INTERNATIONAL JOURNAL OF GEOGRAPHICAL INFORMATION SCIENCE, 2017
VOL. 31, NO. 6, 1245–1271
http://dx.doi.org/10.1080/13658816.2016.1273357

Taylor & Francis
Taylor & Francis Group

A data-synthesis-driven method for detecting and extracting vague cognitive regions

Song Gao[a], Krzysztof Janowicz[a], Daniel R. Montello[a], Yingjie Hu[b], Jiue-An Yang[c], Grant McKenzie[d], Yiting Ju[a], Li Gong[e], Benjamin Adams[f] and Bo Yan[a]

[a]Department of Geography, University of California, Santa Barbara, CA, USA; [b]Department of Geography, University of Tennessee, Knoxville, TN, USA; [c]Department of Geography, San Diego State University, San Diego, CA, USA; [d]Department of Geographical Sciences, University of Maryland, College Park, MD, USA; [e]Institute of Remote Sensing and Geographical Information Systems, Peking University, Beijing, China; [f]Centre for eResearch, The University of Auckland, Auckland, New Zealand

ABSTRACT

Cognitive regions and places are notoriously difficult to represent in geographic information science and systems. The exact delineation of cognitive regions is challenging insofar as borders are vague, membership within the regions varies non-monotonically, and raters cannot be assumed to assess membership consistently and homogeneously. In a study published in this journal in 2014, researchers devised a novel grid-based task in which participants rated the membership of individual cells in a given region and contrasted this approach to a standard boundary-drawing task. Specifically, the authors assessed the vague cognitive regions of *Northern California* and *Southern California*. The boundary between these cognitive regions was found to have variable width, and region membership peaked not at the most northern or southern cells but at substantially less extreme latitudes. The authors thus concluded that region membership is about attitude, not just latitude. In the present work, we reproduce this study by approaching it from a computational *fourth-paradigm* perspective, i.e., by the synthesis of high volumes of heterogeneous data from various sources. We compare the regions which we identify to those from the human-participants study of 2014, identifying differences and commonalities. Our results show a significant positive correlation to those in the original study. Beyond the extracted regions themselves, we compare and contrast the empirical and analytical approaches of these two methods, one a conventional human-participants study and the other an application of increasingly popular data-synthesis-driven research methods in GIScience.

ARTICLE HISTORY
Received 15 August 2016
Accepted 12 December 2016

KEYWORDS
Place; cognitive regions; vagueness; data synthesis; latent Dirichlet allocation

1. Introduction and motivation

In its broadest sense, the concept of a *region* describes a bounded spatial extent characterized by the similarity or invariance of a set of properties. This includes the

CONTACT Song Gao sgao@geog.ucsb.edu
Present affiliation for Benjamin Adams is Department of Geography, University of Canterbury, Christchurch, New Zealand

region defined by the property of *always facing away from the Earth*, i.e., the dark side of the moon, as well as regions defined by convention such as the thoracic anatomical region that encompasses the chest. Geographic information science is typically concerned with regions in geographic space that enable us to differentiate places inside of a region from those outside of it (Montello 2003). This includes *administrative regions* with fiat, institutional boundaries (Smith and Varzi 2000, Galton 2003) where the membership of places is exclusively determined by a binary containment relation (Frank 1996), for example, all counties in the state of California are completely and equally within California. Consequently, such regions do not have a graded structure; Santa Barbara County is not a lesser part of California than Los Angeles County. Interestingly, such administrative regions are generally the only type of regions that can accurately be described by the infinitely thin-line geometries that dominate GIS to date (Couclelis 1992). Instead, geographic regions typically have boundaries that are more or less vague.

Boundary vagueness occurs for one or more of a variety of specific reasons; Montello (2003) listed *measurement, temporal, multivariate, contested*, and *conceptual* vagueness. For example, the boundaries of the Kashmir region are disputed. Nonetheless, India, China, and Pakistan have their own national policies that exactly specify those boundaries (Goodchild 2011); this is contested vagueness. Other types of regions, such as *thematic regions*, are potentially multivariate. For example, the precise boundaries of ecological biomes can neither be acquired by measurement – as this would require an infinity dense mesh of simultaneous observations of all their properties, nor by theoretical considerations – as the concept of a biome is not specified to a degree that would enable the extraction of crisp boundaries (Bennett 2001, Montello 2003). Consequently, thematic regions generally have two-dimensional boundaries and a graded structure. Places near the boundary may be less characteristic of the region than those in the center. In fact, the boundary zone between two regions is often of particular scientific interest, such as in studies of the upper timberline (Galton 2003, Holtmeier 2009). As noted by Mark *et al*. (1999), fiat boundaries are often projected onto physical space without a clear discontinuity of property values, for example, in the case of valleys and their relation to mountains, or by introducing different kinds of barriers (White and Stewart 2015).

Another type of region arises from the complex interaction of individuals, society, and the environment. These *cognitive regions* (Montello 2003) are informal regions that are also characterized by vague boundaries (Bennett 2001) and variable membership functions. Furthermore, the membership of places within a cognitive region may vary non-monotonically; membership strength does not necessarily decrease toward the boundaries, and in theory may vary up and down within the region. Cognitive regions can also vary in extent, shape, and location among groups and individuals, and can be highly specific to a local population; therefore neither homogeneity nor regularity can be assumed. Consequently, cognitive regions and places are difficult to handle computationally, for example, in spatial analysis, cartography, geographic information retrieval, and GIS workflows in general. Interestingly, the spatial properties of cognitive regions are driven by individual and cultural beliefs about thematic properties to such a degree that metric, directional, or mereotopological (Casati and Varzi 1999) properties are relaxed or even ignored. For instance, as will be discussed later in this work, San

Diego is perceived as less *Southern California* than is Los Angeles, despite San Diego being more than 150 km to the south of Los Angeles. We call this a *platial effect* (rather than a *spatial effect*) in this paper, to highlight the fact that thematic and cultural aspects of the landscape can distort or relax spatial properties.

Understanding, assessing, and characterizing cognitive regions and their vague boundaries have been ongoing research activities for years. To give just a few examples, the *egg-yolk* theory proposes the use of concentric subregions to distinguish between an inner (certain) subregion, the *yolk*, and one or more outer, less certain region, called the *white*, those jointly form the *egg* (Cohn and Gotts 1996). In their *Where's downtown* paper, Montello *et al*. (2003) reviewed three strategies to elicit an individual's representation of a region: by sketching the boundary, through a binary regular grid, and by selective binary trial-and-error sampling. Prior to this, Aitken and Prosser (1990) analyzed the cognition of neighborhood continuity and form by their residents. In their most recent work, Montello *et al*. (2014) (MFP, for short) proposed a novel grid-based technique in which participants rated the membership of individual cells at a high resolution. This allows participants to express their beliefs about nonuniform region membership and vague boundaries in detail, and it puts few restraints on the spatial distribution of region membership patterns. For example, it allows membership variation to weaken and strengthen not nonlinearly but even non-monotonically.

In the MFP study, 44 students from UCSB were presented with an outline map of California covered by a hexagonal tessellation of 90 cells (see Figure 1). The students were asked to rate each and every cell on a 1–7 scale, with 1 meaning very *Northern Californian*, 7 meaning very *Southern Californian*, and 4 meaning equally northern and southern Californian. The students were explicitly asked to base their judgment on not just cardinal directions but what people informally mean when they say *Northern California* and *Southern* California, i.e., to take feelings, lifestyles, and so forth into account. Those regions are widely known to locals and colloquially referred to often as *NorCal* and *SoCal*. Participants were asked to take their best guess for cells that they felt unsure about. Each of the 90 hexagons covers an area of approximately 4920 km^2. The tessellation was considered to be a (relatively) high-resolution grid by the study authors, considering that rating 90 cells for the entire state represents much higher spatial resolution than is common when, for example, participants divide the city into two regions of north and south, or three regions of north, south, and central. The statement also implies that rating 90 cells is close to the maximum that can be meaningfully asked of human participants. A detailed description of all studies, the Alberta control study, the study design, and the participants, can be found in the original MFP publication.

Figure 1 shows a slightly altered reproduction of the results of the MFP study. Cells with an asterisk are not part of the original study but have been added by us via linear interpolation in order to fully cover the land area of California and thereby collect postings from these areas. Point-in-polygon analysis has been used to aggregate point observations and assign them to the hexagonal cells; more details are provided in Section 3.3. Interestingly, region membership is not monotonic, i.e., cells south of another cell may be less *Southern* California and cells to the north of another cell may be less *Northern* California. For instance, the hexagon containing the city of El Centro which borders Mexico is considered to be less *Southern* California than the cell

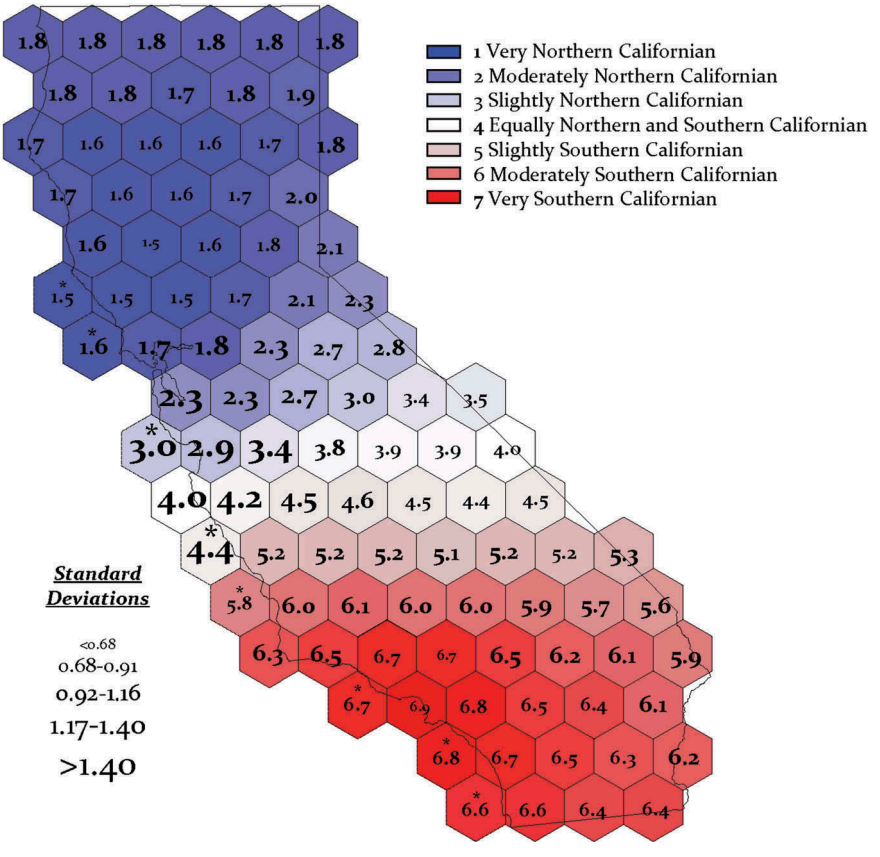


Figure 1. Means and standard deviations of ratings of *Northern* and *Southern* California based on Montello *et al*. (2014); dashed borders and asterisks indicate interpolated cells. Those cells marked with an asterisk were not part of the original study but have been added by us in order to fully cover the land area of California.

containing Santa Barbara which is to the northwest of Los Angeles. Similarly, on average, the cells in the San Francisco Bay Area are considered to be more *Northern* California than the most northern cells on the outline map. Furthermore, there is a clear coast-inland trend by which places at the coast are considered to be more *Northern* or *Southern* California than places to the east at the same latitude. This leads to a vague central boundary between *Northern* or *Southern* California that is of heterogeneous thickness, being thinner to the west and thicker to the east. Finally, the standard deviations across participant rankings are higher for northern than southern core areas of the respective regions. As we noted above, we call such phenomena *platial effects* to highlight the fact that thematic and cultural aspects of the environment can distort or relax spatial properties, including distance, direction, latitude and longitude, size, and so on.

