A Tale of Cities: Urban Biases in Volunteered Geographic Information

Brent Hecht[1] and Monica Stephens[2]

[1]Department of Computer Science and Engineering, University of Minnesota, Minneapolis, MN, USA, [2]Department of Geography, University at Buffalo, State University of New York, Buffalo, NY, USA.

Abstract

Geotagged tweets, Foursquare check-ins and other forms of *volunteered geographic information* (VGI) play a critical role in numerous studies and a large range of intelligent technologies. We show that three of the most commonly used sources of VGI – Twitter, Flickr, and Foursquare – are biased towards urban perspectives at the expense of rural ones. Utilizing a geostatistics-based approach, we demonstrate that, on a *per capita basis*, these important VGI datasets have more users, more information, and higher quality information within metropolitan areas than outside of them. VGI is a subset of user-generated content (UGC) and we discuss how our results suggest that urban biases might exist in non-geographically referenced UGC as well. Finally, because Foursquare is exclusively made up of VGI, we argue that Foursquare (and possibly other location-based social networks) has fundamentally failed to appeal to rural populations.

Introduction

Researchers and practitioners have long known that people who live in rural areas tend to use technology differently than people who live in cities. These differences have been observed for over one hundred years and with technologies ranging from the early telephone (Kline 2002) to MySpace (Gilbert, Karahalios, and Sandvig 2010; Gilbert, Karahalios, and Sandvig 2008).

In the era of user-generated content (UGC), differences in how people use technology have implications far beyond the individual and her immediate network. A person who does not tweet or use Facebook loses more than just the chance to, for instance, broadcast information to her friends. She also effectively removes herself from numerous studies and omits her points-of-view from UGC-based technologies such as enterprise sentiment monitoring applications and other large-scale artificial intelligence systems. The same can be said of someone who does not upload photos to Flickr, check into Foursquare, or participate in other online communities.

This paper demonstrates that differences in technology use in rural and urban areas have led to a systemic bias against rural points of view in volunteered geographic information (VGI), the important subset of user-generated content that is geographically referenced (e.g. has lat/lon tags) (Goodchild 2007). We show that this is true in three separate large-scale sources of VGI – Twitter, Flickr, and Foursquare – each of which frequently appears in academic research and is used in many widely deployed technologies. In doing so, we provide evidence that suggests that at least some of the studies that have been conducted on the three online communities considered here have oversampled urban populations. Our evidence also suggests that the same is true of systems that use information from these communities.

Focusing on the United States' 59 million rural residents and 249 million urban residents (US Census Bureau 2013), this research further demonstrates that urban bias in VGI data can be quite extreme. For instance, in one analysis of a well-known dataset of Foursquare check-ins, we found that there are 24.4 times more Foursquare users *per capita* in urban areas than rural ones. A similar phenomenon exists in Twitter, where we found that there are 5.3 times as many tweets *per capita* in urban areas than rural ones.

In addition to examining basic properties of our corpora such as the number of Foursquare users and the number of tweets, we also analyze lower-level properties that play a key role in specific areas of the VGI literature. We look at, for instance, properties related to Twitter's social network (e.g. median number of followers) and the richness of metadata (e.g. tags per photo), finding urban biases in many cases.

Although this paper focuses on VGI, it likely also has implications for UGC as a whole. We discuss the ways in which this work might be generalized focusing in particular on Foursquare, the online community in which we found some of the strongest urban biases. Since VGI and UGC are synonymous in Foursquare, our work suggests that Foursquare has more or less failed in rural

areas. We argue that, just as with the telephone a century ago (Kline 2002), location-based social networks (and possibly other UGC communities as well) must be fundamentally adapted if they are to meet rural needs.

The structure of this paper is as follows: First, we highlight related work and discuss the datasets we analyze in this paper. We then highlight our methodology, which includes an approach for handling spatial autocorrelation in geospatial datasets, an important property of these datasets that is often ignored. Following the section on methodology, we present our results demonstrating that VGI is biased towards urban perspectives. We close with a discussion of the generalizability of our work to all of UGC, the limitations of this work, and summary of our contributions.

Related Work

Researchers and practitioners in the computing community have focused far more on urban areas than rural ones (Gilbert, Karahalios, and Sandvig 2010). For instance, within the social media space, researchers have inferred characteristics of urban environments with Foursquare check-ins (Cranshaw et al. 2012; Cranshaw and Yano 2010), studied diurnal urban routines using tweets (Naaman et al. 2012), and examined the coverage and growth of peer produced content in urban places (Mashhadi, Quattrone, and Capra 2013; Quattrone et al. 2014). More broadly, much work has been done developing augmented reality technology specifically for cities (e.g. Fischer and Hornecker 2012) and numerous studies have looked at technology use specifically in urban areas (e.g. Kumar and Rangaswamy 2013; Smyth et al. 2010). Outside of the research domain, there has been an equally strong emphasis on cities, e.g. IBM's Smarter Cities initiative (IBM 2013).

Rural areas have not, however, been completely ignored. For instance, Wyche and Murphy (2013) investigated the effectiveness of crank-based charging systems in rural Kenya and Kam et al. (2008) developed mobile games to improve English literacy in rural India. With regard to rural issues in North America, Collins and Wellman (2010) studied the use of technology in a remote Canadian community, finding that technology reduced the isolation of the community significantly.

