Chapter 1

Demographic biases

This is Chapter 1 of:

> Momin M. Malik (2018). "Bias and beyond in digital trace data". PhD thesis. Pittsburgh, PA: Carnegie Mellon University. URL: http://reports-archive.adm.cs.cmu.edu/anon/isr2018/abstracts/18-105.html

It is an updated version of a previously published paper. For referencing, please give citations to both the originally published work and to this thesis.

ACM Momin M. Malik, Hemank Lamba, Constantine Nakos, and Jürgen Pfeffer. 2015. Population Bias in Geotagged Tweets. In *Papers from the 2015 ICWSM Workshop on Standards and Practices in Large-Scale Social Media Research (ICWSM-15 SPSM)*. 18–27.

APA Malik, M. M., Lamba, H., Nakos, C., & Pfeffer, J. (2015). Population bias in geotagged tweets. In *Papers from the 2015 ICWSM Workshop on Standards and Practices in Large-Scale Social Media Research (ICWSM-15 SPSM)* (pp. 18–27).

Chicago Malik, Momin M., Hemank Lamba, Constantine Nakos, and Jürgen Pfeffer. "Population Bias in Geotagged Tweets." In *Papers from the 2015 ICWSM Workshop on Standards and Practices in Large-Scale Social Media Research (ICWSM-15 SPSM)*, 18–27. 2015.

IEEE M. M. Malik, H. Lamba, C. Nakos, and J. Pfeffer, "Population bias in geotagged tweets," *Papers from the 2015 ICWSM Workshop on Standards and Practices in Large-Scale Social Media Research (ICWSM-15 SPSM)*, pp. 18–27, 2015.

MLA Malik, Momin M., Hemank Lamba, Constantine Nakos, and Jürgen Pfeffer. "Population Bias in Geotagged Tweets." *Papers from the 2015 ICWSM Workshop on Standards and Practices in Large-Scale Social Media Research (ICWSM-15 SPSM)*, 2015, pp. 18-27,

BibTeX

```
@inproceedings{malik2015,
    author = {Malik, Momin M. and Lamba, Hemank and
              Nakos, Constantine and Pfeffer, J\"{u}rgen},
    title = {Population bias in geotagged tweets},
    year = {2015},
    booktitle = {Papers from the 2015 ICWSM Workshop on
                 Standards and Practices in Large-Scale Social
                 Media Research},
    pages = {18--27},
    conference = {Ninth International {AAAI} Conference on Web
                  and Social Media},
    series = {ICWSM-15 SPSM},
    url = {http://www.aaai.org/ocs/index.php/ICWSM/ICWSM15/paper/view/10662}
}
```

ACM Momin M. Malik. 2018. Bias and Beyond in Digital Trace Data. Ph.D. Dissertation. Carnegie Mellon University, Pittsburgh, PA. Retrieved from http://reports-archive.adm.cs.cmu.edu/anon/isr2018/abstracts/18-105.html.

APA Malik, M. M. (2018). *Bias and beyond in digital trace data* (Doctoral dissertation, Carnegie Mellon University). Retrieved from http://reports-archive.adm.cs.cmu.edu/anon/isr2018/abstracts/18-105.html.

Chicago Malik, Momin M. "Bias and Beyond in Digital Trace Data." PhD dissertation, Carnegie Mellon University, 2018. http://reports-archive.adm.cs.cmu.edu/anon/isr2018/abstracts/18-105.html.

IEEE M. M. Malik, "Bias and beyond in digital trace data," Ph.D. diss., Inst. Softw. Res., Sch. Comput. Sci., Carnegie Mellon Univ., Pittsburgh, PA, 2018. Available: http://reports-archive.adm.cs.cmu.edu/anon/isr2018/abstracts/18-105.html.

MLA Malik, Momin M. *Bias and Beyond in Digital Trace Data*. 2018. Carnegie Mellon U, PhD dissertation. SCS Technical Report Collection, http://reports-archive.adm.cs.cmu.edu/anon/isr2018/abstracts/18-105.html.

BibTeX

```
@phdthesis{malik2018,
    author = {Malik, Momin M.},
    title = {Bias and beyond in digital trace data},
    year = {2018},
    school = {Carnegie Mellon University},
    address = {Pittsburgh, PA},
    month = {08},
    url = {http://reports-archive.adm.cs.cmu.edu/anon/isr2018/abstracts/18-105.html}
}
```

Summary. Geotagged tweets are an exciting and tremendously popular data source. But, like all social media data, they potentially have biases in who are represented. Motivated by this, I investigated the question, 'are users of geotagged tweets randomly distributed over the US population'? I carry out a statistical test by which I answer this question strongly in the negative, by linking approximately 144 million geotagged tweets within the US, representing 2.6m unique users, to high-resolution Census population data. Utilizing spatial models and integrating further Census data to investigate the factors associated with this nonrandom distribution, I find that, controlling for other factors, population has no effect on the number of geotag users, and instead it is predicted by a number of factors including higher median income, being in an urban area, being further east or on a coast, having more young people, and having high Asian, Black or Hispanic/Latino populations.

Compared to the previously published version, I have an updated literature review, a correction to the main model (previously, the reference category of a categorical variable was incorrectly chosen, it has now been changed to the majority category), and updated figures (plotting skewed distributions as complementary cumulative density functions as is recommended in Clauset et al., 2009, rather than as log-log scatterplots or CDFs).

1.1 Geotagged tweets

'Geotagged' or 'geocoded' tweets, where users elect to automatically include their exact latitude/longitude geocoordinates in tweet metadata, provide data that are:

- High-quality: geotagging is automated, so there are fewer chances of data error such as from user specification (Graham et al., 2014; Hecht, Hong, et al., 2011);[1]
- Precise: geotags are to a ten thousandth of a degree in latitude and longitude;
- Richly contextual: geotags are connected to tweets with all their temporal, semantic, and social content;
- Easily available, through the Streaming API;
- Large: using the Streaming API, a researcher can build a collection of tens of millions of tweets.

Unsurprisingly, this makes them an enormously attractive source for studying a wide range of human phenomena (Hong et al., 2012). Previous to the publication of Malik et al. (2015), works had used geotagged tweets to study

- mobility patterns (Hawelka et al., 2014; Yuan et al., 2013; Cho et al., 2011),
- urban life (Doran et al., 2013; Frias-Martinez et al., 2012),
- transportation (Wang, Al-Rubaie, et al., 2014),
- natural disasters, crises, and disaster response (Morstatter, Lubold, et al., 2014; Lin and Margolin, 2014; Shelton et al., 2014; Sylvester et al., 2014; Kumar et al., 2014), and
- public health (Sylvester et al., 2014; Nagar et al., 2014; Ghosh and Guha, 2013)

as well as the interplay between geography and

- language (Hong et al., 2012; Eisenstein et al., 2010; Kinsella et al., 2011),
- discourse (Leetaru et al., 2013),
- information diffusion and flows (Kamath et al., 2013; Liere, 2010),
- emotion (Mitchell et al., 2013), and
- social ties (Stephens and Poorthuis, 2014; Takhteyev et al., 2012; Cho et al., 2011).

Furthermore, maps of geotagged tweets tend to look remarkably similar to maps of population density (figs. 1.1 and 1.2; see also Leetaru et al., 2013), even if there are differences at a finer scale (figs. 1.3 and 1.4). This naturally leads to the question: are Twitter users who send geotagged tweets (henceforth, ‘geotag users’) randomly distributed over the population? This is a critical question because, if users who elect to geotag are systematically different from people in general, the results of studying geotagged tweets will not have external validity.

[1] Note that through the use of the API, users and services can tag their tweets with arbitrary geocoordinates. We found some evidence of this being used for generating high visibility in a spam-like manner, but only in a few cases. Still, what is most important is that the precise and numerical nature of geotags do not allow users to specify (linguistically) whimsical or ambiguous locations as they can do in the ‘location’ field (and users who whimsically locate their tweets in Antarctica or the middle of the ocean would not be picked up by a geobox around the contiguous United States, see below).

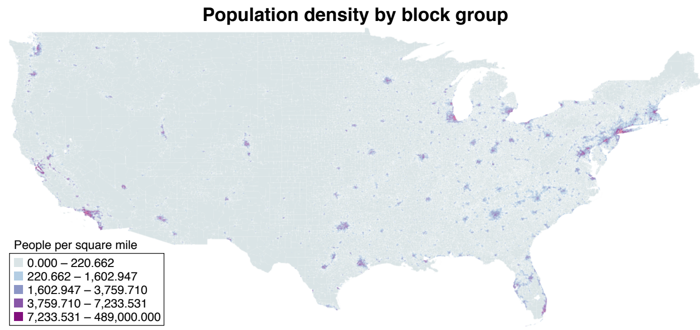

FIGURE 1.1: Quintiles of population per square mile by 'block group' (see below) in the 2010 Decennial Census.

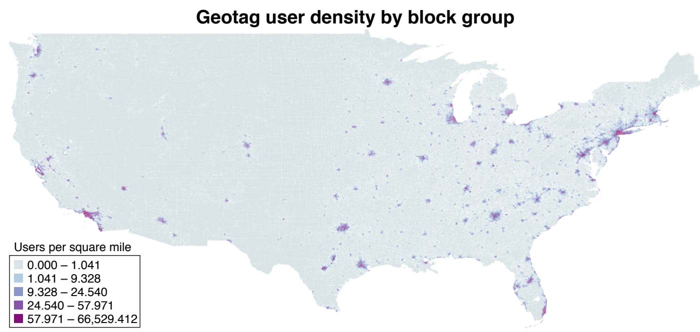

FIGURE 1.2: Quintiles of geotag users, uniquely assigned (see 'mobile users' below) per block group, divided by block group area.

Since this study's publication in 2015, geotagged tweets continue to be used for a variety of substantive purposes, such as studying home alcohol consumption (Hossain et al., 2016), finding vectors of food poisoning (Sadilek et al., 2016), further looking at mobility (Fiorio et al., 2017), making its calls for considering the impact of biases as relevant as ever.

Conversely, this study has contributed to a growing area that seeks to study, understand, and correct for the biases discussed here. Citing my results, Brogueira et al. (2016) did not assume generalizability, and were careful to state results as first and foremost about Twitter users. Brent Hecht, whose 2014 article with Monica Stephens (Hecht and Stephens, 2014) greatly informed the theory of this study, published further work building on this result (Johnson et al., 2016; Thebault-Spieker et al., 2017), including a work looking at how biases affect ultimate results (specifically, how geolocation inference performs more poorly for rural users). Montasser and Kifer (2017) took up weighting schemes to correct for population biases. Citing my result as one motivation, Mowery (2016) looks at the effect of misdiagnoses on attempts to estimate flu prevalence using Twitter. This work is even cited in further survey research (Mellon and Prosser, 2017), showing how new top-down approaches work with survey estimates to illuminate phenomena. In one case, McNeill et al. (2016) found that demographic biases did not significantly affect estimates of local commuting patterns.