The MFP research is a representative example of human participants studies carried out in cognitive and behavioral geography (Montello 2009), spatial cognition, and geographic information science. It demonstrates a new methodology – grid-based interval-level rating – by applying it to an interesting geographic phenomenon. In this

work, we reproduce their study, using the same California example and grid-based interval-level rating. However, we approach data acquisition and study design from a radically different angle, namely from a computational *fourth-paradigm* perspective (Hey *et al*. 2009), i.e., through the synthesis of high volumes of heterogeneous data provided by various online sources (Janowicz *et al*. 2015). In this paper, we discuss the differences in study and task design between the two approaches, present the results of this computational approach, compare them to the original human-participants study, and relate our results to the ongoing debate over the use of social media in GIScience.

The research contributions of our work are as follows:

- We propose an automatable (and thus scalable) framework which can synthesize multiple heterogeneous datasets from different sources to study vague cognitive regions.
- We compare the results from our *data-synthesis-driven* approach with those from a human-participants experiment, and discuss the pros and cons of the two approaches.
- In addition to the grid-based membership study, we also approximate crisp boundaries for the cognitive regions and explore their underlying thematic topics.
- We explore the use of topic modeling to gain further insights into how vague cognitive regions can be represented and delineated.

To date, the literature on data-synthesis-driven approaches to quantitative geographic analysis is very sparse. Online social media records represent a form of secondary archival data[1] (Montello and Sutton 2013), which is not particularly novel in itself. However, the automated filtering and analysis of such data, particularly to analyze cognitive concepts such as cognitive regions, is novel. We introduce the term *data-synthesis-driven* here as an alternative to the popular notion of *data-intensive* science for two reasons. First, the term *data-intensive* could be misunderstood as implying that the MFP work (or any other work along the same lines) is not heavily based on data merely on grounds of the amount of data used. Second, we believe that the real and radical novelty of the fourth paradigm lies in the way data are acquired and handled, and in the role they play in asking certain types of scientific questions (Janowicz *et al*. 2015).

The remainder of this paper is structured as follows. In Section 2, we discuss existing studies related to the present work. Next, Section 3 presents the design of our study, the required data collection, changes that had to be made to the data from the MFP study for comparison to our work, as well as the processing workflow and methods employed. Section 4 presents our results, and compares them to the results of the original MFP study. Section 5 discusses the broader impact of this research, and finally, Section 6 summarizes this work and gives an outlook for future research and technology directions.

2. Related work

Cognitive places are examples of vague places that are also referred to as vernacular places (Hollenstein and Purves 2010, Purves *et al*. 2011), at least when they are concepts shared by groups of people and not idiosyncratic to one person. While typically not included in authoritative gazetteers, vague places are frequently used

in our everyday dialogue, such as when describing locations and asking directions. The intrinsic nature of a vague place is its boundary vagueness, as seen in examples such as *downtown*. Fuzzy-set-based methods have been widely used to extract the intermediate boundaries of vague places in GIScience and spatial cognition (Burrough and Frank 1996, Montello *et al*. 2003). Given their indispensable role in human thought and culture, researchers have conducted studies to acquire a better understanding of vague places. Based on a human-participants study, Davies *et al*. (2009) discussed the user needs and implications for vague place modeling. Jones *et al*. (2008) harvested Web pages related to particular vague places in the UK, and identified their approximate boundaries based on the geo-referenced locations in the pages. Liu *et al*. (2010) proposed a point-set-based region model to approximate vague areal objects and conducted a cognitive experiment to investigate the borders of *South China*. Li and Goodchild (2012) collected geotagged *Flickr* data for studying vague places, and constructed spatial boundaries using kernel density estimations. Recently, Hobel *et al*. (2016) presented a computational framework which employed natural language processing and machine learning techniques to derive the geographic footprint of the cognitive region *historic center of Vienna* based on the TripAdvisor website and OpenStreetMap entries, and validated the results by comparing them with a historical map of the city.

Social media provides an alternative data source for studying the interactions between people and places. While often being criticized for concerns of representativeness (Li *et al*. 2013, Tufekci 2014), social media data nevertheless reflect the behavior of millions of users throughout the world, and therefore have value (Tsou *et al*. 2013, Adams *et al*. 2015, Steiger *et al*. 2016). The wide availability of social media has greatly enriched traditional volunteered geographic information (VGI) approaches, such as OpenStreetMap and Wikimapia (Goodchild 2007, Haklay and Weber 2008, Mummidi and Krumm 2008). Unlike these traditional VGI platforms which focus on online collaborative mapping, geotagged social media data reflect the spatial footprints of people in the real world, and therefore can be employed for studying human behavior. For example, Gao *et al*. (2014b) demonstrated a strong positive correlation between traffic flow in the greater Los Angeles area and geotagged Twitter data. Using geotagged Flickr data, Keßler *et al*. (2009) developed a bottom-up approach to construct place entities that can help enrich official gazetteers. Also based on Flickr data, Hu *et al*. (2015) extracted urban areas of interest (AOI) for six different cities in the past 10 years, and analyzed the spatiotemporal dynamics of the extracted AOI.

3. Study design

In this section, we describe the datasets used, our workflow and methods, preprocessing steps, and the three analysis tasks we performed in order to reproduce the MFP study with a data-synthesis-driven approach.

3.1. *Data collection*

In contrast to the MFP study, we did not collect data by interacting with selected participants but by automatically observing the use of terms in existing data. To do

so, we filtered our data with two sets of keywords. The first grouped the keywords 'SoCal', 'South California', and 'Southern California' into one set, which we call *SoCal*, and the keywords 'NorCal', 'North California', and 'Northern California' into a second set, which we refer to as *NorCal*.

With these two sets of keywords, we collected data from five sources: *Flickr, Instagram, Twitter, Wikipedia*, and *TravelBlog.org*. *Flickr* is a photo sharing portal that stores millions of tagged and geo-referenced pictures. We believe that *Flickr* represents a more tourism-oriented view of California than the other social media sources. *Twitter* and *Instagram* are examples of online social media networks that are popular among both residents and visitors to California. These sources capture daily activities, news, visited points of interest, and so forth. We retrieved geo-referenceable entries from *TravelBlog* that provides trajectory-style data and capture outdoor locations well, including parks. While all these sources provide data and views from individuals, *Wikipedia* provides a consensus truth (broader agreement) about *NorCal* and *SoCal*, as articles containing these terms are the results of edits done by a larger community. As shown in Table 1, we collected 344,475 data entries/postings (203,713 for *SoCal* and 140,762 for *NorCal*) within the contiguous California State boundary (without islands). As for social media postings, the location mentions in the content might be different from where they were generated. We discuss the distinction between *the said place* and *the locale* further in Section 5. However, we only selected those geo-referenced (Twitter and Instagram) postings which were generated from mobile devices and provided the users' GPS coordinates; therefore, we can be confident from where the postings were actually generated. More detailed information about each source is presented below.

(a) **Flickr**: We extracted 41,838 postings contributed by 1338 unique users that contain the keywords (tags) mentioned above for the *SoCal* group and the *NorCal* group from 99.3 million Flickr photos taken from 2004 until 2014 and released by Yahoo Labs (Thomee *et al*. 2015). The photos are either geo-referenced manually or by the built-in positioning technologies in the mobile device or the camera.

(b) **Instagram**: Instagram is an online mobile photo (and video)-sharing social networking service. According to a Pew Research report (Duggan *et al*. 2015), Instagram has grown in popularity with more than half (53%) of Internet-using young adults (age 18–29) using the service. The content shared on Instagram is geo-referenced by built-in positioning technologies on mobile devices or by manually selecting the location from the preloaded Facebook gazetteer. We retrieved a total of 286,632 geo-referenced and *SoCal* and *NorCal* keyword-filtered postings by 79,371 unique users between 2011 and 2015.

(c) **Twitter**: Combining the Twitter Streaming API and Search API, we retrieved a total of 13,670 geo-referenced and *SoCal* and *NorCal* keyword-filtered tweets posted by 8482

Table 1. Data collection counts from five different sources.

Source	SoCal Group	NorCal Group	Total
Flickr	22,132	19,706	41,838
Instagram	169,648	116,984	286,632
Twitter	10,376	3294	13,670
Travel Blogs	107	78	185
Wikipedia	1450	700	2150
SUM	203,713	140,762	344,475

unique users during the winter of 2014–2015. When posted from an Android or iOS application, the locations of the tweets were geo-referenced by the built-in positioning technologies if the user opted in to the location service.

(d) **TravelBlogs.org**: Over 440,000 raw blog entries were downloaded. Each place name was matched to an entry in the GeoNames gazetteer, providing a latitude and longitude. More detailed information about the geoparsing procedure for these unstructured, natural language documents can be found in Adams *et al*. (2015). We extracted 185 travel blogs which mentioned at least one of the keywords from the *SoCal* or *NorCal* sets. Because this is such a small number of travel blogs extracted, we combined them with the Wikipedia articles discussed below for further analysis.

(e) **Wikipedia**: We extracted 2150 articles which contained the *SoCal* or *NorCal* group of keywords, and inside the California State boundary. If the articles were not directly geo-referenced, information from DBpedia was applied for geo-referencing (Bizer *et al*. 2009, Adams *et al*. 2015).

Among the five selected sources, the number of data entries vary substantially, due to API access restrictions, limited geo-referenced content, and so forth. We discuss the differences among these data sources in Section 4.

