Twitter Wars! Social Media Mapping in the Classroom

Barry J Kronenfeld

Department of Geology and Geography, Eastern Illinois University, Charleston, IL USA; bjkronenfeld@eiu.edu

Abstract: The explosion of social media has enabled spatial exploration of hitherto unobservable socio-cultural phenomena, but constructing meaningful maps is challenging. This study reports on key problems and solutions developed for a graduate level classroom activity involving real-time collection and cartographic visualization of tweets on contemporary news topics. Two primary challenges are identified and discussed: (1) obtaining a high volume of accurately geocoded tweets with minimal time and resources, and (2) developing a clear sequence of cartographic transformations to eliminate data/visual biases and turn noisy data into meaningful visual information.

Keywords: Twitter, social media mapping, geocoding, cartographic transformations, cartograms

1. Introduction

The explosion of social media has led to the tantalizing prospect that hitherto unobservable cultural phenomena can be observed, mapped and mined for spatio-temporal patterns. Such efforts, it is hoped, might even reveal something revelatory about the spatial functioning of society, such as variations in perception of environmental hazards (Zhou and Zhang 2016), dispersion of innovation and social movements (Tsou 2015) and spatial distribution of global awareness (Han et al. 2015). Despite this promise, the reality is that many social media mapping projects come up short, revealing less about society and more about the limitations of data and analysis techniques.

The challenge of social media mapping is nowhere more acute than in the classroom. Students are often quite social-media savvy and adept at identifying promising cultural memes for exploration. Yet they are less adept at recognizing fundamental data issues and selecting appropriate cartographic transformations to extract meaning from chaos. In this paper, I attempt to summarize lessons learned in a *Twitter Wars!* social media mapping exercise developed for an introductory GIS course. The exercise, used with both the undergraduate and Master's level students, has two primary goals: (a) to introduce social media mapping as a means for exploring socio-cultural spatial pattern, and (b) to explore cartographic visualization as a series of transformations to turn spatial data into meaningful information. The exercise is performed over a two-week course unit.

The challenges of classroom project development are a microcosm of the challenges faced in more serious research endeavors. Extracting and visualizing meaningful spatial patterns from social media data in two short weeks requires clear exposition of problems and effective but pragmatic solutions. In this paper, I use the *Twitter Wars!* exercise as a jumping point from which to examine two fundamental challenges of social media mapping: (a) obtaining a high volume of accurately geocoded tweets with minimal time and resources, and (b) transforming idiosyncratic data borne from specific spatial, temporal and thematic contexts into meaningful knowledge using general procedures. I argue that for social media-based cartography to be impactful, these challenges must be met with reasonable, pragmatic solutions that are robust to minor data errors and accessible to starting cartographers without extensive programming expertise or financial resources.

2. Project Background and Requirements

Social media has been used to analyze socio-cultural patterns at multiple scales. For example, Twitter data, has been used to investigate the functional role of different locations within a city (Zhou and Zhang 2016), regional variations in perceived risk during a snowstorm (Shook & Turner 2016), and national variations in global awareness (Han et al. 2015). The scale and extent of the *Twitter Wars!* exercise was determined by pragmatic considerations. Twitter provides the ability to obtain tweets in real time filtered by keywords or spatial extent, but not both. To study a particular keyword-based meme, therefore, necessarily involves obtaining a global dataset. However, a majority of tweets originate from the United States. To maximize data within a simple, contiguous area, the contiguous lower 48 U.S. states and District of Columbia was selected as the study region. Implementation was further guided by the following considerations:

- students were not expected to have any programming skills
- money for geocoding was not available
- several thousand tweets are needed to create meaningful maps of the study area
- students are limited to a few hours of tweet acquisition

The starting point for the exercise is a python script created by the instructor that obtains tweets in real time using the Twitter streaming query API. The script contains several editable parameters including search keywords and running time, and is freely available on the author's website (http://castle.eiu.edu/~bjkronenfeld/). Geocoding is performed using the Geonames web service (www.geonames.org), which currently allows 2,000 geocoding operations per hour. To run the script, users must supply their own Twitter and Geonames account information.

For sample data and as the basis of a case study, tweets with the keywords "Trump", "Oscars" and "Moonlight" were collected on Feb. 28, 2017 between 10-11am and 8-9pm EST (7-8am and 5-6pm PST).

