The Generalizability of Twitter Data for Population Research

Guangqing Chi
Department of Agricultural Economics, Sociology, and Education,
Population Research Institute, and Social Science Research Institute
Pennsylvania State University
University Park, PA 16802, USA
Email: gchi@psu.edu; Tel: +1 814 865 5553

Junjun Yin
Population Research Institute and Social Science Research Institute
Pennsylvania State University
University Park, PA 16802, USA

Jennifer Van Hook
Department of Sociology and Criminology
Pennsylvania State University
University Park, PA 16802, USA

Eric Plutzer
Department of Political Science
Pennsylvania State University
University Park, PA 16802, USA

Heng Xu
Kogod School of Business
American University
4400 Massachusetts Avenue, NW
Washington, DC 20016, USA

Abstract

Social media data such as from Twitter have been used in many fields. Demography, the discipline dealing with numbers the most among all social science disciplines, has been slow in taking advantage of the abundance of Twitter data. The biggest concern is the representativeness of Twitter users for population. This study is to evaluate the extent to which Twitter users (mis)represent the population across different demographic groups. We conduct the research at the county level in the U.S. from 2014–2017 using 96% geotagged tweets. The specific aims are to: extend and refine already developed methods for imputing the gender, age, race/ethnicity, and county of residence of each Twitter user; use these imputed values to assess the (mis)representativeness of Twitter samples at the county level; and explain the determinants of biases. If successful, this research will open the door for demographers to take advantage of rich Twitter data.

1. Introduction and Research Objectives

Social scientists heavily rely on surveys and census data to study social problems and phenomena. However, it is challenging to use these data to study the fast-changing world due to two limitations of traditional survey and census data:

- Survey and census data collection is slow, labor intensive, and expensive. This applies to small-scale surveys, large-scale national surveys, and census data.
- Many important population characteristics and behaviors cannot be measured well by traditional surveys employing robust probability sampling. For example, in the minutes after a disaster, crime, or terrorist event, what percentage of the population is fearful? What if we wish to assess the degree of anger immediately after the announcement of a jury verdict in a highly publicized case? Or the

> level of attentiveness to unplanned public health announcements? In each of these examples, it would be difficult to field a probability-based survey in real time, and respondents may not be able to reconstruct how they felt or behaved at the time of the event, even if interviewed just a few days later.

Social media data provide significant opportunities for social scientists to study social problems and advance social sciences by overcoming the limitations of traditional survey and census data. Social media data, including those from social networking services (such as Twitter, Facebook, Instagram, and LinkedIn), and other types of online interaction, are rich sources. Of these, Twitter provides a highly accessible Big Data stream and has drawn interest from scientists in computer science [1], political science [2,3], urban studies [4], public health [5,6], and behavioral science [7]. Because it is possible to follow individual Twitter users over time and across space, demographers have used Twitter data to track population movements [8], suggesting that Twitter data could track refugee flows or migration flows following natural disasters before such population movements can be detected by surveys or censuses.

However, authors of Twitter posts—Twitter users—are not representative of the general population. This diminishes confidence in the findings and limits the ability of such studies to contribute to bodies of literature based on high-quality sample surveys. Therefore, Twitter data has been strongly resisted by social scientists, especially sociologists and demographers, largely because of concerns about the data's representativeness of the population as a whole and because we know little about the demographic characteristics of the users [7,9]. We know that Twitter users tend to be younger. But this is a generalization averaged across the entire nation. In this paper we seek to assess the magnitude of the (mis)representation for all 3,144 US counties. That will provide a preliminary assessment of the extent of bias if Twitter data were applied to problems such as migration, or to highly localized events such as floods, mass shootings or disease outbreaks. Specifically, we seek to quantify the county level departures of Twitter data from Census based data with respect to the distribution of age, sex, race and ethnicity. We do this by applying methods that can impute these traits from geotagged Twitter data. Second, we explore key covariates that we expect will be associated with greater or lesser departures from representativeness.

In some cases the departures from representativeness are large, but potentially within the range that is typically correctable in low-response sample surveys. Thus, an important implication of our research is that it may be possible to apply survey-based methods of post-stratification weighting to make Twitter samples more representative.

