Geosocial capta in geographical research - a critical analysis

Michał Rzeszewski

Faculty of Geographical and Geological Sciences, Adam Mickiewicz University, Poznan, Poland
Dziegielowa 21, 61-680 Poznan, Poland, phone: +48 61 8296171, e-mail: mrz@amu.edu.pl

Abstract: This paper presents a critical view of the use of geodata/capta from social media sources as a research tool in geographic research. It compares three captasets from Instagram, Fickr and Twitter, based upon spatial and descriptive statistics. Observed discrepancies in the different distribution of values are found to be related to differences in the spatial practices of their users and the modes of production of the ambient geosocial information. The results indicate that the interpretation of geosocial media capta can must consider underlying social processes, but such linkages are currently poorly understood. Therefore, caution should be exercised when aggregating capta from more than one social media platform. Geotagged content can represent various interactions and intentions of its creators. At the same time, the observed differences can give more insights into relationships between material space and the production of digital realities.

Keywords: data, capta, geosocial media, critical GIS
Cyberscapes of social media in geographic research

With Web 2.0, social media became an important part of everyday life for a large number of people, providing new forms of communication for leisure and work. Over time, this led to a gigantic amount of data being created every single day. Researchers from various disciplines suddenly found themselves submerged in a flood (or rather an “exaflood”) of “big data” with volume, velocity and variety never experienced before (Kitchin, 2013). Big data energized fields of social sciences that had been relatively data poor, and now are facing new challenges as they change their methodological paradigms (Gonzalez-Bailon, 2013; Ruppert, 2013; Housley et al., 2014). This also challenges areas of geography and GISciences, especially since at least a small part of the modern data stream is not only implicitly connected to material, real world places, but is also connected explicitly by virtue of some form of geotag - additional information about a geographical location embedded into mobile digital content.

Even small collections of the exaflood are sufficient to provide new insights into various fields of research, and some in the field have even called it a renaissance of geographic information (Hudson-Smith et al., 2009). To date, this relatively new phenomenon has been described and theorized in a number of disciplinary-dependent ways. No single term exists for it that has been widely accepted and adopted. When it first started to be studied, two terms—“neogeography” (Turner, 2006; Wilson and Graham, 2013) and “Volunteered Geographic Information” (VGI) (Goodchild, 2007; Elwood, 2011; Elwood, 2008; Goodchild, 2008)—were coined to name a new kind of internet activity that brought together enthusiasts without cartographic training, to undertake initiatives like Open Street Map, YouMap, Wikimapia, etc. Next, the term “spatial media” was proposed by Crampton (2009) in reference to spatio-technical presences (location-based services and interfaces) that encouraged the production of geographic information. Equally popular among geographers is the concept of the “Geoweb” (Scharl and Tochtermann, 2007; Haklay, Singleton and Parker, 2008; Elwood and Leszczynski, 2011) that accounts for the new materialities, as well as the new spatial practices of Web 2.0 and mobile communications.

However, new social media sites tend to provide geographic information to researchers in ways that are often opaque to casual users. In their daily routines, they generate spatial footprints that can be used in various ways, some of which are very distant from the original intent of their creators - and thus the term “volunteered ” seems to be no longer applicable (Poorthuis et al., 2014). Harvey (2013) proposed that volunteered data must be collected under “opt-in” provisions while “opt-out” provisions commonly lead to the creation of contributed geographic information (CGI). A similar observation led
Stefanidis, Crooks and Radzikowski to the conclusion that geosocial media is not a source of geographic information per se, but nevertheless encodes geographic messages in the form of ambient geospatial information (AGI) (Stefanidis, Crooks and Radzikowski, 2013). Equally important are discussions about the relationship between virtual and material realities. Early theoretical models used the “cyberspace” meme that was coined by Gibson (1984) and was assumed to be a separate, aspatial entity. However, it was soon realized, especially by geographers, that there is a strong, dynamic interplay between the virtual and the material, and the term cyberspace is often wrongly used as a deterministic metaphor for technological change that obscures the importance of spatial interactions between ICT (Information and Communication Technologies) and society (Graham 1998, Graham 2013).