In an independent angle of study about the meaning of geotagged tweets, coming some years after the

FIGURE 1.3: Detail of fig. (1.1) for New York.

FIGURE 1.4: Detail of fig. (1.2) for New York.

publication of this study but complementary to it, Tasse et al. (2017) conducted surveys in which they found that geotag tweet users use the tags "consciously and turn geotagging on and off frequently." They suggest thinking of geotagged tweets as "postcards, not ticket stubs": that we should study them not as though they are a record of people's behaviors, but as conscious and selective declarations of having been in a certain place at a certain time. This study, by looking from the perspective of user motivations, provides theoretical reasons that back my finding that geotagged tweets are not representative. This explanation of behavior also explains some of the qualitative results we observed, namely that airports were the heaviest outliers for their ratio of population to geotag tweet users: many people tweet to declare their travels, rather than to identify where they live or spend time.

Hemank Lamba, Constantine Nakos, Jürgen Pfeffer and I used the Twitter API to get a collection of 144,877,685 geotagged tweets from the contiguous US, from which we extracted 2,612,876 unique twitter handles. We uniquely assigned each handle to a *block group*, a geographic designation of the US Census Bureau that is the smallest geographic unit for which Census data is publicly available. We then linked the counts of unique geotag users per block group to the 2010 Decennial Census population counts per block group. I created a statistical test for the null hypothesis that geotag users are randomly distributed over the US population, and found sufficient evidence to reject this null. Using other Census data, I then use a Simultaneous Autoregressive (SAR) model (also known as a 'spatial errors' model) to test some candidate explanatory factors and investigate what is nonrandom about this distribution. This, to my knowledge, was the first paper to use statistical testing to establish population bias along multiple dimensions in geotagged tweets across the entire United States.

1.2 Background and related work

This study followed an increasing body of work about biases in who and what is represented in social media data. The first work with Twitter data was by Mislove et al. (2011), who found an overrepresentation of populous counties and an underrepresentation specifically of the Midwest, an undersampling in counties in the southwest with large Hispanic populations, an undersampling in counties in the south and midwest with large Black populations, and an oversampling of counties associated with major cities with large White populations. However, these findings come from interpretations of distributions and county-level cartograms, rather than from statistical testing, and they rely on the user-defined 'location' field, which has been shown to have many inconsistencies (Graham et al., 2014; Hecht, Hong, et al., 2011). The present study is on the one hand deeper because I use the far higher resolution of block groups and carry out statistical tests, but on the other hand not as general because my findings apply only to characteristics of *geotag users* within the US population rather than to geotag users within the Twitter population, or to Twitter users within the US population. Also worth noting is that Twitter has undergone large changes since the data used by Mislove et al. (2011), both in the governance and management of the platform itself (van Dijck, 2013) and in patterns of user behavior (Liu et al., 2014). Sloan et al. (2013) followed up the work by Mislove et al. (2011) by building a large-scale system for demographics inference in order to make social media data more usable for further sociological research, although they did not look specifically at biases.

More recently, Hecht and Stephens (2014) investigated urban biases across the US, a topic previously investigated on Foursquare by Ishida (2012). Following Goodchild (2007), Hecht and Stephens (2014) adopt the term *Volunteered Geographic Information* (VGI) for this type of data. Collecting 56.7m tweets from 1.6m users over a 25-day period in August and September 2013 and comparing it to Census data, they use a method of calculating a reduced effective sample size in order to correct for spatial dependencies. From this they calculate ratios of users per capita and find a bias towards urban areas, with 5.3 times more geotagged tweets per capita in urban regions as in rural ones, a magnitude even more pronounced in Foursquare data. Longley et al. (2015) investigate biases across a number of factors, focusing on the Greater London area. Using work on forename-surname pairs identifying gender, age and ethnicity, they parse usernames and other profile information to get a collection of estimated names, which they then compare to the 2011 UK Census and find an overrepresentation of young males, an underrepresentation of middle-aged and older females, an overrepresentation of White British users, and underrepresentation of South Asian, West Indian, and Chinese users, although tests of significance are not applied. Theoretically, Blank (2016) makes a similar point, that uneven demographics has implications for what signals are present in Twitter data.

Shelton et al. (2014) carry out a smaller-area case study of geotagged tweets, and do not use statistical modeling, but dramatically illustrate potential harms from relying on biased geotagged tweets. Looking at tweets about Hurricane Sandy in the New York area, they showed that the areas with the most severe disaster relief needs were not necessarily the areas that had the most tweets. Thus, they conclude, a naïve approach of using tweet frequency for directing relief efforts would have disadvantaged people in certain outlying areas, and focused on areas potentially with many complaints but with less dire needs.

Coming from another methodological direction, a nationally representative survey study of smartphone owners ($n = 1,178$) by Pew (Zickuhr, 2013) looks at the demographics of location service users. Overall, 12% of those surveyed reported using what Pew terms 'geosocial' services (which includes geotagged tweets, and excludes informational services like Google Maps). Interestingly, the survey finds the most frequent users of geosocial services are those of low*est* income and middle income; those of low*er* income use it less, and those of upper income use it least. More 18-26 year olds use geosocial services than older users, and almost double the proportion of hispanic smartphone owners (both English- and Spanish-speaking) use geosocial services as compared to non-hispanic white and non-hispanic black smartphone owners. However, out of the respondents who specified which geosocial services they use (n=141), most reported using Facebook (39%), Foursquare (18%) or Google Plus (14%); only 1%, or 1 respondent, used Twitter's geosocial services (i.e., geotagged tweets), such that it is not possible to make inferences about geotag users from the results of this study.

Our paper answered the general call for stronger methodological investigations about the nature of population representation in social media data (Ruths and Pfeffer, 2014; Tufekci, 2014), as well as the specific call for combining geographic data from user-generated sources with non-user-generated sources, such as Twitter data with the Census (Crampton et al., 2013).

1.2.1 Ecological inference

One major limitation of this work that I realized only after publication is the problem of *ecological inference*, inferring individual behavior from group-level data. A canonical illustration given by King et al. (2004) is if we have the number of blacks and whites in voting districts, and we have the number of people in each district who voted, given enough districts can we determine the conditional probabilities of whites voting and blacks voting and not voting? Surprisingly, the answer in general is no; the marginals clearly give bounds on the conditional probabilities, but this turns out to in general not be enough to get the desired point estimates. As O'Loughlin (2000) points out, geographers have tended to skirt the problem of ecological inference by talking about properties of *areas* rather than of individuals within those areas. Since over- or under-representing areas associated respectively with dominant or marginalized demographics may effectively produce the same outcomes as over or under-representing individuals of those demographics (i.e., misrepresentation happening through a mediator of geography), and since ecological inference is far from a solved problem (Freedman et al., 2009a; Freedman et al., 2009b), I take this approach: my results are about properties of *areas* we can predict to be over/underrepresented from using geotagged tweets.

1.3 Method

1.3.1 Data collection

Geo-Coded Twitter Data. From Twitter's Streaming API, we collected 144,877,685 tweets from April 1 to July 1, 2013 using the geographic boundary box $[124.7625, 66.9326]W \times [24.5210, 49.3845]N$. This covers the contiguous US (i.e., the 48 adjoining US states and Washington DC but not Alaska, Hawaii, or offshore US territories and possessions). Consequently, all our tweets are geo-coded with lat/long GPS coordinates. As Morstatter, Pfeffer, Liu, and Carley (2013) report from the Twitter Firehose, about 1.4% of tweets are geotagged; and elsewhere (Morstatter, Pfeffer, and Liu, 2014) they report the Streaming API is more likely to be biased when the response to a query exceeds 1% of the total volume of tweets. Given also that North America accounted for only 22.32% of geotagged tweets in their collection, a fraction consistent with what Liu et al. (2014) report finding in a collection of decahose data covering the time period I consider, it is reasonable to assume that the use of the Twitter API to collect tweets geotagged in the US covers all or nearly all of geotagged tweets within the given time frame and geographic bounds. Similarly, in the 1% sample, Sloan et al. (2013) found 0.85% of the tweets worldwide being geotagged, also less than 1%.

Since the distribution of geotagged tweets over geotag users is characteristically long-tailed (fig. 1.5), with a minority of users sending out the majority of tweets, I decided that the relevant quantity was the number of geotag users rather than the number of tweets. I identified 2,612,876 unique user accounts in our data, which is the basis of my analysis.

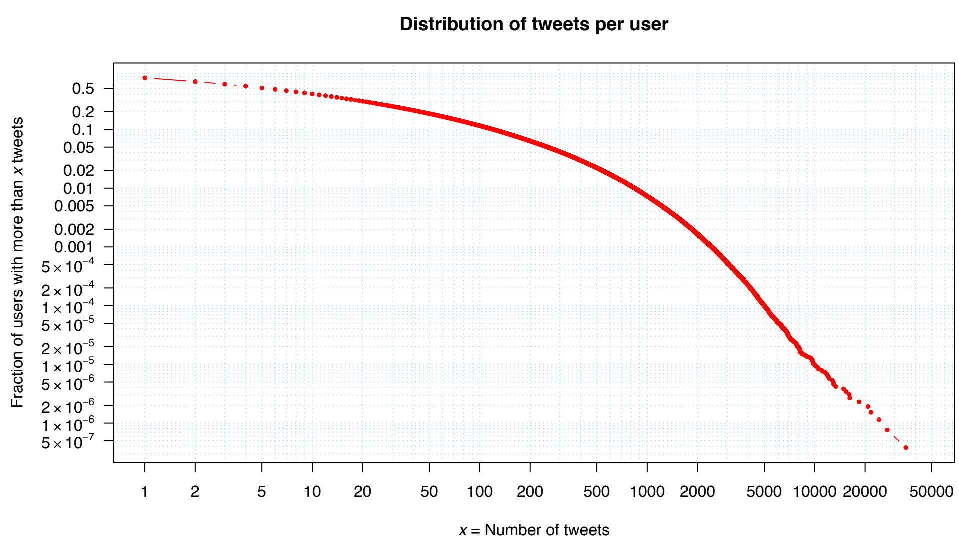

FIGURE 1.5: A long-tailed distribution of the number of users who have tweeted a certain number of tweets, plotted as a survival function (complementary cumulative distribution). Because of this skew, I focus on unique users alone, and ignore the volume of tweets.

Geospatial Data. Each block group has a unique identifier, the 12 digit *FIPS Code*, consisting of identifiers for state (first two digits), county (next three digits), tract (next six digits), and block group (last digit).[2]

The contiguous US plus Washington DC include 215,798 block groups[3] (2010 specification) which range in size from .002 square miles to 7503.21 square miles. Block groups are designed by the Census Bureau to have roughly comparable population sizes. I verified this by noting that, in log scale, the distribution of populations per block group has a symmetric distribution and stable variance (fig. 1.6).