3.2. *Workflow and methods*

The overall analysis procedure for our data-synthesis-driven approach involves (a) extracting data that are frequently tagged with *SoCal* and *NorCal* in social media postings, (b) examining the spatial patterns of these data, and (c) defining the variability and boundary vagueness of *SoCal* and *NorCal*. The most challenging part of this approach is to select a large number of good quality data that meet our criteria from the raw data. We design a standard processing workflow (see Figure 2) to calculate the membership scores for the hexagon-cell-based representation of cognitive regions (**Task I**), to identify and characterize the vague boundaries (**Task II**), and to extract prominent thematic topics tied to cognitive regions from the natural language descriptions (**Task III**).

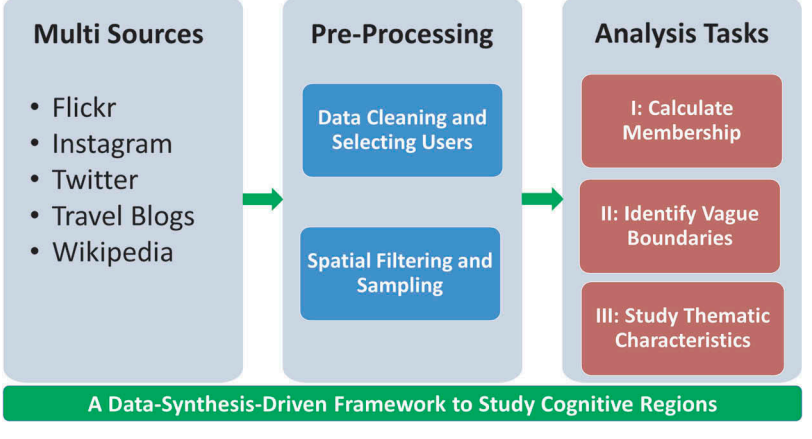


Figure 2. The processing framework for studying cognitive regions using a data-synthesis-driven approach.

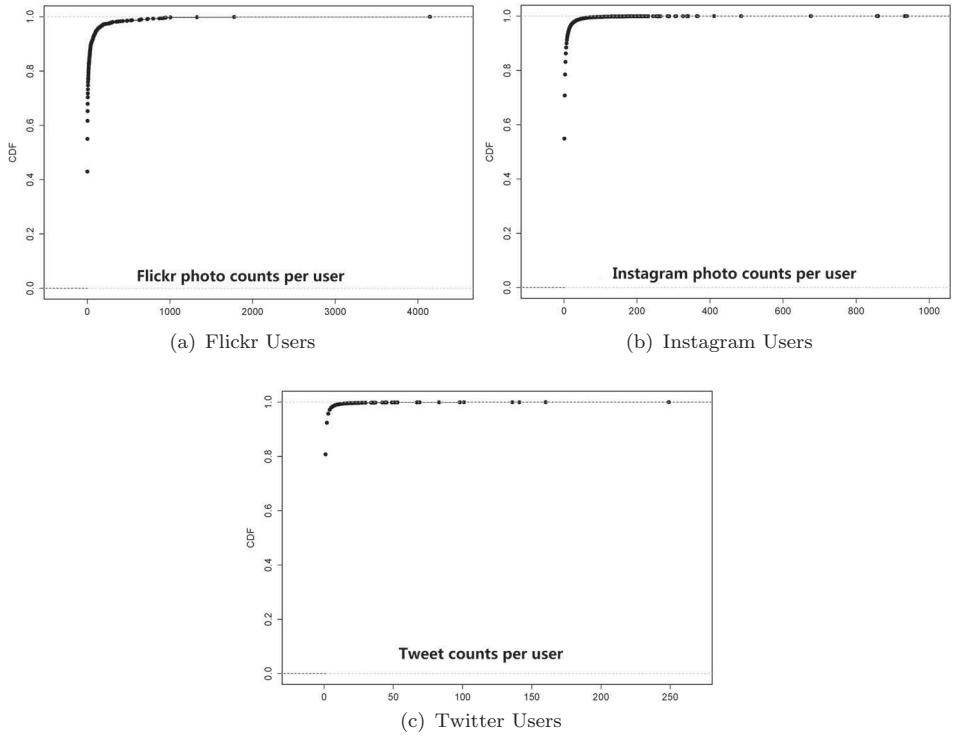
3.2.1. *Preprocessing step 1. Cleaning data and selecting appropriate users and contributed entries*

The information shared on social media and online crowdsourcing platforms usually follows a power-law distribution (Kwak *et al*. 2010, Gao *et al*. 2014a), which means most of the postings are contributed by a few users. In our case, we do not want the resulting patterns to be dominated by the most active users. In order to reduce such effects, we limited the number of entries contributed by each user. First, we calculated a probabilistic cumulative distribution function (CDF) for the posting counts per user (Figure 3) to decide on an appropriate threshold.

Taking Flickr photo postings as an example, the 90th percentile threshold value is 41 photos for the *SoCal* group and 40 for the *NorCal* group. This means that about 90% of the users posted no more than 41 photos for *SoCal* and 40 photos for *NorCal*. For users who contributed less than or equal to the percentile threshold *p*, all photos are kept. For users who contributed more photos, we randomly selected photos up to the threshold.

3.2.2. *Preprocessing step 2. Spatial clustering of entries and sampling for each user*

Second, we limited the number of posts by the same users to avoid having a few users dominate the overall patterns of a specified local region. For a given local region (within a certain search radius), we value contributions from multiple users because they represent a consensus among the general public for this region, which is similar to a



(a) Flickr Users (b) Instagram Users

(c) Twitter Users

Figure 3. The cumulative distribution functions (CDF) of entries contributed per user in Flickr, Instagram, and Twitter.

human-participants test. Therefore, we spatially filtered out repeated postings from a single user within a search radius of 100 m so that we retained only one post per user in this local region.

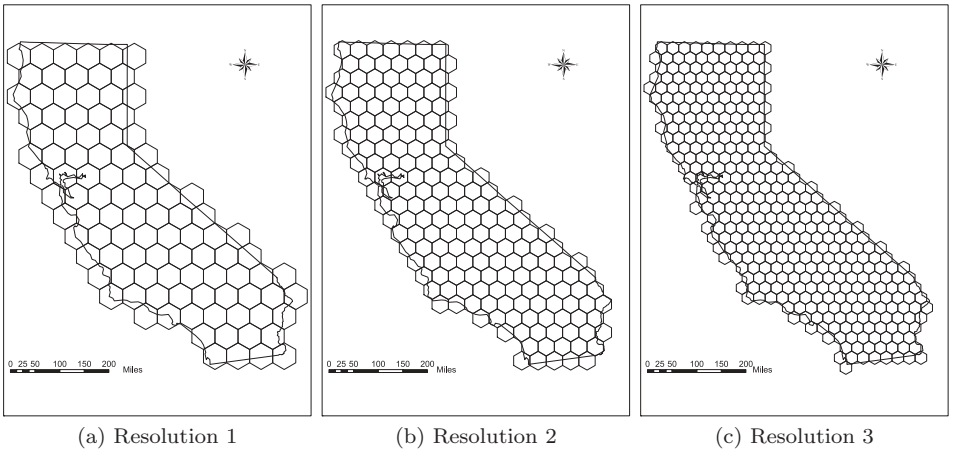
3.3. *Analyzing selected data*

3.3.1. *Task I. Calculating membership scores*

After the data filtering and clustering performed during the preprocessing steps, we applied point-in-polygon analysis to aggregate point observations to three different hexagonal tessellations at three different resolutions (Figure 4). The first level of hexagonal tessellation has the same spatial resolution as used in the MFP study, with each hexagon covering about 4920 km^2. The second-level and the third-level hexagons are at higher resolution, covering a half (2460 km^2) and a quarter (1230 km^2) of the first-level area in each cell, respectively. Varying the spatial resolution in a data-synthesis-driven approach is easy to do, while increasing the resolution is difficult in a traditional human-participants survey, since participants can be overwhelmed by a large number of cells to rate.

After spatially joining the point observations associated with the *SoCal* and *NorCal* group keywords to the hexagon grids, we obtained two occurrence counts in each cell for a given data source. Let S_i^j denote the occurrence counts of *SoCal* mentions and N_i^j as the *NorCal* mentions, where i is the hexagon ID at one of three resolution levels of tessellation grids, and j represents the data sources: Flickr, Instagram, Twitter or Travel Blogs and Wikipedia. Cells with a sum count of N_i^j and S_i^j less than 10 were considered as providing insufficient observations, and were therefore filtered out before the quantitative computation and comparison steps. We created two simple measures to derive the membership value of cells (Equation (1) and (2))

$$M1_i = S_i^j - N_i^j \tag{1}$$



(a) Resolution 1 (b) Resolution 2 (c) Resolution 3

Figure 4. The hexagon-based tessellations at different spatial resolutions.

$$M2_i = S_i^j / S_{max}^j - N_i^j / N_{max}^j, \tag{2}$$

where S_{max}^j represents the maximum occurrence counts of *SoCal* mentions per cell across the whole study area for a given data source *j*; and N_{max}^j is the maximum for *NorCal* mentions. The purpose of *M*1 is to quantify the absolute occurrence differences per cell while *M*2 measures a normalized ratio difference. Next, the cells are classified and rated from 1 to 7 for each data source based on ranking percentiles. From these, the spatial distribution maps of cell memberships for each data source were derived.

We also computed the mean values $M1_i^{mean}$ and $M2_i^{mean}$, as well as standard deviations $M1_i^{sd}$ and $M2_i^{sd}$ for each cell for both measures across all data sources. For both *M*1 and *M*2, the higher the value of the means, the more likely the cell is rated as being a *SoCal* (or *NorCal*) cell.

In order to determine the inter-source agreement of different data sources among the cells, we took each data source as one *rater layer* and index the cells that had sufficient observation counts in all layers with the ranks (1∼k) sorted by their occurrence counts. This results in four sets of tuples [cell-id, rank (1∼k)]. For instance, a cell with the ID 19 may have a value of 2 in Twitter (*moderately NorCal*) but a value of 1 in Instagram (*strongly NorCal*). We use *Kendall's coefficient of concordance (W)* (Kendall and Smith 1939) to assess the agreement among these different rater layers.