Space is often automatically produced (Thrift and French, 2002) and our geographies are software-sorted– influenced by technological systems embedded within modern cities (Graham, 2005). Dodge and Kitchin (2005, 2011) further proposed that code plays a vastly influential role in shaping our spatiality. They differentiated between “coded spaces” and “code/space”, with the latter being entirely dependent on software and computer algorithms. Other authors have further confronted this phenomenon by proposing the notions of “DigiPlace” - composed of virtual (named cyberscapes) and material layers of content (Zook and Graham, 2007) or augmented reality, described as a “material/virtual nexus mediated through technology, information and code, and enacted in specific and individualized space/time configurations” (Graham et al., 2013).

It has become increasingly clear that the virtual and material separation is entirely artificial, as is highlighted in spatial media/tion theory (Leszczynski, 2014). It is therefore evident that digital representations in social media have the power to alter the meaning and perceived fabric of material environments, through visualization and naming. In this way, the virtual has potential to be even more powerful than physical reality (Zook and Graham, 2007). In Harvey’s (1989) terms - the spaces of representation can change spatial practices. Discussions about the intersection of the material and the real remain vigorous and the increasing prevalence of mobile, “smart” and location-aware ICT blurring virtual/material distinctions as never before.

There have been numerous attempts to utilize this exaflood of social media data to explain and understand real world problems, as well as to advance various theories. Geosocial media allows researchers to:

- Delineate city cores (Hollenstein and Purves, 2010)
Gain insights into travel plans and tourism (Xiang and Gretzel, 2010)

Extract crowd behavioral patterns in urban environments (Lee et al., 2013)

Characterize urban landscapes (Frias-Martinez et al., 2012), and

Study global migrations (Hawelka et al., 2014).

One of the more spectacular examples is the way in which social media can be used to enhance responses to natural disasters (Crutcher and Zook, 2009; M. Zook et al., 2010; Vieweg et al., 2010; Crooks et al., 2013; Shelton et al., 2014). The wealth of data has also sparked numerous imaginative ways of reading, visualizing and interpreting the world, such as the Hochman and Manovich (2013) study of social and cultural patterns of cities through the lens of Instagram photos; or the various surprising and creative mash-ups of the Floatingsheep collective (floatingsheep.org). Furthermore, as a recent literature review of spatiotemporal analyses of Twitter data by Steiger et al. (2015) suggests, there is still much room for development, especially within the GIScience discipline.

This paper suggests a more critical view of the use of geodata from social media sources as a research tool in geographic research and highlights issues like representativeness and representations. It presents a spatial and statistical comparison between three datasets from Instagram, Fickr and Twitter. The aim of the paper is to investigate possible links between differences in the datasets and both the spatial practices of social media users and the modes of production of the ambient geosocial information. The analysis illustrates how bias can be introduced in geosocial media analysis if the poorly understood underlying social processes are not accounted for.

A critical view of social media as a data/capta source

There is little doubt that “big data” from social media can create many new and exciting research directions. Recently however, Kitchin (2014) pointed out several reasoning errors behind this “big data” hype, especially the notion that this data is objective, exhaustive, and can therefore “speak for itself”. Social media data is not representative in any way of the general population. Even without considering the technical limitations and black-box nature of its access (Zook and Graham, 2007), this data is still only composed of representations, filtered consciously or unconsciously by human producers. For example, on Twitter we can observe and measure the density of tweets, but we are not able to do the same with the human motivations. We are even unsure about what part of this signal is human in origin, as digital content can be as easily produced by automated measures like Twitter "bots" (Crampton et al.,
2013) and that users can schedule tweets to be issued when they are not online or have moved elsewhere. Such content pose a serious problem for social media studies and a number of researchers have proposed various filtering methods (Guo and Chen, 2014; Tsuo et al., 2015). This lack of knowledge is one of the things that holds back our understanding of the nature of geosocial content as well as of human behavior.

One of the points made by Kitchin (2014) is that the term “data”, stemming from the Latin word “dare”—meaning “to give”—is misleading, because data is always extracted and selected, and never given. Researchers almost always operate with a subset of the great mass of facts related to a specific entity and they must select the categories to which they will pay heed. This subset has been called “capta” by various authors concerned with this issue (Gherardi and Turner, 2002; Checkland and Holwell, 2006; Rob Kitchin and Dodge, 2011; Rob Kitchin, 2014) to stress the process that lies behind it. The word itself comes from the Latin “capere”, meaning “to take”. While “data” is a well-established term both in scientific and business language and will continue in use (Kitchin 2014), in this paper “capta” will be used to highlight the main arguments.