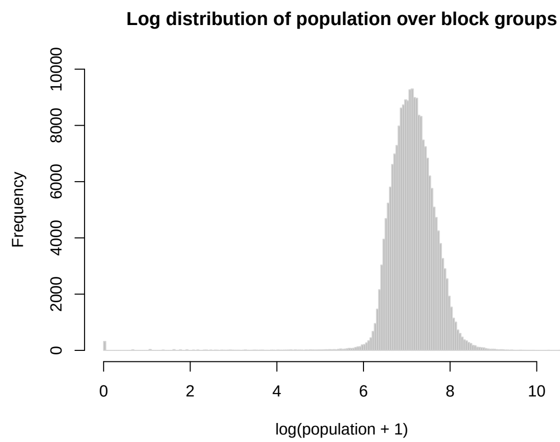

FIGURE 1.6: The Census Bureau designs block groups to enclose population sizes that are comparable. However, it does also allow for block groups with zero population, which is the zero-inflation visible after the add-one smoothing of log(population+1).

For every state, the US Census Bureau provides geographic boundary files ('shapefiles') that includes the GPS coordinates of the borders of every block group within the state. I combined the shapefiles of the 48 contiguous states and the District of Columbia, deleting 364 block groups representing bodies of water (identifiable by being coded as having zero area, and having a FIPS code ending in zero[4]). With Python code (utilizing the `shapely` package) we identified the Census block group into which each tweet fell.

I found 364 block groups with zero area; these also had zero population, and their FIPS codes all ended with 0. These turn out to correspond to bodies of water. While not all have zero geotag users tweeting from within them (for example there are 1,821 users who tweeted from an area of the East River bounded on

[2] Specifically: I use FIPS state codes 01 (Alabama) through 56 (Wyoming), excluding 02 (Alaska) and 15 (Hawaii). The FIPS specification skips 03, 07, 14, 43 and 52 (codes previously allocated for American territories, now depreciated). The District of Columbia is included in the sequence, with FIPS code 11.

[3] Probably due to a rounding error in geographic calculations, I lost three small island block groups (2 in Florida, 1 in New York), such that my n = 215,795.

[4] "Geographic Terms and Concepts - Block Groups", 6 December 2012, United States Census Bureau, `https://www.census.gov/geo/reference/gtc/gtc_bg.html`, accessed 3/2015. Note that this page references block groups "beginning with zero", but since the 'block group' part of a FIPS code is only the last digit, this should be interpreted as, "FIPS codes ending in zero.'

one side by the Brooklyn Bridge), for comparison with (potentially) populated areas I removed these water block groups (tab. 1.1).

FIPS code	Users	Description
06 083 990000 0	2,526	Channel Islands, CA
36 061 002500 0	1,821	Brooklyn Bridge, NY
51 810 990100 0	1,643	Coast off Virginia Beach, VA
36 061 009900 0	1,629	Chelsea Piers, NY
24 003 990000 0	1,373	Coast off Annapolis, MD

TABLE 1.1: Most popular bodies of water for tweeting from.

Socioeconomic Data. While the ideal would be to have rich and timely demographic data about the users who sent the tweets in our data (as attempted in Sloan et al., 2013), this was not realistic to collect for 2.6m users. But by aggregating data at the level of block groups, I can link Twitter data to the enormously rich demographic data the Census Bureau makes available at this level. I primarily use data from the 2010 Decennial Census, which I supplement with median income (not available in the Decennial Census) estimates from the 2009-2013 American Community Survey. For this ACS data, there were 1,224 block groups with missing values for median income, few enough that I filled these out as zeros rather than using imputation or smoothing. I also set 21 block groups with the value "2,500-" to 2,500, and 2,651 block groups with the value "250,000+" to 250,000. The 2009-2013 ACS had 54 block groups in the contiguous US whose boundaries (and FIPS) codes were from the 2000 Census, for which I found equivalent block groups in the 2010 Decennial Census to which to map. While the ACS 1-year estimates are more timely, they are more sparse and only at the county level (U.S. Census Bureau, 2008), and I decided to prioritize the accuracy and completeness of values in the Decennial Census for this analysis. I similarly decided to not use the ACS 2009-2013 estimates for population quantities as there was more missing data, and there was high correlation between the 5-year estimates and 2010 Decennial Census figures across variables (generally around .95). Still, prioritizing timeliness over completeness, and looking at the county level with 2013 ACS 1-year estimates, may be the focus in future analysis.

The Census Bureau also makes estimates of the same quantities at 1-, 3-, and 5-year intervals through the American Community Survey, and there are estimates from 2013; however, 1-year ACS estimates only cover areas with populations over 65,000 and only at the level of counties (U.S. Census Bureau, 2008), only the 5-year estimates cover all population sizes and go to the block group level. The 5-year estimates were only slightly more contemporaneous and I found them to include far more missing data. I thus decided to prioritize resolution and completeness[5] over timeliness for the greater power, and because I assume that shifts in population would not be enough to change the basic dynamic between population and tweets. However, this is a testable assumption, and future work may wish to look at the county level in order to study geotagged tweets with more timely demographic estimates.

[5] "American Community Survey: When to use 1-year, 3-year, or 5-year estimates", 23 March 2015, United States Census Bureau, http://www.census.gov/acs/www/guidance_for_data_users/estimates/, accessed 3/2015.

Mobile users. My construct of interest is the *number of potential geotag users*, for which population is the available proxy; there are cases where there are more geotag users than population, which points to tourists or, more generally, mobile users, as a complicating factor (Hecht and Stephens, 2014). I counted 18,835,284 distinct user-block group instances (i.e., if I were to use the number of unique users appearing within each block group, I would have inflated the user count by six times).

Hecht and Stephens (2014) provide a useful review of techniques to uniquely assign users to a single geographic region. They identify two candidate techniques: temporal, where a user must send at least two tweets a set number of days apart in a region for the user to be located uniquely in that region, and 'plurality rules,' where the most frequently tweeted-from region is taken as the unique location of the user. Checking the 'location' field fails because of the low quality of the information there (Hecht, Hong, et al., 2011). As one other option, Wang, Chen, et al. (2014) use the location of the first geotagged tweet sent by a user as the location of the user. This is the simplest, but also has no motivation beyond convenience.

Despite the drawbacks of plurality not accounting for people local to two regions, my comparison is with the US Census which also does not account for this possibility. However, another problem is that foreign tourists are not counted in the US Census (unlike domestic tourists, who reside in some US block group), and of which there were 70m in the US in 2013[6]. This is substantial when compared to the total 2013 US population of 316m[7] (of which 307m are counted in the block groups I use). If many foreign tourists send geotagged tweets, it would introduce unaddressed bias; since our data collection only had geotagged tweets in the US, short of massive additional data collection I am unable to identify foreign tourists (such as by looking at the proportion of geotagged tweets outside of the US). This is a potential problem in my analysis that may be a topic for clarification in future work.

Additionally, I filter users by the number of tweets, considering only those with a certain number of tweets.[8] As the distribution of tweets per user (fig. 1.5) is smooth and has no natural break point, I arbitrarily pick 5 and 10 as cutoffs to use alongside all users.

1.3.2 Statistical models

Random distribution over population. The basic relationship in which I am interested is between population and geotag users. In order to make a concrete test for random distribution, I suggest a model where there is a linear relationship between the population count and the number of users, i.e., users are drawn from the population at a constant rate subject to some noise. We can imagine the noise is heteroskedastic, which suggests the following data-generating process over population P, users U, and mean-zero noise term ε:

$$U = \alpha P + \varepsilon P \tag{1.1}$$

[6]"2013 Monthly Tourism Statistics: Table C - Section 1: Total Visitation, Canada, Mexico, Total Overseas, Western Europe Non-Resident Visitation to the U.S. By world region/country of residence 2013", n.d., `http://travel.trade.gov/view/m-2013-I-001/table1.html`, accessed 3/2015.

[7]"Population, total", 2015, The World Bank, `http://data.worldbank.org/indicator/SP.POP.TOTL`, accessed 3/2015.

[8]I thank an anonymous reviewer for this fruitful suggestion.

I transform both users and population to stabilize their variances, so this then becomes

$$\log U = \log \alpha + \log P + \log\left(1 + \frac{\varepsilon}{\alpha}\right) \tag{1.2}$$

Then, consider the linear model

$$\log U = \beta_0 + \beta_1 \log P + \varepsilon' \tag{1.3}$$

If eqn. (1.1) described the true data-generating process, from eqn. (1.3) we should get that $\hat{\beta}_1 = 1$, and then $\exp(\hat{\beta}_0)$ would estimate the value of the proportion α. That is, the $\log \alpha$ term is the intercept of the regression of $\log P$ onto $\log U$, and $\log\left(1 + \frac{\varepsilon}{\alpha}\right)$ is a mean zero error term now independent of P, and we have a null hypothesis $H_0 : \beta_1 = 0$. While this may seem unrealistic as a null model, other quantities that we would believe are randomly distributed proportional to population indeed match this. For example, I regressed log population onto log males and found it to be meaningful (presented below under results). With this validation, I argue that the model of eqn. (1.1) is a reasonable way of representing a quantity being randomly distributed over the population. Note that my interest is not in fitting this specific model and interpreting the parameters, but just having a way to test the null hypothesis of random distribution. Note also that I originally sought to compare log population density to log geotag user density as a way of treating measures on different block groups as equivalent (given that block groups are already designed to somewhat control for the variance in population density), but found that it produced excellent fits that did not disappear when the data was shuffled, suggesting that the dividing by area created artifactual relationships.

Model specification For comparison with analyses of race and Hispanic populations (Mislove et al., 2011; Zickuhr, 2013), I use Census variables[9] P0030001 through P0030008 and P0040001 through P0040003. For comparison with analyses by age (Longley et al., 2015; Zickuhr, 2013), I use P0120003 through P0120049 and aggregate across gender into the same age bins as in Zickuhr (2013). Existing analyses by sex (Longley et al., 2015; Zickuhr, 2013; Mislove et al., 2011) is based on name-based inference or survey data; I decided that, while the Census does have sex data, the even distribution of sex across the US means that the sex ratio of a block group is not a meaningful proxy for geotag users who live there. For comparison with analyses of urban and rural populations (Hecht and Stephens, 2014; Zickuhr, 2013), I use P0020002 through P0020005.[10]

Thus, in total, I include terms for populations, the black population, the Asian population, the Hispanic/Latino population, the rural population, and respective populations of people ages 10-17, 18-29, 30-49, 50-64, and 65+. For all of these, I stabilize variance with a log transformation with add-one smoothing. I include median income (Zickuhr, 2013), and test for a northern/eastern effect by including the (demeaned) latitudes and longitudes of block group centroids, and for a coastal effect by including terms for latitude and longitude squared.

[9]"Census Data API: Variables in /data/2010/sf1/variables", 2010, `http://api.census.gov/data/2010/sf1/variables.html`, accessed 3/205.

[10]The Census API returned zero values for these, so I manually downloaded the variables of "P2. URBAN AND RURAL" for each state individually from `factfinder.census.gov`.