We will conduct the research at the county level in the United States from January 2014 to December 2017, primarily replying on the tweets that we have been collecting since July 2013 (40 TB of data so far). Our database includes 96% of all geotagged tweets (the rest were lost mostly due to Internet disconnections). Ground truth data come from the U.S. Census Bureau's American Community Survey (ACS) estimates. It should be noted that this study is not to develop new methods for estimating Twitter user demographics. Rather, it is to utilize the best methods/tools that have already been developed, largely by computer scientists, and evaluate the extent to which Twitter users (mis)represent the population as a whole and corresponding demographic groups in different regions. It should also be noted that this manuscript reports preliminary results using 2014 data only; we are conducting analyses using 2014–2017 data and will report the results in a later version.

This study has two major contributions. First, we provide the *necessary* evaluation of generalizability of Twitter data for population-related research before social scientists accept the use of Twitter data for their research. Our research findings will tell whether and how the Twitter data can be generalized to the population by gender, age, race/ethnicity and migration status, using replicable procedures. Second, we seek to understand the extent to which the Twitter data are biased relative to population data, and the determinants of the biases. Knowing the generalizability and replicability of Twitter data for population *is* and *should be* the first step for almost all population research involving Twitter data. This study focuses on the evaluation. If successful, this research will significantly advance population science. It will open the door for demographers and sociologists to take advantage of rich Twitter data. This project will also strengthen research in many other social science disciplines that use demographic data.

2. Background

Research on demographic processes, population health, crime, and many other topics rely heavily on government and government-funded survey research, including highly focused studies for specific purposes (“designed data”) [11]. These data, the result of rigorous research designs intended to maximize precision and minimize bias, are employed at multiple scales to address many key research and policy questions. However, the data collection process is often labor intensive and expensive, and the release of the collected data typically occurs more than a year later. In response, researchers have for decades advocated for supplementary data sources that are more nimble [12].

2.1. Promises of Big Data

In recent years, Internet connections and smartphones have become ubiquitous, and Location Based Services have advanced. These changes mean that massive amounts of user-generated digital information are now being generated and becoming increasingly accessible for analysis. In comparison to designed data, digital information is directly collected from a large group of individuals with no predefined criteria and is thus deemed "organic." Such data come from a variety of platforms and media, such as archived newspapers, web search histories, social media outlets such as Facebook and Twitter, personal blogs, Wikipedia entries, and others and are often referred to as Big Data [13]. Big Data have spurred the development of advanced information-technology-based novel scientific methods and theories among computer scientists, which in turn may enable social scientists to integrate Big Data into their domain-specific knowledge and research topics [14]. Demographers have begun to utilize Big Data, as reflected in recent publications in general science journals [15–17], demography, methods, and public health journals [18–22], including a recent special issue on Big Data [23]. The trend is likely to continue, as suggested by the many Big Data papers presented at recent annual meetings of the Population Association of America (e.g., [24–28]).

The potential of Big Data lies in the ability to collect massive amounts of information from a large group of individuals. With the global adoption of social media, social media user populations have expanded to an unprecedented level. It is estimated that among approximately 2.5 billion non-unique users, Facebook, Google+, and Twitter account for over half of the users [29]. The large N allows the potential for high resolution classification and the possibility of generating samples of individuals in small or hard-to-reach populations, assisting in the real-world study of various population dynamics. In addition, Big Data allow researchers to track changes in populations very quickly, in real time. For example, researchers have used mobile phone data for dynamic population mapping and estimation [15] and have performed near-real-time assessments of population displacement following disasters [30]. Google search query data were used to detect influenza epidemics [31], and Twitter emotion data were used to predict stock market changes [32].

Among the various sources of Big Data, Twitter has attracted special attention because it is accessible to researchers [1,33,34]. Due to privacy concerns, the mobile phone call data are privileged and are not publicly accessible; Google search query data are also criticized for a lack of transparency. Proprietary data are not ideal for replicability in scientific findings or for conducting comparative studies across different regions [1,35]. That said, Twitter data are accessible. Twitter is arguably the most popular information source for the scientific research community due to its accessibility [36]. Twitter API allows researchers to access its data with certain restrictions [37]. For example, researchers can access Twitter data by downloading real-time tweets (with a limit of 1% of the data in the data stream) via the streaming API or by purchasing the entire Twitter posts (tweets) archive via the enterprise API [38].