The work that provided the initial inspiration for the study presented here is Gilbert et al.'s research on rural MySpace users (Gilbert, Karahalios, and Sandvig 2010; Gilbert, Karahalios, and Sandvig 2008). Gilbert and colleagues found that there were extensive differences between the rural and urban MySpace user populations, with rural users having, for instance, many fewer friends than urban users. In this work, we seek to understand the effect of these differences on the geographically referenced content produced by online communities (i.e. VGI), and do so through a multi-community lens. Our work also affords the opportunity to compare some of Gilbert et al.'s findings from MySpace in 2007 to other sites that belong to today's vastly changed social media landscape.

In research that also helped to motivate this paper, Mislove et al. (Mislove et al. 2011) found an association between Twitter adoption rates and the total population in a U.S. county. Similarly, Zielstra and Zipf (Zielstra and Zipf 2010) reported that OpenStreetMap coverage decreased as distance increased from 10 German cities and an analogous finding was identified in the London area by Mashhadi, Quattrone, and Capra (2013). These papers are not specifically interested in the rural/urban divide, which is significantly more complex than raw populations and distance from significant cities (e.g. see the census discussion below). We additionally build on this work in both breadth and depth by (1) taking a cross-site approach to the analysis of Foursquare, Twitter, and Flickr, which allows us to examine and reason about VGI holistically and (2) examining each VGI source at a level of detail greater than simple adoption rates and content coverage.

A number of studies have examined the relationship between demographics more broadly and participation rates in various online communities that are associated with volunteered geographic information. For instance, Hargittai and Litt identified that African-Americans and people with higher Internet skills are more likely to use Twitter (Hargittai and Litt 2011). Similarly, Stephens explored the large gender divide in contributions to OpenStreetMap (OSM) and examined its effects on OSM's content (Stephens 2013).

Lastly, one of the most important areas of related work is the significant body of research that either (1) leverages VGI as a way to study human behavior or (2) utilizes VGI as world knowledge for AI systems. Examples of the former include Cheng et al.'s use of Foursquare to understand human mobility patterns (2011), Hecht and Gergle's (2010) use of Wikipedia and Flickr to understand geospatial patterns in contributions to UGC communities, and many studies of language use in social media (e.g. Poblete et al. 2011; Hecht et al. 2011). As we describe in the discussion section, our work suggests that these studies were done on a sample biased heavily towards urban populations. In other words, the results below indicate that these are less studies of human behavior than studies of *urban* human behavior.

Research that leverages VGI as world knowledge for AI systems includes, for instance, work on identifying representative labels across geographic space from tags on georeferenced Flickr photos (e.g. Kennedy et al. 2007; Moxley, Kleban, and Manjunath 2008) and modeling lexical variation using tweets (e.g. Eisenstein et al. 2010;

Kinsella, Murdock, and O'Hare 2011). We argue below that our results suggest that these systems are gaining a perspective on the world that is biased towards the urban point-of-view.

Data

The datasets considered in this research fall into one of two categories: large repositories of VGI and statistics from government agencies.

Volunteered Geographic Information

Twitter: Not all tweets have geographic references. In fact, ignoring the problematic data in the location field of user profiles (Hecht et al. 2011), only around 1-3% of all tweets have latitude and longitude geotags (Morstatter et al. 2013; Broniatowski, Paul, and Dredze 2013). For these 1-3% of tweets, the geotag is generated automatically when a Twitter user has opted into this process. We analyzed a corpus of automatically geotagged tweets that we downloaded via Twitter's Streaming API over a 25-day period in August and September 2013. The corpus contains 56.7 million tweets from 1.6 million users. Recent work (Morstatter et al. 2013) has shown that, due our exclusive focus on geotagged tweets, this corpus is roughly identical to that which would have been generated using the "Firehose" API.

Flickr: The dataset of Flickr photos we collected consists of all geotagged photos uploaded to Flickr before November 2012 whose geotags indicate that they were taken in the United States. Geotagged photos typically come from two sources: photos that are automatically geotagged on mobile devices and photos from digital cameras that have been manually geotagged by their photographers. We filtered out photos with lower quality geotags by requiring that each geotag be accurate to approximately the city level (accuracy = 10 in Flickr's API). In total, our Flickr corpus contains 52.0 million photos from approximately 522,000 users.

Foursquare: Foursquare check-ins are not public by default, but can be shared widely if a user connects her/his account to Twitter. As such, following standard practice in the literature (e.g. Cheng et al. 2011; Cranshaw et al. 2012), we looked at "check-in tweets", or tweets that are automatically generated when a user who has connected her Twitter and Foursquare accounts checks into a location. The dataset of check-in tweets we analyzed is a subset of that used in Cheng et al. (2011) and contains 11.1 million check-ins from approximately 122,000 users.