Spatial autocorrelation. Discretization into uneven geographic units (as block groups certainly are) can cause statistical artifacts. Specifically, if the divisions do not correspond to the contours of the underlying spatial process (and there is little reason to believe they would), there will be dependencies between proximate geographic areas, and not accounting for this can inflate the R^2 statistic, shrink standard errors, and give misleadingly significant results. I use the standard statistic for measuring spatial autocorrelation, Moran's I,

$$I = \frac{n}{\sum_i \sum_j w_{ij}} \frac{\sum_i \sum_j w_{ij}(X_i - \overline{X})(X_j - \overline{X})}{\sum_i (X_i - \overline{X})^2} \tag{1.4}$$

This is the empirical covariance, appropriately normalized, of the values of variable X between geographic units i and j. $W = [w_{ij}]$ is an $n \times n$ matrix of weights, discussed below. Rather than exploring autocorrelation in individual variables, I look for spatial autocorrelation in the residuals of a linear model (Anselin and Rey, 1991). For management of spatial data and implementation of computation and estimation for spatial models, I used the R package `spdep` (Bivand and Piras, 2015; Bivand, Hauke, et al., 2013).

Moran's I is well-investigated in terms of its asymptotic and theoretical properties (Gaetan and Guyon, 2012). It is tested under a null hypothesis of zero autocorrelation, either using assumed normality along with analytic forms of the higher moments of the statistics under normality or else permutation testing, which requires no distributional assumptions and which may be approximated by MCMC methods (Gaetan and Guyon, 2012). As I found that my variables and residuals were approximately normally distributed, I used tests based on asymptotic normality, for which the higher moments have analytics forms, rather than MCMC methods that make no distributional assumptions (Gaetan and Guyon, 2012) as the number of block groups made permutation testing computationally expensive. Fortunately, most of my variables had symmetric distributions with stable variance in the log scale.

Spatial autocorrelation is not inevitable, and indeed evidence of spatial autocorrelation may be due to model specification that can be eliminated by adding additional controls (Bivand, Pebesma, et al., 2013); alternatively, if spatial autocorrelation is not a quantity of interest, including it in a regression is itself a control. While we may test for spatial autocorrelation in the variable of interest if spatial dependencies are of explicit interest, a way more appropriate to my bivariate model is to look for spatial autocorrelation in the residuals of a linear model (Anselin and Rey, 1991). For management of spatial data and implementation of computation and estimation for spatial models, I used the R package `spdep` (Bivand and Piras, 2015; Bivand, Hauke, et al., 2013). I have found little work applying spatial models developed in econometrics, epidemiology and ecology to geographically dispersed social media data (an exception is Sylvester et al., 2014)), and hoped to bring such models to wider attention as thematically well-suited for analyzing issues of bias and representation (although, since the publication of this article, I have not seen this happen).

Weights matrix. Measuring spatial autocorrelation requires a 'weights matrix' of adjacencies between geographic units. There are multiple ways to generate this, and the choice of how to do so represents a substantive decision based on the problem at hand (Gaetan and Guyon, 2012). However, given that we do not know in advance the form of the spatial autocorrelation, in practice we can test for autocorrelation over different choices of weights matrices to see which is most appropriate (Anselin, Sridharan, et al., 2007). Thus, I consider the following weights matrices:

- Queen contiguity (regions sharing a corner or edge are adjacent, equivalent to 8-connectivity in image processing);
- Rook contiguity (regions sharing an edge are adjacent, equivalent to 4-connectivity in image processing)
- k-nearest-neighbors for $k = \{2, 3, 4, 5, 6, 7, 8\}$, calculated from the midpoints of block groups.

For the contiguity cases, I consider both row-normalized (which normalizes the 'effect' of each neighboring unit such that they sum to one) and binary (which gives greater possibility for autocorrelation between a unit and its neighbors for units with more neighbors). In the row-normalized case, I also employ Lagrange Multiplier tests developed in that contest (Anselin, 2002).

Spatial errors model. I model the relationship between population and geotag users using a Simultaneous Autoregressive (SAR) model, which is where one or more terms in the regression are correlated with itself. The main autoregressive model assumes that the residuals of unit i are correlated with the residuals of those units j adjacent to i, which is known in econometrics literature as a spatial errors model. The adjacencies are indexed exactly by the terms of the weights matrix. This gives the following two equations,

$$\mathbf{Y} = \mathbf{X}\beta + \mathbf{u} \tag{1.5}$$

$$\mathbf{u} = \lambda \mathbf{W}\mathbf{u} + \varepsilon \tag{1.6}$$

where u are the correlated residuals, $\varepsilon \sim \mathcal{N}(\mathbf{0}, \sigma^2 \mathbf{I})$ are the uncorrelated error terms, and the coefficient λ is the 'spatial multiplier' that captures the strength of the spatial autocorrelation (Anselin, 2002). We can rewrite these in a single form as either

$$\mathbf{Y} = \mathbf{X}\beta + (\mathbf{I} - \lambda\mathbf{W})^{-1}\varepsilon \tag{1.7}$$

or, substituting eqn. (1.5) back into eqn. (1.6),

$$\mathbf{Y} - \lambda\mathbf{W}\mathbf{Y} = \mathbf{X}\beta - \lambda\mathbf{W}\mathbf{X}\beta + \epsilon \tag{1.8}$$

The terms $\lambda\mathbf{W}\mathbf{Y}$ and $\lambda\mathbf{W}\mathbf{X}\beta$ are known as spatial lags. While there are other SAR models, I use spatial errors as the simplest to interpret and the most appropriate for my purpose.

A spatial errors model lags the explanatory and response variables by the same multiplier. Other SAR models use lags differently; a different coefficient on the spatial lag for $\mathbf{Y}$ and the spatial lag for $\mathbf{X}\beta$ yields a spatial Durbin model, $Y = \rho\mathbf{W}Y + \mathbf{X}\beta + \mathbf{X}\beta\gamma + \mathbf{u}$, and if we only include a spatial lag on $\mathbf{Y}$, it becomes a spatial lag model, $Y = \rho\mathbf{W}Y + \mathbf{X}\beta + \mathbf{u}$. Estimation of the models results in different numerical issues, with the spatial errors model being the most straightforward to compute and to interpret (Bivand, Pebesma, et al., 2013), and the most appropriate as I only seek to account for spatial autocorrelation and not necessarily to measure it.

I originally sought to compare log population density to log geotag user density as a way of treating measures on different block groups as equivalent (given that block groups are already designed to somewhat control for the variance in population density), which I found generated extremely good fits and extremely significant coefficients. However, when I shuffled the data to break the relationship (I both tried shuffling densities, and shuffling counts and dividing them by shuffled areas), the estimated coefficients had the same values, and the R^2 remained the same, suggesting that the model fit in the case of density is an artifact of how transformation combines the underlying densities. In contrast, for my current model, I found a shuffle test broke the significance of the slope term, which is what should happen (in which case, the estimated intercept becomes the logarithm of the mean of the response).

Zero values. Zero values frequently cause problems, especially when transforming to log scale. I considered removing all block groups with zero population, and all block groups with zero geotag users, as these required padding that caused some data artifacts (visible in plots below). However, I found that excluding them only improved measures of model fit, such that including them (via add-one smoothing) leads to a more conservative estimate.

1.4 Results and discussion

1.4.1 Observational results

The block groups with the highest number of distinct users (before users are assigned uniquely) are major international airports and major tourist attractions (table 1.2).[11] The inclusion of several international airports on the list suggests that geotagging tweets during the process of travel is a common user behavior. There were some areas with zero population but nonzero users; out of these, the ones with the highest counts of distinct users are mostly the same: major airports and parks.[12]

Conversely, there were only 67 block groups from which nobody sent geotagged tweets; only 30 of these also had no population (these were national forests, minor airports, areas off highways, etc.). Of those that did have a population, the most populous was a block group with a population of 4,854 within San Quentin State Prison in California. The second-most populous block group is also a Corrections Department building in Texas, and third is a state prison in California (although not all prisons lack geotag tweet users; the block group of Rikers Island in New York has geotagged tweets from 22 users).

Out of the 2,612,876 unique users I identified, 2,216,219 (84.82%) had a single block group from which they tweeted most frequently. The others had ties for which block group was the highest; for these users, I uniquely assigned them to one of their block groups by randomization. I tried analyses on just the 84.82% as well, but found it made little substantive difference in the results.

[11] Block groups may be looked up by their FIPS code at `http://www.policymap.com/maps`.

[12] Interestingly, Central Park has a nonzero population (of 25), as do some airports. Some other tourist attractions (e.g., Universal Studios) also appear.

TABLE 1.2: Block groups from which the most users have sent geotagged tweets.

FIPS code	Users	Description
32 003 006700 1	28,280	Las Vegas Strip
06 037 980028 1	23,100	Los Angeles Int'l Airport
32 003 006800 4	16,748	McCarran Int'l Airport
13 063 980000 1	15,481	Atlanta Int'l Airport
12 095 017103 2	15,392	Walt Disney World
36 081 071600 1	15,067	JFK Int'l Airport
11 001 006202 1	14,906	National Mall
36 061 014300 1	14,605	Central Park
06 059 980000 1	14,576	Disneyland
17 031 980000 1	13,610	Chicago Int'l Airport

In the terminology of Guo and Chen (2014), the most active accounts belong to 'non-personal users.'[13] In this case, the most active tweeter (44,624 tweets) seems to be a commercial service for travel, the second-most active (35,025) is an automatic news updater in Florida, etc. Starting from the 13th most active tweeter, with 12,922 tweets, there were accounts that appeared on inspection to be personal ones. As for number of block groups traversed, the top 'traveler' (23,547 block groups) is the same as the top tweeter, and others are similarly non-personal users. Across block groups, it is not until the 18th most mobile user, traversing 1,209 block groups, that there is a personal user.

How much mobility is there between units? Figures 1.7 and 1.8 show respectively that while there is minimal mobility between states, with only 22.39% of users sending geotagged tweets from more than one state and only 7.83% send from more than 2. However, there is a great deal of mobility between (possibly neighboring) block groups, with 65.24% of users sending geotagged tweets from more than one block group.

How well does unique assignment do? As one check, I consider the ratio of geotag users to population; there are 509 block groups where this ratio is greater than 1 (for users with 5 or more tweets only, there are 353, and for users with 10 or more tweets only, there are 290), indicating either the failure of population as proxy for potential geotag users or of the method of assigning mobile users. As I found the block groups with the largest ratios to be airports, it seems to be a case of the latter.

The largest ratio is in the block group containing Los Angeles International Airport, 1365.5 to 1 (558.25 to 1 and 287.25 to 1 for the two respective filter levels). The second-highest ratio is the block group in Manhattan containing Bryant Park, and the remainder of the top five are more major airports. This points to the method of unique assignment unsuccessfully handling tourist destinations even with filtering for a minimum number of tweets.

[13]They find that only 2.6% of geotag users are non-personal. This should be small enough to have no effect on results, so I did not employ filtering. However, this may be considered in a future work.