A second strength of Twitter data is that the geolocations of tweets either are geotagged or could be inferred. Because of the popularity of accessing Twitter via smartphones, one of the important and unique data products from Twitter are geotagged tweets, which are tweets tagged with real-world locations derived from Twitter users’ smartphones with integrated GPS or Wi-Fi positioning. Compared to non-geotagged tweets, geotagged tweets are significantly lower in terms of data volume. Around 3% of all tweets worldwide were geotagged in 2012 [39]. However, recent research has developed methods to infer Twitter user locations based on users’ profiles or from their social networks because less than 10% of Twitter user profiles are made not publicly accessible [40]. A geocoder based on a simple major cities gazetteer and relying on the user-provided Location and Profile fields is able to geo-locate more than one-third of all tweets with high accuracy when measured against the GPS-based baseline [39]. A critical evaluation of state-of-the-art network-based methods for performing geolocation inference was conducted by Jurgens et al. [41], who tested nine geolocation inference techniques at the global scale, all presented at top-tier conferences [42–50]. They also released the implementations of these methods in an open-source geo-inference package [41]. Combined with the capability of continuously monitoring a large group of Twitter users, such data offer a great potential for long-term migration and short-term human behavior studies. Geotagged Twitter data have already been used to integrate the dimensions of internal and international migration [8] and to study mobility patterns at global [51] and national scales [52,53].

A third strength of Twitter data is that its contents can be mined for valuable information. Although tweets are composed and disseminated for the purpose of social networking, the data provide more information by representing various forms of interaction, such as social ties embedded in the friendship network; attitudes and discourses toward certain political, cultural, and even personal topics; and behavior through the message content or movement across space. Therefore, Twitter data have been applied in a number of social science domains, such as social trend/movement, public health, migration, and mobility. Twitter data were used to examine political behavior [54] and in tracking drug abuse trends, including problem drinking [55], nonmedical use of Adderall among college students [56], and marijuana concentrations across the U.S. [57]. Complex and time-sensitive studies of migration activities also benefit from using geotagged Twitter data. For example, several studies use Twitter to explore migration flow patterns [58], estimate refugee migration patterns [59] and to test existing migration theories, such as the relationships between short-term mobility and long-term migration [60], and some have suggested that Twitter data can serve as a barometer for migration flows, providing timely migration information before the availability of the official statistics [8].

The merits of social media and Twitter data just discussed excite some in the research community. They believe that social media data provide researchers with a useful tool to study public opinion in times and places in which surveys are unavailable [61]. Twitter data can also provide a useful complement to existing household survey data and even potentially replace survey data if none are available [62].

2.2. Challenges of Using Twitter Data

However, others in the research community are apprehensive about using Big Data, including Twitter data. One of the most serious concerns is centered on the characteristics and representativeness of the data [63]. When using social media data for population research it is critical to be able to generalize the results to the whole population. But this ability is limited by some unknowns of Twitter data. When using Twitter API, data sampling is controlled by Twitter and little is known about its sampling methodology. More importantly, selection bias exists in Twitter data because the demographics of the Twitter users are unknown, and certain demographic groups are known to be over- or underrepresented. For example, researchers found Twitter users are not a representative sample of the population as a whole, tending to skew towards young, urban, minority individuals [64]. Other research suggests that more social media users tend to be younger [65], socioeconomically advantaged [66,67], and not proportional in terms of male users and female users, leaving the results biased and invalid without any calibration [68]. For example, the age bias was found to affect attempts to predict political elections from Twitter sentiments [69]. Almost all the studies that use Twitter data for migration study purposes also documented this limitation [8,52,58–60]. Thus, without addressing the problem of selection bias, generalization from populations of social media users may lead to unreliable results. Multiple research lines began looking into addressing such issues by proposing methods to estimate the characteristics of Twitter user demographics, such as gender and race/ethnicity estimation based on names [70–72] or on Twitter user profile images via image recognition services or even by crowdsourcing efforts using humans doing the estimation [68], but such efforts have not fully evaluated their proposed methods.

Still other concerns about using Twitter data in research include: the hardships of dealing with slang, sarcasm, and unconventional forms of written expression, including hashtags, emoticons, and acronyms [63]; dealing with the fact that not all Twitter users are humans but bots, which can distort the results; and managing the cost of obtaining, storing, and cleaning the massive datasets from Twitter [63,73].