The use of check-in tweets has important advantages and disadvantages in the context of this research. The primary advantage is that the use of check-in tweets makes our findings directly applicable to the existing Foursquare literature. This is an important benefit as one goal of our work is to demonstrate that many existing VGI studies and VGI-based systems have a built-in urban sampling bias. The primarily disadvantage, a disadvantage that is inherent to all studies of check-in-tweets, is that Foursquare users that do not use the automatic tweet feature are excluded.

Data from Federal Agencies

We utilized data from several agencies of the United States federal government in this research. This data primarily consists of the polygonal outlines for all U.S. counties (and equivalents, e.g. "boroughs" in Alaska), population data for these counties, and the specific urban-rural properties of these counties. The county geometries and population data we employed were provided by the United States Census Bureau. The Census Bureau is also the source of the key indicator by which we determine the "urbanness" of each county: the percent of the population in the county that lives in an urban area (*PCT_POPURB)*. This percentage, which was updated in 2010 for all counties, is calculated using the Census Bureau's detailed definition of an urban area (US Census Bureau 2013), which includes both big cities and towns of population 2,500 or more. For example, *PCT_POPURB* for San Francisco County is 100.0 and it is 0.0 for Bristol Bay Borough (Alaska).

Where a discrete classification of each county along the rural-urban spectrum was needed, we utilized the National Center for Health Statistics' (NCHS) urban-rural classification scheme for counties (Ingram and Franco 2012). This classification scheme places every U.S. county on a discrete scale from 1 ("large central metro") to 6 ("noncore", or not part of any metropolitan or micropolitan statistical area). These codes can roughly be interpreted as ranging from core urban counties (1) to entirely rural counties (6).

Methodology

At a high-level, our methodological approach consisted of two steps. First, the number of tweets, check-ins, and other VGI attributes in each U.S. county was counted or otherwise summarized (e.g. medians were calculated). Where appropriate, these summary statistics were normalized by the population of each county. Second, the spatial distributions of these county-level summary statistics were compared with the spatial distribution of the urban/rural population ratio (i.e. *PCT_POPURB)*. As noted above, the set of attributes considered for each VGI repository consists of attributes that are utilized in VGI-based research and systems.

Executing this two-step approach involved addressing key challenges related to (1) properly "locating" a unit of

VGI in the context of this study and (2) accounting for *spatial autocorrelation*, a property of geospatial information that can lead to incorrect conclusions about an effect's significance. Each of these challenges is addressed in turn below.

Because our research is interested in the *perspectives* of urban and rural users – not simply the location of users at the time they contribute VGI – assigning each unit of VGI to a county was a non-trivial process. Consider for example a tourist from San Francisco who spends five days in Yosemite National Park, tweeting about her vacation at a high rate (with geotagging enabled). If these tweets – and the many others like them – were naively assigned to one of the very rural counties that contain Yosemite, we might falsely conclude that rural *perspectives* are over-represented on Twitter on a per-capita basis, when in fact we were misinterpreting tweets from urban tourists as tweets from locals of rural counties.

The challenge of accounting for contributor mobility (e.g. vacations to Yosemite) has been recognized by a number of researchers executing related studies and has been addressed in several ways. First, many studies have utilized information entered into the "location field" in users' online profiles. However, this approach was problematized by Hecht et al. (2011), who found that a non-trivial percentage of the information in location fields is not of a geographic nature and that location field information is of limited granularity. Another approach that has been employed to address this issue is to require that a user submit VGI at least n days apart in a given region (e.g. county) before considering the user to be local to the region (Li, Goodchild, and Xu 2013; Popescu and Grefenstette 2010). Finally, other researchers have utilized a "plurality rules" approach in which all of a user's contributions are associated with the single region in which they were the most active (e.g. Musthag and Ganesan 2013).

Neither the temporal nor the plurality approach has emerged as a best practice, and each approach has its limitations. The temporal approach is biased towards power users, as even a user that does not travel at all must contribute at least two units of VGI during the study period to be counted as a local in her home region (and these units must appear in the sample). On the other hand, the plurality approach does not account for users who are indeed locals to two or more regions (e.g. a native of a rural county who has moved to an urban area but commutes relatively frequently). Given that each approach has benefits and drawbacks, we performed all analyses and report all results using both approaches. Below, we label the temporal approach *n-days* and, following Li et al. (2013), set $n = 10$. We label the plurality rules approach *plurality*.

The second major challenge in our methodological approach involved addressing the spatial autocorrelation in our VGI and percent urban population spatial distributions. Spatial autocorrelation describes the tendency for measurements nearby each other in space to be correlated and is a general property of "variables observed across geographic space" (Legendre 1993). Broadly speaking, it is the necessary quantitative result of Tobler's First Law of Geography, which states "everything is related to everything else, but near things are more related than distant things" (Tobler 1970). Although significant autocorrelation does not occur with all spatial phenomena, when nearby measurements in spatial data are indeed correlated, the measurements cannot be considered independent. This violates the independence assumptions of many statistical tests.

