level of detail into data for the identification of spatial patterns. Whilst this is to some extent true, the selection of an appropriate weighting surface and values for w and s can lead to an improvement in the representivity of the dataset, making it reflect a likely spatial distribution of the phenomenon in question which, across a large dataset, can provide a better representation of reality than the "raw" input data. It could be argued that it is not possible to quantify the effect of this algorithm, as the "true" location for each data point is unknown, and the locations returned from the geocoder are known to be incorrect due to the inherent geometric simplification. As previously discussed, however, this approach may be seen as analogous to other cartographic techniques such as dasymetric mapping, whereby resulting data may not necessarily be "correct," but any errors are offset by the information that the technique reveals [26]. Nevertheless, it is possible to demonstrate the effect of the algorithm with an exercise such as is demonstrated in Figure 7. These data represent the circa 1% of the Tweets collected associated with the Royal Wedding that were directly georeferenced using a GPS receiver. Figure 7 demonstrates the effect of "reducing" the location of each individual data point to the centroid of their respective country (top row) and county (bottom row) in order to simulate the locational precision of PG data, and then redistributing those data using the algorithm described in this paper (using the same input data and parameters as per Figure 5). It is clearly apparent in both cases that the redistributed data represent the true patterns in the data more effectively than do the simulated PG data, thus demonstrating the benefit of this algorithm for the visualization of spatially ambiguous point data.

Though the proposed algorithm is intended as a solution to the visualization of passively georeferenced data, the authors consider that there is further work that could be done in this area. For instance, there could be some benefit to an investigation into the relationship between the input variables and the size and shape of administrative areas, and how this could be incorporated into the algorithm without sacrificing usability. This is beyond the scope of this work, but could certainly lead to further developments in the application of this algorithm. Other questions to ask of this work in the future include whether or not these visualization techniques could be usefully applied to areas beyond cartographic visualization, such as decision support for example. It is likely that the patterns produced by this algorithm will be of some use in the identification of patterns in phenomena across a whole population or dataset, for example, which could contribute towards decision-making, but such approaches would need to be validated prior to application.

Whilst it may be argued that an increasing amount of data is now being "geotagged," rendering techniques such as this outdated, it is important to note that, thus far, uptake of geotagging appears to be slow, with only circa 1% of the Twitter data collected for this project (in 2011) associated with a coordinate pair (a similar figure was reported by Craglia et al. [3] and Dredze et al. [7]). To use the geotagged data alone therefore, would result in the discarding of circa 99% of the dataset, which could be viewed as an unacceptable amount for a representative visualization. Given the likelihood that a mobile-enabled platform and pioneer of geotagging such as Twitter will see uptake far before other more traditional forms of content, and that this uptake will likely introduce spatial biases of its own, tech-