3. Estimate Twitter User Demographics at the County Level

As mentioned earlier, one of the major concerns of social scientists in using social media data in population research is selection bias. Twitter users are not representative of the overall population [74]. To address the selection bias, the very first step in using Twitter data for population research is to understand the demographic characteristics of Twitter users. Therefore, Research Aim 1 is to utilize a combination of existing best methods and tools to estimate Twitter user demographics by gender, age, and race/ethnicity. Although the methods and tools we use already exist, to our best knowledge no studies have ever used them collectively to achieve the best estimates. Also, no studies have estimated Twitter user demographics using such massive data as we propose using.

The age of a Twitter user will be estimated by using a facial recognition service provided by Microsoft Azure. The gender information of a Twitter user will be estimated based on the first name extracted from the user profile matching to a first name database from Facebook profile pages [75]; if the first name information is absent, it will be estimated based on the profile image using the facial recognition service of Microsoft Azure. The race/ethnicity of a Twitter user will be estimated based on the user's last name extracted from the user profile matching to the U.S. Census Bureau's surname database for race/ethnicity and facial recognition services as provided by Face++ [76] and Kairos [77]. We have three race/ethnicity groups: Hispanic, non-Hispanic White, and non-Hispanic Black. Figure 1 illustrates the overall flow of demographic estimates. The following subsections detail the components and methods to be used. We have conducted preliminary analysis with the geotagged tweets dating from January 1 to December 31, 2014, in the contiguous U.S. [78]. We report our preliminary findings in corresponding subsections below.

