The Geography and Importance of Localness in Geotagged Social Media

Isaac L. Johnson*, Subhasree Sengupta*, Johannes Schöning†, Brent Hecht*
*GroupLens Research, Department of Computer Science, University of Minnesota,
†Expertise Center for Digital Media, Hasselt University – tUL -iMinds
ijohnson@cs.umn.edu, sengu025@umn.edu, johannes.schoening@uhasselt.be,
bhecht@cs.umn.edu

ABSTRACT
Geotagged tweets and other forms of social media volunteered geographic information (VGI) are becoming increasingly critical to many applications and scientific studies. An important assumption underlying much of this research is that social media VGI is "local", or that its geotags correspond closely with the general home locations of its contributors. We demonstrate through a study on three separate social media communities (Twitter, Flickr, Swarm) that this *localness assumption* holds in only about 75% of cases. In addition, we show that the geographic contours of localness follow important sociodemographic trends, with social media in, for instance, rural areas and older areas, being substantially less local in character (when controlling for other demographics). We demonstrate through a case study that failure to account for non-local social media VGI can lead to misrepresentative results in social media VGI-based studies. Finally, we compare the methods for determining localness, finding substantial disagreement in certain cases, and highlight new best practices for social media VGI-based studies and systems.

Author Keywords
Geotagged social media; volunteered geographic information; localness; user-generated content;

ACM Classification Keywords
H.5.m. Information interfaces and presentation (e.g., HCI): Miscellaneous;

INTRODUCTION
Social media volunteered geographic information (VGI) such as geotagged tweets, geotagged photos, and "check-ins" provides an unprecedented real-time lens into many important spatiotemporal processes. As has been noted recently in *Science* [37], this lens has been a game changer when it comes to the study of many of these processes. Indeed, researchers in HCI (e.g., [1,9,45,50]), the social sciences (e.g., [10,14,21,39,46,51]), and even the natural sciences (e.g., [38,44,49]) now regularly use social media VGI to better understand phenomena of interest ranging from social unrest and emergencies [6,24,43] to disease tracking [22,27,42].

A common thread in studies of social media VGI is the reliance on a simple assumption we call the *Localness Assumption*. Under this assumption, which is almost always adopted implicitly, *a unit of social media VGI always represents the perspective or experience of a person who is local to the region of the corresponding geotag*. Put more simply, the localness assumption presumes that social media users can be considered locals from everywhere they post geotagged content. For example, adopting the localness assumption, one can assume that a person who posts a geotagged tweet about a political candidate is doing so from her or his home voting district, thereby affording applications like election forecasting and political preference monitoring.

Human mobility is the major potential confound to the localness assumption. Studies that adopt the localness assumption implicitly argue that people who post geotagged social media while on business trips, vacations, and other forms of travel are not a significant factor across large datasets of social media VGI. This consideration of human mobility and the localness assumption more generally dates back to the origins of the term "volunteered geographic information": in the foundational paper on VGI, well-known geographer Michael Goodchild argued that the core value of VGI is that it tends to come from locals. However, Goodchild was writing in a time largely before smartphones and social media, a time when it might be reasonable to assume that human mobility may be dampened in VGI datasets.

In this paper, we present the first systematic examination of the validity of the localness assumption in social media VGI. Analyzing four datasets across three distinct types of social media VGI, we find that, due to human mobility, *the localness assumption does not hold for approximately 25% of social media VGI.* Additionally, we identify extensive geographic variation in the localness of social media VGI. In other words, while Goodchild's "localness ideal" – in which VGI contains high-quality local information – holds

CHI'16, May 07-12, 2016, San Jose, CA, USA
 ISBN 978-1-4503-3362-7/16/05 $15.00
DOI: http://dx.doi.org/10.1145/2858036.2858122

somewhat true in certain areas, there are other areas where the connection between the population contributing social media VGI and local population is much more tenuous. Of particular concern, we find that the degree of localness in an area tends to compound previously identified population biases in social media VGI [20,28,31], with rural and older areas not only having a diminished voice overall in social media VGI, but (as our results show) that voice is diluted by outsiders at a disproportionate rate.

Through a case study focusing on recent work that assesses the "geography of happiness" in the U.S. through geotagged tweets [32], we explore the direct effect of the localness assumption on social media VGI-based studies. We replicated the approach employed in [32] and compared its output to that of several versions of the approach that explicitly filter out non-local tweets (thereby accounting for human mobility and not adopting the localness assumption). We found this filtering process resulted in small shifts in the happiness geography of the United States overall, but that there were significant shifts in certain key types of regions, highlighting the importance of filtering out non-local content when doing social media VGI-based research.

This paper also makes an important methodological contribution: this paper is the first to aggregate and compare the various methods for determining VGI localness in the small literature that has explicitly considered localness, finding important differences between these methods. We discuss the implications of these results and outline best practices for filtering out non-local content in studies that use social media VGI.

To summarize, this paper makes the following contributions:

- This is the first paper to characterize the amount of non-local content in multiple social media VGI repositories, finding that approximately 25% of geotagged content on average is non-local to an area.
- We find that the geographic contours of localness follow important sociodemographic properties, with, for example, more urban and younger areas having consistently greater proportions of local content.
- We examine the impact of non-local VGI in one of the many studies that adopts the localness assumption. We show that the presence of non-local VGI can significantly alter algorithmic determination of regional properties (e.g., "happiness" retrieved from tweets) for certain areas.
- We also make an important methodological contribution: we characterize the various definitions of localness in use by the research community, find clear differences in their results, and outline a series of corresponding best practices for social media VGI studies.

Below, we first cover related work and detail the social media VGI datasets we consider. We then present our methods and results, with this discussion structured around four research questions. Each of these questions corresponds to one of the contributions listed above. We close by highlighting the broader implications of our work, a discussion that includes a series of localness best practices for social media VGI researchers.

RELATED WORK

This research is primarily motivated by three areas of prior work: (1) research on localness in VGI, (2) research on population biases in social media, and (3) studies that use social media VGI. Below, we address each area in more detail.

Localness in VGI

Key inspiration for this paper came from the work of Sen et al. [41] that demonstrated that, at a country-to-country scale, there is extensive geographic variation in the localness of geographic content in Wikipedia (peer production VGI) and that this variation corresponds to global socioeconomic contours. This paper can very broadly be thought of as an extension of Sen et al.'s work to the social media VGI domain. Social media VGI has a fundamentally different "spatial content production model" [18] than peer production VGI, which suggests that its localness dynamics will be substantially different (i.e. to post a geotagged tweet, one has to be at the location of the geotag, but one can write a Wikipedia article about anywhere in the world from anywhere in the world). This work also addresses the call in Sen et al. for localness work that considers localness at a spatial scale more granular than that of the country (we study localness at the U.S. county scale).