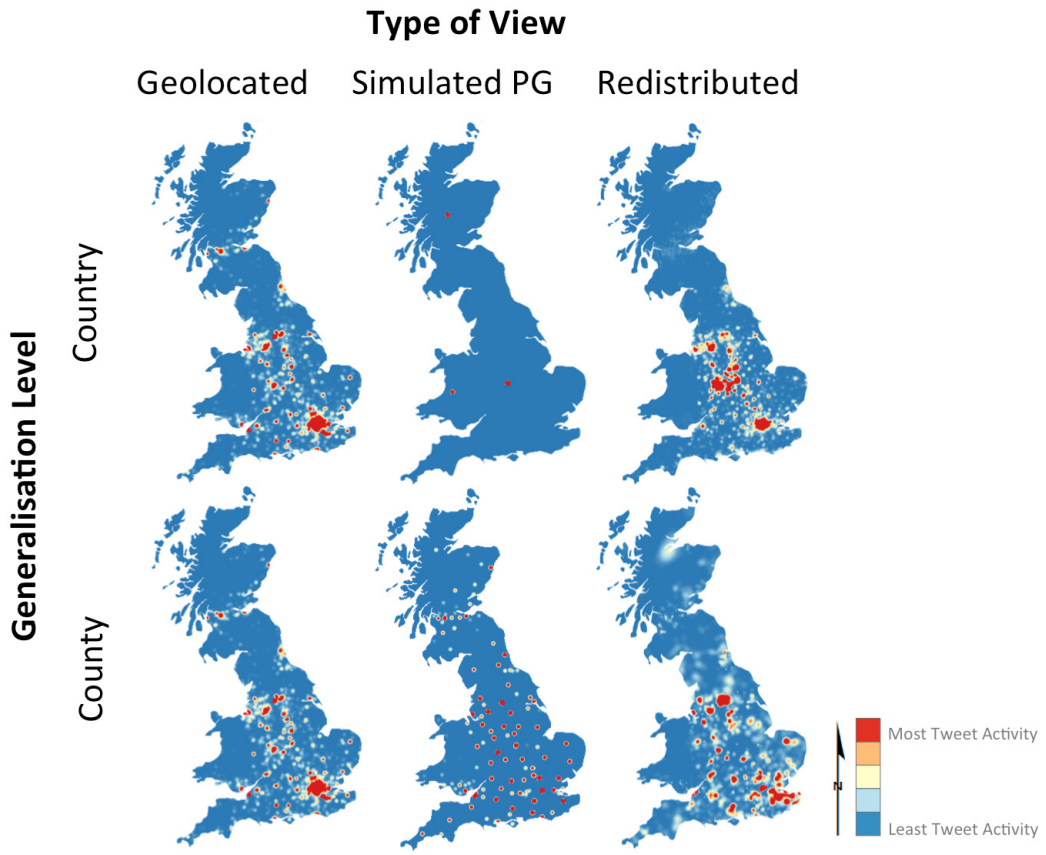

Figure 7: An illustrative exercise, whereby directly georeferenced data are generalized to the centroid of their respective country and county administrative area centroids in order to replicate the effect of PG data, and then redistributed using the algorithm described within this paper. This demonstrates the benefit of the redistributed data over the generalized. Maps in (c) are displayed using a 5-color, 2% clip histogram stretch, and the "Geolocated" and "Simulated PG" figures were produced using an 8 km $\times$ 8 km density kernel at 2 km resolution.

niques such as Weighted Redistribution are invaluable to enable the researcher to identify and explore spatial patterns in this data.

Furthermore, even if the vast majority of data is eventually geotagged at the point of creation (including web-based information, social media, books, media content, and so on) there will still be an enormous amount of historical data, including much of that produced today, that would still require methods such as Weighted Redistribution in order that patterns in the data might be revealed. Weighted Redistribution represents a new approach to the processing and visualization of PG data, preventing the requirement for researchers to choose between an accurate yet very small (circa 1%) and potentially unrepresentative sample, and a much larger (circa 99%) sample prone to "false hotspots." As such, this algo-

rithm provides new opportunities for researchers to visualize the spatial patterns in such datasets, allowing a greater level of understanding of those patterns than was previously achievable.

Acknowledgments

The authors thank Mark Lochrie (formerly of Lancaster University School of Computing and Communications) for assistance with the data collection infrastructure, and the four reviewers of this work for their suggested improvements to this manuscript.

References

[1] AMITAY, E., HAR'EL, N., SIVAN, R., AND SOFFER, A. Web-a-where: Geotagging web content. In *Proc. 27th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval* (New York, NY, USA, 2004), SIGIR '04, ACM, pp. 273–280. doi:10.1145/1008992.1009040.

[2] CHENG, Z., CAVERLEE, J., LEE, K., AND SUI, D. Z. Exploring millions of footprints in location sharing services. In *Proc. fifth International AAAI Conference on Weblogs and Social Media* (2011), pp. 81–88.

[3] CRAGLIA, M., OSTERMANN, F., AND SPINSANTI, L. Digital Earth from vision to practice: Making sense of citizen-generated content. *International Journal of Digital Earth 5*, 5 (2012), 398–416. doi:10.1080/17538947.2012.712273.