3. Pragmatic Geocoding

Tweets can have location information in at least three forms (Hecht et al. 2011). Tweets are *geotagged* using GPS if this option is enabled by the Twitter user. Text-based information may also be contained in the location field associated with each user's account. In some cases, lat/lon values are automatically *embedded* into this location field when the user account is created through certain clients. More commonly, users are able to enter their own *freeform* text-based location description.

It has been estimated that only 1.23% of tweets are geotagged (Liu et al. 2014). This rate would be too low to support the *Twitter Wars!* exercise. For example, at 250 tweets per minute (typical for popular, trending topics), with 50% of tweets coming from the USA and 1.23% of these geocoded, over 32 hours of tweet collection would be required to obtain 3,000 tweets. The number of tweets with *embedded* coordinates has been reported as high as 11.5% (Hecht et al. 2011), but in preliminary experiments I found typical rates to be much lower. A much larger proportion of tweets contain freeform text-based location descriptions, but the exact proportion depends on how quality is assessed. For example, Liu et al. (2014) found 42.4% of location strings could be interpreted by Bing with "high confidence". However, they did not validate these results manually. Based on human assessment, Hecht et al. (2011) found 54.5% of the tweets they examined to containing valid geographical descriptions. On the other hand, they also found that 34% of users entered fake locations, indicating that care must be taken to avoid false geocoding results.

3.1 Quality Assessment

Although some research has been conducted on the data quality of Twitter location descriptions as noted above, I am not aware of any study rigorously assessing the spatial vagueness, ambiguity and granularity of these descriptions. Such effort may be useful to validate pragmatic approaches to obtaining a reasonable quantity of high quality data efficiently. To provide a preliminary assessment, a random subset of 500 location descriptions from the sample data were examined. Among these, 77 were deemed not to contain any usable location information. An additional 94 locations outside of the USA were excluded from analysis. The remaining 329 descriptions of USA locations were

categorized in two ways. First, each location description was assigned a granularity level based on the most specific location described. Second, descriptions were marked if they contained one of six types of obstacles to automated geocoding. Results are shown in Table 1.

Table 1. Freeform descriptions of 329 tweet locations in the USA, classified according to spatial granularity and six types of obstacles to automated geocoding. Some tweets contained more than one obstacle.

Granularity	**Tweets**	**Pct**	**Geocoding Obstacles**							
			multiple locations	***ambiguous name(s)***	***spatial vagueness***	***colloquial/ historical***	***unusual format***	***spatial/ nonspatial***	***total affected***	***pct***
country	37	11%	0	0	0	6	2	2	8	22%
multistate	8	2%	0	0	4	3	0	2	8	100%
state	80	24%	1	0	0	6	3	4	11	14%
intrastate	7	2%	0	0	6	1	0	1	6	86%
county	3	1%	1	0	0	0	1	0	1	33%
metro area	10	3%	0	0	3	2	4	0	8	80%
city	167	51%	7	4	1	9	8	11	26	16%
intracity	11	3%	1	0	0	2	0	1	4	36%
ambiguous	6	2%	1	6	1	0	0	0	4	100%
all combined	*329*	*100%*	*11*	*10*	*15*	*29*	*18*	*21*	*76*	*23%*

Approx. half of all location descriptions were at the city level, and another ¼ were at the state level. Notably, descriptions at these levels of granularity contained the fewest obstacles to geocoding.

Based on this analysis, a pragmatic high-volume high-quality (HVHQ) strategy for tweet collection in the USA would be to focus on the city level first. Such a strategy was implemented in the python script using a simple text-based filter with the following three patterns:

<any text> <state name>
<any text> <state abbreviation>
<large, unambiguous city name>

Only descriptions conforming to one of these patterns are passed to the Geonames geocoding service. To support this strategy, lists were compiled of the 50 U.S. states and abbreviations, as well as 128 major cities with unambiguous names. Cities with ambiguous names were excluded from the list, meaning (for example) that 'Portland, ME' would be geocoded but 'Portland' would not.

This strategy was assessed on the sample data shown in Table 1, yielding geocoding results for 141/329 (43%) of USA tweets with location information. Among these, 131 (93%) were city-level descriptions and another 4 (3%) were sub-city-level descriptions that were correctly geocoded. Five of the remaining six (4%) ("Sonoma County" (x2), "San Francisco Bay Area", "southern California", "West Texas") were geocoded to a seemingly arbitrary point. None of these were entirely erroneous, although the latter was geocoded to a city center even though it more likely indicates a large, vaguely defined region. One description ("Dallas, PA") was geocoded to a small city in Pennsylvania, but might refer to two locations instead of one.

