Structural Causes of Bias in Crowd-derived Geographic Information: Towards a Holistic Understanding

Isaac Johnson and Brent Hecht

GroupLens Research, Department of Computer Science and Engineering, University of Minnesota

joh12041@cs.umn.edu, bhecht@cs.umn.edu

Abstract

A critical subset of human-generated content is that which is *geographically referenced*. The spatial context of these data has enabled a new class of observational studies and technologies. Unfortunately, researchers have identified a number of biases in these datasets (e.g. urban/rural) that threaten their utility. However, these findings have been highly specialized, focusing on single datasets and dimensions of bias. This extended abstract seeks to begin the process of synthesizing a cohesive understanding of the structural causes of bias in geographically-referenced human content. We outline five cross-cutting causational factors, as well as introduce a novel framework that aids in understanding these factors. In doing so, we hope to initiate a discussion that moves the literature on bias in geographic content towards one focused on systemic issues. This would allow the anticipation of bias in unseen datasets and, importantly, enable research towards systemic solutions.

Introduction

Geotagged tweets, Wikipedia articles about places, mobile query logs, and other forms of crowd-derived geographic information have become critical sources of information for many systems and studies. Indeed, thanks to the rapid democratization of positioning technologies (e.g. GPS), crowd-derived geographic information is now critical to systems in fields ranging from disaster management to natural language processing and studies in domains across the natural sciences and the humanities.

While crowd-derived geographic information has proven tremendously useful to researchers and practitioners, a growing stream of literature has identified that this information is beset with biases along a number of dimensions. For instance, OpenStreetMap coverage has been found to be much lower in more impoverished areas (e.g. Mashhadi, Quattrone, and Capra 2013) and, along with Wikipedia, demonstrate a male bias (e.g. Stephens 2013). Similarly, geotagged tweets and photos have been shown to be biased towards wealthier and more highly educated areas (e.g. Li, Goodchild, and Xu 2013). However, this research has thus far has been highly focused, usually concentrating on a single dataset (e.g. Twitter) and single dimensions of bias (e.g. income). This has created a literature with many silos and few overarching conclusions.

Our goal with this extended abstract is to initiate a discussion that moves the growing literature on bias in crowd-derived geographic information from one focused on single instances of bias to one focused on systemic issues. Looking forward, our hope is that this discussion will eventually allow researchers and practitioners to anticipate bias-related problems in unseen datasets *a priori*. Importantly, the identification of systemic issues is also the first step towards systemic solutions, solutions that could simultaneously lead to both more equitable and more effective innovation in this space.

The main mechanism by which we begin the aforementioned discussion – and the primary contribution of this extended abstract – is a synthesis of the literature into a first-pass understanding of the structural causes of bias in crowd-derived geographic information. We identify five cross-cutting causational factors ranging from population density to gender. We also introduce a novel framework for surfacing and understanding this bias. This framework extends beyond simple population bias to incorporate other biases that are reflections of the geospatial nature of crowd-derived geographic information (e.g. per-area coverage biases and biases related to place identity).

Below, we first consider the challenges that arise from the diverse terminology used in the relevant literatures. Next, we introduce a novel framework through which the diverse structural biases in crowd-derived geographic information can be understood. Using this framework, we then enumerate a number of structural causes of bias that have been identified in the diverse crowd-derived geographic information bias literature, as well as some that are emergent from a holistic reading of this literature.

Terminology in a Balkanized Literature

Crowd-generated data that has a geographic reference has been described by a number of terms in a number of different domains, with these terms often referring to overlapping, somewhat ambiguously-defined concepts. For instance, in geography, the term *volunteered geographic information* (VGI) is paramount (e.g. Goodchild 2007). However, geographers have recently problematized this term (Sieber and Haklay 2015), with the crowd-generated origins of this information – not whether the data is volunteered – often being the more important characteristic of the information. Similarly, within computer science and related fields, the terms *geographic user-generated content* (Hecht and Gergle 2010) and *geotagged social media* (Eisenstein et al. 2010) have been used. These terms have been considered equally problematic: user-generated content is an ambiguous term in and of itself (Wunsch-Vincent and Vickery 2007) and drawing boundaries around social media ignores its many similarities with other types of data, e.g. geographically-referenced peer-produced information.