Sen et al. is not the only work to consider localness in a peer production VGI context. Other, more peripherally related research includes work establishing that local peer-produced VGI is of higher quality than non-local contributions [11,52], modeling Wikipedia contributions as a spatial process [17], and examining global core/periphery dynamics between Wikipedia editors and the geographic articles that they edit [16].

Additional core motivation for this work is derived from Hecht and Gergle [18], which examined the distance between content contributors and the subjects of their contributions in Wikipedia and Flickr. To our knowledge, this is the only other research to directly consider localness in a social media VGI context, finding that while the computer vision community had often assumed that that Flickr photos primarily came from tourists (the opposite of the localness assumption that is pervasive in social media VGI research), many Flickr photos come from locals (a dynamic beautifully visualized in the maps created by Fischer [12]). This research builds on that of Hecht and Gergle by directly targeting the localness assumption that is central to so many studies that utilize social media VGI, characterizing the degree of localness across three separate types of social media VGI, its geographic and

sociodemographic variation, and its effects on studies. Moreover, this is the first work to problematize the operationalizations of localness that have appeared in the literature (including that in the work of Hecht and Gergle [18]) and identify corresponding best practices.

Population Bias in Social Media

Another area of work related to this research is the body of literature on population bias in geotagged social media, a problem that has also been observed in social media more generally [37]. For instance, Li et al. [28] found that the density of tweets per capita and geotagged Flickr photos per capita are positively correlated with sociodemographic factors like income, youth, and education (when comparing county-aggregated values in California). Similar work with Twitter was done by Malik et al. [29], but across the United States. Hecht and Stephens [20] focused specifically on the rural/urban divide, finding that in Foursquare, Twitter, and Flickr, there was far more content per capita in urban areas than rural areas. As we will see below, when it comes to geographic variations in localness, the same types of areas that tend to be advantaged when it comes to raw quantity of social media VGI (e.g., urban areas) also tend to be advantaged in terms of the their VGI's localness.

Social Media VGI Based Studies

Like has been the case for social media more generally [37], geotagged social media has provided tremendous new opportunities for research in a wide variety of domains and across a broad swath of disciplines. Venerandi et al. [47] and Wood et al. [49] provide good overviews of aspects of this work. As noted above, much of the research has adopted the localness assumption. For instance, this has occurred in work that seeks to describe diurnal patterns in sentiment [15], predict public health measures [9], and measure human mobility [5], among other studies (e.g., [1,8,50]). Below, we show through a case study on the work of Mitchell et al. [32] what can occur if non-local VGI is filtered out of these studies. Notably, localness likely does not pertain to most studies that use geotagged social media for sensing natural events such as earthquakes or disasters (e.g., [24,38]).

Several studies have made intentional efforts to filter out non-local tweets, thus explicitly eschewing the localness assumption. For instance, the work of Li et al. [28] and Hecht and Stephens [20] fall into this class of research. However, as we will see below, these studies and others that have separated local and non-local tweets take unique approaches to doing so, with each approach operationalizing very different understandings of what it means to be a local.

Finally, it is important to note that a much larger body of work unintentionally works around the localness assumption by using data from the location fields of social media users' profiles rather than geotags [3,26,34]. While this data source has serious problems that geotags do not (e.g., [19,48]), when valid, associating a user's social media with their self-described home location rather than the locations of its geotags is one technique to filter out non-local social media. As such, when comparing approaches to quantifying localness in social media VGI, we consider the use of the location field as one such technique.

DATASETS

In order to gain a broad, ecologically valid understanding of localness phenomena in social media VGI, we look at VGI from three different types of popular social media communities: a microblogging platform (Twitter), a photo-sharing community (Flickr), and a check-in based location-based social network (Swarm). In this section, we describe our data from each of these communities (and other sources) in more detail.

Twitter

We analyze two datasets of geotagged tweets: both were gathered through the public Streaming API and were restricted only to tweets with geotags (latitude and longitude coordinates). The first dataset, which we shall refer to as *T-51M*, was gathered from October 19, 2014, through November 19, 2014. It contains 51.2 million tweets in the contiguous United States from 1.6 million users. The second dataset, *T-11M*, was gathered from May 27, 2015 through August 19, 2015. It contains 10.8 million tweets from 964,000 users in the contiguous United States[1]. We combine these two datasets for our study of happiness to improve robustness and incorporate data from different times of the year. The combined dataset has 61.9 million tweets from 2.2 million unique users. For all Twitter datasets, we remove organizational accounts per best practices for social media research [37] prior to analysis through use of the classifier described by McCorriston et al. [30]. Organizational tweets comprised 6.3% of the total dataset.

Flickr

For Flickr, we analyze the YFCC100M dataset, which we shall refer to as *F-15M*. It contains 15.4 million geotagged Creative-Commons-Licensed photos from 73,797 thousand users in the contiguous United States. The YFCC100M dataset was publicly released by Yahoo Labs and Flickr in June 2014 and has been used in research such as predicting location based on photo tags [25].

Swarm

Swarm (formerly Foursquare) is a location-based social network in which users check in to locations and broadcast this information to their social network. The Swarm API does not allow public access to check-in data, but some users choose to publicly tweet their check-ins. Swarm check-ins shared via Twitter have been used extensively in the past (e.g. [13,35]). We analyze the dataset collected by Cheng et al. [5] from September 2010 through January

[1] The size difference in these datasets arises from using a contiguous United States and global bounding box respectively.

2011 consisting of 7.8 million check-ins from 89 thousand users for the United States, which we shall call *S-8M*.

Sociodemographic Statistics

Several of our analyses involve comparing sociodemographic information to the percentage of VGI in a U.S. county that is local (U.S. counties are second-order administrative districts, right below states). All of the sociodemographic variables we examine relate to population biases that have been detected in prior work with social media VGI: from the 2010 US Census, we examine urban/rural (*% Urban)* [20,29,31], race (% White, Non-Latino or *%WNL*) [28,31], and age (median age or *MedAge*) [28,29]. From the 2009-2013 American Communities Survey, we examine income (household median income or *HMI*) [28,29] and % Management, Business, Science, and Art occupations (*% MBSA*) [28].

Because these data are limited to the United States, we focus in this paper on social media with geotags that fall within U.S. borders. Additionally, due to the requirements and assumptions of our spatial modeling approaches, analyses are limited to the contiguous United States (i.e. the "lower 48").