3.2 Case Study

A total of 32,440 tweets were collected during the two 1-hour periods, of which 8,241 (25.4%) were geolocated by the HVHQ strategy (Table 2).

Table 2. Number of tweets collected meeting criteria for analysis.

Group	**Count**
Total	**32,440**
not geolocated	*24,199*
geotagged	***23***
embedded lat/lon	***21***

geolocated freeform description	***8,197***
Total Geolocated	**8.241**
outside USA	*18*
Alaska	*16*
Hawaii	*23*
Contiguous USA	**8,184**
both keyword groups	*121*
neither keyword group	*1,748*
"Oscars" or "Moonlight"	***1,079***
"Trump"	***5,236***
Total Tweets Used in Analysis	**6,315**

Surprisingly, only a tiny proportion of tweets were geotagged (23, 0.07%) or contained embedded lat/lon coordinates (21, 0.06%); the remainder were freeform descriptions geolocated using the HVHQ strategy. Since this strategy relies on lists of U.S. places, the vast majority of geolocated tweets (8,223, 99.8%) were located in the USA. Of these, 16 in Alaska and 23 in Hawaii were removed, and an additional 1,869 were eliminated because they either didn't contain direct references to the target keywords (these often contained indirect references, such as urls linking to Trump's personal tweets) or contained references to both keyword groups. Two tweets ("Monterey Bay, CA" and "Manhattan Beach, CA") were located in the ocean (per a 1:500,000 census TIGER dataset) and had to be manually included.

4. Information through Transformation

At the time tweets were collected, the movie *Moonlight* had just won the Best Picture "Oscar" award. The objective of the case study was to determine where this was most discussed on Twitter. The data, however, reflects a myriad of unrelated, idiosyncratic circumstances including the time(s) of day, frequency and timing of Twitter activity, and heterogeneous population density. A series of five cartograph transformations was developed to elicit meaningful spatial information from this noisy base. Each transformation is not new, but by placing them in a sequence I hope to illustrate how cartographic visualization of social media data should proceed as a set of structured transformations to eliminate specific sources of noise and bias and to reveal meaningful pattern.

As a starting point, Fig. 1 shows the exact locations of the 6,315 tweets in the case study. This map looks messy and is messier than it looks. The inset map illustrates a point with 87 coincident tweets, while nearby points represent only one tweet. If the meaning of a basic point symbol is so uncertain, is there any hope to elicit meaning from such a map?

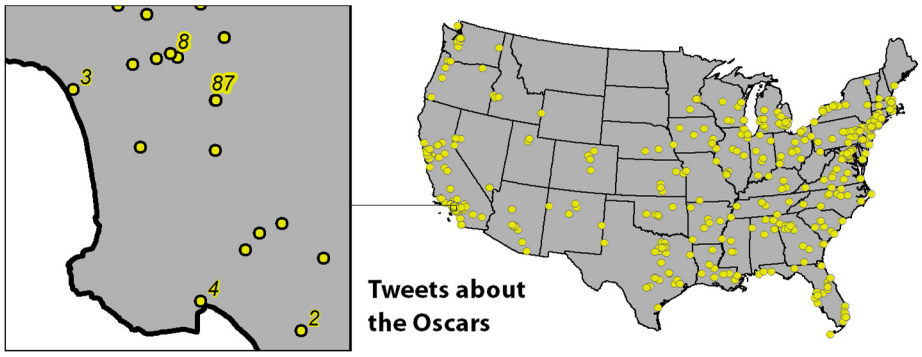

Fig. 1. Distribution of geolocated tweets used in case study. Labels on inset map shows number of coincident points.

4.1 Density Transformation

The first transformation turns events into event rates per unit land area, thereby avoiding the problem of overlapping symbols. An example of a kernel density transformation is illustrated in Fig. 2.