To do this, assume there are *m* sources rating *n* subjects in rank order from 1 to k. Let $r_{i,j}$ represent the rating a source *j* gives to a subject *i*. Let R_i be the total ranks given to the subject *i* (i.e., $\sum_{j=1}^{m} r_{i,j}$) and $\bar{R}$ be the mean of R_i, the sum of the squared deviations *S* can be calculated as by Equation (3). Then the Kendall's *W* is defined as given by Equation (4).

$$S = \sum_{i=1}^{n} (R_i - \bar{R})^2 \tag{3}$$

$$W = \frac{12S}{m^2(n^3 - n)}. \tag{4}$$

3.3.2. *Task II. Extracting continuous boundaries of cognitive regions*

Task I employed a discrete approach based on a hexagonal grid to calculate the membership score of each individual cell. In the second task, we aimed at determining the core regions of *NorCal* and *SoCal* using a continuous approach by approximating the boundaries of these two cognitive regions. While perceived borders of vague regions often vary among individuals (Montello 2003), our goal here is to extract the core regions which are agreed upon by most people.

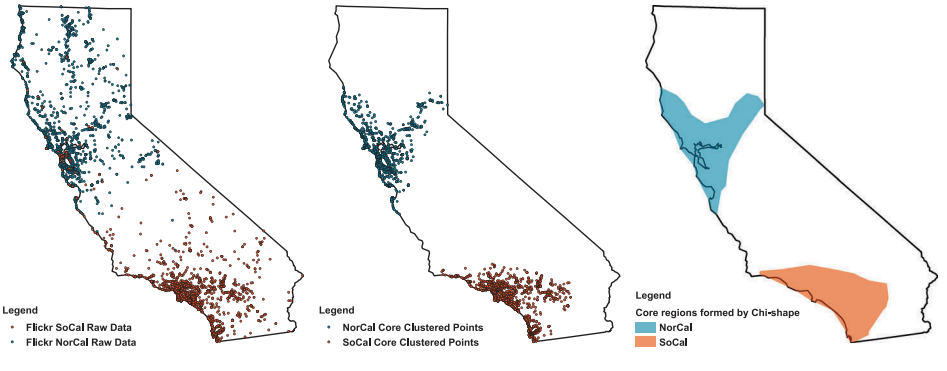
We use three social media sources, namely Flickr, Twitter, and Instagram, to identify the core regions. Using multiple sources helps ensure that the identified regions are not artifacts of one particular data source. In addition, it also reduces the potential bias introduced by the different user demographics of different social media platforms.

We applied a two-step workflow to extract the approximate regional boundaries for *NorCal* and *SoCal*. In step 1, we performed spatial clustering and identified point clusters based on geo-referenced social media data. This step considers each mention, for

example, a tweet about *NorCal* or *SoCal*, as a *vote* for the corresponding region and identify as those core areas that are agreed upon by a significant number of people. In step 2, we constructed polygons from the identified point clusters. While such polygons may not be completely consistent with the understanding of each individual, they can provide intuitive delineations of the general areas. In addition, these constructed polygons can be used to support spatial queries, for example, *show me all the hotels in SoCal*. Figure 5 illustrates this workflow, where Figure 5(a,b) shows the clustering process, and Figure 5(b,c) demonstrates the polygon construction.

To identify point clusters from geo-referenced social media postings, we use DBSCAN which is a density-based spatial clustering algorithm (Ester *et al*. 1996). Compared with distance-based clustering methods such as K-means or K-medoid, DBSCAN has two advantages which make it more suitable for our task. First, DBSCAN can identify clusters with any arbitrary shape. In this research, the shapes of the potential cognitive regions are unknown, and DBSCAN can help discover their perceived boundaries. Second, DBSCAN is robust to noise which commonly exist in social media data. Clustering methods, such as K-means, will classify noise observations into clusters and therefore can distort the derived regions.

DBSCAN requires two parameters, namely ε and *MinPts*. ε defines the search radius while *MinPts* specifies the minimum number of data points within the said search radius. The two parameters together define a density threshold; clusters are identified at the locations whose density values are higher than the defined threshold. To find a proper ε value, we performed a nearest neighbor analysis on the three social media datasets as suggested by Ester *et al*. (1996). We assumed 1% of the data were noise, and found the 99th percentile of the nearest neighbor distance (NND) in each dataset. Accordingly, the 1% of data points which were further away from the vast majority of observations were considered as noise. We calculated the average of the 99th NND percentiles for the three data sources and used the averaged value for ε. For *MinPts*, we cannot use a single absolute value (e.g., 4) as in traditional DBSCAN applications, since the number of data entries from different sources varied significantly. For example, the number of Instagram

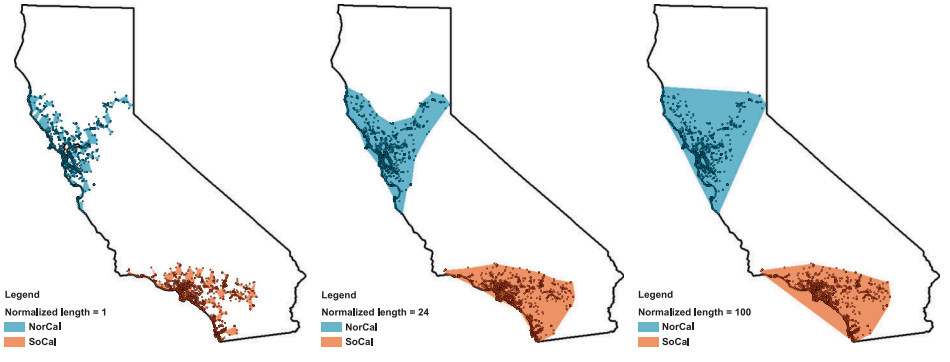


(a) Raw data (b) Clustered points (c) Constructed polygons

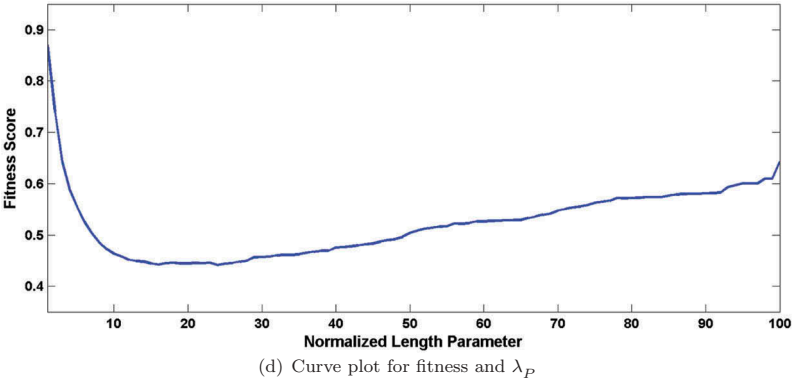
Figure 5. The workflow for extracting continuous boundaries for the cognitive regions of *NorCal* and *SoCal* (the visualized dataset is based on Flickr).

postings was much larger than those of the other sources (see Table 1). Consequently, it would have been much easier for Instagram observations to form clusters than for the other two sources, if a single value was used for *MinPts*. To address this issue, we used percentages instead of absolute counts for *MinPts*, namely 1%, 2%, and 3% of the total number of postings per data source, to model the vague nature of the cognitive regions. Other settings could be explored in future work with larger values shrinking the core region.

With point clusters identified, the second step was to construct polygons to approximate the boundaries of the cognitive regions. A *convex hull* approach has been used in many studies to represent the minimum bounding shape for a group of points (Preparata and Hong 1977). Such a hull, however, is unable to accurately delineate for the shapes of point clusters. The chi-shape algorithm, proposed by Duckham *et al*. (2008) computes a *concave hull* for a set of points. The chi-shape algorithm requires a normalized length parameter λ_P, which ranges from 1 to 100. A value of 1 creates polygons which are closest to the original point set but may generate spiky edges (Figure 6(a)). A larger value of λ_P will create smoother boundaries and also generate more empty space within the polygon. When λ_P is set to 100, the constructed polygon is equivalent to a convex hull (Figure 6(c)). Recent work by Akdag *et al*. (2014) proposes a



(a) $\lambda_P = 1$ (b) $\lambda_P = 24$ (c) $\lambda_P = 100$



(d) Curve plot for fitness and λ_P

Figure 6. Constructed polygon representations for the cognitive regions of *NorCal* and *SoCal* using different λ_P values.

fitness function which balances the complexity and the emptiness of the constructed polygon. Based on their work, we iterated λ_P from 1 to 100 and identified the optimal λ_P value (which is 24 in our experiment) that achieves the minimum value for the fitness function. Figure 6(d) shows the resulting curve plot. The polygon generated with $\lambda_P = 24$ is shown in Figure 6(b), respectively.

3.3.3. *Task III. Inferring thematic characteristics via topic modeling*

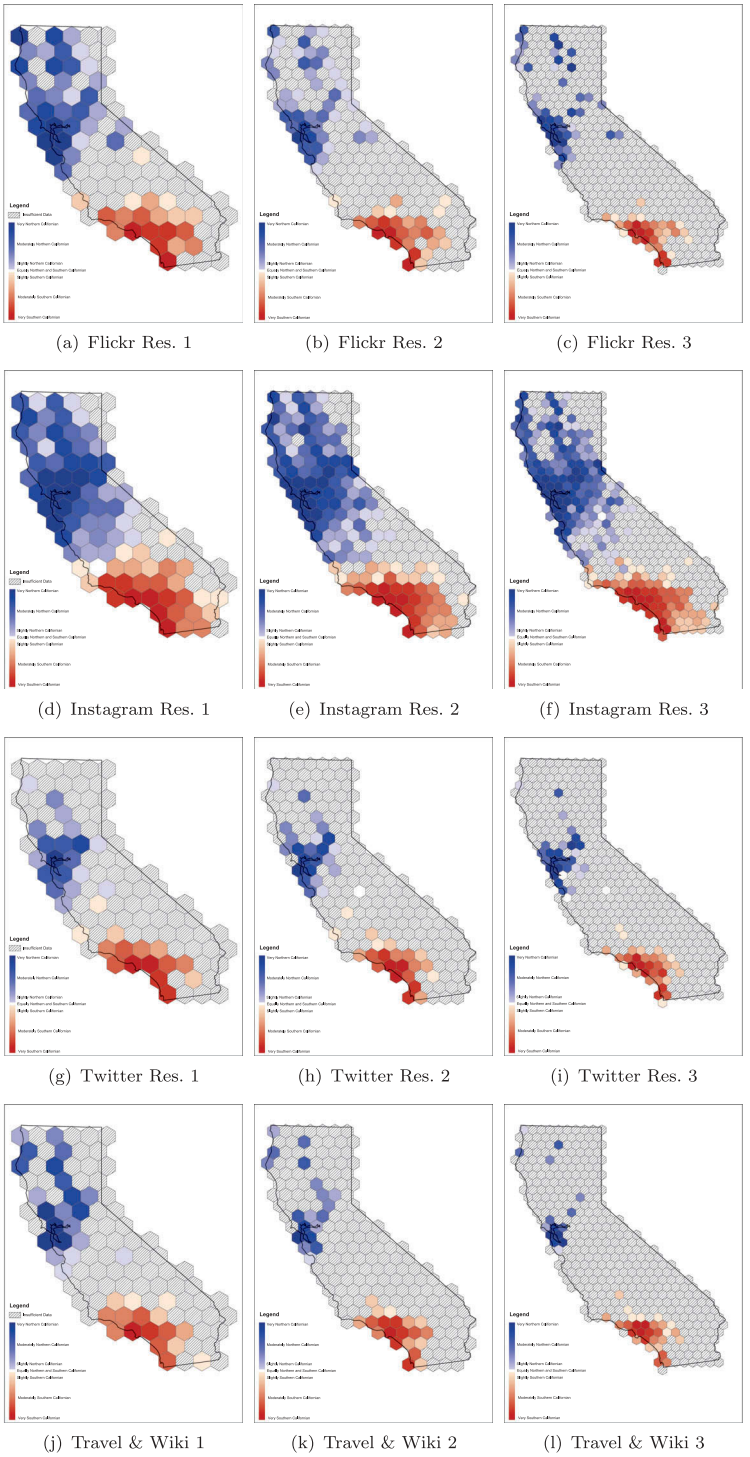
Having identified and delineated regions, we explored what these regions have in common with each other and how they differ. To do this, we use *topic modeling* over social media. We selected the *Resolution 3* (Figure 4(c)) spatial data layer as our basis for topic modeling given that it offers the most detailed depiction of California that we assessed in our experiments, allowing for nuanced changes in topics to have an impact. Each of the social layers (Flickr, Twitter, Instagram) was spatially intersected with the Resolution 3 hexagons, and the unstructured textual data were grouped and aggregated to the individual hexagon level. Next, the data were cleaned to remove standard English stop-words, non-alphabetic characters and words consisting of less than three characters. The words for each hexagon were then stemmed[2] and place names were removed via DBPedia Spotlight[3] and manual extraction.