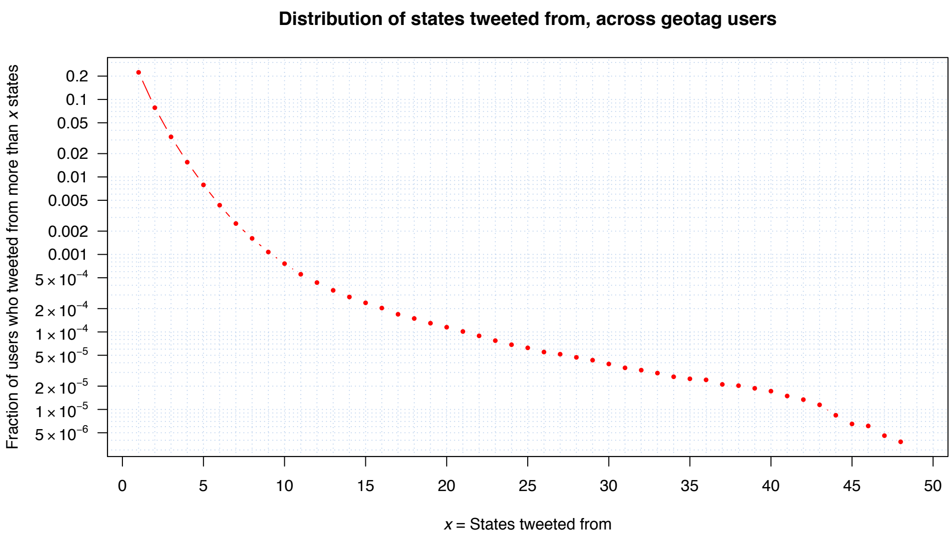

FIGURE 1.7: A full 77.61% of geotag users in our set tweeted only from one state, and having tweeted from 5 or fewer states accounts for 99.21% of users.

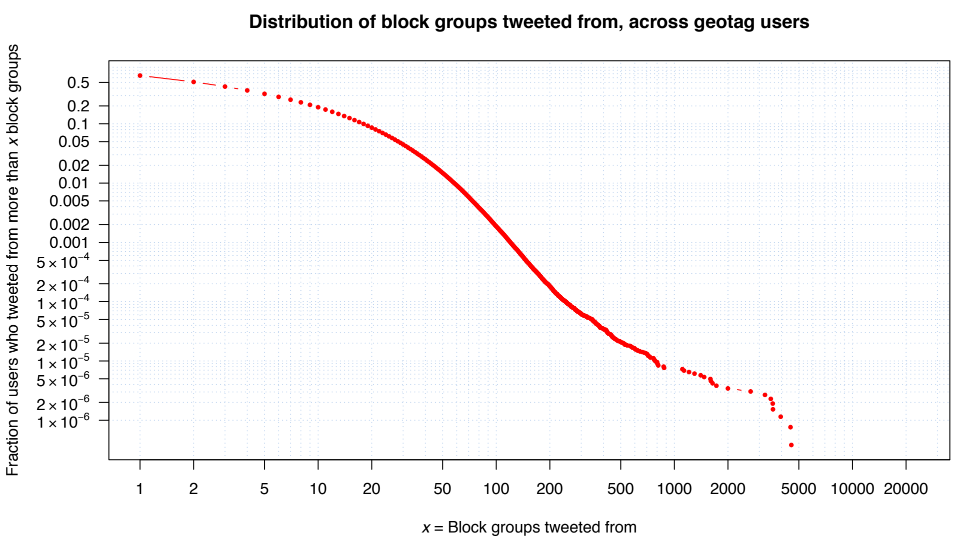

FIGURE 1.8: 34.76% of geotag users tweeted only from one block group. 27 or fewer block groups were 95%, 50 or fewer block groups were 99%. One outlier at 23,547 excluded.

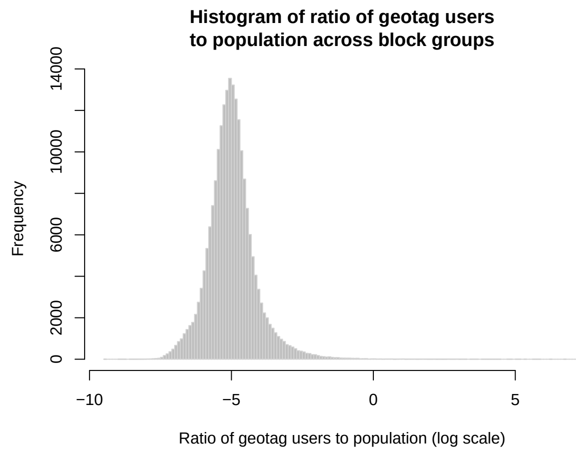

FIGURE 1.9: Ratios above zero are obvious failures of the metric, as the number of uniquely assigned geotag users should not exceed the population, but the distribution is smooth and symmetric.

1.4.2 Bivariate regression model

I first test my null hypothesis of a linear regression yielding a coefficient of 1 to the logarithm of the population. Looking at the plot of the relationship of the logarithm of the two (fig. 1.11), there is a faint linear relationship, although the slope does not appear to be 1. An OLS regression fits slope $\hat{\beta}_1$ = .4916 (.002996) and intercept $\hat{\beta}_0$ = -1.219 (.02143),[14] although recall that the standard errors are not reliable under spatial autocorrelation.

Compare this plot to the plot of the test case mentioned earlier, the distribution of males over the population, pictured in fig. (1.10). The true ratio of males to total population across the block groups we consider is .4915; according to my model, the exponential of the intercept should be this, and the coefficient of the log population term should be 1. Indeed, log(.4915) is within the 95% confidence interval (log(.4914), log(.4962)), and 1 is just outside the 95% confidence interval (.9980, .9994), but this is without accounting for how spatial autocorrelation shrinks estimated standard errors. The R^2 value of this model is also impressive at .975, although under spatial autocorrelation R^2 is inflated thereby not interpretable. Overall, my model fits the relationship of males to population exactly as we would expect it to fit to something randomly distributed over the population.

Using this as a validation of my statistical test, we can strongly reject the null hypothesis that $\hat{\beta}_1 = 1$ even without correcting for spatial autocorrelation. And the R^2 value for this regression is a paltry .109, too small to worry about being inflated. Thus, we can conclude that geotag users are not randomly distributed over the US population, and indeed that the population count is not very informative about the number of geotag users.

[14]Filtering for only those users who have 5 or more tweets and for those users with 10 or more tweets, the respective fitted slopes are .5192 (.002932) and .5136 (.002786).

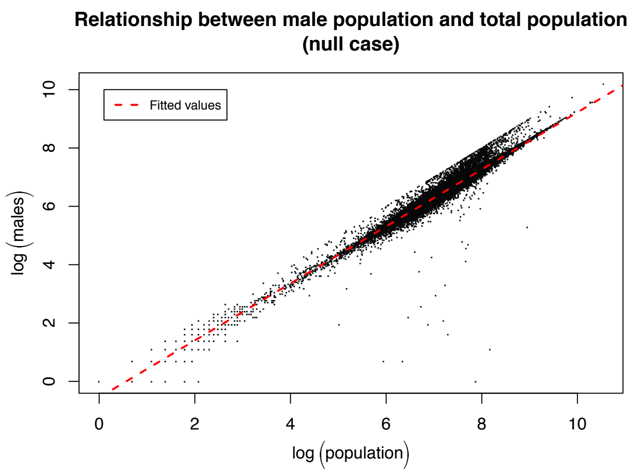

FIGURE 1.10: The relationship between males and total population behaves exactly as we expected of a quantity randomly distributed over the population, making it an effective null model against which to compare the observed distribution of geotag users.

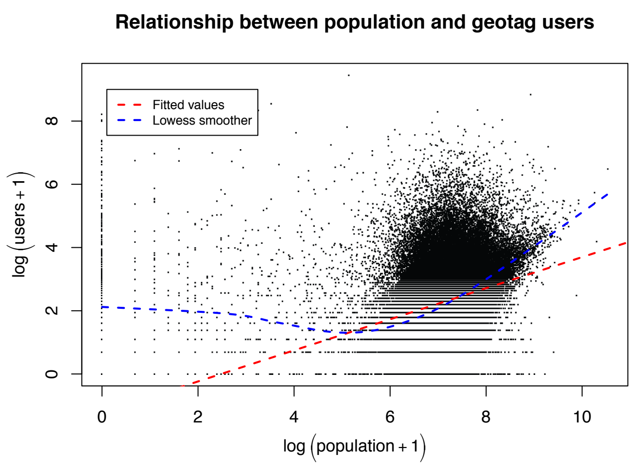

FIGURE 1.11: Eliminating zero-count observations reduces the artifacts visible at $x = 0$ and $y = 0$ but does not substantially change the fit.

1.4.3 Weights matrix and spatial autocorrelation

Testing the residuals in my basic model for spatial autocorrelation using Moran's I against all weights matrices considered above, I find the results reported in table (1.3).

I found identical results of Moran's I for binary weights matrices and row-normalized weights matrices in the k-nearest neighbor case. For the two contiguity cases, row normalization made a difference, and I list both values. In all cases, an asymptotic test against the expected value of 0 was significant at $p < .0001$. The

TABLE 1.3: Selected Values of Moran's I in residuals

	Population vs Users	Population vs Male
2nn	.3699	.2336
4nn	.3550	.2142
6nn	.3398	.1996
8nn	.3270	.1883
Rook	.4166 (b)	.2125 (b)
	.3992 (rn)	.2201 (rn)
Queen	.4151 (b)	.2097 (b)
	.3919 (rn)	.2154 (rn)

For the Rook contiguity case and the Queen contiguity case, binary (b) and row-normalized (rw) weights gave different values.

autocorrelation in the population-user model is stronger than in the 'null' population-male model. It appears, then, that the spatial autocorrelation is strong enough that the choice of weights matrix is not critical. For the population to user model fit on counts of users with 5 or more tweets, or 10 or more tweets, the spatial autocorrelation was similar (generally lower, but still higher than the autocorrelation of population vs. male).

1.4.4 Spatial errors model

The maximum likelihood method of fitting a SAR model involves computing the log determinant of the $n \times n$ matrix $|I - \lambda W|$, which is infeasible at my n of over 200,000. An alternative method finds the log determinant of a Cholesky decomposition of $(I - \lambda W)$, although this then requires W to be a symmetric matrix (Bivand, Pebesma, et al., 2013). Since all of the candidate weights matrices picked up spatial autocorrelation at a significant level, I use a binary contiguity weights matrix. I tried both Rook and Queen, and they gave comparable fits, so I report only for Rook (1.4).

The spatial multiplier term is significant, although neither the coefficients nor the standard errors are substantively different than the previous model. However, calculating Moran's I on the residuals of this model gives a value of -.02367, with a p-value of 1, meaning we have successfully controlled for spatial autocorrelation.

I then investigate the full model specified above. I interpret this model in the standard way: for a log transformed explanatory variables X_i, a 1 percent change is associated with a β_i percent change in $\mathbf{Y}$. I present the results of the regression on counts of only those users with 5 or more tweets. This is shown in table (1.5).