RESEARCH QUESTIONS

In this research, we posed four separate research questions, each of which corresponds to one of the contributions enumerated above. Specifically, we asked:

- **RQ0:** What precisely does it mean for a unit of social media VGI to be local to a given region?
- **RQ1:** What percent of social media VGI is local? In other words, to what extent is the localness assumption true for social media VGI?
- **RQ2:** What is the geography of localness within social media VGI? Does variation in localness follow the same socioeconomic contours that are seen in other types of volunteered geographic information?
- **RQ3:** How does the inclusion of non-local contributions impact the results of research and algorithms that leverage social media VGI?

In the following sections, we present our methods and results associated with each research question. This is followed by a holistic discussion of the implications of our results.

RQ0: WHAT IS LOCAL?

Methods

To address the challenge of defining localness (and what it means for geotagged social media to be local), we turned to the limited social media VGI literature that has made efforts to separate local and non-local information. Conducting the first survey of techniques for distinguishing local from non-local content based on geotags, we identified four approaches in this literature, with each quantifying a different definition of localness. We implemented all four of these approaches, and use all four throughout this paper.

In order to gain a better understanding of each localness approach and how it relates to the others, we classified every unit of social media in all four repositories as either local or non-local according to each approach and compared and contrasted the results. Below, we first describe the four localness approaches in more detail (as well as cover the key role that spatial scale plays in all of them). Following that, we discuss the results of our comparative analysis.

"n-days" Localness Metric

The *n-days* localness metric [20,28] takes all of a user's contributions (e.g., tweets, photos, check-ins) and assigns the user as local to a given region (e.g., county, city) if they made contributions in that region at least *n* days apart. In order to be considered a local, this metric thus requires a person to demonstrate that they have either spent at least *n* days in a particular region or returned there at a later date at least *n* days after they initially contributed. A sufficiently large choice of *n* must be used to filter out people who are just traveling through a region. The choice of *n* has varied, but both Hecht and Stephens [20] and Li et al. [28] use 10 days as the minimum length of time. We follow suit and set *n* equal to 10 days. Note that the *n-days* metric operationalizes an idea of localness in which a person can be local to between 0 and *m* regions, where *m* is the number of regions being studied – e.g., the number of counties in the U.S. – although in practice the number of local regions tends to be low.

"Plurality" Localness Metric

The *plurality* metric [20,33] assigns a user as local to the region in which they contributed the most social media VGI in a given repository. Uniquely, this algorithm ensures that even users that do not make frequent contributions (e.g., who for example might be filtered out by the *n*-days algorithm for sheer lack of content) will still be included in the analysis. *Plurality* assigns each user as local to exactly one place, except in ties in which case all regions at that level of contribution are local.

"Geometric Median" Localness Metric

The *geometric median* metric [23] has been most commonly used in the geolocation inference literature to assign a home location to users. We implement the multivariate L1-median definition used, for example, by Jurgens [23] and Compton [7], which defines the median of a set of points as the point in space that minimizes the distance between it and all of the points in the set. We further require, per Jurgens [23], that users have a minimum of five VGI points and that the median absolute deviation of the user's points to their geometric median be no greater than 30 km (i.e. half of the user's points must be within 30 km of the geometric median).

"Location Field" Localness Metric

The *location field* metric (e.g., [19,23,36]) has been used heavily for expanding social media VGI datasets beyond just explicitly geotagged social media, which often make up

Repo	n-days (10)	Plurality	Geo. Med.	Loc. Field
T-51M	60.1%	100.0%	49.5%	34.4%
T-11M	65.9%	100.0%	24.7%	37.3%
F-15M	67.9%	100.0%	33.9%	31.1%
S-8M	88.3%	100.0%	69.1%	15.9%

Table 1. The recall of each metric, or the percentage of users who were assigned as local to at least one county.

a small overall percentage of these datasets [3,26,34]. As noted above, this approach uses the self-reported location information in the "Location" field in users' social media profiles, which exists for all three social media communities considered here. The accuracy and completeness of location field data has been problematized by Hecht et al. [19], but its use continues as location field data is one of the only ways to geolocate the large percentages of social media users who do not geotag their content.

In order to turn a textual location into a machine-readable latitude/longitude coordinate, a *geocoder* is necessary. We used Jurgens et al.'s Geonames-based geocoder [23], which builds on the Creative-Commons-Licensed Geonames places dataset and handles noisy text through a series of regular expressions and common replacements (e.g., St. and Saint). We further validated the implementation by comparing our results to those achieved by use of Wikipedia redirects as implemented in the WikiBrain library [40] and described in [19]. For location field entries that both tools could geocode (~54% of the Geonames results), there was 90% agreement.

Choosing the Correct Spatial Scale

A final source of variation in how localness has been operationalized in the literature occurs in the spatial scale of the localness definition. For example, Sen et al. [41] define a local Wikipedia editor as someone who edits an article about a place in the editor's home *country*, whereas Li et al. [28] define local at the U.S. *county* scale. In this paper, we focus on the county-scale for two reasons: (1) it is a common scale at which social media VGI research is done (e.g., [9,20,28]) and (2) it is a scale at which the sociodemographic information we need to address RQ2 is available.

Because our definition of localness is at the U.S. county-scale, this means that the "regions" operated on by *n-days* and *plurality* are U.S. counties. For instance, if a person tweets predominately from places within Cobb County, GA, under *plurality*, all tweets from this person with geotags within Cobb County will be considered local, and those outside Cobb County will be considered non-local. Unlike *n-days* and *plurality*, the *geometric median* and *location field* metrics map users to a point. In these cases, we use simple point-in-polygon operations to assign the user as local to the county that contains the point.

Putting it all Together: Calculating Localness

To make the process of calculating localness more concrete, let us consider the case of a tweet whose geotag refers to a point in Philadelphia County, PA. This tweet would be classified either as local or non-local depending on whether or not the user who posted the tweet is considered to be a local of Philadelphia County. More specifically, this is how each metric would make its localness assessment:

- *n-day*: If the user tweets multiple times in Philadelphia County over a span of at least 10 days, the tweet would be considered local.
- *plurality*: If the user had posted more (or equal) tweets in Philadelphia County than any other county, the tweet would be considered local.
- *geometric median*: If the user had posted at least five tweets and enough of them were centered in the Philadelphia area for the median to be in Philadelphia County and within 30km of half of the user's points, the tweet would be considered local.
- *location field*: If the user had written in her Twitter location field "Philadelphia" or "Philly" (or a similar variant) and that text was successfully geocoded to a lat/lon in Philadelphia County, the tweet would be considered local.

Results: Comparing Localness Definitions

Running all four localness metrics against the same datasets affords us a unique ability to compare and contrast the definition of localness each metric encodes. Overall, three trends emerge: (1) some localness metrics fail to identify a single local county for many users, (2) though we see substantial agreement in localness determinations for the users bridging our two Twitter datasets, there is a large minority for whom results vary, and (3) although there is not strong agreement between any of the metrics, *n-days*, *plurality*, and *geometric median* agree far more often with each other than any of the three do with *location field*.