We applied *Latent Dirichlet Allocation (LDA)* (Blei *et al*. 2003) for topic modeling using the MALLET toolkit (McCallum 2002). *LDA* is a generative, unsupervised model that takes a bag-of-words approach to constructing topics. In this case, the corpus consists of all hexagons in California while the textual references within each hexagon make up a single document. The topics are constructed by exploring the co-occurrence of words in each document. Provided these topics, each hexagon could then be thematically defined as a distribution across all topics. For this research, and in line with previous work (Griffiths and Steyvers 2004, Adams and Janowicz 2015, McKenzie *et al*. 2015), we used 60 topics. The resulting topic distributions were then assigned back to the hexagons allowing for visual and statistical representation through thematic layers.

4. Results and discussions

4.1. *Membership variability and comparisons with survey*

Figure 7 depicts the spatial distributions and membership values of the *SoCal* and *NorCal* cognitive regions from the aforementioned data sources at three different resolutions. The cells rated as most *NorCal* are color-coded as blue, whereas *SoCal* cells are colored red. Darker colors represent a higher degree of membership. The core region of *NorCal* is around San Francisco (the Bay Area) which is roughly 300 miles south of the northern border of California. On the other side, cells that are most highly rated as being *SoCal* are around the greater Los Angeles area, which is more than 100 miles north of the southern border of the state. The boundary between *NorCal* and *SoCal* is vague and quite similar for the different data sources with respect to its shape, width, and location. However, the data do show different boundary transition patterns due to varying number of postings across data sources. The Instagram dataset has adequate observations in the transition zone between *NorCal* and *SoCal*, while other sources don't have sufficient data. All data sources reveal intensity values that are higher for both regions at the coasts and



(a) Flickr Res. 1 (b) Flickr Res. 2 (c) Flickr Res. 3

(d) Instagram Res. 1 (e) Instagram Res. 2 (f) Instagram Res. 3

(g) Twitter Res. 1 (h) Twitter Res. 2 (i) Twitter Res. 3

(j) Travel & Wiki 1 (k) Travel & Wiki 2 (l) Travel & Wiki 3

Figure 7. The spatial distribution of membership measures derived from different data sources at three different resolutions.

decrease toward the east, mostly failing to cross our minimum threshold to be considered part of either vague region. These results confirm the existence of a *platial effect* that distorts and relaxes spatial properties such as cardinal directions, substantially altering its monotonic variation across the landscape.

The average pattern across the social media sources is shown in Figure 8; each cell contains the mean of classified ranking percentiles (on a 1–7 scale) across all data sources. Figure 8 makes it evident that the cognitive regions we derive from our data-synthesis approach are highly similar to those from the original human-participants survey by Montello *et al*. (2014), although not identical. Both empirical approaches show that the *NorCal–SoCal* distinction is mostly relevant to the west coast, including the coast ranges, beach communities, and metropolitan areas of San Francisco–Oakland–San Jose, Los Angeles, and San Diego. Indeed, the data-synthesis approach leaves several cells unclassified as either *NorCal* or *SoCal*, especially cells in the middle and eastern parts of the state (more below). Thus, both approaches show clearly that the boundary between the two cognitive regions is not homogeneous but wedge-shaped, being much narrower toward the west coast and broader toward the east; the data-synthesis boundary area is even a trapezoid or truncated wedge. Both approaches show that the locations of core intensity for the cognitive regions of *NorCal* and *SoCal* are not at the northern and southern state borders, respectively, but considerably south of the northern border and north of the southern border. The two approaches identify a Southern California core that is virtually identically located – encompassing downtown

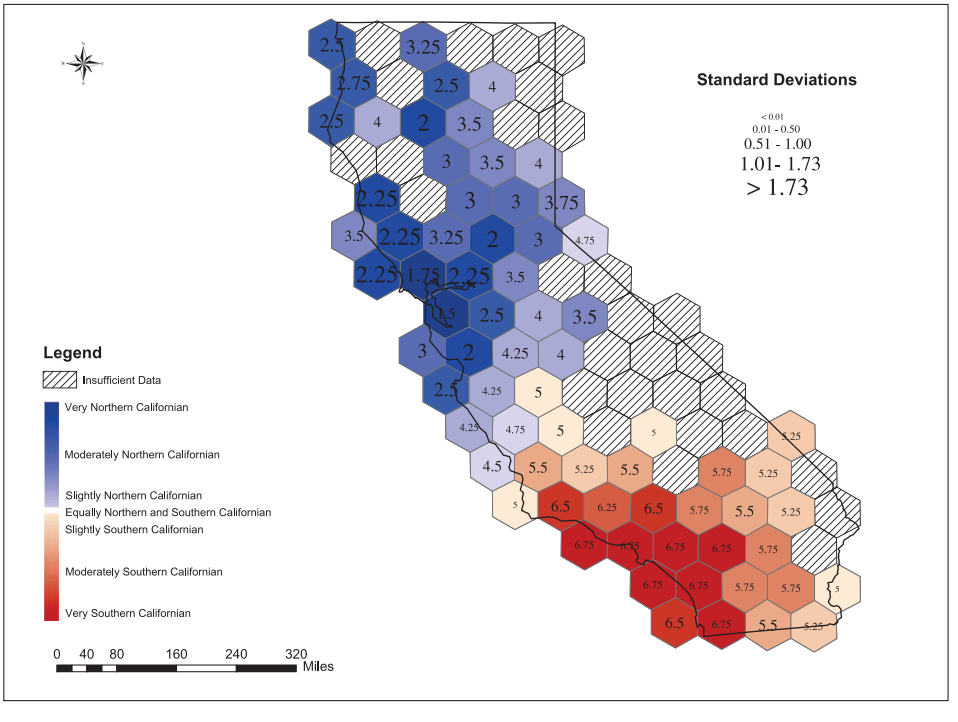


Figure 8. The results of identifying *SoCal* and *NorCal* cognitive regions using the data-synthesis-driven ranking percentiles.

Los Angeles and the west side, including the coast. The core of *NorCal*, however, is identified by our data-synthesis approach as quite a bit further south than it is by the human-participants approach of MFP. It is essentially the San Francisco Bay Area for the data-synthesis approach, while it is north of that for the MFP approach, around the confluence of the counties of Lake, Colusa, Yolo, Napa, and Sonoma. This difference aside, the data-synthesis approach agrees with the human-participants approach that the concepts of *NorCal* and *SoCal* are not merely latitudinal but attitudinal (i.e., both reveal platial effects).

To compare our results with the MFP results quantitatively, we computed Spearman's rank correlation ρ between the four layers from social media sources and the single layer from the human-participants survey, for the 69 cells which had sufficient social media data. As Table 2 shows, the correlations are uniformly very high for each of the four sources with the human-participants data; averaging across all four sources, the correlation with the human-participants data is 0.870 for scoring function $M1$, based on absolute occurrence differences, and 0.882 for scoring function $M2$, based on normalized ratio differences. All these high correlations are significant at p-value < 0.001 (df = 67), indicating that our automated approach generated membership results for these cognitive regions that closely approximate those of direct human raters. Moreover, Kendall's rank correlation τ is 0.712 for $M1$ and 0.721 for $M2$, respectively, which also implies a positive ordinal association between our approach and the human-participants approach.

As shown in Table 3, the value for Kendall's W (0.953, p-value <0.001) shows a high agreement among our four data sources with respect to the membership rankings of all cells. Kendall's W remains very high (0.929, p-value <0.001) even after adding the survey ranks from the MFP study as the fifth source, demonstrating a consistency between our data-synthesis-driven approach and the human-participants survey. In other words, the effects we see are not merely artifacts of a specific data source (and its user community).

Figure 8 also shows the standard deviations (SDs) for each cell, as were presented by MFP (Figure 1). Our pattern of SDs is starkly different than that for MFP's results. MFP found the least variability – the greatest consensus – for cells at and near the core of *NorCal* and *SoCal*. The boundary cells between the cores show the greatest variability. This would perfectly fit a pattern of statistical range restriction near the extremes of the

Table 2. Correlation between the data-synthesis-driven results and the human-participants results from the original MFP study.

Source	ρ (M1)	ρ (M2)	τ (M1)	τ (M2)
Flickr	0.881	0.880	0.721	0.719
Instagram	0.867	0.856	0.711	0.701
Twitter	0.874	0.838	0.714	0.673
TravelBlogs and Wikipedia	0.897	0.878	0.747	0.718
Means	0.870	0.882	0.712	0.721

Table 3. Kendall's coefficient of concordance *W*.