This change of emphasis from “data” to “capta” is significant. As Checkland and Howell (2006) observe, the transformation of data into capta is a process that is almost transparent to us, as it has become familiar and is often overlooked in scientific inquiries. Poore and Chrisman (2006) observed that the refinement of information from raw data to knowledge is rarely made explicit and the information itself is actively transformed by its recipients. There are various networks of power and social relations that are fundamental in establishing of meaning. Researchers struggling with understanding increasingly large and complex “big data” sets should critically question the origins of the data, the purpose of its collection, the amount of pre-processing involved, and the methods with which this was done (Kitchin, 2014). These questions are even more important when various sources are used. Frameworks that are constructed for the analysis of information from geosocial media often integrate more than one source of capta (e.g. Stefanidis et al., 2013). Integrating heterogeneous data sources can improve spatial and temporal coverage and can enrich data analysis with more varied content. On the other hand, users of social media services and the content they produce and distribute on the web are very heterogeneous. There are different tools in the geolocated media “produsers” arsenal (Coleman et al. 2009) that are used to create content. These tools are platform dependent and for some services adding geographical localization is easier than for others. The service provider’s philosophy is also important: Is geotagging entirely optional, encouraged, or silently enforced? In other words - there is a distinct possibility that the
geolocated capta points from two social media services are incomparable by virtue of how they were generated and processed.

The integration of captasets is relatively easy when the capta source relies on geographical coordinates to provide a location. However, this can lead to an erroneous assumption that all geolocated capta represent phenomena taking place in material space in a similar way or for similar reasons. But a geotagged tweet, a digital photo posted on a photo-sharing site, a blog post, or a Foursquare check-in can be the results of many different, incomparable processes. Each act of posting geolocated content is mediated not only by spatiality, but also by the available technology and by social and economic mechanisms. The users are different, and the capta generated by them is shaped by their differences. This must be acknowledged prior to any analysis that makes use of geosocial information. Researchers must always question themselves how their agenda influences their research methods, particularly in terms of choosing what data is captured and how it is processed.

This is an important aspect of GIS viewed as a social technology with social impacts (Sheppard 1995) and therefore very vulnerable to the problems of representativeness and research ethics (Curry 1995, Crampton 1995). Assumptions made working with Big Data are very similar to ones that accompanied geography's quantitative revolution (Barnes 2013). As Burns (2015) pointed out Big (spatial) Data is an epistemology that can promote knowledge of privileged people, who tend to be more technically savvy in the case of geosocial media. Given that there is no standard, well-established methodology for geosocial media analysis and that accessing large amount of data/capta is relatively easy, it is tempting for researchers to ask questions for which there are no answers or too many answers without having to take responsibility for methodology.

Methods

The problems accompanying social media capta will be illustrated here with the use of a small capta sample from a relatively small area of one Polish city - Poznan. In Poland, internet penetration rates are lower than in the country’s more developed neighbors, a characteristic shared with other post-socialist countries. This is also true for social media use, although the ratio between geolocalized and other content is similar to that observed in other parts of the world (Rzeszewski, 2015). In such settings, one can expect to observe mechanisms associated with a digital divide or rather divides, that separate different groups of users according to their economic status and how and why they access digital content. Other biases
inherent to social media are also present e.g. the “tyranny of the loud” where a small but extremely vocal minority produces a disproportionately large amount of content. Capta was gathered (or rather captured) from three social media services - Twitter, Instagram and Flickr - through the use of their respective APIs, with custom-made Python scripts, over a a period of one year (2014). According to the documentation, Flickr and Instagram APIs permits downloading of all the capta for a given timeframe. This was confirmed by requesting separate queries using both different bounding boxes and other non-geographical filters like user names - in all the cases resulting captasets were almost identical with differences not greater than 1% of the whole volume. Those discrepancies can be attributed to the dynamic nature of both capta sources. Users can delete images thus there is a distinct possibility that some of the content will be missing unless capta gathering is continuous. In the case of the Twitter API there are many disclosed and undisclosed limits on the amount of content that can be gathered using publicly accessible endpoints. However, Morstatter et al. (2013) showed that when geographic bounding boxes are used, the collected capta are almost the complete set of geotagged tweets and therefore can be used for analytical purposes with a large degree of trust.