The balkanization of the terminology used to refer to data generated by similar processes is both another indicator of the need for a more holistic understanding in this research space and a major obstacle to the gaining of this understanding. As such, while we are loathe to add to the alphabet soup in this literature, we address the inherent nomenclatural challenge to this extended abstract by adopting the term **crowd-derived geographic information** (CdGI). This term is sufficiently general to incorporate geographic information from the crowd that is and is not volunteered, that is and is not social media, and that is and is not user-generated content.

A CdGI Framework for Understanding Biases

To support our discussion of the structural causes of bias in crowd-derived geographic information, we first introduce a simple framework for CdGI that affords straightforward understanding of these causes. In this framework, every CdGI-based study and system[1] can be classified according to three properties: (1) its mapped process, (2) its coverage type and (3) the importance it places on *localness*. The value of each property for a given study or system defines (in part) the potential biases of the study or system, as well as the degree to which these can be addressed.

The **mapped process** is simply the geographic phenomenon a researcher is attempting to capture or model through CdGI. For example, the mapped process of a study seeking to estimate the "Gross National Happiness" (Kramer 2010) of a region using social media is the geographic variation in the expressed sentiment in the region. As we will see below, the definition and mis-definition of the mapped process can result in important biases.

Coverage type refers to the unit of normalization along which a CdGI-based study or system can be said to have achieved representative or complete coverage of a mapped process. Although coverage type is almost never explicitly addressed in the literature, it can play a critical role in the structural biases to which a given study or system is exposed. We have identified two coverage types in CdGI literature: **per-capita coverage** and **per-area coverage**.

The goal in the case of per-capita coverage studies and systems is to better understand a *human population* using CdGI. In these studies, ideal coverage amounts to either representative or complete sampling of the population using CdGI. For instance, studies that seek to measure geographic variations in sentiment on Twitter operationalize a per-capita definition of coverage, as they are seeking to gain a representative sample of the population's emotions. The goal in the case of per-area coverage is to better understand an *area* using CdGI, not a human population. For instance, many citizen science projects adopt a per-area coverage model, as they are seeking to, for instance, understand the distribution of species in a region of interest.

The importance of localness is equally significant for the bias that may be present in a given CdGI-based study or system. Broadly speaking, CdGI-based studies and systems can divided into those that are **localness-important** and **localness-agnostic.** Studies and systems for which the information about a given area must be generated by people who are local to that area are localness-important, with this not being true of those that are localness-agnostic. An example of a localness-important study is one that seeks to predict election results using social media. For this study, it is critical that tweets, Facebook posts, etc. that are geotagged to a specific region come from an eligible voter in that region. On the contrary, an example of a localness-agnostic study is one that examines human mobility using taxi cab activity traces for the purposes of urban planning (or the equivalent using Swarm or Facebook check-ins). In this case, regardless of whether a cab passenger (or checked-in user) is a local or a tourist, her/his movement patterns are, for the most part, equally relevant. Like is the case with coverage type, the importance of localness (and the corresponding biases that may occur) are rarely if ever explicitly considered in CdGI-based research.

[1] Note: Like the rest of this extended abstract, we view this framework as a starting point for discussion rather than a finalized schema. Other dimensions besides the three considered here will likely be needed to full describe the potential biases in CdGI.

Structural Causes of CdGI Biases

We now turn our attention to a discussion of the structural causes of bias in CdGI that have been identified across the literature. Each causational factor is discussed in the context of the framework described above.

Population Density

Population density has been shown to introduce biases for both per-capita and per-area mapped processes. The per-capita biases related to population density are relatively straightforward. Across multiple CdGI datasets, people who live in urban areas have been found to be more active participants in CdGI-generating activities than people who live in rural areas (Hecht and Stephens 2014; Gilbert, Karahalios, and Sandvig 2010).

In the case of processes for which per-area coverage is necessary, these population density biases are significantly compounded. Put simply, *in rural areas, there are many fewer people per unit area to map a process*. This means that even if there were equal participation in rural and urban areas, each rural individual would still have a much higher burden to provide coverage on par with urban areas. Unfortunately, recent work in our group suggests that this is a structural bias that will be very difficult to overcome even with massive increases in local participation or the extensive "importing" of labor from urban areas (which would severely reduce localness). More generally, it appears that CdGI may not be an effective data source for per-area mapped processes in rural regions.