Source	Four raters	Five raters
Kendall's *W*	0.953	0.929
Chi-sq	259	316
p-value	<0.001	<0.001

scale (i.e., floor and ceiling effects), except that the MFP participants agreed a great deal that the eastern cells making up the boundary were neither *NorCal* nor *SoCal*. Our data-synthesis SDs show a complex pattern. They clearly do not reveal any statistical range restriction, perhaps understandable given the cell values are not based on a direct numerical rating scale. References to *NorCal* are highly variable for cells making up the core of that region, while they are very consistent for cells making up the core of *SoCal*. Apparently users of the selected social media sources agree more strongly about the spatial reference of *SoCal* than they do about *NorCal*. In general, we find high variance for cells in the northern half of the state and low variance for cells in the southern half.

4.2. *Sharpening the boundaries*

Like the MFP results, we generated boundaries for *NorCal* and *SoCal* that are vague or approximate. In our results, that is because social media references to *NorCal* and *SoCal* terms do change abruptly at or for some precise location on the landscape. For the MFP results, people do not express the belief that there is a precise transition location for these regions. In other words, whether considered a cultural phenomenon or a mental phenomenon (or both), these cognitive regions are *conceptually* vague (Montello 2003).

However, aside from the basic research motivation of understanding the nature of vague cognitive regions, we can apply our understanding to improving the functionality of various geographic information technologies. In several contexts, such as geographic information retrieval, this functionality will be increased by sharpening (also called hardening) the vague boundaries. The data-synthesis approach can be used to do this, even though we recognize that the cognitive boundary as such remains conceptually vague.

Here, we discuss our results from applying DBSCAN clustering and the chi-shape algorithm to Flickr, Twitter, and Instagram results to 'precisify' the vague cognitive regions by computing crisp boundaries for their core areas.

We do this by varying the threshold of reference density we require to include a cell as being in one of the two regions (Figure 9). For each of the subfigures, graduated colors (from light to dark) represent the extracted polygons based on minimum density thresholds of 1%, 2%, and 3% of the total number of data observations, respectively. Naturally, the region hulls shrink as we increase the threshold density. This is due to the fact that the 3% threshold puts a higher DBSCAN requirement for point clusters to be formed than the 2% threshold. However, the boundaries formed by the 3% threshold are also more reliable since they are derived from more observations.

In Figure 9(d), we overlap the results of all three data sources to identify the *common core* regions for *NorCal* and *SoCal*. These identified common cores can be combined in different ways to fit specific applications. For example, a GIS which requires high *precision* for its spatial query results can employ the overlapped core region (i.e., those in the darkest color). In contrast, an application that needs high *recall* for its retrieved result can use a spatial union of the 1% polygons derived from the three sources. As can be seen in Figure 9 and Figure 7, there is a substantial overlap between the regions derived from the three different datasets. This consistency indicates that these regions are not mere artifacts of a particular dataset, but reflects a broader and shared understanding.

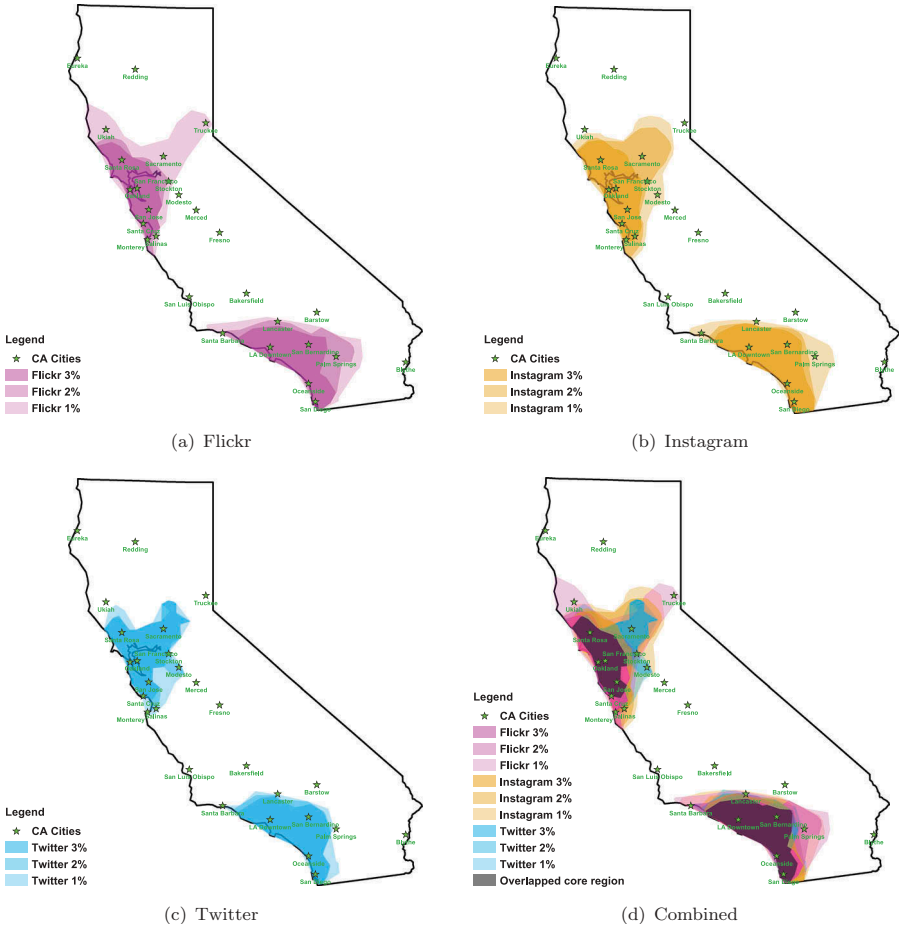


Figure 9. Core regions of *NorCal* and *SoCal* extracted using different datasets.

4.3. *Thematic characteristics*

For Task III, we modeled topics associated with *NorCal* and *SoCal* social media postings using *LDA*. This topic modeling approach considers the co-occurrence of words in a document and constructs topics from those words often occurring together. Upon examination, one can see that these topics are often thematically related and coalesce around properties such as those related to *Nature, Food*, or *Hiking*. Figure 10 shows three examples of the total of 60 topics generated via the topic model. Each topic is shown as a map of California with the color of each hexagon determined by the probability value of that topic appearing in that cell. The word cloud associated with each map shows the top terms contributing to that topic.

Figure 10(a,b) depicts topics related to physical features in the environment and the outdoors. Words such as *Mountain, Park,* and *Tree* contribute highly to both topics. There is a clear geospatial difference in the topics, however, with Figure 10 (a) showing high density in the southern interior, and Figure 10(b) presenting higher probability values in the center and northern parts of the state. These are examples of

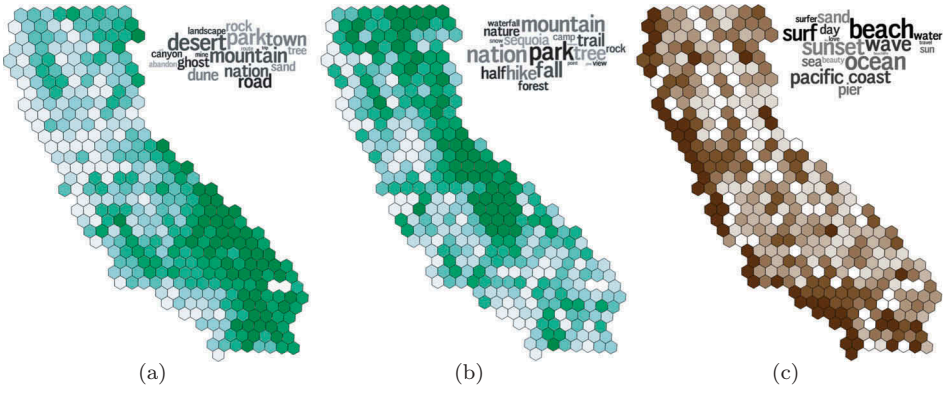


Figure 10. Three topics mapped to California along with their related word clouds. The darker the chromatic hue, the more prominent are the topics of terms in the postings from a particular cell.

topics that are clearly influenced by the linguistic characteristics of individuals contributing data from either *NorCal* or *SoCal*. In contrast, Figure 10(c) presents a topic that is split east and west rather than north and south. This topic lists the highest probabilities along the coast, consisting of words such as *Beach, Ocean,* and *Surf*. Both *NorCal* and *SoCal* are equally represented in this map showing that social data contributors mention words related to this topic regardless of the northern/southern California split.

From a purely visual representation in Figure 10, one could assume that there is no clear topic-wise distinction between the two cognitive regions of *SoCal* and *NorCal*. However, this is not the case. To demonstrate this, we selected 10 prototypical *SoCal* and 10 *NorCal* hexagons based on the membership intensity values reported in the original MFP paper. We extracted topic distributions for these hexagons and calculated the *Kullback–Leibler divergence (KLD)* (Kullback and Leibler 1951) for hexagons within *NorCal, SoCal,* and between both. *KLD* is a measure of the difference between two probability distributions. Low values indicate similar distributions while higher values suggest dissimilar distributions. Figure 11 shows these *KLD* values plotted as a smoothed histogram.

The core hexagons for *SoCal* are highly similar in terms of their distribution of topics. Core hexagons in *NorCal* are also quite similar to each other though slightly less than those for *SoCal*. This reflects the less cohesive data for *NorCal* we have discussed previously. When comparing interregion hexagons, we find a peak *KLD* value that is much larger, indicating substantially greater topic dissimilarity between *NorCal* and *SoCal* cells than between cells within each region separately. In short, the intra-region topic similarities are substantially higher than the interregion similarities. This means that while no single topic on its own is sufficient to distinguish the two cognitive regions from social media posts, the 60 topics can jointly distinguish between *SoCal* and *NorCal*. This is an important finding, as it suggests that everyday conversation is 'geo-indicative' (Adams and Janowicz 2012) to a degree where it can likely be used to discriminate regions and other geographic properties and entities (Louwerse and Benesh 2012).

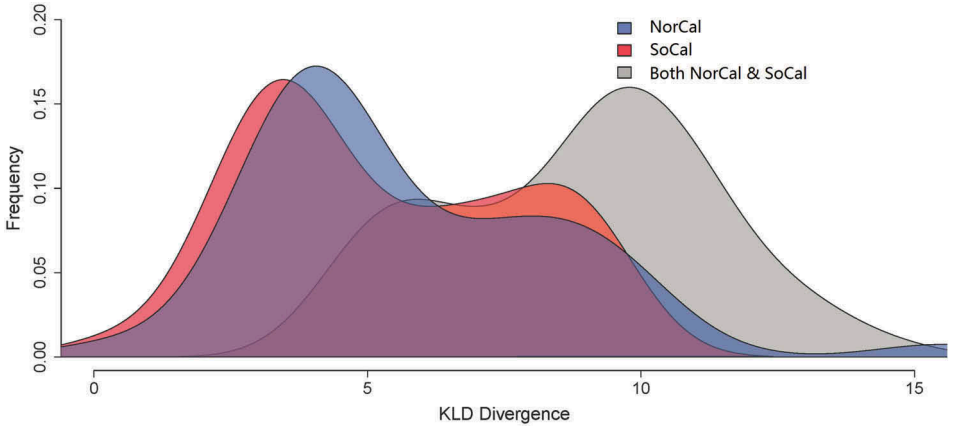


Figure 11. Kullback–Leibler divergence showing similarity of topics within *SoCal* hexagons and similarity of topics within *NorCal* hexagons, and dissimilarity of topics between both.