[4] CRESSWELL, T. *Place: A short introduction*. John Wiley & Sons., 2004.

[5] DAVIS JR., C. A., PAPPA, G. L., DE OLIVEIRA, D. R. R., AND DE L. ARCANJO, F. Inferring the location of Twitter messages based on user relationships. *Transactions in GIS 15*, 6 (2011), 735–751. doi:10.1111/j.1467-9671.2011.01297.x.

[6] DEMIRBAS, M., BAYIR, M., AKCORA, C., YILMAZ, Y., AND FERHATOSMANOGLU, H. Crowd-sourced sensing and collaboration using Twitter. In *Proc. IEEE International Symposium on a World of Wireless Mobile and Multimedia Networks (WoWMoM)* (2010), pp. 1–9. doi:10.1109/WOWMOM.2010.5534910.

[7] DREDZE, M., PAUL, M. J., BERGSMA, S., AND TRAN, H. Carmen: A Twitter geolocation system with applications to public health. In *Proc. AAAI Workshop on Expanding the Boundaries of Health Informatics Using AI (HIAI)* (2013), Citeseer, pp. 20–24.

[8] EARLE, P., GUY, M., BUCKMASTER, R., OSTRUM, C., HORVATH, S., AND VAUGHAN, A. OMG earthquake! can Twitter improve earthquake response? *Seismological Research Letters 81*, 2 (2010), 246–251. doi:10.1785/gssrl.81.2.246.

[9] EICHER, C. L., AND BREWER, C. A. Dasymetric mapping and areal interpolation: Implementation and evaluation. *Cartography and Geographic Information Science 28*, 2 (2001), 125–138. doi:10.1559/152304001782173727.

[10] FIELD, K., AND O'BRIEN, J. Cartoblography: Experiments in using and organising the spatial context of micro-blogging. *Transactions in GIS 14* (2010), 5–23. doi:10.1111/j.1467-9671.2010.01210.x.

[11] FINK, C., KOPECKY, J., BOS, N., AND THOMAS, M. Mapping the Twitterverse in the developing world: An analysis of social media use in Nigeria. In *Social Computing, Behavioral-Cultural Modeling and Prediction*. Springer, 2012, pp. 164–171.

[12] GIBSON, C. C., OSTROM, E., AND AHN, T. The concept of scale and the human dimensions of global change: a survey. *Ecological Economics 32*, 2 (2000), 217–239. doi:10.1016/S0921-8009(99)00092-0.

[13] GOLDBERG, D. W. Advances in geocoding research and practice. *Transactions in GIS 15*, 6 (2011), 727–733. doi:10.1111/j.1467-9671.2011.01298.x.

[14] GOLDBERG, D. W. Improving geocoding match rates with spatially-varying block metrics. *Transactions in GIS 15*, 6 (2011), 829–850. doi:10.1111/j.1467-9671.2011.01295.x.

[15] GOLDBERG, D. W., WILSON, J. P., AND KNOBLOCK, C. A. From text to geographic coordinates: The current state of geocoding. *URISA journal 19*, 1 (2007), 33–46.

[16] GOODCHILD, M. F. Citizens as sensors: The world of volunteered geography. *GeoJournal 69*, 4 (2007), 211–221.

[17] GOODCHILD, M. F., AND PROCTOR, J. Scale in a digital geographic world. *Geographical and environmental modelling 1* (1997), 5–24. doi:10.1007/s10708-007-9111-y.

[18] HUCK, J., WHYATT, D., AND COULTON, P. Challenges in geocoding socially-generated data. In *Proc. GIS Research UK 20th Annual Conference* (Lancaster, 2012), D. Whyatt and B. Rowlingson, Eds., vol. 1, Lancaster University, pp. 39–45.

[19] JONES, C. B., PURVES, R. S., CLOUGH, P. D., AND JOHO, H. Modelling vague places with knowledge from the Web. *International Journal of Geographical Information Science 22*, 10 (2008), 1045–1065. doi:10.1080/13658810701850547.