Highly-varied Recall

With regard to the first theme, efforts to filter out non-local geotagged social media have not considered recall as an issue. However, Table 1, which shows the percentage of users for which each metric was able to find at least one local county, suggests that this is an important factor to consider. If a localness metric is not able to find a local region (e.g., county) for a given user, that user can have no local social media, thereby removing him/her from social media VGI studies that filter for localness (a practice strongly supported by other results in this paper). In cases when data is not plentiful (e.g., for analyses at very granular spatial scales), limited recall could be a major problem.

Looking at Table 1 in more detail, we see that while *plurality*, by definition, succeeds for all users, *location field* sits at the opposite end of the spectrum, failing to identify a local country for well over half of all users in every case. *Location field*'s low recall is most likely attributable to two factors: (1) not all users fill out their location fields

Repo.	n-days (10)	Plurality	Geo. Med.	Loc. Field
T-51M	84.0%	90.1%	91.2%	57.9%
T-11M	77.1%	76.9%	85.0%	51.1%
F-15M	78.4%	52.9%	70.7%	40.7%
S-8M	88.2%	70.1%	73.0%	1.1%

Table 2. Relative percentage of social media VGI classified as local. Geotagged social media from users for whom no local county could be identified are excluded from these figures.

(especially on Swarm) and (2) location field entries are often non-geographic in nature, which will result in the geocoder returning no value (in the ideal case) [19].

Varying Longitudinal Consistency

Examining the set of tweets from the 389,635 users who appeared in both our *T-51M* and *T-11M* datasets, we found a very high consistency for the *location field* metric (91%) as well as *geometric median* metric (74%). In other words, for users we could identify in both datasets, the counties for which they were considered local were frequently the same using *location field* and *geometric median*, even though there is a 7-month gap between the data collection periods. However, the same is not true for *plurality* (54%) and *n-days* (48%), suggesting that there is a trade-off between recall (*plurality* and *n-days* both have high recall) and longitudinal consistency. This is a point to which we return in the discussion section.

Different Localness Definitions, Different Results

Each localness metric operationalizes a different idea of localness, and, as such, it is not a surprise that they frequently disagree as to whether an individual piece of VGI can be considered a local to a county. *Plurality, n-days* and *geometric median* agreed the most, but, for instance, their output agreed that a given tweet was local only 76.9% of the time for *T-51M*, and that is the highest agreement of any of the four repositories[2]. *Location field* rarely came to the same conclusions as any of the other three metrics, and, as such, the repository for which there was the most agreement across all four metrics (*F-15M*) still had only 16.3% agreement (with agreement defined the same way as above).

Because of the diversity in localness operationalization across the four metrics, the remainder of our studies below use at least two metrics, and usually use all four so as to establish robustness across varying definitions of localness. In the discussion section, we outline how this approach is likely a best practice for social media VGI research more generally.

[2] Revised 7/17/17: Original sentence: "…agreed the most, but for users for which all three could determine at least one local county, their output overlapped by at least one county only 76.9% of the time for T-51M…"

RQ1: HOW LOCAL IS SOCIAL MEDIA VGI?

In this section, we discuss our research on assessing the degree to which social media VGI is local. In other words, in asking this question, we are inquiring whether the localness assumption is valid.

Methods

Once we had completed our work for RQ0, addressing RQ1 was very straightforward: we simply calculated the percentage of overall social media units that are local according to each algorithm.

Results

Table 2 shows the localness of each social media VGI repository according to each of the localness metrics. The picture of social media VGI localness that emerges from Table 2 is that while *the majority of VGI appears to be local according to most metrics, a large minority is non-local*. For instance, we see that according to *n-days*, the localness of our four repositories ranges from 77.1% (*T-11M*) to 88.2% (*S-8M*), with the *T-51M* and *F-15M* repositories' localness between these two values. With the median localness percentage across all four datasets and all four metrics being only 75%, *it is difficult to make the argument that the localness assumption holds true in social media VGI*.

RQ2: DOES LOCALNESS VARY GEOGRAPHICALLY?

In this section, we describe how we addressed our research question related to potential geographic variation in the localness results we reported above. We focus this investigation on whether any variation corresponds to important sociodemographic contours.

Methods

As the first step in addressing RQ2, we calculated the percent of social media VGI in each county that is local to that county. This is analogous to our approach outlined above for RQ1, but instead of identifying the share of social media that is local in entire repositories, we did so on a county-by-county basis. We then analyzed the localness ratio in each county (for each repository) in the context of key sociodemographic statistics of the county (see the Datasets section).

This analysis was conducted using a multivariate regression with percent local as the dependent variable and the sociodemographic statistics as independent variables. We first test for spatial autocorrelation (if none is found, traditional OLS would be appropriate) and then adjust for the presence of spatial autocorrelation as discussed in [9,29] by running either a spatial error or lag model from the R library package spdep [4]. We make the specific choice of model based on Lagrangian Multiplier measures of fit, which is considered a best practice in the field of spatial econometrics [2]. The dependent and independent variables are log-transformed as necessary to achieve normality and all variables are Z-score standardized so that we can relate all coefficients to changes in standard deviations and compare relative effect sizes accordingly.