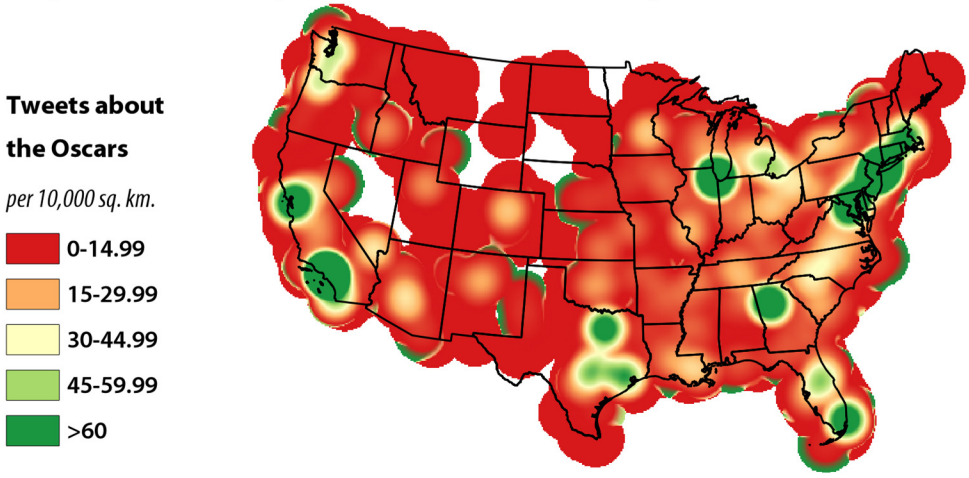

Fig. 2. Density of tweets containing the keyword "Oscars" or "Moonlight".

The map is appealing enough to lure many an intelligent cartography student into complacency, but actually reveals depressingly little. High-density regions include all of the major cities (Los Angeles, Chicago, Atlanta, etc.), and one imagines a similar pattern could be elicited from almost any keyword. If the pattern of *Oscars* tweets differs in a meaningful way from that of other topics, it is difficult to ascertain from this map!

4.2 Normalization

Next is the well-known process of *normalization*, which transforms the data by comparing the events of interest to the population from which they are likely derived. Since Twitter users are people, it is tempting to normalize to population. However, tweet rates differ from one location to another, geocoding success rates vary (e.g. fewer successfully geo-coded tweets from Portland than Seattle), and tweets are more frequent during certain times of the day (here, collection was at 8pm in Boston but 5pm in Los Angeles). Since these effects are impossible to quantify precisely, the only good solution for keyword-based Twitter data is to normalize against other keyword-based Twitter data collected at the same time using exactly the same techniques. In the case study, this was accomplished by acquiring tweets containing the keyword "Trump" as well as the Oscars. This is illustrated in Figure 3.

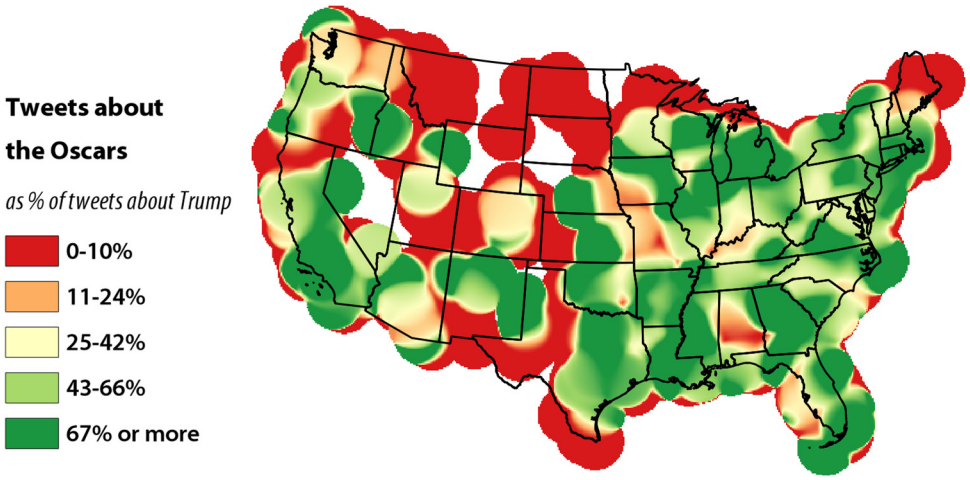

Fig. 3. Ratio of density of tweets containing Oscars keywords to density of tweets containing "Trump". Tweets on both keywords were collected concurrently in the same set of Twitter API requests.

4.3 Anchor Scaling Transformation

While data normalization is essential to derive meaningful information, it may be difficult to find an anchor for interpretation. What ratio of *Oscars* tweets to *Trump* tweets is "normal"? To answer this visually, data values are often compared to an average over the entire study area. This is illustrated in Fig. 4. Using a diverging color scheme, the middle (yellow) color provides a visual anchor from which it can be determined where relative interest in the Oscars is higher or lower than the national average.