One straightforward approach to understanding the strength and direction of the relationships between the "urbanness" of counties (i.e. *PCT_POPURB*) and the properties of the VGI they "contain" (e.g. median number of Twitter followers in each county) is to use correlation measures. However, spatial autocorrelation violates the assumption of the independence of observations of well-known correlation measures, such as Pearson and Spearman's correlation coefficients. To address this issue, we leverage a method used in the natural sciences (e.g. Grenyer et al. 2006) and introduced by Clifford et al. (1989). The Clifford et al. approach involves calculating a (reduced) effective sample size to address the "redundant, or duplicated, information contained in georeferenced data" (Griffith and Paelinck 2011) that is the result of spatial autocorrelation. The significance of the correlation coefficients shown below is calculated using the Clifford et al. method and its notion of effective sample size, thereby accounting for the autocorrelation in our data. We note that nearly every correlation below is significant at $p < 0.001$ using traditional approaches, but this is not true after applying the Clifford et al. correction.

Finally, in addition to exhibiting spatial autocorrelation, the nature of many of our VGI-derived attributes required that we use a correlation coefficient that does not make any distribution assumptions and that is robust against outliers. As such, we describe the relationships between rurality and the VGI-derived attributes using Spearman's rank correlation coefficient (calculated using the Clifford et al. approach). Because we execute a number of significance tests, statistical significance is reported using Bonferroni correction.

Results

Twitter

We begin the discussion of our results by focusing on our Twitter corpus. Table 1 describes the Spearman's

correlations between the percent urban population (*PCT_POPURB)* in a county and properties of the Twitter VGI assigned to each county according to the procedures described above. For instance, the first row of Table 1 contains the result of calculating the Spearman's correlation coefficient between the Twitter users per capita in each county and that county's *PCT_POPURB*. For all results in this section, we only include in tables correlations with population-normalized attributes and correlations with attributes for which this is not necessary (e.g. medians).

Immediately visible in Table 1 is the series of positive correlation coefficients for critical properties like users per capita, tweets per capita, and so on. The story here is clear: the more urban the county, the more Twitter activity it has *per person*. For instance, let us return to the "Users per Capita" row (first row). Here we see that for both methods of assigning VGI to counties there is a strong positive correlation between the number of Twitter users per capita in a county and the percent of the population in the county that lives in an urban area. In other words, the more urban a county, the higher the percentage of people in that county that are Twitter users (and geotag their tweets). Additional insight on the relationship between users per capita and "urbanness" can be obtained through the use of the NCHS urban/rural county classifications discussed above. "Core urban" counties (NCHS category 1) have *3.5 times* more Twitter users per capita than "entirely rural" counties (NCHS category 6) using *plurality* and 2.7 times more using *n-days*.

Property	n-days	plurality
Users per Capita	0.46***	0.54***
Number of Total Tweets per Capita	n/a	0.53***
Sample Period Tweets per Capita	0.49***	0.50***
Median Total Tweets	n/a	0.28***
@ Mentions per Tweet	0.19***	0.21***
URLs per Tweet	0.10**	0.12**
Median # of Followers	-0.14**	0.11†
Median # of Users Followed	0.05 (n.s.)	0.06 (n.s.)
Mean Length of Tweets	-0.15**	-0.14**
Hashtags per Tweet	-0.16**	-0.07 (n.s.)

Table 1: Attributes of Twitter VGI and their correlation with the percent of a population that lives in a rural area. Significance is calculated using the Clifford et al. "effective sample size" method that controls for spatial autocorrelation in spatial datasets. † (marginally) significant at p < .10; * significant at p < .05; ** significant at p < .01 *** significant at p < .001 (with Bonferroni correction)

Turning our attention from users to content, we assessed the quantity of tweets in a given region in two ways: (1) we counted the number of tweets in our sample that were assigned to a county ("Sample Period Tweets") and (2) we summed together the total number of tweets posted by all users assigned to a county in the entire history of their Twitter accounts ("Total Tweets"). The total number of tweets posted by a user is available through the Twitter API. We only looked at "Total Tweets" using the *plurality* method because in the *n-days* method, users can be considered local to multiple counties and one cannot infer what percent of their total tweets came from each county.

Table 1 shows that in every case, as *PCT_POPURB* goes up, so does tweets per capita. This is true for "Sample Period" tweets using both *n-days* and *plurality* as well as "Total Tweets" using *plurality*. Indeed, applying the NCHS classifications in the same fashion as above, we find for instance that the number of total tweets per capita in core urban counties is 5.3 times higher than in entirely rural counties.

Table 1 also reveals interesting biases and lack of (significant) biases in the social network properties of Twitter. For instance, the table shows that "@" mentions have a moderate positive correlation with urban population percentage. "@" mentions have been identified as an important means of information diffusion on Twitter (Yang and Counts 2010) and our results suggest that rural areas participate less in this process.

The low (and often non-significant) correlations between *PCT_POPURB* and "Median # of Followers" and "Median # of Accounts Followed" are quite interesting in light of the work of Gilbert et al. (2010; 2008). Working in 2007, Gilbert and colleagues found that the median number of friends for rural MySpace users was less than half that of urban users. On the other hand, Table 1 reveals little to no relationship between the "ruralness" of a county and the median number of Twitter friends and followers in that county. The extent to which rural users have caught up with urban users in terms of articulated connectivity or whether this is an effect of online community type is an important open question.