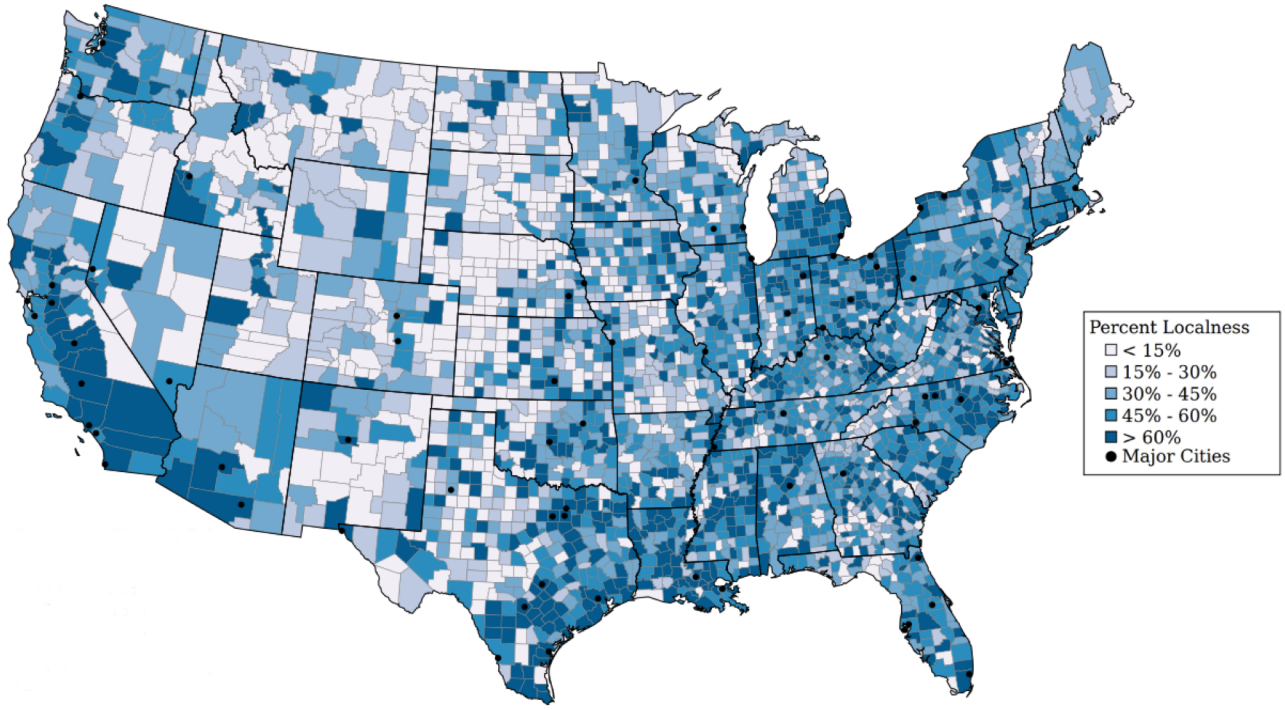

Figure 1. Map of Percent Local Content in T-11M according to the *geometric median* metric.

Repo.	% Urban	MedAge	HMI	%WNL	%MBSA
T-51M	0.29***	-0.18***	-0.06**	0.15***	-0.05**
T-11M	0.40***	-0.14***	-0.04	0.02	-0.04*
F-15M	0.28***	-0.06**	-0.01	0.14***	0.07***
S-8M	0.39***	-0.18***	0.01	0.16***	0.02

Table 3. Summary of % Localness Multivariate Regressions for *n-day* (10) localness filter. * is p<0.001, ** is p<0.01, and * is p <0.05.**

Results

Though we see, on average, that approximately 75% of content is local, the standard deviation for most metrics and datasets is around 20%, suggesting that there is noticeable geographic variation in the degree to which content is local. The results of our spatial regressions, which describe the percentage of local content in a county as a function of its sociodemographic factors, can be seen in Table 3. We report only the results for *n-day* calculations, but we ran the spatial regressions for each localness algorithm and found that all four algorithms had very similar results within each repository[3].

Table 3 reveals that the county-level variation in localness percentages is certainly not random; we see significant and consistent effects for a number of our independent variables. Across the board, there appear to be moderate increases in localness with increases in *%WNL* and increased youth (i.e. decreased median age) and much larger increases in localness with increased *% Urban Pop*.

Examining the effect sizes from our regression, we see that with all else held equal, for every standard deviation increase in the percent of the population that is urban (+31.6%), there is a 15-40% absolute increase in the localness of social media. This relationship results in social media VGI of substantially different character at opposite ends of the *% Urban Pop* spectrum. For instance, for counties with a *% Urban Pop* > 90% (e.g., San Francisco County, CA; Cook County, IL), 82% of *T-51M* tweets whose geotag is in the county come from a local user according to *n-day*. The corresponding value for counties with *% Urban Pop* < 10% (e.g., Twin Falls County, ID) is only 63%. Figure 1 shows a similar trend occurring with geometric median across the *T-11M* dataset, with the very rural Great Plains region containing much lower proportions of local content than more heavily populated areas. A similar trend occurs with *Median Age*, but with a smaller magnitude. Unpacking the normalized effect sizes in Table 3, there is a 6% decrease in localness in *T-51M* as the median age shifts from 32 (e.g., Newport News, VA) to 47 (e.g., Hernando County, FL).

A striking trend in Table 3 is the extent to which the table largely mirrors known findings about overall population bias in our three social media communities. It appears that not only is there more social media VGI per capita in urban areas [20,29], but our results suggest that urban social media is also far more local. The same is true with regard to

[3] The sole exception was location field for Swarm, which had mostly insignificant coefficients for the regression due to low recall.

Ranking	Unfiltered	*n-day*	*Plurality*
1	Montana	Montana	Montana
2	Vermont	Vermont	Maine
3	Maine	Maine	Vermont
•••		•••	
47	Delaware	Mississippi	Mississippi
48	Maryland	Maryland	Louisiana
49	Louisiana	Louisiana	Maryland

Table 4. The happiest and saddest states using an unfiltered dataset (i.e. no localness metric applied), *n-day*, and *plurality*.

County	Unfiltered	*n-day*	Change
St. Louis County, Missouri	162	657	-495
Baltimore County, Maryland	983	1251	-268
San Francisco, California	139	309	-170
Mercer County, West Virginia	475	296	+179
Iroquois, Illinois	479	289	+190

Table 5. Counties with some of the largest shifts in happiness ranking after filtering out non-local tweets.

population biases in age; older areas have less social media VGI per capita, and it is less local. This result is a point to which return in the discussion section below.

The results in Table 3 also suggest that the importance of filtering for localness is not geographically uniform. It appears that adopting the localness assumption will reduce the accuracy of studies that use social media VGI in rural areas more than it will reduce the accuracy in urban areas. The same is true of older areas vs. younger areas, areas with a smaller WNL population vs. a larger one (see discussion section), and so on.

RQ3: IMPACT OF NON-LOCAL VGI

While RQ1 and RQ2 sought to characterize localness in social media VGI, RQ3 seeks to directly understand its effects on social media VGI-based research. To do so, we adopt a case study approach, focusing on an analysis performed by Mitchell et al. [32]. This analysis used their algorithm from [10] applied to geotagged tweets to calculate the geography of happiness in the United States.

Methods

Using the data and code for their algorithm provided by Dodds et al.[4] and a combined version of our two Twitter datasets (*T-51M + T-11M*), we computed the geography of happiness in the United States at both the state-level and the county-level. We performed this computation three times: once under the localness assumption (in which we did no filtering for non-local tweets), once filtering out non-local tweets using *n-days* and once doing the same with *plurality*[5]. By comparing the results of these three computations, we can gain an understanding of the effects of the localness assumption (as well as effects of filtering by each localness metric).

We include the 48 contiguous states as well as Washington D.C. in our states calculation but limit the counties results to only those counties containing at least 3000 total tweets to ensure a sufficient sample size of tweets for the happiness algorithm.