5. Broader implications

This study revisits the work of Montello *et al*. (2014) using a very different data-synthesis-driven approach to obtaining data instead of a human-participants survey. We demonstrated that a data-synthesis-driven approach can be successfully used to reproduce cognitive regions and membership like those established with a direct study of human research participants. We have also demonstrated how the used data and methods can be applied to go beyond previous work by extracting hardened hulls to represent these regions and how to study their thematic topics via topic modeling. Using the example of the informal cognitive regions of *SoCal* and *NorCal*, our work proposes an approach to deriving human conceptions of places, including regions, from social media data sources. The approach potentially captures not only spatial patterns but also the semantics of cognitive regions.

Our results suggest it is possible to reproduce the results of a direct human-participants study by mining existing social media postings from the Web. While we do not argue that such a data-synthesis-driven approach can or should entirely replace human-participants testing, the data-synthesis approach has the clear advantage that it can be repeated for a wide set of cognitive regions at flexible spatial scales without running into the limitations of participants testing, for example, the limited number of participants, limited attention span, variable knowledge of local geography, and so forth.

This research raises some further issues about the data-synthesis-driven approach. First is the difference between *the said place* which a person tags or mentions in a social-media entry and *the locale* where the person is located when posting the entry. *The said place* is not necessarily the same as *the locale*, since people can post any message about any place no matter where they are. In fact, we assume this might happen fairly often. In our data, for instance, the tag *SoCal* was sometimes mentioned in a small number (about 2%) of entries posted from the Bay Area, a core part of *Northern California*; while about 1% of the tag *NorCal* mentions were posted from the core part of *Southern California*. This is the nature of crowdsourced data. Researchers must pay attention to this issue when interpreting the experiment results. Different types of location inferences and insights can be

extracted from the social Web (Ikawa *et al*. 2013, Ajao *et al*. 2015). For this reason, we used two membership measures *M1* and *M2* to focus on relative differences and proportions instead of raw counts of place mentions. The results validated our proposed metrics. In future work, natural language processing techniques (e.g., place name disambiguation, preposition, and contextual analysis) can be employed in analyzing social media entries to better differentiate the said place and the locale.

There are also some arguments (e.g., sampling bias) with regard to the data-intensive paradigm in scientific research. The results of this study, however, suggest that user-generated social media data at least partially do reflect people's experiences, focus, opinions, and interests in places. Thus, these rich datasets can be synthesized as *social sensors* to support the study of vague cognitive regions in geography and GIScience.

An advantage of this data-synthesis-driven processing and geocomputation framework is the flexibility with which one can change the spatial resolution of hexagons or any other polygonal tessellation used to discretize the landscape. This includes not only finer resolutions but also coarser, more aggregated resolutions. If there was a theoretical argument to do so, one could even create a tessellation with multiple scales in a single layer. For example, the cognitive regions of *NorCal* and *SoCal* appear to apply much more to coastal California than the Central Valley, the Sierra Nevada, or the eastern deserts; thus, one might want to tessellate the state with a higher resolution in the coastal areas. Besides the potential for resolution of nearly unlimited fineness, we recognize the general appropriateness of matching the scale of one's analysis to the scale of the phenomenon one studies (Montello 2013). More generally, we recognize that the analytic possibilities of the data-intensive approach may create phenomena that are not psychologically plausible and can thus be misleading. We were able to analytically 'harden' (sharpen) the boundaries of our cognitive regions, but individual people typically do not have this ability and their conceptions of informal regions likely do not have such precise boundaries.

The human-participants approach asks individual people to directly express the degree to which they believe a particular place should be considered Northern or Southern California. This means that data relevant to the concept or feature of interest (*NorCal* and *SoCal*) are generated for all locations within the study framework (California). A limitation of the data-synthesis-driven approach is that cells lacking sufficient observations have to be filtered out, which means that comprehensive spatial coverage is lost, unlike a human-participants survey. These missing data cells are places with small numbers of residents and visitors, including areas within national forests, large water bodies, or mountain ridges. Alternatively, another way to look at this is that when people make the *NorCal–SoCal* distinction (as cognitive regions), they are referring only to western, coastal California, maybe mostly to just the San Francisco Bay Area versus the Los Angeles Area. In that case, the human-participants approach might be misleading because it required people to apply a distinction to every location within the state, even if the person never thinks of that distinction as being relevant to places like the Sierra Nevada, the northeastern Modoc Plateau, or the southern deserts. Alternatively, one could allow human participants to rate only cells that relate to the regional distinction as they understand it.

The human-participants approach asks directly for expressions of one's beliefs about informal regions, including both their spatial properties and their thematic

associations. The data-intensive approach is indirect, collecting communications that include a verbal reference to *NorCal* or *SoCal* but not asking anyone explicitly what they actually think about these regions. As a case in point, modeling the topical references in the social media postings showed us that they can statistically segregate the two regions, but it told us nothing about the thematic content of themes related to *NorCal* and *SoCal*. That is, it did not tell us what thematic associations come to mind when people use one of the two region terms rather than the other; a human-participants study could presumably do this directly. The same can arguably be done with topic modeling in a future study but may require additional data sources. The data-synthesis approach will often tap into cultural conventions that may or may not correspond closely to the beliefs of individuals. Presumably, such reference occur in social media on many occasions when the creator of the message is not thinking at all about the characteristics of places or the regions of California. Considering all of these issues though, we find it even more impressive how much agreement we find between our approach and that of MFP.

Going back to geographic information retrieval as one of the application areas of research on vague cognitive regions, there is one interesting question that we have not addressed so far. Although highly problematic for large areal features, the vast majority of geographic features, be it museums or mountains, is still represented by point coordinates. Google Maps, for instance, includes such point features for both *NorCal* and *SoCal*[4]. How representative are these locations with respect to the identified regions in both the original study and our replication? Interestingly, the *SoCal* point coordinates from Google Maps are located in the middle of the desert between interstates I-15 and I-40, more precisely at about (34.96, −116.42). This puts Google's *SoCal* marker about 180 km to the northeast of the centroid (33.81, −117.68) computed for the 3% common core region (near Anaheim, CA). Google's *NorCal* marker (38.84, −120.9) is placed near Garden Valley, CA northeast of Sacramento, CA. This is about 160 km to the northeast of the centroid identified by our work (37.96, −122.21) which is located in the broader Bay Area. In other words, the map markers for both regions differ substantially from the result obtained by MFP and our work. They also do not follow the west-east trend, where membership intensity values to both regions are higher at the coasts.

6. Conclusions

In this research, we investigated using a data-intensive approach to determining vague cognitive regions. We compared them to the corresponding MFP study based on human participants which validated our proposed approach. Using data sources from social media including Flickr, Instagram, Twitter, Travel Blogs, and Wikipedia pages, we derived region membership scores for cells within the state of California that correlated significantly to those in the original study, both in terms of Spearman's as well as Kendall's rank correlation statistics. Overall, the shapes of *NorCal* and *SoCal* were quite similar for the two empirical approaches, including the non-monotonicity of the two regions and the heterogeneity of their vague boundaries. Most importantly, our work showed the same *platial effects* observed in the original study. Furthermore, our work examined the implications of increasing the spatial resolution of the tessellations on the cognitive regions that result.

In addition to assessing membership scores within the hexagons, we further explored the continuous boundaries and the core regions for *NorCal* and *SoCal*. A two-step workflow based on the DBSCAN clustering method and the chi-shape algorithm was designed to generate approximate boundaries for the cognitive regions. Experiments were conducted to select optimal parameters for the workflow, and we observe consistency among the polygon representations that are derived from the different datasets.

We also explored thematic associations for *NorCal* and *SoCal* with the help of topic modeling. This generated various topics most often associated with different regions of California on our social media sources. Comparing the topic distributions of prototypical *NorCal* and *SoCal* hexagons shows high similarity within each region and a lower similarity between the two regions.

In sum, our paper is about the prospects for utilizing multiple social media sources to apply a data-synthesis approach to extracting and characterizing informal geographic concepts and features, such as the cognitive regions of *NorCal* and *SoCal*. Our study sheds light on differences in the methodology of traditional human-participants approach and the increasingly popular data-synthesis approach, suggests advantages and limitations of both approaches, and points to future avenues for research and system design in GIScience.

Acknowledgments

We would like to thank the editor Dr. May Yuan and three anonymous referees for their valuable comments and suggestions.

Disclosure statement

No potential conflict of interest was reported by the author.

Notes

1. From a broad research-methods perspective, such as that in Montello and Sutton (2013), social media records are not data until they are coded for content – they are sources of data. In this paper, however, we follow the convention of the data-synthesis ('big data') research community and refer to the collected records as data.
2. Using the Snowball stemming method http://snowball.tartarus.org
3. http://spotlight.dbpedia.org
4. The interface will accept both of these terms and map them to 'Northern California' and 'Southern California' respectively.

References

Adams, B. and Janowicz, K., 2012. On the geo-indicativeness of non-georeferenced text. *In*: *Proceedings of the international AAAI conference on weblogs and social media (ICWSM 2012)*, Dublin, Ireland, 375–378.

Adams, B. and Janowicz, K., 2015. Thematic signatures for cleansing and enriching place-related linked data. *International Journal of Geographical Information Science*, 29 (4), 556–579. doi:10.1080/13658816.2014.989855

Adams, B., McKenzie, G., and Gahegan, M., 2015. Frankenplace: interactive thematic mapping for ad hoc exploratory search. *In*: *Proceedings of the 24th international conference on World Wide Web, WWW '15*, Florence, Italy. Republic and Canton of Geneva, Switzerland: International World Wide Web Conferences Steering Committee, 12–22.

Aitken, S.C. and Prosser, R., 1990. Residents' spatial knowledge of neighborhood continuity and form. *Geographical Analysis*, 22 (4), 301–325. doi:10.1111/j.1538-4632.1990.tb00213.x

Ajao, O., Hong, J., and Liu, W., 2015. A survey of location inference techniques on Twitter. *Journal of Information Science*, 41 (6), 855–864. doi:10.1177/0165551515602847

Akdag, F., Eick, C.F., and Chen, G., 2014. Creating polygon models for spatial clusters. *In*: T. Andreasen, *et al*., eds. *Foundations of intelligent systems.ISMIS 2014*, Lecture Notes in Computer Science, vol 8502. Cham: Springer, 493–499.