A further analysis was conducted in R and QGIS with only the most basic preparations: removal of duplicates and points with obviously erroneous coordinates i.e. located outside bounding box of 20 km buffer of city boundary. Only the points tagged with precise latitude and longitude coordinates were selected. Apart from this, no capta was altered prior to the visualization and calculation of the summary statistics. This means that there are various issues regarding quality of the captasets. It is rarely appropriate to present and analyze capta in their raw form (Poorthuis and Zook, 2015), but in this case it was necessary since it allowed for the detection of variations and differences that are also a function of quality. Capta from two sources that are not standardized for a population should, for example, mimic the underlying demographics in a similar way i.e. be biased towards city centres and other highly populated areas. If this is not the case, then one can assume that one or more other processes are involved in shaping the observed signal. Any further manipulation of the captasets should be specific to a given social media platform and could introduce an uncontrolled bias. However, since the raw captasets can hide more subtle spatial patterns an odds ratio approach proposed by Poorthuis et al. (2014) was adopted. For any given cell an odds ratio (OR) was calculated using the following equation:

\[ OR = \frac{p_1/p}{r_1/r} \]
where $p_i$ is the number of capta points in cell from a given social media platform and $p$ is the total number of points from a given social media platform, $r_i$ is the number of “social media population” points in cell and $r$ is the grand total of that population. The “social media population” was constructed by aggregating capta from Twitter, Instagram and Flickr for the same period of time and serves as a proxy for the amount of “buzz” generated in each particular location. For the analysis only OR values above 1 that were also statistically significant (within 95% confidence interval) were used.

Density maps presented in this paper use hexagonal grids that overlays the area of study. This approach is useful because it addresses some of the problems associated with visualization and analysis of large point captasets, for example over plotting, that inhibit or prevent useful insights, or when variance in number of points in different parts of the study area is high. Hexagon maps are also better than rectangular grids in communicating spatial patterns of the phenomenon to map readers because they are less distracting (Carr et al., 1992) and offer higher representational accuracy (Scott, 1985). However, as with all cells that aggregate phenomenon, care must be taken to minimize the potential effect of the Modifiable Areal Unit Problem (see, for example, Wilson 2013). The set of hexagons was determined by comparing different cell sizes. Based on visual comparison the largest size was chosen that still retains all the spatial patterns - when smaller cells were used some of the patterns have changed. For more detailed discussion about using hexagon maps in visualization of spatial media capta see Shelton 2014 and Poorthuis and Zook 2015.

**Results**

Spatial distributions of the three captasets were visualized by displaying density estimates. The first visual clue as to the differences between the three sets can be seen in Figure 1. This is a relatively raw image, but it displays the main concentrations of digital content and therefore activity well. Images of this kind are also used in literature for visualizing the spatial patterns of urban populations (Turner and Malleson, 2012) or detecting attractive locations (Hochmair, 2010; Mirković et al., 2010). The density distributions exhibit similarities as well as discrepancies. The city center is the focal point of activity in all three captasets. However, the Twitter users tended to post their content around Poznan Glowny (the main train station) and Stary Browar (a famous mall) while Flickr and Instagram users preferred the Old Market, the banks of the Warta River, and Cathedral Island–the oldest parts of the city with the high concentration of historical monuments and tourist attractions. Also, the Twitter and Flickr posts had
similar secondary clusters in the southern part of the city outskirts—in green areas; while Instagram appears to be an entirely urban phenomenon. This indicates that the differences are likely to come from more factors than a simple contrast between photo-sharing and microblogging sites. Other service-specific elements of the user experience like the simplicity of geolocation may as well play important roles.

Figure 1. Comparison between the raw geolocated capta point density of different social media platforms

However, such simple comparison of raw capta without taking into account differences in population hides more subtle patterns. To mitigate this, an odds ratio approach was adopted (Poorthuis et al., 2014) where digital social activity of each service was normalized using the whole “social-media population” i.e combined capta-stream of Twitter, Instagram and Flickr. This corrects for substantial variation in the total number of points and highlights even minor differences. The resulting maps (Figure 2) show places where each service is significantly (p < 0.05) stronger represented than expected in the
context of the whole geosocial media environment - assuming Twitter, Instagram and Flickr are representative in this case. It can be observed than Flickr is even more different from the other services than in Figure 1. It dominates outskirts of the city with no apparent preference for any given region. Instagram and Twitter are more similar to each other but still there is a visible preference for the former in the southeastern and for the latter in the northwestern part of the Poznan. This may indicate differences in user-base since those are two large housing areas with different demographics.