Results

At the scale of states (e.g., *n-days* and *plurality* would group contributions from Los Angeles together with those from Yosemite National Park), we see no major shifts in the rankings of happiest states when filtering on *n-days* or *plurality* (i.e. when filtering out non-local tweets according to those metrics). The top three happiest and top three saddest states for each implementation are shown in Table 4. The one notable shift that we see is that tourism-heavy Nevada moves from a relatively happy rank 15 out of 49 in the unfiltered dataset all the way down to middle-of-the-road rank 25 when non-local tweets are filtered out, by far the largest shift of any state.

At the scale of counties, the difference in ranks between the unfiltered computation and those using *n-days* and *plurality* is larger, but not tremendously so. For *n-days*, the median absolute change in ranking from the unfiltered rankings is 32, but there are 105 counties that see shifts of more than 10% (126 rankings) up or down the list. *Plurality* had very similar results (median change 32; 96 moved by 10%).

The high-level results, however, obscure an interesting phenomenon in which several counties experienced very large jumps up and down the ranks when localness filtering was introduced. Examples of some of these counties are shown in Table 5. These are counties where the signal coming from the local populace is being overwhelmed by a very different signal from the non-local population. The greatest single change occurs in St. Louis County, Missouri, with a drop of 495 rankings when non-locals were filtered out. Baltimore County, Maryland, also moved substantially (10^{th} most) with a drop of 268 rankings. During the data collection periods, these two counties were experiencing fallout from the deaths of Michael Brown and Freddie Gray, respectively. In both cases, it appears that while local sentiment declined precipitously, this decline was obscured in the unfiltered dataset by travelers (non-locals), a group that did not experience the drop in sentiment.

The reverse appears to be happening in Mercer County, Arkansas, and Iroquois County, IL, which are both rural areas that happen to lie on major interstate highways. In

[4] https://github.com/andyreagan/labMT-simple

[5] Following the approach implemented by Hecht and Stephens, instead of removing non-local tweets as we have in RQ0 through RQ2, we instead assign them to their local counties/states.

both instances, non-locals who were driving through likely caused the drop in happiness in the unfiltered dataset relative to the filtered one.

DISCUSSION AND IMPLICATIONS

Best Practices for Social Media VGI Research

Ruths and Pfeffer [37] recently called for "higher methodological standards" for "large-scale studies of human behavior in social media." In doing so, they laid out a framework of best practices for working with social media, e.g., accounting for population biases, filtering for nonhuman accounts, and showing results from multiple platforms or temporally-separated datasets[6].

The research presented above points to an extension of Ruths and Pfeffer's framework that is specific to geotagged social media. Specifically, our results suggest the following five best practices:

Best Practice #1: As we found that a large minority of geotagged social media is not local, *studies that utilize geotagged social media should not adopt the localness assumption*. Doing so will result in significant and sociodemographically-biased side effects that, as we saw in our work related to RQ3, can alter study results.

Best Practice #2: We revealed clear differences between the localness metrics that have been used in the literature, indicating that *researchers need to think carefully about how to best operationalize localness for their research questions*. The decision of how to operationalize localness should involve both semantic and practical considerations, considerations that we unpack below.

Best Practice #3: With regard to semantic considerations, researchers should carefully examine which localness metric best fits the needs of their study. Each metric we identified in the literature defines localness differently, with metrics like *n-days* assuming a user can be local to many counties and metrics like geometric median assigning users to a single lat/lon coordinate. These definitional differences were likely a major cause of the deviations between the outputs of each localness metric, indicating that choosing an incorrect definition of localness may be costly.

Best Practice #4: With regard to practical considerations, according to the needs of a given study, *researchers must negotiate the trade-off of the consistency of metrics such as geometric median with the recall of metrics like plurality*. While geotagged social media occurs at massive scales at a high-level, if one is trying to study phenomena that occur in rural areas or at very granular spatial scales (or using a repository that does not make its data as available as Twitter, e.g., Swarm), a 56% (geometric median) or 70% (location field) reduction in the amount of data that can be considered (let alone that can be assured to be local) is highly problematic.

Best Practice #5*: Wherever possible and appropriate, researchers should consider multiple definitions of localness in parallel*, as we have done here. Given the differences in what is considered local and non-local by each metric, using multiple metrics can ensure that findings are robust against different definitions of localness. To ease this burden, future work could explore whether a single ensemble metric, such as a combination of *n-days* and *plurality*, is more robust to dataset variation.

While there are geotagged social media best practices outside the localness domain that also likely need to be encoded (e.g., guidelines for geocoding location field information as in [19]), our results make clear that a smart and intentional approach to handling localness is an important step towards a robust social media VGI study.

To promote these best practices and provide full access to our study, we have released our implementation of the localness metrics, R code for spatial regressions, additional maps, and full results[7].

Localness Compounds Population Bias

Population bias is a known concern with social media research. As we have shown, localness also tends to be lower in these areas already known to be disadvantaged in social media representation (e.g., older and more rural), compounding the existing population biases. In particular, this indicates that there are further barriers to robust research on rural areas or other underrepresented populations using geotagged social media. There is an established thread of geolocation inference research that seeks to locate social media without explicit location information. Our work motivates the need to develop these tools specifically with the goal of translating the non-geotagged social media of these underrepresented populations into VGI.

LIMITATIONS AND FUTURE WORK

Scale- and Time-Dependence

While we are the first paper to survey the localness metrics in use within the literature and directly compare their results, similar studies should be conducted varying the spatial scale at which localness is defined instead of the metric. It has long been known in geography that processes that operate one way at one scale may not operate in the same way at a different scale. It is possible that localness has a scale-dependent component. We began to explore this question with our *Happiness* case study, where we saw that the inclusion of non-local content at the state scale impacted the rankings to a lesser degree than at the county scale. Along with spatial scale, future work should look at

[6] We implement most of these suggestions in this work (i.e. removing non-human accounts in Twitter, testing on multiple repositories and datasets, and not relying on a single algorithm).

[7] https://github.com/joh12041/chi-2016-localness

the relationship between time and localness. Localness is not a time-invariant measure – that is, the degree to which one is local to an area degrades not only with distance, but also with time.

New Demographic Contours Should be Considered

While doing early work on the relationship between gender and localness, we noticed an interesting new difference between our localness metrics: Leveraging well-known gender inference approaches [9,31], we saw a consistent-but-small female skew in the users for whom *n-days*, *plurality*, and *geometric median* were able to assign at least one local county. However, this skew flipped and strengthened significantly for *location field*, to the point that *location field* is 23% more likely to be able to assign a local county to a male user than a female user. This result has potentially important implications for localness research, as well as for use of the location field on Twitter more generally (e.g., Does the use of this field induce a population bias?), implications that need to be explored in future work.