Per-area systems and studies in domains ranging from citizen science (e.g. mapping bird populations across the United States) to natural language processing (e.g. geographically-referenced text models) are at extensive risk for this population density-induced bias. Let us consider, for example, the case of a geographic text model built for the purpose of inferring the location of Twitter users (a very active area of research, e.g. (Eisenstein et al. 2010)). In these text models, latitude and longitude grid cells (graticules) are used as "documents", with the geotagged tweets in each cell serving as the "content" of the documents. Because graticules are of similar (though not identical) size across most study/application regions, graticules in low-population density areas are likely to have many fewer tweets. This will result much sparser "documents" about rural areas than about urban areas. Evidence from our own in-progress work shows that this sparsity can have a tremendous effect on geolocation accuracy. Even after controlling for mapping population bias in rural and urban areas using resampling, we observed that a state-of-the-art text-based geolocation algorithm (Priedhorsky, Culotta, and Del Valle 2014) still has significantly worse accuracy in rural areas than urban areas.

Our research has identified similar effects in Wikipedia's and OpenStreetMap's efforts to describe per-area geographic phenomena, e.g. write articles about every incorporated hamlet in the world or map all roads, even in very rural areas. We observed that where there are likely to be fewer local experts – e.g. in low-population density regions – content quality tends to be substantially lower. In fact, without automated content generation agents, it is likely that much of this content would not exist in the first place. This is something that we observed in rural China, where automated content generation agents do not have access to important data due to government restrictions.

Socioeconomic Status

There is ample and growing evidence that socioeconomic status (SES) is a significant cause of bias in per-capita CdGI studies and systems. For instance, recent research has found that there are more tweets per capita in richer areas (Li, Goodchild, and Xu 2013) and that poorer areas have lower coverage in OpenStreetMap when controlling for other factors (Mashhadi, Quattrone, and Capra 2013). Similarly, Thebault-Spieker et al. found that participation in TaskRabbit was effectively non-existent in poorer areas of Chicago (Thebault-Spieker, Terveen, and Hecht 2015).

The mechanism behind SES bias may be partially attributable to access to technologies like broadband Internet and cell phone coverage. Recent research has found that at least in the Wikipedia context, these factors represent necessary – but not sufficient – conditions for good coverage (Graham, Straumann, and Hogan 2015).

Community Practices and Norms

Many CdGI datasets are associated with online communities, with the community's practices and norms often defining the CdGI it generates (i.e. the processes the community maps). Unfortunately, community practices and norms have been identified as the source of some troubling CdGI biases. For instance, Stephens (2013) showed that while OpenStreetMap's important community-defined typology of mapped processes distinguishes between strip clubs and brothels (which, as Stephens argues, are more associated with male interests), as of 2011, childcare was grouped together with kindergarten (more associated with female interests). This effectively removes from the map distinctions between childcare centers and kindergartens.

Along the same lines, the snapshots of human mobility provided by geotagged social media such as Swarm (i.e. Foursquare) and Facebook check-ins are colored by the practices and norms in each corresponding community. If one mistakenly assumes one can use these check-ins for a study or system that relies on accurate human mobility as its mapped process, it will falsely presume that almost no one goes to socially-charged places like abortion clinics.

Gender

Community practices are one source of gender-related bias in CdGI, but gender is a factor in other contexts as well. In particular, gendered differences in location disclosure behavior and related safety concerns have been found to lead to male bias in per-capita studies and systems.

For instance, recent in-progress research has shown that women fill out their Twitter location fields in a fashion that is significantly less "geolocatable" (e.g. they are less likely to populate the location field and, when they do, they are more likely to input non-geographic information, e.g. "Tomorrowland"). Since Twitter location field data plays a critical role in boosting the number of tweets that can be georeferenced (only 1-3% are explicitly geotagged), this difference in location disclosure behavior may be generating significant gender bias in many Twitter-based per-capita studies and systems. More explicitly on the safety front, Thebault-Spieker et al. (2015) found evidence that in mobile crowdsourcing systems (e.g. TaskRabbit), women were much less willing to accept tasks in areas they perceived to be unsafe.