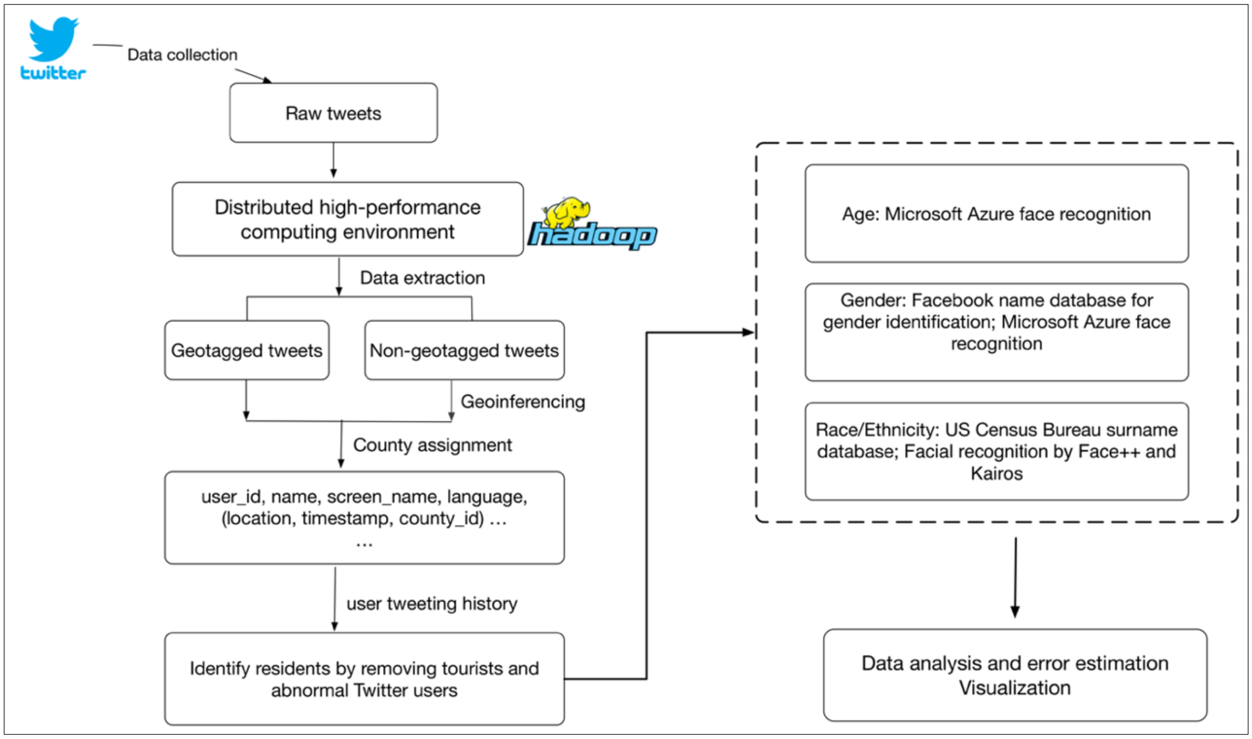

twitter
Data collection
Raw tweets
Distributed high-performance computing environment
hadoop
Data extraction
Geotagged tweets
Non-geotagged tweets
Geoinferencing
County assignment
user_id, name, screen_name, language, (location, timestamp, county_id) …
…
user tweeting history
Identify residents by removing tourists and abnormal Twitter users
Age: Microsoft Azure face recognition
Gender: Facebook name database for gender identification; Microsoft Azure face recognition
Race/Ethnicity: US Census Bureau surname database; Facial recognition by Face++ and Kairos
Data analysis and error estimation Visualization

Figure 1. The Flow Chart for Estimating Twitter User Demographics by Gender, Age, and Race/Ethnicity

3.1. Twitter Data Collection and Extraction

Twitter data are categorized into geotagged and non-geotagged tweets. The difference is that geotagged tweets are tagged with a location. Such a location is represented by a pair of latitude and longitude coordinates, which are usually derived from location-based service enabled smartphones with integrated GPS or Wi-Fi positioning. The geographic locations are considered to have high spatial resolution down to 10 meters [52]. It is worth mentioning that the accuracy of locations estimated from mobile phone call records is in several kilometers [15]. Thus, the use of geotagged tweets for research provides better geographical accuracy than mobile phone-based data, which may be important when finer geographic scale is important. The location information of non-geotagged tweets can be derived from user profiles with self-reported addresses or inferred from the content history the user's tweets [64,79] or based on user relationships [42].

Our team has been collecting geotagged tweets over the entire globe since June 2013 by using the publicly accessible Twitter streaming API [80]. To overcome the 1% data download limit [51], where the number of downloaded tweets cannot exceed 1% of the total number of tweets in the data stream, we divided the entire globe into multiple regions and collected data over these regions simultaneously. In this way, we downloaded approximately 96% of all geotagged tweets worldwide. The current global data collection is over 40 TB in size. To manage and process such a large volume of data with efficiency, we have been utilizing the cyberinfrastructure resources and high-performance computing environment allocated and supported by the Bridges supercomputer at Pittsburgh Supercomputing Center [81].

We have already developed a suite of MapReduce [82] programs to process the dataset in a parallel fashion based on distributed computing environment provided by the Bridges supercomputer. We will open source these programs available to the public. First, we extract several fields of information from each tweet, specifically, the unique user id (*user_id*), username (*name*), screen name (*screen_name*), language of the message (*lang*), location of the tweet (*loc*), and timestamp of the tweet (*t*) (in its local time zone). For non-geotagged tweets, we first extract the location from users' profiles and then derive the social networks for each user, which are used to refine and determine a user's location based on geo-inference techniques from aforementioned studies. Based on the given location, second, we use a program [53] to extract all the geotagged tweets that fall in the United States. Third, we will assign each geotagged tweet to a corresponding U.S. county (there are 3,144 counties or equivalents in the U.S.) using the same program. At this stage, each tweet is represented by a tuple $\langle user\_id, name, screen\_name, lang, loc, t, county\_id \rangle$. By matching the unique user id, all the tweets that are posted by the same user are appended together and sorted based on the timestamp in chronological order as this user's historical tweet collection, which is illustrated below:
$User_i \; \forall \; \{user\_id_i, name_i, screen\_name_i, lang_i, \langle loc_1, t_1, county\_id_1 \rangle \big| \langle loc_2, t_2, county\_id_2 \rangle \ldots | \langle loc_j, t_j, county\_id_j \rangle\}$
where i represents the i^{th} Twitter user in the list, and *j* refers to the j^{th} county.