Figure 2. Comparison between the odds ratio of different social media platforms in relation to the combined captaset.
To further analyze the spatial distribution and density of capta we restricted the points to only one per unique user per hex cell (Figure 3). We did this to filter out concentrations that resulted from actions performed by overactive users. The change was almost negligible for Instagram but was very visible for Flickr and Twitter, whose capta points cover large parts of the city. In all three images, there is also a stronger contrast between the central and peripheral areas. Surprisingly, from this modified perspective, the three services’ captasets are much more similar to each other. Each of the three captasets can also be described in statistical terms (Table 1). During the year of the capture, the service with the most geotagged points was Twitter, which was to be expected given its microblogging character, and the smallest number was extracted from Instagram. However, despite having the smallest number of geolocated points, Instagram had the largest number of unique users, which led to very small mean number of points per user. On the other end of the spectrum was Flickr with only one fifth as many users as Instagram, but with an almost three times larger captaset. Twitter was located between these two values. As can be seen from the empirical distribution function curve (Figure 4), the statistically described user interaction shows significant dissimilarities between the three services, with Instagram standing the most apart from the other two. However, the differences are located at the head of the distributions, with the medians for the number of points per user being between 1-4 for all three services. The discrepancies between the mean number of points per user means that, in the case of Instagram, almost all the content was produced by infrequent, casual users; whereas for Flickr the cyberscape was strongly influenced by so-called “power users”, who are capable and willing to create large amounts of capta points. This confirms the differences seen in density maps in Figure 3. This is even more drastically visible when we compare the maximum values recorded in the captured period. While 164 photos for one Instagram user seems very large when the mean value is only about 2, it pales in comparison with the 6247 points for one Twitter user, and especially with the 16203 points (pictures per day) recorded for Flickr. The last value is even more striking because it constitutes almost half of all the posts. At this point in the study it was necessary to acknowledge the possibility that the content of these two accounts was automatically produced. It is a common sense that the assumption that all content on the web is created by the human hand is erroneous. However, closer inspection revealed that, the Flickr account seemed to belong to a single human user who was spamming the service with photos of trains, the transport infrastructure, stations, etc. On the other hand, the Twitter account was at least in part generated automatically - thanks to the web service Unfollowers.com that among other automation features periodically tweets user stats.
without the need for human intervention. Nonetheless, the majority of the content was still personal in nature. This single account was partially responsible for the secondary cluster in the southern part of the city in Figure 1.
disproportionately large degree of attention to the outliers, but in this analysis this was acceptable since it provided a glimpse at the spatial characteristics of the most prolific users. The mean values were larger for the photo-sharing services than for Twitter - almost twice as large in the case of Flickr. However, this can be attributed to the nature of the service they provide, and the fact that in most cases microbloggers (such as Twitter users) tend to tweet from their home - or rather from some kind of cyber anchor point (Couclelis et al., 1987).

Figure 4. ECDF plot of capta points per user

Table 1. User statistics

<table>
<thead>
<tr>
<th></th>
<th>FLICKR</th>
<th>INSTAGRAM</th>
<th>TWITTER</th>
</tr>
</thead>
<tbody>
<tr>
<td>points in year</td>
<td>33398</td>
<td>12995</td>
<td>142955</td>
</tr>
<tr>
<td>number of unique users</td>
<td>275</td>
<td>6107</td>
<td>4338</td>
</tr>
<tr>
<td>mean number of points per user</td>
<td>121</td>
<td>2</td>
<td>32</td>
</tr>
<tr>
<td>maximum value of points per user</td>
<td>16203</td>
<td>164</td>
<td>6247</td>
</tr>
<tr>
<td>mean Standard Distance Deviation [m]</td>
<td>1120</td>
<td>851</td>
<td>664</td>
</tr>
</tbody>
</table>