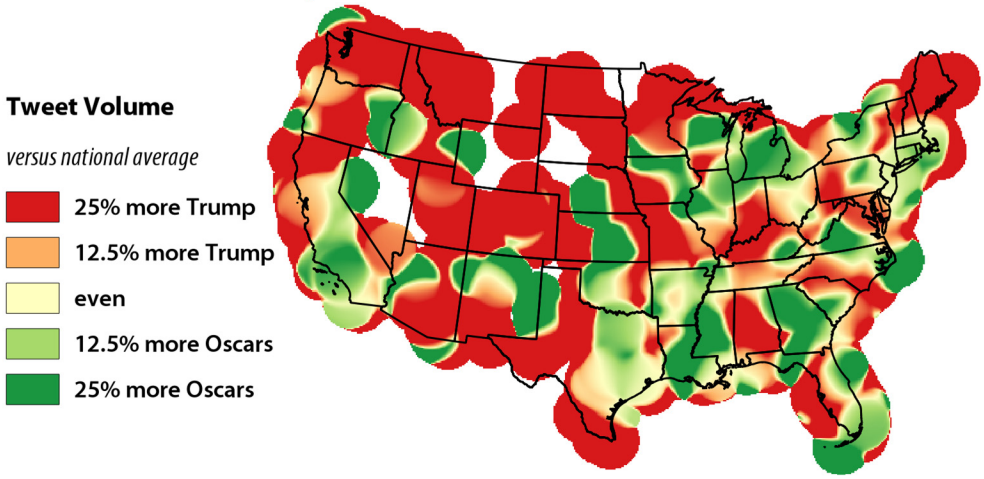

Fig. 4. Ratio of values shown in Fig. 3 to national average.

4.4 Color Blending

Individual values in Fig. 4 are accurate and informative, but visual emphasis is biased because locations with very different tweet volumes are treated the same. For example, a single tweet in northeastern Wyoming is visually equivalent to nearly 500 southern California tweets. One way to avoid this is to use graphical blend modes to de-emphasize areas with fewer tweets. In Fig. 5, total tweet density is used to *screen* the colors from Fig. 4. The result is a bivariate map. Matching the lighter colors produced by the *screen* blend mode with the map background conveys the general sense that data is lacking.

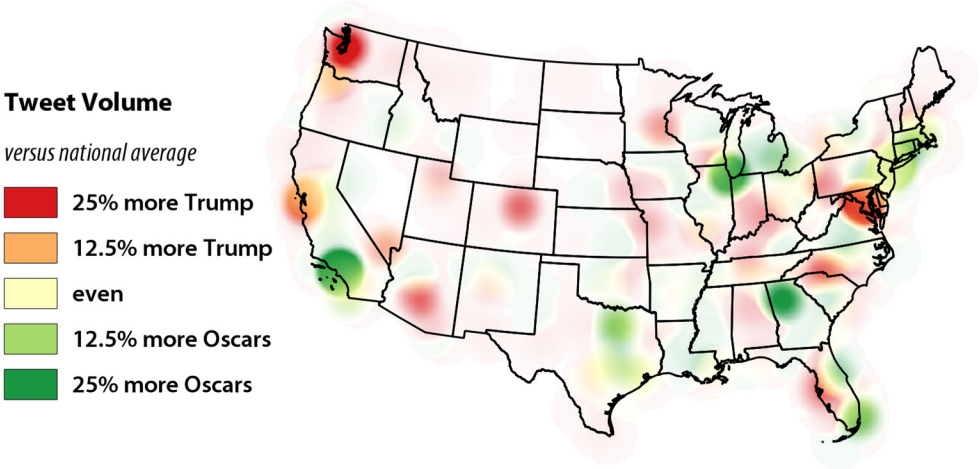

Fig. 5. Values in Fig. 4 screened by total tweet density.

4.5 Space Transformation

There is a lot of white in Fig. 5. Once the bias in visual emphasis is removed, it is revealed that all along most of the information was crammed into a small portion of the map. This limits the amount of detail that can be seen and also results in visual bias, as data from any location near another more populous location will be effectively hidden. One solution, illustrated in Fig. 6, is to place the data on a population cartogram. Here, the cartogram is not used for "shock value" but instead serves as a base map, allowing visual inferences to be made per person rather than per population (Sui and Holt 2008). The population is effectively spread out, allowing patterns to be seen within. For example, Baltimore (more *Oscars* tweets) and Philadelphia (more *Trump* tweets) are now distinct from Washington D.C. and New York, and seven distinct tweet centers can be identified in California.