Because we are interested in studying the demographics of Twitter users and their relations with the actual population, we need to remove non-human Twitter user accounts (i.e., bots). This task is non-trivial [83]. Many approaches were developed to address this issue, such as using machine learning approach [84] and looking at the ratio between the number of a user's followers and followings [85]. For geotagged tweets, a simple yet popular heuristic is based on the speed of the displacement between two consecutive tweets, i.e., a user with tweets with a reallocation speed greater than 1000 km/hour (i.e., the typical speed of an airplane) is removed from the study [51–53]. Further, to focus on the resident population rather than tourists, we will impose a criterion that a Twitter user is considered to be a "resident" when the user is observed to have a time interval between the first and last tweet of more than 30 days in the same county [86]. Note that this one-month criterion is subjectively defined but strict to ensure that observed Twitter population are actively observed in their home county (at least over a month). We will explore shorter time intervals and other possible methods such as a frequency-based approach for distinguishing residents from tourists in future research.

In our preliminary analysis, we started with all tweets collected in 2014—over 2.9 billion tweets generated from 27.3 million Twitter accounts worldwide, over 3 TB in file size. Based on the given location, we extracted all the geotagged tweets that fall in the contiguous United States, which reduces the data collection to approximately 1.2 billion tweets from over 6.4 million Twitter accounts (counted based on unique Twitter user ids). Further, we assigned each geotagged tweet to a corresponding U.S. county (of 3,105 counties in the contiguous U.S.). After removing non-human Twitter user accounts and tourists, the total geotagged Twitter user population in the preliminary study was reduced to approximately 835 million tweets produced by 3.78 million unique Twitter users.

3.2. Twitter User Gender Identification Based on First Name Database and Image Recognition Techniques

To identify the gender of Twitter users, a popular method used in recent research is to probabilistically determine a user's gender by using the first name of the user and matching its occurrence in a first name database. In the database, each first name is associated with a probability of being a female or male [70–72]. Such first name databases vary in different studies, such as those based on electoral registers and telephone directories in the UK [87] or generated from social security card applications in the U.S. [88]. In addition, researchers have generated a first name database from Facebook profile pages with a collection of 23,363 total first names [89]. This first name database utilizes a probabilistic approach similar to that of the social security first name database, where each name in the collection has the counts of name occurrences labeled as male or female. This Facebook first name database is found to achieve 96.3% accuracy of gender estimation when applied to Facebook users [89]. Another advantage of this first name database is that it provides an additional database for gender identification, where the name list contains nicknames and the counts of occurrences of the nickname labeled as male or female. Considering the aforementioned advantages of this first name database and the fact that Twitter and Facebook have a highly overlapping user group, where over 90% of Twitter users also use Facebook [90], we employ this first name database for Twitter user gender identification.

To ensure more accurate results using the Facebook first name database, we perform several rounds of data cleaning on the first names derived from our Twitter dataset. First, we remove special characters, such as emoji, from the username and the screen name. If the name has a prefix, such as "A. John Doe" or "Mr. John Doe," the middle name is assigned as the first name of this user. The derived first name is then sent to the first name database to find its match. Note that in cases where a user has provided a full name with first name, middle name, and surname, both the first name and middle names are kept to query the name database when there is no match for the first name. If the name appears in the first name database, we calculate the gender probability based on the fraction of occurrences that were labeled as male or female. If there is no match in the first name database, we continue the search on the nickname database in the same manner. Finally, the name is assigned the gender with the higher probability. In our preliminary analysis, the gender information of 70% of users can be identified based on the combination of first name and screen name.

Although the first name-based approach for Twitter user gender estimation is promising, in many cases Twitter usernames and screen names are missing, incomplete, or intentionally misspelled, or there are simply no matches in the first name database. Indeed, there are approximately 30% of Twitter users in our preliminary study whose gender cannot be determined based on their first names. To estimate gender for these Twitter users and to improve the overall accuracy in Twitter user gender identification, we also analyze Twitter users' profile images using the Microsoft Azure Face API [91]. Specifically, a URL to a Twitter user's profile image is generated based on the user's screen name, which is then sent to the Face API. The Face API returns the gender information for the person in the image. Although many Twitter users provide unusable profile images, such as cartoon avatars or scenery pictures, the Face API does improve the overall gender estimation, where the gender recognition accuracy is 90.86% [21].

We use the gender estimates collectively from the first name-based approach and the Face API approach. If the two estimates agree, the user's gender information is retained. If they disagree, we use the results from the Face API to replace the gender estimated by first names with a probability value less than 0.8. If there are two or more persons in the image supplied to the Face API, we compare the gender information estimated from the Face API to that from the first name-based estimate; we retain the gender estimated from the first name-based approach as long as there is one person from the Face API with the same gender. Overall, with the combined efforts from the first name and the Face API approaches, 80% of all Twitter users can be identified with a gender in our preliminary analysis.