Flickr

As was the case with our Twitter results, our Flickr results (Table 2) reveal a bias towards urban areas. For instance, as the proportion of the population that lives in urban areas goes up, so does photos per Flickr user and photos per capita. Returning to the NCHS classifications and using *plurality*, there are 2.0 times more photos per capita in core urban areas than entirely rural areas and the median number of photos per user is also 2.0 times higher. The equivalent ratios for *n-days* are 2.4 and 2.2, respectively.

Property	n-days	cluster
Median Number of Photos Per User	0.41***	0.38***
Tags per Photo	0.11***	0.26***
Photos per Capita	0.20***	0.26***
Users per Capita	-0.05 (n.s.)	0.10 (n.s.)

Table 2: Spearman's correlations between the percent urban population in a county and properties of our Flickr data assigned to that county.

The moderate positive correlation coefficient between tags per photo and *PCT_POPURB* is interesting at a lower level. Flickr tags have been used in a large variety of contexts ranging from identifying representative labels for a given region (e.g. Kennedy et al. 2007; Moxley, Kleban, and Manjunath 2008) to mining vernacular (i.e. "colloquial") regions (Thomee and Rae 2013) to georeferencing photographs without geotags (Crandall et al. 2009). Our results suggest that there may be a risk of undersampling rural points of view in these endeavors, which could lead to the omission of rural opinion about regions and decreased accuracy in rural areas.

There is one exception to the overall trend towards positive associations in Table 2: users per capita. The correlation with the number of Flickr users per capita is non-significant for both *plurality* and *n-days*. Looking at the raw data for each county, a possible explanation quickly emerged: although the methods we used to assign users to regions were effective, they were not perfect. For instance, nine of the ten counties with the most Flickr users per capita using both *plurality* and *n-days* contain well-known national parks or related natural attractions (e.g. Denali County, AK with Denali National Park; Grand County, UT with Arches National Park). In other words, a "tourist signal" remains in our Flickr corpus, even after applying *n-days* and *plurality*. However, when employing the naïve approach of using simple geotags rather than *plurality* or *n-days*, this effect gets even more extreme. San Juan County, Colorado – a very low-population county with a scenic highway – has 1.19 Flickr users per capita using this method.

It is important to note that the "tourist signal" should not be considered pure noise. There are interesting research questions to be asked with regard to the dominance of outsider perspectives vs. local perspectives in VGI about rural areas that have a very high tourist/local ratio. While the "localness" of UGC (using VGI as a proxy) has been studied in general (Hecht and Gergle 2010), future work is needed examining these issues specifically in rural areas.

Attribute	n-days	cluster
Check-Ins Per Capita	0.61***	0.63***
Foursquare Users per Capita	0.51***	0.61***
Median Number of Check-Ins Per User	0.51***	0.43***

Table 3: Spearman's correlations coefficients between the percent urban population in a county and properties of the Foursquare data assigned to each county.

Foursquare

Our Foursquare results tell a very consistent story of strong urban bias. The positive correlation coefficients between the percent of a county's population that lives in an urban area and all three Foursquare attributes we considered can be seen in Table 3. As the percent urban population goes up, so does the number of Foursquare users per capita, the number of check-ins per capita, and the median number of check-ins per user.

An analysis of our Foursquare data using the NCHS schema reveals similar biases. Using *plurality,* core urban counties have *24.4 times* the number of Foursquare users per capita as entirely rural counties and *23.1 times* the number of check-ins per capita. The equivalent ratios for *n-days* are 8.7 and 18.4, respectively.

While a ratio of 8.7 still represents extensive bias, the discrepancy between *n-days* and *plurality* in this case is interesting. It is by far the largest such discrepancy we observed (with the check-ins per capita discrepancy being second). One possibility is that Foursquare power users are more likely to submit VGI (check in) when they travel, including when they travel to rural areas. If they travel to/through the same county more than once (and do so more than 10 days apart), they would be treated as "locals" to these areas according to *n-days*. Our future work involves comparing *n-days*, *plurality*, and location fields in detail and for a number of different applications, and these discrepancies will be a subject of study in this research.

Discussion

This paper has articulated numerous ways in which commonly used corpora of volunteered geographic information are biased towards urban perspectives. Below, we discuss the implications of these findings for (1) user-generated content as a whole and (2) research that utilizes VGI as either trace data or as input to AI systems.

Implications for User-generated Content

As noted above, VGI is defined as the subset of user-generated content (UGC) with a geographic reference

(Goodchild 2007). While using VGI as a proxy for all of user-generated content is a relatively common practice in the computer science and geography literatures (e.g. Graham and Zook 2013; Hecht and Gergle 2009; Stephens 2013), we have taken a cautious approach here and framed our concrete contributions around VGI specifically rather than UGC as a whole.

There are, however, several reasons to believe that our results above will apply more generally. First, while location tagging in social media has seen uneven adoption among different populations (Zickuhr and Smith 2011), a growing body evidence suggests that location tagging has become much more of a mainstream activity (Zickuhr 2013). Second, there are several data points above that point to a bias towards urban perspectives in non-geotagged UGC. Primarily, in the analysis of our Twitter corpus, we looked at the total number of tweets– geotagged and non-geotagged – posted by users that appeared in our corpus ("Total Tweets") and found that the aggregate and median number of total tweets were highly correlated with *PCT_POPURB* (as was also the case with tweets that actually appeared in the corpus, or "Sample Period Tweets"). While a user had to geotag at least one tweet to appear in our corpus, this gives us some hint as to what may be occurring in Twitter more broadly.