On the cartogram, variation in color value created by the blend mode conveys new, useful information. Screening was based on tweet volume per "area", but since "area" on a cartogram is proportional to population, the result is that deeper colors on the cartogram represent more tweets per population. Thus, for example, it is revealed that people are tweeting more in Texas than in Michigan, while New Yorkers are conspicuously silent.

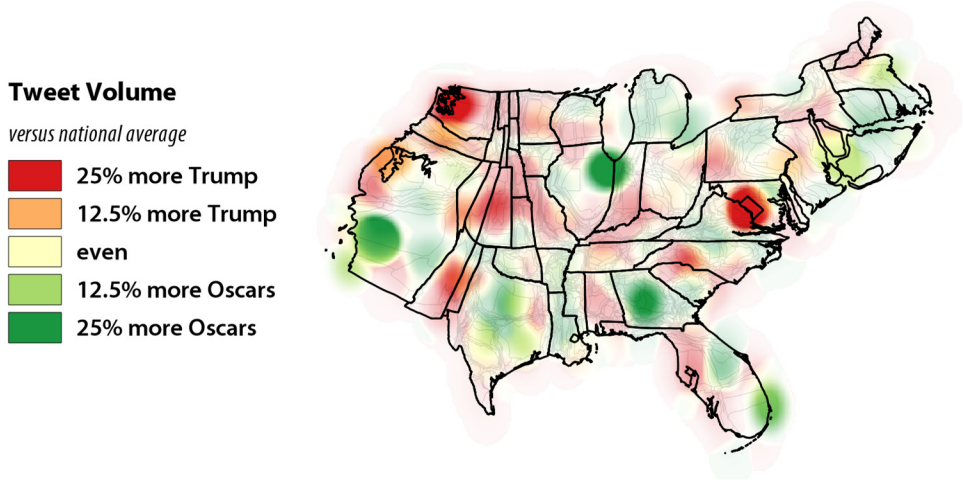

Fig. 6. Data from Fig. 5 shown on 2010 county-level population cartogram.

5. Discussion

This study examined pragmatic issues related to obtaining geolocated tweets and constructing meaningful twitter maps. By focusing on city descriptions, the HVHQ geocoding strategy appears to be an effective way to collect a reasonably high volume of accurately geocoded tweets in the USA, but assessment of a larger sample is needed and different strategies may work better in other regions. A series of five transformations (density, alternate keyword basis, anchor scaling, color blending and space) was proposed to provide alternative cartographic representations that systematically remove biases to reveal meaningful patterns. All but the last transformation can be readily performed in most GIS, with many possible variations.

References

Han, S.Y., Tsou, M-H., Clarke, K.C. (2015). Do global cities enable global views? Using Twitter to quantify the level of geographical awareness of U.S. cities. *PLOS One*, 10(7): e0132464. doi:10.1371/journal.pone.0132464

Hecht, B., Hong, L., Suh, B., Chi, E. H. (2011). Tweets from Justin Bieber's heart: the dynamics of the location field in user profiles. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, Vancouver, BC, Canada, May 07 - 12, 2011. Pages 237-246. NY, NY: ACM.

Liu, Y., Kliman-Silver, C., Mislove, A. (2014). The Tweets They Are a-Changin': Evolution of Twitter Users and Behavior. *International AAAI Conference on Weblogs and Social Media (ICWSM)* 13: 55.

Shook, E., Turner, V.K. (2016). The socio-environmental data explorer (SEDE): a social media–enhanced decision support system to explore risk perception to hazard events. *Cartography and Geographic Information Science*, 43(5), 427-441, doi:10.1080/15230406.2015.1131627

Sui, D.Z., Holt, J.B. (2008). Visualizing and analysing public-health data using value-by-area cartograms: Toward a new synthetic framework. *Cartographica*, 43(1):3-20, doi: 10.3138/carto.43.1.3

Tsou, M-H. (2015). Research challenges and opportunities in mapping social media and Big Data, *Cartography and Geographic Information Science*, 42(sup1), 70-74, DOI: 10.1080/15230406.2015.1059251

Zhou, X., Zhang, L. (2016). Crowdsourcing functions of the living city from Twitter and Foursquare data. *Cartography and Geographic Information Science*, 43(5), 393-404, doi:10.1080/15230406.2015.1128852