[20] JUNG, C., KARCH, D., KNOPP, S., LUXEN, D., AND SANDERS, P. Engineering efficient error-correcting geocoding. In *Proc. 19th ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems* (2011), ACM, pp. 469–472. doi:10.1145/2093973.2094050.

[21] KINSELLA, S., MURDOCK, V., AND O'HARE, N. I'm eating a sandwich in Glasgow: Modeling locations with Tweets. In *Proc. 3rd International Workshop on Search and Mining User-Generated Contents* (2011), ACM, pp. 61–68.

[22] LEETARU, K., WANG, S., CAO, G., PADMANABHAN, A., AND SHOOK, E. Mapping the global Twitter heartbeat: The geography of Twitter. *First Monday 18*, 5 (2013).

[23] LIU, S. B., AND PALEN, L. The new cartographers: Crisis map mashups and the emergence of neogeographic practice. *Cartography and Geographic Information Science 37*, 1 (2010), 69–90. doi:10.1559/152304010790588098.

[24] LOCHRIE, M., AND COULTON, P. Mobile phones as second screen for TV, enabling inter-audience interaction. In *Proc. 8th International Conference on Advances in Computer Entertainment Technology* (2011), ACM, p. 73.

[25] LONGLEY, P. A., GOODCHILD, M. F., MAGUIRE, D. J., AND RHIND, D. W. *Geographic Information Systems and Science*, 3rd ed. John Wiley and Sons, 2011.

[26] OPENSHAW, S., AND OPENSHAW, S. The modifiable areal unit problem. Geo Abstracts University of East Anglia.

[27] PHUVIPADAWAT, S., AND MURATA, T. Breaking news detection and tracking in twitter. In *Proc. IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology (WI-IAT)* (2010), vol. 3, IEEE, pp. 120–123.

[28] RAHIMI, S., YANG, H., COBB, M., ZHOU, H., ALI, D., AND PETRY, F. E. An inexact inferencing strategy for spatial objects with determined and indeterminate boundaries. In *Proc. 12th IEEE International Conference on Fuzzy Systems (FUZZ'03)* (2003), vol. 2, IEEE, pp. 778–783.

[29] ROBINSON, A. H., MORRISON, J. L., MUEHRCKE, P. C., KIMERLING, A. J., AND GUPTILL, S. C. *Elements of Cartography*, 6th ed. John Wiley & Sons, 1995.

[30] ROONGPIBOONSOPIT, D., AND KARIMI, H. A. Comparative evaluation and analysis of online geocoding services. *International Journal of Geographical Information Science 24*, 7 (2010), 1081–1100. doi:10.1080/13658810903289478.

[31] RUSHTON, G., ARMSTRONG, M. P., GITTLER, J., GREENE, B. R., PAVLIK, C. E., WEST, M. M., AND ZIMMERMAN, D. L. Geocoding in cancer research: A review. *American Journal of Preventative Medicine 30*, 2 (2006), S16–S24. doi:10.1016/j.amepre.2005.09.011.

[32] SILVÁN-CÁRDENAS, J. L., WANG, L., AND ZHAN, F. Representing geographical objects with scale-induced indeterminate boundaries: A neural network-based data model. *International Journal of Geographical Information Science 23*, 3 (2009), 295–318. doi:10.1080/13658810801932021.

[33] TAKHTEYEV, Y., GRUZD, A., AND WELLMAN, B. Geography of Twitter networks. *Social Networks 34*, 1 (2012), 73–81. doi:10.1016/j.socnet.2011.05.006.

[34] TOBLER, W. Geocoding theory. In *Proc. National Geocoding Conference* (Washington DC, 1972), vol. 1, Department of Transportation.

[35] WORBOYS, M. Imprecision in finite resolution spatial data. *GeoInformatica 2*, 3 (1998), 257–279. doi:10.1023/A:1009769705164.