Finally, due to its nature as a location-based social network (LBSN), UGC and VGI are one and the same in Foursquare. This means that our results above shine a complete light on the urban and rural biases in that community (at least with respect to the properties we examined). Since these results show Foursquare activity *per capita* in rural areas to be 4% (*plurality*) - 11% (*n-days*) of what it is in urban areas, they raise the question of how Foursquare might better appeal to rural communities. Is the POI/check-in model not well suited to rural areas, where the number of commercial "venues" accessible to a potential Foursquare user is small? If not, could "continuous", non-venue-based models of location sharing – e.g. the approach taken in Google Latitude – be more successful? Another possibility is that Foursquare's "meritocratic" "mayor"-based incentive system is not compatible with the social structure of rural areas. If this were the case, there would be precedent: the telephone clashed with rural social structure when it was first introduced (Kline 2002). Lastly, another possibility is that Foursquare's role in the search/discovery process is significantly reduced in rural areas due to the limited number of venue options. If this has an effect on participation rates, it would shed new light on the discussion related to the ecosystem of purposes for location sharing in LBSNs (e.g. Lindqvist et al. 2011).

Implications for VGI-based Research

Our work has important implications for both research that uses VGI as trace data to study human behavior and research that uses VGI as world knowledge for AI systems. With regard to the former, consider studies like Cheng et al. (2011), which used Foursquare data to study general human mobility. While this work provided a number of important insights – e.g. that consecutive check-ins follow a Lévy Flight pattern – our work helps to put these insights in the context of the underlying dataset. That is, Foursquare affords the study of *urban* mobility patterns, not mobility patterns in general (or rural mobility patterns specifically). Similarly, Hecht and Gergle's (2010) results regarding the "localness" of user-generated content should likely be re-examined looking at urban and rural areas separately to avoid urban bias in the overall VGI sample, an approach that can also likely be applied in other VGI studies. Finally, our work also suggests that VGI-based methods that have been demonstrated to be highly effective in urban areas (e.g. Cranshaw et al. 2012) may have to be significantly altered in order to be effective in rural areas, especially methods relying on Foursquare data.

The results above also indicate that when world knowledge for AI systems is derived from Twitter, Foursquare, and Flickr, these systems likely become biased towards urban perspectives. For example, consider a system like those in recent work (e.g. Eisenstein et al. 2010) that can build a geographic topic model for all first-order administrative districts (e.g. states, polygons). Our results suggest that these topic models will disproportionately consist of topics discussed by urbanites in each district. If we assume that our results generalize to all of UGC, this issue becomes exponentially larger, with all topic models built using tweets, Flickr tags, and so on being biased in the same way. One way to address this issue – at least for systems that only use VGI – is to do stratified sampling across urban and rural counties.

Limitations and Future Work

While we have discussed a number of limitations to our analyses above (e.g. related to the nature of our Foursquare dataset), there are several additional issues worthy of discussion. First and foremost, this paper focused on the urban/rural divide in a single country. While differences in how rural and urban people use technology generally speaking exist around the world (e.g. Kam et al. 2008; Wyche and Murphy 2013), their manifestation in volunteered geographic information may vary geographically.

This research examined biases in volunteered geographic information with an urban/rural lens, a lens that has ample precedent in the computing literature (e.g.

Gilbert et al. 2008; Gilbert et al. 2010; Wyche and Murphy 2013; Quattrone et al. 2014; Collins and Wellman 2010) and that has immense value to researchers and practitioners in a wide range of contexts. Entire journals in the social sciences are dedicated specifically to rural issues (e.g. the *Journal of Rural Studies)*, and the same is true for numerous government agencies and non-profits (e.g. USDA Rural Development). Our work has immediate value for these constituencies by providing insight on rural technology adoption rates and the extent to which these rates are reducing the volume of rural voices. That said, our work also suggests that it may be useful employ other lenses in similar studies of bias in VGI and UGC more generally. For instance, rural populations tend to be older and poorer (e.g. Glasgow, Berry and Oh 2013), indicating that a study identical to this one but targeted specifically at these demographic variables may reveal interesting trends that would supplement existing, "smaller *n*" work on the relationship between age, income and technology use in the United States (e.g. Zickuhr 2013; Zickhur and Smith 2011).

Following the literature, we have conceptualized the notions of rural and urban as binary. While useful at a broad level, these categories are more naturalistically conceived along a continuum. In particular, considering the notions of suburbs may reveal additional insights. Examining counties that are "fringe counties" of large metropolitan areas according to the NCHS schema (category 2) revealed interesting trends. For instance, the median number of Flickr photos per user in category 2 counties was 42.0% higher than that for core urban counties. On the other hand, other VGI attributes more smoothly decrease from urban core to suburb to exurb to rural area. Focusing on the urban/suburban divide is a topic of future work.