As before, testing for spatial autocorrelation finds no significant amount, with a p-value of 1.

I considered using the youngest ages (ages 0-9) as the omitted category in order to accord with how the Pew study Zickuhr, 2013 does not cover usage by children. However, it is more appropriate to exclude ages 18-29, as it theoretically may be considered the baseline category. Furthermore, Pew does have an earlier

TABLE 1.4: Spatial errors basic model, binary Rook contiguity

	Dependent variable:
	log(user + 1)
log(population + 1)	.4401*** (.002655)
Intercept	−1.138*** (.01890)
$\hat{\lambda}$:	.1107***
LR test value:	73,375
Numerical Hessian $\widehat{\text{se}}(\hat{\lambda})$:	8.4241e−06
Log likelihood:	−222,020.8
ML residual variance (σ^2):	.4206
Observations:	215,795
Parameters:	4
AIC:	444,050
Note:	***p<.0001

study on geosocial service usage by children ages 12-17 Zickuhr, 2012, finding that teenagers and adults used geosocial services at the same rates, about 18% in 2012. Ideally we would aggregate the Census data into age bins of 0-11 and 12-17 to correspond to those of Pew; unfortunately this is impossible from Census data, as the Census provides counts for ages 0-4, 5-9, 10-14, and 15-17. Coding 0-9 and 10-17 is the closest we can get.

Controlling for other factors, population still has a significant, positive, and large effect.[15] The hypothesis test I built, by which I rejected a random distribution over the population, is still valid; the revision is about population having an effect versus not having an effect, but either way it is not the only effect. Formally, adding in other factors indeed improve the model fit: running a bivariate spatial errors model with > 5 users against only population, I get an AIC of 444,000 (versus 423,530), and a likelihood ratio test of the bivariate (i.e., restricted) model against the full model rejects at the p<.0001 level the null that the restricted model is correct.

The term for area included as a control is significant, with a one percent rise in block group area associated with a 16.56% rise in geotag users. It seems here that size overcomes the effects of population density (as mentioned above, block group population has stable variance only in log scale even though block groups are designed to enclose populations of roughly comparable size). Consistent with survey findings (Zickuhr, 2013), a 1% larger Hispanic/Latino population is associated with 3.78% more geotag users. However, the

[15] In the original paper, which used ages 0-9 as the reference category, I had found that population only lost its significance for users with 5 or more geotagged tweets; I theorized that this was an appropriate cutoff (to exclude users who only tried geotags and to not include only power users) and thus privileged the model outputs for this dependent variable. However, with this result, and generally how population is significant and large across different variations of the model, I revise my previous conclusion. Also notable is that now the effect of median income is no longer significant. Given that its effect size was weak before, this is not too much of a change, but to see no significant effect (not even a weak effect) is still surprising.

TABLE 1.5: Spatial errors full model with ages 18-29 as the omitted category, binary Rook contiguity, users with >5 tweets only. Revised from published version.

	Dependent variable:	
	log(user + 1)	s.e.
log(population + 1)	.4277***	(.006479)
log(area)	.1656***	(.001809)
log(asian + 1)	.1249***	(.001603)
log(black + 1)	.06130***	(.001483)
log(hispanic + 1)	.03787***	(.002112)
latitude (demeaned)	.02522**	(.007352)
longitude (demeaned)	.01962***	(.002864)
latitude^2	-.0003490**	(.00009910)
longitude^2	.00006872***	(.00001475)
median income ($10K)	.001234	(.00068035)
log(rural + 1)	-.05791***	(.001119)
log(ages 00-09 + 1)	-.01104	(.005509)
log(ages 10-17 + 1)	-.05442***	(.005389)
log(ages 30-49 + 1)	-.05466***	(.007479)
log(ages 50-64 + 1)	-.1793***	(.007126)
log(ages 65 and up + 1)	-.2585	(.003857)
Intercept	.1497	(.1998)
$\hat{\lambda}$:	.1039***	
LR test value:	39,934	
Num. Hessian $\widehat{\mathsf{se}}(\hat{\lambda})$:	.0003735	
Log likelihood:	-211,745	
ML resid. var. (σ^2):	.3871	
Observations:	215,795	
Parameters:	19	
AIC:	423,530	
Note:	**p<.001; ***p<.0001	

effect size is smaller than either that of the Asian population (a 1% rise is associated with a 12.49% rise in geotag users) and, in contrast to survey findings, that of the Black population (a 1% rise is associated with 6.13% rise in geotag users). This might point to the Pew sample not including enough Twitter users, as there is an active Black community on Twitter that had gained scholarly attention even when this article was published (Clark, 2014; Florini, 2014; Sharma, 2013).

The latitude, both in linear and quadratic terms, is significant at the p<.001 level. Thus, after controlling for population size and longitude, being further north or towards the mean latitude of the US is associated with more geotag users. While the effect size of latitude is larger than those of longitude, so are the standard errors (hence being significant at a lower level), hence the true effect of latitude is not necessarily stronger.

While I tried to test for nonlinearity in income, inclusion of a squared term for median income made the matrix computationally singular; however, inspecting the bivariate relationship did not yield any evidence for a nonlinear effect, and the linear effect is weak and nonsignificant (a $10,000 rise in the median income is associated with a 0.12% rise in the number of geotag users).

Consistent with findings about urban biases (Hecht and Stephens, 2014), I find that a 1% higher rural population is associated with a 5.79% decrease in the number of geotag users.

There is a negative effect from having a higher population of any other age group except for ages 30-49, which surprisingly is associated with slightly more geotag users.[16] Under this choice of omitted category, the size of the population of ages 65 and up and of the population of ages 0-9 are no longer significant. In contrast to Zickuhr (2012), I find that there is a significant difference between the number of teenage users and the number of adult users.[17] However, the different age bins make these results not exactly comparable, since it is certainly possible that children of ages 10-11 use geotags at a far lower rate than those of ages 12-17, dragging down the mean of a category that combines the two groups. It is surprising that the negative effect from population of ages 65 and up is not significant; as I speculated before, this might be due to mixed populations, for example places that are popular for retirement also being popular for tourism.

As is usual with logarithmic dependent variables, the intercept is not particularly interpretable as it would be a prediction for a block group at the center of the US with a population of 1.

Running the SAR model using all users, instead of just those with 5 or more tweets, produces similar results, except that log population is significant with coefficient -.04196 (.007858); this suggests a nonlinear effect, and indeed, an added squared term for the log population came out as significant and positive at .06329 (.0008394). This points to some noise for those people who only 'try out' geotagged tweets but do not adopt their use that disappears if we maintain a minimum tweet threshold. When running the model on only those users with 10 or more tweets, results are again similar except the longitude squared term is no longer significant (p = 0.1870), and the latitude term becomes significant (p = 0.02017). This might be from the coasts having more users who try out geotagged tweets for a longer period of time before choosing not to continue. These subtle differences point to opportunities for modeling the demographics of different types of users (as determined by number of geotagged tweets or other factors), although I do not explore them more here.

[16] This is inconsistent with how using ages 0-9 as the omitted category in the original paper gave a larger coefficient for 18-29 year olds than for 30-49 year olds, pointing to possible issues with model misspecification.

[17] Zickuhr (2012) compares 12-17 with 18 and up. To mimic this, we-coded ages into only three categories of ages 0-9, ages 10-17, and ages 18+, and re-ran the model using 18+ as the omitted category. The significance and direction of the coefficients for ages 0-9 and 10-17 were identical to the full model.

1.5 Conclusion

Geotag users are not representative of the US population. Despite the volume of geotagged tweets and their impressive coverage (there were only 67 block groups out of 215,795 with no geotagged tweets), the users who send geotagged tweets are nonrandomly distributed over the population in subtle ways. These include predicable and already established biases towards younger users, users of higher income, and users in urbanized areas, as well as surprising biases towards Hispanic/Latino users and Black users that, in the latter case, have not been seen in large-scale survey research. I also demonstrate an unsurprising but previously unreported coastal effect, where being located on the east or west coast of the US is associated with more geotag users. Geotag users may not be a random sample of the population of any given block group, but given the fine level of detail and large-scale demographic variability, the demographics of a block group is a reasonable proxy for the demographics of geotag users located in that block group. Certainly, even with complications of uniquely assigning mobile users, it is enough to establish the nonrandom distribution of geotag users, and some candidate biases.

While from this study, I am unable to say whether or not geotag users are representative of the *Twitter* population; they are a self-selecting group, and my analysis is further not able to say anything about why certain demographic profiles would be more likely to select in (or what other causal features there may be behind the decision of some people of a given demographic to use geotagged tweets but not others). But the interesting question that can be addressed with the given data is whether geotagged tweets can be a useful proxy for the *general* population within the US. This is a critical question because geotagged Tweets are an enormously popular source of data for studying a wide variety of social and human phenomena. For future work, I emphasize that findings using geotagged tweets should not be assumed to generalize, and conclusions should be restricted only to geotag users with their population biases.

Future Work There are a number of directions for future work. The most obvious is to update the data and models with more recent ACS estimates, and geotagged tweets collected in the same year. In terms of model terms, in cases where it is possible to measure differences in usage by gender, there are strong reasons to hypothesize that fewer women than men use geotagged tweets, based on the larger and more severe harassment received by women (Matias et al., 2015; Hess, 2014; Meyer and Cukier, 2006) and how abusers use knowledge of physical location to make explicit or implicit threats (Matias et al., 2015; Megarry, 2014). Indeed, one recommendation for targets of abuse is to turn off geolocation.[18] Furthermore, there are strong theoretical reasons to consider interaction effects between race and gender (Clark, 2014; Dixon, 2014).[19]

Other directions are to see the effect of filtering out non-personal users, and to build ways to filter out foreign tourists and better uniquely place geotag users in the block group that is likely to be their residence. Modeling demographic differences between users of different levels of use is also possible with this data. I have applied one spatial model, but spatial modeling is a rich area with many other available techniques. For example, there are also relevant disease mapping models that break down incidence by various demographic

[18]Recommendations for 'Social Media Safety' from the Rape, Abuse & Incest National Network, `https://rainn.org/sexual-assault-prevention/social-media-safety`.

[19]I thank Amanda Jean Stevenson (2014) for pointing this out to me.

strata (Bivand, Pebesma, et al., 2013) that would be appropriate here, as well as nonparametric models that might better capture irregular effects. Furthermore, I elected to not consider the temporal aspect; there is work on spatio-temporal modeling (Longley et al., 2015; Sylvester et al., 2014; Nagar et al., 2014; Kamath et al., 2013) but it tends to be in the short-term window of a day or week. With reliable spatio-temporal models of how the prevalence of geotagged tweets per block group changes over longer periods of time and a better understanding of the demographic characteristics towards which geotag users are biased, I may be able to create models to provide a rapid and high-resolution proxy for demographic changes such as processes of gentrification, or urbanization, or urban decay; that is, utilize the very biases of social media data to make inferences about larger phenomena. This was already done on a smaller scale, within the city of London, using a combination of Twitter and Foursquare data, by Hristova et al. (2016); they find correlations between properties of networks on those sites and measures of gentrification via the UK's Index of Multiple Deprivation. It may be possible to scale this up, making use of the geographic span of Twitter usage.