Bennett, B., 2001. What is a forest? On the vagueness of certain geographic concepts. *Topoi*, 20 (2), 189–201. doi:10.1023/A:1017965025666

Bizer, C., *et al*., 2009. DBpedia-A crystallization point for the web of data. *Web Semantics: Science, Services and Agents on the World Wide Web*, 7 (3), 154–165. doi:10.1016/j.websem.2009.07.002

Blei, D.M., Ng, A.Y., and Jordan, M.I., 2003. Latent dirichlet allocation. *The Journal of Machine Learning Research*, 3, 993–1022.

Burrough, P.A. and Frank, A., 1996. *Geographic objects with indeterminate boundaries*. Vol. 2. London: CRC Press, Taylor & Francis.

Casati, R. and Varzi, A.C., 1999. *Parts and places: the structures of spatial representation*. Cambridge, MA: MIT Press.

Cohn, A.G. and Gotts, N.M., 1996. The 'egg-yolk'representation of regions with indeterminate boundaries. *Geographic Objects with Indeterminate Boundaries*, 2, 171–187.

Couclelis, H., 1992. People manipulate objects (but cultivate fields): beyond the raster-vector debate in GIS. *In*: A.U. Frank, I. Campari, and U. Formentini, eds. *Theories and methods of spatio-temporal reasoning in geographic space*. Berlin: Springer, 65–77.

Davies, C., *et al*., 2009. User needs and implications for modelling vague named places. *Spatial Cognition & Computation*, 9 (3), 174–194. doi:10.1080/13875860903121830

Duckham, M., *et al*., 2008. Efficient generation of simple polygons for characterizing the shape of a set of points in the plane. *Pattern Recognition*, 41 (10), 3224–3236. doi:10.1016/j.patcog.2008.03.023

Duggan, M., *et al*., 2015. *Social media update 2014*. Technical report. Washington, DC: Pew Research Center.

Ester, M., *et al*., 1996. A density-based algorithm for discovering clusters in large spatial databases with noise. *In*: *The Second international conference on knowledge discovery and data mining (KDD-96)*. Portland, OR: AAAI Press, 226–231.

Frank, A.U., 1996. The prevalence of objects with sharp boundaries in GIS. *In*: Burrough and Frank, eds. *Geographic objects with indeterminate boundaries*, vol. 2. London: CRC Press, Taylor & Francis, 29–40.

Galton, A., 2003. On the ontological status of geographical boundaries. *In*: Duckham *et al*., eds. *Foundations of geographic information science*. London: Taylor & Francis, 151–171.

Gao, S., *et al*., 2014a. Constructing gazetteers from volunteered big geo-data based on Hadoop. *Computers, Environment and Urban Systems, 61,* 172–186.

Gao, S., *et al*., 2014b. Detecting origin-destination mobility flows from geotagged Tweets in greater Los Angeles area. *In*: *Proceedings of the eighth international conference on geographic information science*, Vienna, Austria, 1–4.

Goodchild, M.F., 2007. Citizens as sensors: the world of volunteered geography. *GeoJournal*, 69 (4), 211–221. doi:10.1007/s10708-007-9111-y

Goodchild, M.F., 2011. Looking forward: five thoughts on the future of GIS. *Esri ArcWatch, February 2011*.

Griffiths, T.L. and Steyvers, M., 2004. Finding scientific topics. *Proceedings of the National Academy of Sciences*, 101 (Suppl 1), 5228–5235. doi:10.1073/pnas.0307752101

Haklay, M. and Weber, P., 2008. Openstreetmap: user-generated street maps. *IEEE Pervasive Computing*, 7 (4), 12–18. doi:10.1109/MPRV.2008.80

Hey, A.J., *et al*., 2009. *The fourth paradigm: data-intensive scientific discovery*, Vol. 1. Redmond, WA: Microsoft Research Redmond.

Hobel, H., Fogliaroni, P., and Frank, A.U., 2016. Deriving the geographic footprint of cognitive regions. *In*: Sarjakoski *et al*., eds. *Geospatial data in a changing world: selected papers of the 19th AGILE conference on geographic information science*. Switzerland: Springer, 67–84.

Hollenstein, L. and Purves, R., 2010. Exploring place through user-generated content: using Flickr tags to describe city cores. *Journal of Spatial Information Science*, 1, 21–48.

Holtmeier, F.K., 2009. *Mountain timberlines: ecology, patchiness, and dynamics*. Vol. 36. Berlin: Springer Science & Business Media.

Hu, Y., *et al*., 2015. Extracting and understanding urban areas of interest using geotagged photos. *Computers, Environment and Urban Systems*, 54, 240–254. doi:10.1016/j.compenvurbsys.2015.09.001

Ikawa, Y., *et al*., 2013. Location-based insights from the social web. *In*: *Proceedings of the 22nd international conference on World Wide Web* (WWW '13), Rio de Janeiro, Brazil, May 13-17, 1013–1016.

Janowicz, K., *et al*., 2015. Why the data train needs semantic rails. *AI Magazine*, 36 (1), 5–14.

Jones, C.B., *et al*., 2008. Modelling vague places with knowledge from the Web. *International Journal of Geographical Information Science*, 22 (10), 1045–1065. doi:10.1080/13658810701850547

Kendall, M.G. and Smith, B.B., 1939. The problem of m rankings. *The Annals of Mathematical Statistics*, 10 (3), 275–287. doi:10.1214/aoms/1177732186

Keßler, C., *et al*., 2009. Bottom-up gazetteers: learning from the implicit semantics of geotags. *In*: K. Janowicz, M. Raubal, and S. Levashkin, eds. *GeoSpatial semantics*. Berline: Springer, 83–102.

Kullback, S. and Leibler, R.A., 1951. On information and sufficiency. *The Annals of Mathematical Statistics*, 22, 79–86. doi:10.1214/aoms/1177729694

Kwak, H., *et al*., 2010. What is Twitter, a social network or a news media?. *In*: *Proceedings of the 19th international conference on World Wide Web* (WWW '10), Raleigh, North Carolina, USA, April 26–30. ACM, 591–600.

Li, L. and Goodchild, M.F., 2012. Constructing places from spatial footprints. *In*: *Proceedings of the 1st ACM SIGSPATIAL international workshop on crowdsourced and volunteered geographic information*, (GEOCROWD '12), Redondo Beach, California, November 06. ACM, 15–21.

Li, L., Goodchild, M.F., and Xu, B., 2013. Spatial, temporal, and socioeconomic patterns in the use of Twitter and Flickr. *Cartography and Geographic Information Science*, 40 (2), 61–77. doi:10.1080/15230406.2013.777139

Liu, Y., *et al*., 2010. A point-set-based approximation for areal objects: A case study of representing localities. *Computers, Environment and Urban Systems*, 34 (1), 28–39. doi:10.1016/j.compenvurbsys.2009.05.001

Louwerse, M.M. and Benesh, N., 2012. Representing spatial structure through maps and language: lord of the rings encodes the spatial structure of middle earth. *Cognitive Science*, 36 (8), 1556–1569. doi:10.1111/cogs.2012.36.issue-8

Mark, D.M., Smith, B., and Tversky, B., 1999. Ontology and geographic objects: an empirical study of cognitive categorization. *In*: C. Freksa and D.M. Mark, eds. *Spatial information theory. Cognitive and computational foundations of geographic information science*. Berlin: Springer, 283–298.

McCallum, A.K., 2002. *MALLET: a machine learning for language toolkit*. Technical report, University of Massachusetts Amherst.

McKenzie, G., *et al*., 2015. POI pulse: a multi-granular, semantic signatures-based approach for the interactive visualization of big geosocial data. *Cartographica: The International Journal for Geographic Information and Geovisualization*, 50 (2), 71–85. doi:10.3138/cart.50.2.2662

Montello, D.R., 2003. Regions in geography: process and content. *In*: Duckham *et al*. eds. *Foundations of geographic information science*. London: Taylor & Francis, 173–189.

Montello, D.R., *et al*., 2003. Where's downtown?: behavioral methods for determining referents of vague spatial queries. *Spatial Cognition & Computation*, 3 (2–3), 185–204. doi:10.1080/13875868.2003.9683761

Montello, D.R., 2009. Cognitive geography. *International Encyclopedia of Human Geography*, 2, 160–166.

Montello, D.R., 2013. Scale in geography. *In*: J. Wright, ed. *The international encyclopedia of social and behavioral sciences*. 2nd ed. Oxford: Elsevier.

Montello, D.R., Friedman, A., and Phillips, D.W., 2014. Vague cognitive regions in geography and geographic information science. *International Journal of Geographical Information Science*, 28 (9), 1802–1820. doi:10.1080/13658816.2014.900178

Montello, D.R. and Sutton, P., 2013. *An introduction to scientific research methods in geography and environmental studies*. 2nd ed. Thousand Oaks, CA: SAGE.

Mummidi, L.N. and Krumm, J., 2008. Discovering points of interest from users' map annotations. *GeoJournal*, 72 (3–4), 215–227. doi:10.1007/s10708-008-9181-5

Preparata, F.P. and Hong, S.J., 1977. Convex hulls of finite sets of points in two and three dimensions. *Communications of the ACM*, 20 (2), 87–93. doi:10.1145/359423.359430

Purves, R., Edwardes, A., and Wood, J., 2011. Describing place through user generated content. *First Monday*, 16, 9. doi:10.5210/fm.v16i9.3710

Smith, B. and Varzi, A.C., 2000. Fiat and bona fide boundaries. *Philosophical and Phenomenological Research*, 60, 401–420. doi:10.2307/2653492

Steiger, E., Westerholt, R., and Zipf, A., 2016. Research on social media feeds–a GIScience perspective. *In*: C. Capineri, *et al*., eds. *European handbook of crowdsourced geographic information*. London: Ubiquity Press, 237–254.

Thomee, B., *et al*., 2015. The new data and new challenges in multimedia research. *arXiv preprint arXiv:1503.01817*.

Tsou, M.H., *et al*., 2013. Mapping social activities and concepts with social media (Twitter) and web search engines (Yahoo and Bing): a case study in 2012 US presidential election. *Cartography and Geographic Information Science*, 40 (4), 337–348. doi:10.1080/15230406.2013.799738

Tufekci, Z., 2014. Big questions for social media big data: representativeness, validity and other methodological pitfalls. *arXiv preprint arXiv:1403.7400*.

White, E. and Stewart, K., 2015. Barrier dynamics for GIS: a design pattern for geospatial barriers. *International Journal of Geographical Information Science*, 29 (6), 1007–1022. doi:10.1080/13658816.2014.995103