Conclusion

In this paper, we have provided evidence that volunteered geographic information (VGI) tends to be biased towards urban perspectives and away from rural ones. We demonstrated that this bias exists in terms of adoption rates (users per capita), quantity of information (content per capita), and quality of information (e.g. tags per photo) and across three separate commonly used sources of VGI: Twitter, Flickr and Foursquare. Our work suggests that researchers and practitioners utilizing VGI (and perhaps user-generated content more generally) should take care to account for urban biases in their datasets, especially when leveraging VGI as a way to understand broader human phenomena or as a form of world knowledge for AI systems. Moving forward, our work provides strong motivation for the development of VGI technologies better suited to a rural audience. In other words, we have established the problem; now it is time to try to fix it.

Acknowledgements

We would like to thank James Caverlee (and the Texas A&M InfoLab), Brian Keegan, Shilad Sen, Chip Weinberger, and Ate Poorhuis for their assistance on this project and our anonymous reviewers for their invaluable feedback.

References

Broniatowski, David A., Michael J. Paul, and Mark Dredze. 2013. "National and Local Influenza Surveillance through Twitter: An Analysis of the 2012-2013 Influenza Epidemic." *PLoS ONE*.

Cheng, Zhiyuan, James Caverlee, Kyumin Lee, and Daniel Z Sui. 2011. "Exploring Millions of Footprints in Location Sharing Services." In *ICWSM '11*.

Clifford, Peter, Sylvia Richardson, and Denis Hemon. 1989. "Assessing the Significance of the Correlation between Two Spatial Processes." *Biometrics* 45 (1).

Collins, Jessica L., and Barry Wellman. 2010. "Small Town in the Internet Society: Chapleau Is No Longer an Island." *American Behavioral Scientist* 53 (9): 1344–66.

Crandall, David J., Lars Backstrom, Daniel Huttenlocher, and Jon Kleinberg. 2009. "Mapping the World's Photos." In *WWW '09*.

Cranshaw, Justin, Raz Schwartz, Jason I Hong, and Norman M Sadeh. 2012. "The Livehoods Project: Utilizing Social Media to Understand the Dynamics of a City." In *ICWSM '12*.

Cranshaw, Justin, and T. Yano. 2010. "Seeing a Home Away from the Home: Distilling Proto-Neighborhoods from Incidental Data with Latent Topic Modeling." In *NIPS Workshop on Computational Social Science and the Wisdom of Crowds*.

Eisenstein, Jacob, Brendan O'Connor, Noah A. Smith, and Xing, Eric P. 2010. "A Latent Variable Model for Geographic Lexical Variation." In *EMNLP '10*.

Fischer, Patrick Tobias, and Eva Hornecker. 2012. "Urban HCI: Spatial Aspects in the Design of Shared Encounters for Media Facades." In *CHI '12*.

Gilbert, Eric, Karrie Karahalios, and Christian Sandvig. 2008. "The Network in the Garden: An Empirical Analysis of Social Media in Rural Life." *CHI '08*.

Gilbert, Eric, Karrie Karahalios, and Christian Sandvig. 2010. "The Network in the Garden: Designing Social Media for Rural Life." *American Behavioral Scientist* 53 (9).

Glasgow, N., Berry, E.H., and Oh, E.J.V. Rural aging in 21st century America. Springer, Dordrecht; New York, 2013.

Goodchild, Michael F. 2007. "Citizens as Sensors: The World of Volunteered Geography." *GeoJournal* 69 (4).

Graham, Mark, and Matthew Zook. 2013. "Augmented Realities and Uneven Geographies: Exploring the Geolinguistic Contours of the Web." *Environment and Planning A* 45 (1).

Grenyer, Richard, C. David L. Orme, Sarah F. Jackson, Gavin H. Thomas, Richard G. Davies, T. Jonathan Davies, Kate E. Jones,

et al. 2006. “Global Distribution and Conservation of Rare and Threatened Vertebrates.” *Nature* 444 (7115).

Griffith, Daniel A, and Jean HP Paelinck. 2011. “Understanding Correlations Among Spatial Processes.” In *Non-Standard Spatial Statistics and Spatial Econometrics*, 75–95. Springer.

Hargittai, Eszter, and Eden Litt. 2011. “The Tweet Smell of Celebrity Success: Explaining Variation in Twitter Adoption among a Diverse Group of Young Adults.” *New Media & Society*.

Hecht, Brent, and Darren Gergle. 2009. “Measuring Self-Focus Bias in Community-Maintained Knowledge Repositories.” In *Communities and Technologies 2009*.

Hecht, Brent, and Darren Gergle. 2010. “On The ‘Localness’ of User-Generated Content.” In *CSCW ’10*.

Hecht, Brent, Lichan Hong, Bongwon Suh, and Ed H. Chi. 2011. “Tweets from Justin Bieber’s Heart: The Dynamics of the ‘Location’ Field in User Profiles.” In *CHI ’11*.

IBM. 2013. “Smarter Cities: Building and Carrying out Ways for a City to Realize Its Full Potential.” *IBM – Smarter Planet*. http://www.ibm.com/smarterplanet/us/en/smarter_cities/overview/index html.