Bibliography

Anselin, Luc (2002). "Under the hood: Issues in the specification and interpretation of spatial regression models". *Agricultural Economics* 27 (3), pp. 247–267. DOI: 10.1111/j.1574-0862.2002.tb00120.x.

Anselin, Luc and Serge Rey (1991). "Properties of tests for spatial dependence in linear regression models". *Geographical Analysis* 23 (2), pp. 112–131. DOI: 10.1111/j.1538-4632.1991.tb00228.x.

Anselin, Luc, Sanjeev Sridharan, and Susan Gholston (2007). "Using exploratory spatial data analysis to leverage social indicator databases: The discovery of interesting patterns". *Social Indicators Research* 82 (2), pp. 287–309. DOI: 10.1007/s11205-006-9034-x.

Bivand, Roger S., Jan Hauke, and Tomasz Kossowski (2013). "Computing the Jacobian in Gaussian spatial autoregressive models: An illustrated comparison of available methods". *Geographical Analysis* 45 (2), pp. 150–179. DOI: 10.1111/gean.12008.

Bivand, Roger S., Edzer Pebesma, and Virgilio Gómez-Rubio (2013). *Applied spatial data analysis with R*. 2nd ed. Springer, NY. URL: http://www.asdar-book.org/.

Bivand, Roger S. and Gianfranco Piras (2015). "Comparing implementations of estimation methods for spatial econometrics". *Journal of Statistical Software* 63 (18), pp. 1–36. DOI: 10.18637/jss.v063.i18.

Blank, Grant (2016). "The digital divide among Twitter users and its implications for social research". *Social Science Computer Review* 35 (6), pp. 679–697. DOI: 10.1177/0894439316671698.

Brogueira, Gaspar, Fernando Batista, and Joao Paulo Carvalho (2016). "Using geolocated tweets for characterization of Twitter in Portugal and the Portuguese administrative regions". *Social Network Analysis and Mining* 6 (1). DOI: 10.1007/s13278-016-0347-8.

Cho, Eunjoon, Seth A. Myers, and Jure Leskovec (2011). "Friendship and mobility: User movement in location-based social networks". *Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining*. KDD '11, pp. 1082–1090. DOI: 10.1145/2020408.2020579.

Clark, Meredith D. (2014). "To tweet our own cause: A mixed-methods study of the online phenomenon 'Black Twitter'". PhD thesis. The University of North Carolina at Chapel Hill, School of Journalism and Mass Communication. URL: http://search.proquest.com/docview/1648168732.

Clauset, Aaron, Cosma Rohilla Shalizi, and Mark E. J. Newman (2009). "Power-law distributions in empirical data". *SIAM Review* 51 (4), pp. 661–703. DOI: 10.1137/070710111.

Crampton, Jeremy W., Mark Graham, Ate Poorthuis, Taylor Shelton, Monica Stephens, Matthew W. Wilson, and Matthew Zook (2013). "Beyond the geotag: Situating 'big data' and leveraging the potential of the geoweb". *Cartography and Geographic Information Science* 40 (2), pp. 130–139. DOI: 10.1080/15230406.2013.777137.

Dixon, Kitsy (2014). "Feminist online identity: Analyzing the presence of hashtag feminism". *Journal of Arts and Humanities* 3 (7), pp. 34–40. URL: https://www.theartsjournal.org/index.php/site/article/view/509.

Doran, Derek, Swapna Gokhale, and Aldo Dagnino (2013). "Human sensing for smart cities". *Proceedings of the 2013 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining*. ASONAM '13, pp. 1323–1330. DOI: 10.1145/2492517.2500240.

Eisenstein, Jacob, Brendan O'Connor, Noah A. Smith, and Eric P. Xing (2010). "A latent variable model for geographic lexical variation". *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing*. EMNLP '10, pp. 1277–1287. URL: http://dl.acm.org/citation.cfm?id=1870658.1870782.

Fiorio, Lee, Guy Abel, Jixuan Cai, Emilio Zagheni, Ingmar Weber, and Guillermo Vinué (2017). "Using Twitter data to estimate the relationship between short-term mobility and long-term migration". *Proceedings of the 2017 ACM on Web Science Conference*. WebSci '17, pp. 103–110. DOI: 10.1145/3091478.3091496.

Florini, Sarah (2014). "Tweets, tweeps, and signifyin': Communication and cultural performance on 'Black Twitter'". *Television & New Media* 15 (3), pp. 223–237. DOI: 10.1177/1527476413480247.

Freedman, David A., Stephen P. Klein, Michael Ostland, and Michael R. Roberts (2009a). "On 'solutions' to the ecological inference problem". *Statistical models and causal inference: A dialogue with the social sciences*. Cambridge University Press, pp. 83–96.

– (2009b). "Rejoinder to King". *Statistical models and causal inference: A dialogue with the social sciences*. Cambridge University Press, pp. 97–104.

Frias-Martinez, Vanessa, Victor Soto, Heath Hohwald, and Enrique Frias-Martinez (2012). "Characterizing urban landscapes using geolocated tweets". *Proceedings of the 2012 ASE/IEEE International Conference on Social Computing and 2012 ASE/IEEE International Conference on Privacy, Security, Risk and Trust*. SOCIALCOM-PASSAT '12, pp. 239–248. DOI: 10.1109/SocialCom-PASSAT.2012.19.

Gaetan, Carlo and Xavier Guyon (2012). *Spatial statistics and modeling*. Springer Series in Statistics. New York: Springer. DOI: 10.1007/978-0-387-92257-7.

Ghosh, Debarchana (Debs) and Rajarshi Guha (2013). "What are we 'tweeting' about obesity? Mapping tweets with topic modeling and Geographic Information System". *Cartography and Geographic Information Science* 40 (2), pp. 90–102. DOI: 10.1080/15230406.2013.776210.

Goodchild, Michael F. (2007). "Citizens as sensors: The world of volunteered geography". *GeoJournal* 69 (4), pp. 211–221. DOI: 10.1007/s10708-007-9111-y.

Graham, Mark, Scott A. Hale, and Devin Gaffney (2014). "Where in the world are you? Geolocation and language identification in Twitter". *The Professional Geographer* 66 (4), pp. 568–578. DOI: 10.1080/00330124.2014.907699.

Guo, Diansheng and Chao Chen (2014). "Detecting non-personal and spam users on geo-tagged Twitter network". *Transactions in GIS* 18 (3), pp. 370–384. DOI: 10.1111/tgis.12101.

Hawelka, Bartosz, Izabela Sitko, Euro Beinat, Stanislav Sobolevsky, Pavlos Kazakopoulos, and Carlo Ratti (2014). "Geo-located Twitter as proxy for global mobility patterns". *Cartography and Geographic Information Science* 41 (3), pp. 260–271. DOI: 10.1080/15230406.2014.890072.

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". *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. CHI '11, pp. 237–246. DOI: 10.1145/1978942.1978976.

Hecht, Brent and Monica Stephens (2014). "A tale of cities: Urban biases in volunteered geographic information". *Proceedings of the Eighth International AAAI Conference on Weblogs and Social Media*.

ICWSM-14, pp. 197–205. URL: http://www.aaai.org/ocs/index.php/ICWSM/ICWSM14/paper/view/8114.

Hess, Amanda (2014). "Why women aren't welcome on the Internet". *Pacific Standard Magazine* 7 (1). URL: https://psmag.com/social-justice/women-arent-welcome-internet-72170.

Hong, Liangjie, Amr Ahmed, Siva Gurumurthy, Alexander J. Smola, and Kostas Tsioutsiouliklis (2012). "Discovering geographical topics in the Twitter stream". WWW '12, pp. 769–778. DOI: 10.1145/2187836.2187940.

Hossain, Nabil, Tianran Hu, Roghayeh Feizi, Ann Marie White, Jiebo Luo, and Henry A. Kautz (2016). "Precise localization of homes and activities: Detecting drinking-while-tweeting patterns in communities". *Proceedings of the Tenth International AAAI Conference on Web and Social Media*. ICWSM-16, pp. 587–590. URL: http://www.aaai.org/ocs/index.php/ICWSM/ICWSM16/paper/view/13118.

Hristova, Desislava, Matthew J. Williams, Mirco Musolesi, Pietro Panzarasa, and Cecilia Mascolo (2016). "Measuring urban social diversity using interconnected geo-social networks". *Proceedings of the 25th International Conference on World Wide Web*. WWW '16, pp. 21–30. DOI: 10.1145/2872427.2883065.

Ishida, Kazunari (2012). "Geographical bias on social media and geo-local contents system with mobile devices". *Proceedings of the 45th Hawaii International Conference on System Sciences*. HICSS '12, pp. 1790–1796. DOI: 10.1109/HICSS.2012.292.

Johnson, Isaac L., Subhasree Sengupta, Johannes Schöning, and Brent Hecht (2016). "The geography and importance of localness in geotagged social media". *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*. CHI '16, pp. 515–526. DOI: 10.1145/2858036.2858122.

Kamath, Krishna Y., James Caverlee, Kyumin Lee, and Zhiyuan Cheng (2013). "Spatio-temporal dynamics of online memes: A study of geo-tagged tweets". *Proceedings of the 22nd International Conference on World Wide Web*. WWW '13, pp. 667–678. DOI: 10.1145/2488388.2488447.

King, Gary, Ori Rosen, and Martin A. Tanner (2004). *Ecological inference: New methodological strategies*. Cambridge, UK: Cambridge University Press.

Kinsella, Sheila, Vanessa Murdock, and Neil O'Hare (2011). "'I'm eating a sandwich in Glasgow': Modeling locations with tweets". *Proceedings of the 3rd International Workshop on Search and Mining User-Generated Contents*. SMUC '11, pp. 61–68. DOI: 10.1145/2065023.2065039.

Kumar, Shamanth, Xia Hu, and Huan Liu (2014). "A behavior analytics approach to identifying tweets from crisis regions". *Proceedings of the 25th ACM conference on Hypertext and social media*. HT '14, pp. 255–260. DOI: 10.1145/2631775.2631814.

Leetaru, Kalev, Shaowen Wang, Guofeng Cao, Anand Padmanabhan, and Eric Shook (2013). "Mapping the global Twitter heartbeat: The geography of Twitter". *First Monday* 18 (5). URL: http://firstmonday.org/ojs/index.php/fm/article/view/4366.

Liere, Diederik van (2010). "How far does a tweet travel? Information brokers in the Twitterverse". *Proceedings of the International Workshop on Modeling Social Media*. MSM '10, 6:1–6:4. DOI: 10.1145/1835980.1835986.