Language

Research has shown that language can be a powerful cause of bias in CdGI. There is a coverage and localness dimension to this bias. For example, with respect to per-capita coverage processes, if a study or system uses English-only sentiment analysis algorithms to understand the geographic variation in Twitter sentiment in the United States, this algorithm will inherently exclude the millions of people in the United States who tweet in languages other than English (e.g. Spanish).

Localness biases due to language are related to the corresponding per-capita coverage biases. Namely, the per-capita biases with regard to language can be sufficiently strong that non-local perspectives become the *dominant perspective* in a given CdGI dataset. For instance, Sen et al. identified that, in a given Wikipedia language edition, articles about places in countries where the corresponding language is not spoken can have miniscule amounts of local content (Sen et al. 2015). This relationship between localness and per-capita coverage likely exists in other contexts as well: in cases when per-capita coverage is very low, the content about a given area may often be generated by non-locals.

Discussion and Conclusion

The goal of this extended abstract has been to begin the process of synthesizing the structural causes of biases in crowd-derived geographic information (e.g. volunteered geographic information, citizen science, activity traces). The above framework and enumeration of biases can both (1) inform designers of CdGI-based studies and systems and (2) motivate future work to address these biases, an area in which we are doing current work and that we hope to discuss at the symposium.

Before closing, it is important to reiterate that we do not view the above schemas as complete. For instance, additional possible structural biases include national culture factors (c.f. Quattrone et al. (2015)) and those related to distance (c.f. Hecht and Gergle (2010)).

References

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

Gilbert, Eric, Karrie Karahalios, and Christian Sandvig. 2010. "The Network in the Garden: Designing Social Media for Rural Life." *American Behavioral Scientist* 53 (9): 1367–88.

Goodchild, Michael F. 2007. "Citizens as Sensors: The World of Volunteered Geography." *GeoJournal* 69 (4): 211–21.

Graham, Mark, Ralph K. Straumann, and Bernie Hogan. 2015. "Digital Divisions of Labor and Informational Magnetism: Mapping Participation in Wikipedia." *Annals of the AAG*.

Hecht, Brent, and Darren Gergle. 2010. "On The 'Localness' of User-Generated Content." In *CSCW '10*.

Hecht, B., and M. Stephens. 2014. "A Tale of Cities: Urban Biases in Volunteered Geographic Information." In *ICWSM '14*.

Kramer, A.D.I. 2010. An unobtrusive behavioral model of gross national happiness. *CHI '10: 28th ACM Conference on Human Factors in Computing Systems* (2010), 287–290.

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

Mashhadi, A., G. Quattrone, and L. Capra. 2013. "Putting Ubiquitous Crowd-Sourcing into Context." In *CSCW '13*.

Priedhorsky, R., A. Culotta, and S. Y. Del Valle. 2014. "Inferring the Origin Locations of Tweets with Quantitative Confidence." In *CSCW '14*.

Quattrone, Giovanni, Licia Capra, and Pasquale De Meo. 2015. "There's No Such Thing as the Perfect Map: Quantifying Bias in Spatial Crowd-Sourcing Datasets." In *CSCW '15*.

Sen, Shilad W., Heather Ford, David R. Musicant, Mark Graham, Oliver S.B. Keyes, and Brent Hecht. 2015. "Barriers to the Localness of Volunteered Geographic Information." In *CHI '15*.

Sieber, Renée E, and Mordechai Haklay. 2015. "The Epistemology(s) of Volunteered Geographic Information: A Critique." *Geo: Geography and Environment*.

Stephens, Monica. 2013. "Gender and the GeoWeb: Divisions in the Production of User-Generated Cartographic Information." *GeoJournal* 78 (6): 981–96.

Thebault-Spieker, Jacob, Loren G. Terveen, and Brent Hecht. 2015. "Avoiding the South Side and the Suburbs: The Geography of Mobile Crowdsourcing Markets." In CSCW '15.

Wunsch-Vincent, Sacha, and Graham Vickery. 2007. "Participative Web: User-Created Content.". (OECD Directorate for Science, Technology, and Industry.