Users with multiple accounts (power users)

| mean number of points per user | 112   | 4.5     | 74      |
| mean Standard Distance Deviation [m] | 1065 | 635     | 1766    |
One can hypothesize that the most active users (the power users) are more likely to have a presence on more than one social media platform. Following the assumptions made in the introduction, this means that they will automatically exert a stronger influence on the shape of the cyberscape. Although it is difficult to perfectly identify such users with the data content alone, an attempt was made by the very simple method of comparing the user names registered on different platforms. It was assumed that an exact match in the account names meant that they belonged to a single person. A group of 181 users were selected for this process that shared an account name between at least two services, with 24 names present in all three of them. The values for both the SDD and the mean number of points per user were much larger in the case of Twitter. For Instagram this was true for the second parameter, but not for the first; whereas for Flickr both values were slightly smaller, but in this case the mean values for all the users were strongly skewed by a single account. Removing the outlier revealed that the power users produced twice as much content, with a similar SDD. This comparison suggests that users with multiple accounts, on average, produce more capta points in more locations than other users. But still, there are differences between the services, and the ratios are not the same as for the general population of users. Spatial distribution of power user capta (Figure 5) reveals that points are less dispersed and more visibly concentrated in the central parts of the city - this is especially evident in Twitter. There are no hot spots in the outskirts. But what is most evident is that the differences between services are even more pronounced. It seems that even frequent users who use several platforms use them in very different ways - at least in the case of geotagged content. This suggests that the observed spatial patterns may as well be dependent on the characteristics of the given service and not on spatial behaviors of its users.

The differences observed were also present in the content of the capta points. A comparison between the three services, based on the most popular tags, is presented in Table 2. It should be noted that tagging mechanisms in Flickr and Instagram are similar while Twitter hashtags are somewhat different, having more conversational and less descriptive nature. However, all tags are connected to geographical location and therefore it can be assumed that they were used, at least to some extent, to characterize material places. The tags from these three services were very different, both in character and in percentages. It initially seemed that Flickr was much more homogeneous than the other two, but a closer look revealed that this image was distorted by a single prolific user tagging all his or her photos with identical phrases. After the removal of this outlier, the list of tags changed. Still, the percentage of repeated tags was much greater than for posts to Instagram, and drastically greater than posts to Twitter.
Table 2 suggests that, of the two photo sites, Instagram is used more frequently as a social media platform (e.g. #love, #friends, #selfie), while Flickr’s role is more of a web photo gallery of interesting places, events and photography in general (e.g. #iphoneography, #squareformat, #polishcup).

Table 2. Most popular tags in Flickr, Instagram and Twitter

<table>
<thead>
<tr>
<th>FLICKR – outlier removed</th>
<th>FLICKR – all users</th>
<th>INSTAGRAM</th>
<th>TWITTER</th>
</tr>
</thead>
<tbody>
<tr>
<td>tag</td>
<td>% of all points</td>
<td>tag</td>
<td>% of all points</td>
</tr>
<tr>
<td>Poznan</td>
<td>36</td>
<td>Poland</td>
<td>56</td>
</tr>
<tr>
<td>Square</td>
<td>28</td>
<td>Polska</td>
<td>52</td>
</tr>
<tr>
<td>iPhoneography</td>
<td>28</td>
<td>Wielkopolska</td>
<td>48</td>
</tr>
<tr>
<td>InstagramApp</td>
<td>28</td>
<td>Canon</td>
<td>48</td>
</tr>
<tr>
<td>Squaredformat</td>
<td>28</td>
<td>Wielkopolskie</td>
<td>45</td>
</tr>
<tr>
<td>Uploaded: by=instagram</td>
<td>27</td>
<td>Railroad</td>
<td>44</td>
</tr>
<tr>
<td>505sailing</td>
<td>27</td>
<td>Rail</td>
<td>44</td>
</tr>
<tr>
<td>Poznan</td>
<td>21</td>
<td>PKP</td>
<td>44</td>
</tr>
<tr>
<td>Poland</td>
<td>16</td>
<td>Station</td>
<td>44</td>
</tr>
<tr>
<td>Dinghysailing</td>
<td>18</td>
<td>Greaterpoland</td>
<td>44</td>
</tr>
<tr>
<td>Polishcup</td>
<td>18</td>
<td>Puszczykowo</td>
<td>35</td>
</tr>
<tr>
<td>Spichlerz</td>
<td>9</td>
<td>Luboń</td>
<td>26</td>
</tr>
<tr>
<td>Bridge</td>
<td>8</td>
<td>Poznań</td>
<td>20</td>
</tr>
<tr>
<td>Street</td>
<td>7</td>
<td>InstagramApp</td>
<td>18</td>
</tr>
<tr>
<td>Polska</td>
<td>7</td>
<td>iPhoneography</td>
<td>14</td>
</tr>
<tr>
<td>People</td>
<td>5</td>
<td>Building</td>
<td>14</td>
</tr>
</tbody>
</table>