Lin, Yu-Ru and Drew Margolin (2014). "The ripple of fear, sympathy and solidarity during the Boston bombings". *EPJ Data Science* 3 (31). DOI: 10.1140/epjds/s13688-014-0031-z.

Liu, Yabing, Chloe Kliman-Silver, and Alan Mislove (2014). "The tweets they are a-changin': Evolution of Twitter users and behavior". *Proceedings of the Eighth International AAAI Conference on Weblogs and Social Media*. ICWSM-14, pp. 305–314. URL: http://www.aaai.org/ocs/index.php/ICWSM/ICWSM14/paper/view/8043.

Longley, Paul A., Muhammad Adnan, and Guy Lansley (2015). "The geotemporal demographics of Twitter usage". *Environment and Planning A* 47 (2), pp. 465–484. DOI: 10.1068/a130122p. URL: http://www.envplan.com/abstract.cgi?id=a130122p.

Malik, Momin M. (2018). "Bias and beyond in digital trace data". PhD thesis. Pittsburgh, PA: Carnegie Mellon University. URL: http://reports-archive.adm.cs.cmu.edu/anon/isr2018/abstracts/18-105.html.

Malik, Momin M., Hemank Lamba, Constantine Nakos, and Jürgen Pfeffer (2015). "Population bias in geotagged tweets". *Papers from the 2015 ICWSM Workshop on Standards and Practices in Large-Scale Social Media Research*. ICWSM-15 SPSM, pp. 18–27. URL: http://www.aaai.org/ocs/index.php/ICWSM/ICWSM15/paper/view/10662.

Matias, J. Nathan, Amy Johnson, Whitney Erin Boesel, Brian Keegan, Jaclyn Friedman, and Charlie DeTar (2015). "Reporting, reviewing, and responding to harassment on Twitter". Women, Action, and the Media. URL: http://womenactionmedia.org/twitter-report/.

McNeill, Graham, Jonathan Bright, and Scott A. Hale (2016). "Estimating local commuting patterns from geolocated Twitter data". eprint: https://arxiv.org/abs/1612.01785.

Megarry, Jessica (2014). "Online incivility or sexual harassment? Conceptualising women's experiences in the digital age". *Women's Studies International Forum* 47 (Part A), pp. 46–55. DOI: 10.1016/j.wsif.2014.07.012.

Mellon, Jonathan and Christopher Prosser (2017). "Twitter and Facebook are not representative of the general population: Political attitudes and demographics of British social media users". *Research & Politics* 4 (3). DOI: 10.1177/2053168017720008.

Meyer, Robert and Michel Cukier (2006). "Assessing the attack threat due to IRC channels". *Proceedings of the International Conference on Dependable Systems and Networks*. DSN '06, pp. 467–472. DOI: 10.1109/DSN.2006.12.

Mislove, Alan, Sune Lehmann, Yong-Yeol Ahn, Jukka-Pekka Onnela, and J. Rosenquist (2011). "Understanding the demographics of Twitter users". *Proceedings of the Fifth International AAAI Conference on Weblogs and Social Media*. ICWSM-11, pp. 554–557. URL: https://www.aaai.org/ocs/index.php/ICWSM/ICWSM11/paper/view/2816.

Mitchell, Lewis, Morgan R. Frank, Kameron Decker Harris, Peter Sheridan Dodds, and Christopher M. Danforth (2013). "The geography of happiness: Connecting Twitter sentiment and expression, demographics, and objective characteristics of place". *PLOS ONE* 8 (5), e64417. DOI: 10.1371/journal.pone.0064417.

Montasser, Omar and Daniel Kifer (2017). "Predicting demographics of high-resolution geographies with geotagged tweets". *Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence*. AAAI-17, pp. 1460–1466. URL: http://aaai.org/ocs/index.php/AAAI/AAAI17/paper/view/14818.

Morstatter, Fred, Nichola Lubold, Heather Pon-Barry, Jürgen Pfeffer, and Huan Liu (2014). "Finding eyewitness tweets during crises". *Proceedings of the ACL 2014 Workshop on Language Technologies and Computational Social Science*. ACL LACSS 2014. Baltimore, MD, USA: Association for Computational Linguistics, pp. 23–27. URL: http://www.aclweb.org/anthology/W14-2509.

Morstatter, Fred, Jürgen Pfeffer, and Huan Liu (2014). "When is it biased? Assessing the representativeness of Twitter's Streaming API". *Companion to the Proceedings of the 23rd International Conference on World Wide Web*. WWW Companion '14, pp. 555–556. DOI: 10.1145/2567948.2576952.

Morstatter, Fred, Jürgen Pfeffer, Huan Liu, and Kathleen Carley (2013). *Is the sample good enough? Ccomparing data from Twitter's streaming API with Twitter's firehose*. URL: http://aaai.org/ocs/index.php/ICWSM/ICWSM13/paper/view/6071/6379.

Mowery, Jared (2016). "Twitter influenza surveillance: Quantifying seasonal misdiagnosis patterns". *Online Journal of Public Health Informatics* 8 (3). DOI: 10.5210/ojphi.v8i3.7011.

Nagar, Ruchit, Qingyu Yuan, C. Clark Freifeld, Mauricio Santillana, Aaron Nojima, Rumi Chunara, and S. John Brownstein (2014). "A case study of the New York City 2012-2013 influenza season with daily geocoded Twitter data from temporal and spatiotemporal perspectives". *Journal of Medical Internet Research* 16 (10), e236. DOI: 10.2196/jmir.3416.

O'Loughlin, John (2000). "Can King's ecological inference method answer a social scientific puzzle: Who voted for the Nazi Party in Weimar Germany?" *Annals of the Association of American Geographers* 90 (3), pp. 592–601. DOI: 10.1111/0004-5608.00213.

Ruths, Derek and Jürgen Pfeffer (2014). "Social media for large studies of behavior". *Science* 346 (6213), pp. 1063–1064. DOI: 10.1126/science.346.6213.1063.

Sadilek, Adam, Henry Kautz, Lauren DiPrete, Brian Labus, Eric Portman, Jack Teitel, and Vincent Silenzio (2016). "Deploying nEmesis: Preventing foodborne illness by data mining social media". *Proceedings of the Twenty-Eighth Innovative Applications of Artificial Intelligence Conference*. IAAI-16, pp. 3982–3989. URL: https://www.aaai.org/ocs/index.php/IAAI/IAAI16/paper/view/11823.

Sharma, Sanjay (2013). "Black Twitter? Racial hashtags, networks and contagion". *new formations: a journal of culture/theory/politics* 78 (1). URL: http://muse.jhu.edu/journals/new_formations/v078/78.sharma.html.

Shelton, Taylor, Ate Poorthuis, Mark Graham, and Matthew Zook (2014). "Mapping the data shadows of Hurricane Sandy: Uncovering the sociospatial dimensions of 'big data'". *Geoforum* 52 (0), pp. 167–179. DOI: 10.1016/j.geoforum.2014.01.006.

Sloan, Luke, Jeffrey Morgan, William Housley, Matthew Williams, Adam Edwards, Pete Burnap, and Omer Rana (2013). "Knowing the tweeters: Deriving sociologically relevant demographics from Twitter". *Sociological Research Online* (18). DOI: 10.5153/sro.3001.

Stephens, Monica and Ate Poorthuis (2014). "Follow thy neighbor: Connecting the social and the spatial networks on Twitter". *Computers, Environment and Urban Systems* 53, pp. 87–95. DOI: 10.1016/j.compenvurbsys.2014.07.002.

Stevenson, Amanda Jean (2014). "Finding the Twitter users who stood with Wendy". *Contraception* (90), pp. 502–507. DOI: 10.1016/j.contraception.2014.07.007.

Sylvester, Jared, John Healey, Chen Wang, and William M. Rand (2014). "Space, time, and hurricanes: Investigating the spatiotemporal relationship among social media use, donations, and disasters". Research Paper No. RHS 2441314, Robert H. Smith School. DOI: 10.2139/ssrn.2441314.

Takhteyev, Yuri, Anatoliy Gruzd, and Barry Wellman (2012). "Geography of Twitter networks". *Social Networks* 34 (1), pp. 73–81. DOI: 10.1016/j.socnet.2011.05.006.

Tasse, Dan, Zichen Liu, Alex Sciuto, and Jason I. Hong (2017). "State of the geotags: Motivations and recent changes". *Proceedings of the Eleventh International AAAI Conference on Web and Social Media*. ICWSM 2017, pp. 250–259. URL: https://aaai.org/ocs/index.php/ICWSM/ICWSM17/paper/view/15588.

Thebault-Spieker, Jacob, Loren Terveen, and Brent Hecht (2017). "Toward a geographic understanding of the sharing economy: Systemic biases in UberX and TaskRabbit". *ACM Transactions on Computer-Human Interaction* 24 (3), 21:1–21:40. DOI: 10.1145/3058499.

Tufekci, Zeynep (2014). "Big questions for social media big data: Representativeness, validity and other methodological pitfalls". *Proceedings of the Eighth International AAAI Conference on Weblogs and Social Media*. ICWSM-14, pp. 505–514. URL: https://www.aaai.org/ocs/index.php/ICWSM/ICWSM14/paper/view/8062.

U.S. Census Bureau (2008). *A compass for understanding and using American Community Survey data: What general data users need to know*. Washington, DC: U.S. Government Printing Office. URL: https://www.census.gov/library/publications/2008/acs/general.html.

van Dijck, José (2013). *The culture of connectivity: A critical history of social media*. New York, NY: Oxford University Press.

Wang, Di, Ahmad Al-Rubaie, John Davies, and Sandra Stinĉić Clarke (2014). "Real time road traffic monitoring alert based on incremental learning from tweets". *Proceedings of the 2014 IEEE Symposium on Evolving and Autonomous Learning Systems*. EALS '14, pp. 50–57. DOI: 10.1109/EALS.2014.7009503.

Wang, Rui, Fanglin Chen, Zhenyu Chen, Tianxing Li, Gabriella Harari, Stefanie Tignor, Xia Zhou, Dror Ben-Zeev, and Andrew T. Campbell (2014). "StudentLife: Assessing mental health, academic performance and behavioral trends of college students using smartphones". *Proceedings of the 2014 ACM International Joint Conference on Pervasive and Ubiquitous Computing*. UbiComp '14, pp. 3–14. DOI: 10.1145/2632048.2632054.

Yuan, Nicholas Jing, Fuzheng Zhang, Defu Lian, Kai Zheng, Siyu Yu, and Xing Xie (2013). "We know how you live: Exploring the spectrum of urban lifestyles". *Proceedings of the First ACM Conference on Online Social Networks*. COSN '13, pp. 3–14. DOI: 10.1145/2512938.2512945.

Zickuhr, Kathryn (2012). "Geosocial services". Pew Internet and American Life Project, Pew Research Center. URL: http://www.pewinternet.org/2012/05/11/geosocial-services/.

– (2013). "Location-based services". Pew Internet and American Life Project, Pew Research Center. URL: http://www.pewinternet.org/2013/09/12/location-based-services/.