Establishing a Ground Truth

Finally, although we explored in detail the four major families of localness metrics in the literature, one metric that has yet to be considered but could lead to important innovation in this area is that which involves a traditional ground truth. By asking social media users where they believe they are local (outside the social context of the location field and without its particular constraints as outlined by Hecht et al [19]), it should be possible to construct a learned model (that uses the four localness metrics as features) of more robust self-reported localness. Moreover, adopting ground truth approaches could enable models that operationalize highly diverse understandings of localness, including those that take time into account (in a weighted or an unweighted fashion), implement a fuzzy definition of localness, and so on.

CONCLUSION

In this paper, we performed the first focused exploration of the extent to which geotagged social media can be considered to be from a local to the region of its geotag, an assumption that underlies many studies that utilize geotagged social media. We find that this assumption does not hold for about 25% of geotagged social media, although the exact percentage varies from social media community to social media community. We also saw that the degree of localness varies extensively geographically, and it does so in a fashion that mirrors existing sociodemographic contours (e.g., more rural areas are less local). Through a case study, we demonstrated that including non-local social media in research studies can lead to incorrect conclusions in certain cases, and we outlined a series of best practices to help researchers avoid this outcome and further strengthen their work.

ACKNOWLEDGEMENTS

We would like to thank Jacob Thebault-Spieker, Andrew Hall, Max Harper, Loren Terveen, and the rest of our GroupLens colleagues for brainstorming with us. This project was supported by NSF IIS-1526988, NSF IIS-1421655, a Google Research Faculty Award, FWO K207615N, and a University of Minnesota College of Science and Engineering Graduate Fellowship.

REFERENCES

1. Sofiane Abbar, Yelena Mejova, and Ingmar Weber. 2015. You Tweet What You Eat: Studying Food Consumption Through Twitter. ACM Press. http://doi.org/10.1145/2702123.2702153
2. Luc Anselin. 2005. *Exploring Spatial Data with GeoDa: A Workbook*. Center for Spatially Integrated Social Science.
3. Hila Becker, Mor Naaman, and Luis Gravano. 2011. Beyond Trending Topics: Real-World Event Identification on Twitter. *ICWSM*. http://doi.org/10.1.1.221.2822
4. Roger Bivand and Gianfranco Piras. Regional Research Institute.
5. Zhiyuan Cheng, James Caverlee, Kyumin Lee, and Daniel Z. Sui. 2011. Exploring Millions of Footprints in Location Sharing Services. *ICWSM* 2011.
6. Ryan Compton, Chi-Kwan Lee, Tsai-Ching Lu, Lakdeepal de Silva, and Michael Macy. 2013. Detecting future social unrest in unprocessed twitter data: "emerging phenomena and big data." *ISI*, IEEE, 56–60.
7. Ryan Compton, Craig Lee, Jiejun Xu, et al. 2013. Using publicly visible social media to build detailed forecasts of civil unrest. 1–11.
8. Justin Cranshaw, Jason I Hong, and Norman Sadeh. 2012. The Livehoods Project: Utilizing Social Media to Understand the Dynamics of a City. *ICWSM*: 58–65.
9. Aron Culotta. 2014. Estimating county health statistics with twitter. *JSM Proceedings*, ACM Press, 1335–1344. http://doi.org/10.1145/2556288.2557139
10. Bertrand De Longueville, Robin S. Smith, and Gianluca Luraschi. 2009. Omg, from here, i can see the flames!: a use case of mining location based social networks to acquire spatio-temporal data on forest fires. *Proceedings of the 2009 international workshop on location based social networks*, ACM, 73–80.
11. Melanie Eckle. Quality Assessment of Remote Mapping in OpenStreetMap for Disaster Management Purposes. Retrieved September 24, 2015 from http://iscram2015.uia.no/wp-content/uploads/2015/05/5-1.pdf
12. Eric Fischer. Locals and Tourists. Retrieved from https://www.flickr.com/photos/walkingsf/sets/72157624209158632/