Ingram, DD, and SF Franco. 2012. “Rural Classification Scheme for Counties.” Vital Health Stat. National Center for Health Statistics. http://www.cdc.gov/nchs/data_access/urban_rural htm#documentation.

Kam, Matthew, Aishvarya Agarwal, Anuj Kumar, Siddhartha Lal, Akhil Mathur, Anuj Tewari, and John Canny. 2008. “Designing E-Learning Games for Rural Children in India: A Format for Balancing Learning with Fun.” In *DIS’ 08*.

Kennedy, Lyndon, Mor Naaman, Shane Ahern, Rahul Nair, and Tye Rattenbury. 2007. “How Flickr Helps Us Make Sense of the World: Context and Content in Community-Contributed Media Collections.” In *ACM MM’ 08*.

Kinsella, Sheila, Vanessa Murdock, and Neil O’Hare. 2011. “‘I’m Eating a Sandwich in Glasgow’: Modeling Locations with Tweets.” In *Proceedings of the 3rd International Workshop on Search and Mining User-Generated Contents*.

Kline, Ronald R. 2002. *Consumers in the Country: Technology and Social Change in Rural America.* Baltimore: Johns Hopkins Univ Press.

Kumar, Neha, and Nimmi Rangaswamy. 2013. “The Mobile Media Actor-Network in Urban India.” In *CHI ‘13*.

Legendre, Pierre. 1993. “Spatial Autocorrelation: Trouble or New Paradigm?” *Ecology* 74 (6).

Li, Linna, 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.

Lindqvist, Janne, Justin Cranshaw, Jason Wiese, Jason Hong, and John Zimmerman. 2011. “I’m the Mayor of My House: Examining Why People Use Foursquare - a Social-Driven Location Sharing Application.” In *CHI ’11*.

Mashhadi, Afra, Quattrone, G., and Capra, L. 2013. “Putting Ubiquitous Crowd-Sourcing into Context.” In *CSCW ’13*.

Mislove, Alan, Sune Lehmann, Yong-Yeol Ahn, Jukka-Pekka Onnela, and J. Niels Rosenquist. 2011. “Understanding the Demographics of Twitter Users.” In *ICWSM ’11*.

Morstatter, Fred, Jürgen Pfeffer, Huan Liu, and Kathleen M. Carley. 2013. “Is the Sample Good Enough? Comparing Data from Twitter’s Streaming API with Twitter’s Firehose.” In *ICWSM ’13*.

Moxley, Emily, Jim Kleban, and B. S. Manjunath. 2008. “Spirittagger: A Geo-Aware Tag Suggestion Tool Mined from Flickr.” In *MIR ’08*.

Musthag, Mohamed, and Deepak Ganesan. 2013. “Labor Dynamics in a Mobile Micro-Task Market.” In *CHI ’13*.

Naaman, Mor, Amy Xian Zhang, Samuel Brody, and Gilad Lotan. 2012. “On the Study of Diurnal Urban Routines on Twitter.” In *ICWSM ’12*.

Poblete, Barbara, Ruth Garcia, Marcelo Mendoza, and Alejandro Jaimes. 2011. “Do All Birds Tweet the Same?: Characterizing Twitter around the World.” In *CIKM ’11*.

Popescu, Adrian, and Gregory Grefenstette. 2010. “Mining User Home Location and Gender from Flickr Tags.” In *ICSWM ’10*.

Quattrone, Giovanni, Afra Mashhadi, Daniele Quercia, C. Smith, and Licia Capra. 2014. “Modelling Growth of Urban Crowd-Sourced Information.” In *WSDM ‘14*.

Smyth, Thomas N., Satish Kumar, Indrani Medhi, and Kentaro Toyama. 2010. “Where There’s a Will There’s a Way: Mobile Media Sharing in Urban India.” In *CHI ’10*.

Stephens, Monica. 2013. “Gender and the GeoWeb: Divisions in the Production of User-Generated Cartographic Information.” *GeoJournal*: 1–16.

Thomee, Bart, and Adam Rae. 2013. “Uncovering Locally Characterizing Regions within Geotagged Data.” In *WWW ’13*.

Tobler, Waldo R. 1970. “A Computer Movie Simulating Urban Growth in the Detroit Region.” *Economic Geography* 46: 234–40.

US Census Bureau. 2013. “2010 Census Urban and Rural Classification and Urban Area Criteria.” http://www.census.gov/geo/reference/ua/urban-rural-2010 html.

Wyche, Susan P., and Laura L. Murphy. 2013. “Powering the Cellphone Revolution: Findings from Mobile Phone Charging Trials in off-Grid Kenya.” In *CHI ’13*.

Yang, Jiang, and Scott Counts. 2010. “Predicting the Speed, Scale, and Range of Information Diffusion in Twitter.” In *ICWSM ’10*.

Zickuhr, Kathryn. 2013. “Location Tagging Among Social Media Users”. Pew Internet & American Life Project.

Zickuhr, Kathryn, and Aaron Smith. 2011. “28% of American Adults Use Mobile and Social Location-Based Services”. Pew Internet & American Life Project.

Zielstra, Dennis, and Alexander Zipf. 2010. “A Comparative Study of Proprietary Geodata and Volunteered Geographic Information for Germany.” In *AGILE ’10*.