Twitter tag analysis on the other hand presents a very different picture, where there are no dominant phrases - most popular tag #poznan is used just in 1 of 100 tweets. This may suggest a mechanism where geolocated content is used in a very different way - maybe as an argument in conversation or as a byproduct of habitual use of geolocation in a mobile device. What motivates users to geolocate themselves is not well understood in geosocial media analysis and should be investigated before conducting inferential analysis. However, apart from the differences between tags from the three services, similar words are present that are associated with Poznan and Poland and in the case Flickr even with specific regions and places. Geotagging therefore can be viewed as leading to at least some degree of unification between captasets. One thing that is also worth noting is that, among Flickr tags, there were
several indicating that those particular photos were taken with Instagram software. This suggests another set of questions about the differences between these two services, and the choices that were made by their users.

![Figure 5. Comparison between the raw geolocated capta point density of the power users](image)

**Conclusions and discussion**

The results of this study clearly show that captasets taken from different social media platforms vary substantially with regard to their content and spatial distribution. This is visible even when they come from an identical geographical and temporal extent. One might attribute these differences to the nature of the given social media service, but the presented case with Flickr, Instagram and Twitter shows that this explanation is too simplistic, and the distinction between a "microblogging service" and a "photo-sharing site" is not enough. Although there are many similarities, social media platforms seem to create their own
"ecosystems" of users with unique behavior. Accounting for platform-specific behavior is a challenge that geosocial media research cannot ignore. Cyberscapes will look different depending on the data stream gathered by the researcher – with its choice of place, time and type of social service, as well as the usual biases imposed by the cultural and social context. On top of this, there is also the black box behavior of most APIs – where the amount and the type of filtering are not disclosed to developers or customers. In the case of Twitter, it has been shown that a free public data stream that represents at most 1% of all tweets is certainly a viable source of data, but at the same time it cannot be considered a statistically random sample (Morstatter et al., 2013). These limitations must be taken into account in the process of any research design. For example, it would be sensible to gather two independent, slightly temporally separated but spatially overlapping Twitter captasets, and check them for any biases.

However, the main source of discrepancies between geotagged captas gathered from different social media platforms lay in the differences between the spatial practices of their users. We currently have very little knowledge about who or what produces geotagged content, why and when it is produced, and in what situations. This study’s findings suggest that the behavior of users can vary between platforms, which in turn can introduce biases at the interpretation stage of any geosocial research. It must be acknowledged that captas points from different sources can also relate to a material space in different ways, e.g., Instagram users often report social gatherings and Flickr users tend to build digital representations of famous places. Captas are often reduced to points in databases, and even when some further transformations or normalizations are applied, this can lead to misconceptions about significance. Because geotagging social media content differs among different services, countries, social classes or even individuals, it poses a challenge to researchers, but at the same time offers opportunities to delve more deeply into the nature of human relationships in the digital realm. Some of the questions raised are: Why do people add geographical locations to the content they produce? Is this a conscious behavior? How is this information utilized by search engines? How can this information change the perception of a material space/place? For which purposes and extents can we rely on this data in research? How can we aggregate different data sources? We need to go “beyond the geotag” (Crampton et al., 2013).

In GIScience, this means that perhaps we must make more use of social science methods to supplement our explanations when dealing with ambient geosocial information. A similar approach has already been postulated in the field of human movement studies (Raanan and Shoval, 2014; Kotus and Rzeszewski, 2015; Rzeszewski and Kotus, 2014). But first and foremost we need to recognize that AGI
captasets are nowhere near as simple to analyze as it is sometimes proposed, and we should apply caution when tempted to interpretate these new and rich streams of capta.

Acknowledgements

I would like to thank the four anonymous reviewers for their invaluable help – substantial improvements were made to this paper thanks to their comments. I also would like to express my gratitude to the Editor dr Nick Chrisman, that was extremely helpful and encouraging during the whole process of paper submission.

This work was supported by the Polish National Science Center under Grant UMO-2015/17/D/HS4/00272.

References


doi:10.1145/2789187.2789196


http://dl.acm.org/citation.cfm?id=1753486.