13. Huiji Gao, Jiliang Tang, and Huan Liu. 2012. Exploring Social-Historical Ties on Location-Based Social Networks. *ICWSM*.
14. Song Gao, Linna Li, Wenwen Li, Krzysztof Janowicz, and Yue Zhang. 2014. Constructing gazetteers from volunteered Big Geo-Data based on Hadoop. *Computers, Environment and Urban Systems*. http://doi.org/10.1016/j.compenvurbsys.2014.02.004
15. S.A. Golder and M.W. Macy. 2011. Diurnal and Seasonal Mood Vary with Work, Sleep, and Daylength Across Diverse Cultures. *Science* 333, 6051: 1878–1881. http://doi.org/10.1126/science.1202775
16. Mark Graham, Ralph K. Straumann, and Bernie Hogan. 2015. Digital Divisions of Labor and Informational Magnetism: Mapping Participation in Wikipedia. *Annals of the Association of American Geographers* 0, 0: 1–21. http://doi.org/10.1080/00045608.2015.1072791
17. Darren Hardy, James Frew, and Michael F. Goodchild. 2012. Volunteered geographic information production as a spatial process. *International Journal of Geographical Information Science*: 1–22. http://doi.org/10.1080/13658816.2011.629618
18. Brent Hecht and Darren Gergle. 2010. On the "localness" of user-generated content. *CSCW*: 229. http://doi.org/10.1145/1718918.1718962
19. Brent Hecht, Lichan Hong, Bongwon Suh, and Ed H. Chi. 2011. Tweets from Justin Bieber's heart: the dynamics of the location field in user profiles. *CHI*, ACM.
20. Brent Hecht and Monica Stephens. 2014. A Tale of Cities : Urban Biases in Volunteered Geographic Information. *ICWSM*.
21. Nadav Hochman and Raz Schwartz. 2012. Visualizing instagram: Tracing cultural visual rhythms. *ICWSM-SocMedVis*, 6–9.
22. Myung-Hwa Hwang, Shaowen Wang, Guofeng Cao, Anand Padmanabhan, and Zhenhua Zhang. 2013. Spatiotemporal transformation of social media geostreams: a case study of Twitter for flu risk analysis. ACM Press, 12–21. http://doi.org/10.1145/2534303.2534310
23. David Jurgens, Tyler Finethy, James McCorriston, Yi Tian Xu, and Derek Ruths. 2015. Geolocation prediction in twitter using social networks: A critical analysis and review of current practice. *ICWSM*.
24. Marina Kogan, Leysia Palen, and Kenneth M. Anderson. 2015. Think Local, Retweet Global: Retweeting by the Geographically-Vulnerable during Hurricane Sandy. *CSCW*, ACM Press, 981–993. http://doi.org/10.1145/2675133.2675218
25. Giorgos Kordopatis-Zilos, Symeon Papadopoulos, and Yiannis Kompatsiaris. 2015. Geotagging Social Media Content with a Refined Language Modelling Approach. In *Intelligence and Security Informatics*, Michael Chau, G. Alan Wang and Hsinchun Chen (eds.). Springer International Publishing, Cham, 21–40.
26. Juhi Kulshrestha, Farshad Kooti, Ashkan Nikravesh, and P. Krishna Gummadi. 2012. Geographic Dissection of the Twitter Network. *ICWSM*.
27. Vasileios Lampos and Nello Cristianini. 2010. Tracking the flu pandemic by monitoring the social web. *Cognitive Information Processing (CIP), 2010 2nd International Workshop on*, IEEE, 411–416.
28. Linna Li, Michael F. Goodchild, and Bo Xu. 2013. Spatial, temporal, and socioeconomic patterns in the use of Twitter and Flickr. *Cartography and Geographic Information Science* 40, 2: 61–77. http://doi.org/10.1080/15230406.2013.777139
29. Momin M. Malik, Hemank Lamba, Constantine Nakos, and Jürgen Pfeffer. 2015. Population Bias in Geotagged Tweets. *ICWSM*.
30. James McCorriston, David Jurgens, and Derek Ruths. 2015. Organizations Are Users Too: Characterizing and Detecting the Presence of Organizations on Twitter. *ICWSM*.
31. Alan Mislove, Sune Lehmann, Yong-yeol Ahn, Jukka-pekka Onnela, and J Niels Rosenquist. Understanding the Demographics of Twitter Users. *ICWSM*: 554–557.
32. Lewis Mitchell, Morgan R Frank, Kameron Decker Harris, Peter Sheridan Dodds, and Christopher M Danforth. 2013. The geography of happiness: connecting twitter sentiment and expression, demographics, and objective characteristics of place. *PloS one* 8, 5. http://doi.org/10.1371/journal.pone.0064417
33. Mohamed Musthag and Deepak Ganesan. 2013. Labor dynamics in a mobile micro-task market. *CHI*, ACM, 641–650.
34. Mor Naaman, Amy Xian Zhang, Samuel Brody, and Gilad Lotan. 2012. On the Study of Diurnal Urban Routines on Twitter. *ICWSM*: 258–265.
35. Anastasios Noulas, Salvatore Scellato, Neal Lathia, and Cecilia Mascolo. 2012. Mining user mobility features for next place prediction in location-based services. *ICDM*, IEEE, 1038–1043.
36. Barbara Poblete, Ruth Garcia, Marcelo Mendoza, and Alejandro Jaimes. 2011. Do all birds tweet the same?: characterizing twitter around the world. *CIKM*, ACM, 1025–1030. http://doi.org/10.1145/2063576.2063724
37. Derek Ruths and Jürgen Pfeffer. 2014. Social media for large studies of behavior. *Science* 346, 6213: 1063–1064.

38. Takeshi Sakaki, Makoto Okazaki, and Yutaka Matsuo. 2010. Earthquake shakes Twitter users: real-time event detection by social sensors. *WWW*, ACM, 851–860.

39. Raz Schwartz and Germaine R Halegoua. 2014. The spatial self: Location-based identity performance on social media. *New Media & Society*: 1–18. http://doi.org/10.1177/1461444814531364

40. Shilad Sen, Toby Jia-Jun Li, WikiBrain Team, and Brent Hecht. 2014. WikiBrain: Democratizing Computation on Wikipedia. *OpenSym*, ACM, 27:1–27:10. http://doi.org/10.1145/2641580.2641615

41. Shilad W. Sen, Heather Ford, David R. Musicant, Mark Graham, Oliver S.B. Keyes, and Brent Hecht. 2015. Barriers to the Localness of Volunteered Geographic Information. *CHI*, ACM Press, 197–206. http://doi.org/10.1145/2702123.2702170

42. Alessio Signorini, Alberto Maria Segre, and Philip M. Polgreen. 2011. The Use of Twitter to Track Levels of Disease Activity and Public Concern in the U.S. during the Influenza A H1N1 Pandemic. *PLoS ONE* 6, 5: e19467. http://doi.org/10.1371/journal.pone.0019467

43. Anthony Stefanidis, Andrew Crooks, and Jacek Radzikowski. 2013. Harvesting ambient geospatial information from social media feeds. *GeoJournal* 78, 2: 319–338. http://doi.org/10.1007/s10708-011-9438-2

44. Andrea H. Tapia, Nicolas Lalone, Elizabeth MacDonald, Michelle Hall, Nathan Case, and Matt Heavner. 2014. AURORASAURUS: Citizen Science, Early Warning Systems and Space Weather. *HCOMP*.

45. Sho Tsugawa, Yusuke Kikuchi, Fumio Kishino, Kosuke Nakajima, Yuichi Itoh, and Hiroyuki Ohsaki. 2015. Recognizing Depression from Twitter Activity. *CHI*, ACM Press, 3187–3196. http://doi.org/10.1145/2702123.2702280

46. Andranik Tumasjan, Timm Oliver Sprenger, Philipp G. Sandner, and Isabell M. Welpe. 2010. Predicting Elections with Twitter: What 140 Characters Reveal about Political Sentiment. *ICWSM* 10: 178–185.

47. Alessandro Venerandi, Daniele Quercia, and Diego Saez-trumper. 2015. Measuring Urban Deprivation from User Generated Content. *CSCW*.

48. Ting-Yu Wang, F. Maxwell Harper, and Brent Hecht. 2014. Designing Better Location Fields in User Profiles. *Proceedings of the 18th International Conference on Supporting Group Work*, ACM, 73–80. http://doi.org/10.1145/2660398.2660424

49. Spencer a Wood, Anne D Guerry, Jessica M Silver, and Martin Lacayo. 2013. Using social media to quantify nature-based tourism and recreation. *Scientific reports* 3: 2976. http://doi.org/10.1038/srep02976

50. Amy X. Zhang and Scott Counts. 2015. Modeling Ideology and Predicting Policy Change with Social Media: Case of Same-Sex Marriage. *CHI*, ACM Press, 2603–2612. http://doi.org/10.1145/2702123.2702193

51. Yan-Tao Zheng, Zheng-Jun Zha, and Tat-Seng Chua. 2012. Mining Travel Patterns from Geotagged Photos. *ACM TIST* 3, 3: 1–18. http://doi.org/10.1145/2168752.2168770

52. Dennis Zielstra, Hartwig Hochmair, Pascal Neis, and Francesco Tonini. 2014. Areal Delineation of Home Regions from Contribution and Editing Patterns in OpenStreetMap. *ISPRS International Journal of Geo-Information* 3, 4: 1211–1233. http://doi.org/10.3390/ijgi3041211