3.3. Twitter User Age Estimation with Image Recognition Techniques

The age of a Twitter user is estimated based on the user's profile image using the Microsoft Azure Face API, which is an image-based facial recognition service. A URL to a Twitter user's profile image is generated based on this user's screen name, which is then sent to the Face API. The Face API returns the corresponding age estimation for the person in the image. It can provide fairly accurate age estimation from the provided image, where the mean absolute error for real age estimation is 7.62 [92], much more accurate than age estimation using first names [72]. Age estimation based on first names needs to classify the first names into a collection of probabilities across different age groups, which is highly unreliable due to the recurrence of first names in different age groups. In our preliminary research, 45% of geotagged Twitter users can be estimated with an age on the basis of facial characteristics. The remaining users have profile images that are considered invalid, such as cartoons, animals, and scenery pictures; the Microsoft Azure Face API can recognize these images but return a Null value for age estimation.

3.4. Twitter User Race/Ethnicity Identification Based on Surname and Facial Recognition Techniques

Surnames carry rich information related to a person's geographic, social, and demographic background [87]. The U.S. Census Bureau provides a surname database with frequently occurring surnames from the 2010 census [93]. This database contains 162,255 names, and each surname is associated with a self-reported race/ethnicity probability, i.e., each surname in the database is provided with a probability of being the corresponding race/ethnicity group. In this study, we focus on three race/ethnicity groups: Hispanic, non-Hispanic White, and non-Hispanic Black. We match the derived surnames in the name database and get corresponding probabilities for the three race/ethnicity groups, which is similar to the approach used for estimating a user's gender based on first names. The derived probability of each surname belonging to one of the three race/ethnicity groups is represented as $\langle P_h, P_w, P_b \rangle$, where P_h is the probability of being Hispanic, P_w is for non-Hispanic White, and P_b is for non-Hispanic Black. Note that if the language of the Twitter message is labelled as "es" (i.e., Spanish), the probability of this person being Hispanic is 100%. In our preliminary study, 52% of geotagged Twitter users are identified with a race/ethnicity category. We will further implement facial recognition services as provided by Face++ and Kairos, which provide the capability of estimating the race/ethnicity information of a user detected in an image. However, these facial recognition services are still in early stage of development and their accuracy has not been well tested and documented. We will test the accuracy of these facial recognition services, and if reasonable, use them for enhancing our racial/ethnic estimates.

Although the preliminary study is based on geotagged tweets, the methods used in the study to estimate Twitter user demographics by gender, age, and race/ethnicity are independent of the geographical locations of the tweets, instead relying on information from a user's profile, i.e., name, screen name, and image. This means that the same methods can be applied to non-geotagged Twitter data and even other social media data with similar user profile structures.

4. Evaluate the Representativeness of Twitter User Demographics

Based on the estimated Twitter user demographics from Research Aim 1, Aim 2 is to evaluate the (mis)representativeness of Twitter user demographics by comparing them with the demographics of census ACS estimates. Aim 2 also involves an explanatory analysis of the biases. As census estimates are aggregated over predefined areal units, such as census tract, county, state, etc., it is necessary to prepare the Twitter data to be aggregated to the same areal unit, i.e., the county level in this study.

4.1. Aggregate Twitter Users to the County Level

For each individual Twitter user, the resident county is assigned by determining the most frequently tagged county in each year in the user's tweet collection. Twitter users, with their demographic characteristics (gender, age, and race/ethnicity), are aggregated to their corresponding counties. We also obtain county-level population estimates from the ACS for each county, including estimates for total population, different age groups, genders, and each of the three racial/ethnic groups.

In our preliminary study using 2014 geotagged Twitter data in the continental U.S., we estimated the demographics of each Twitter user, and then aggregated them to the county level by 11 demographic groups: females, males, age groups (20–24, 25–34, 35–44, 45–54, 55–64, and 65+), Hispanics, non-

Hispanic Whites, and non-Hispanic Blacks. We calculated the percentages of each Twitter user demographic group, and compared to the corresponding percentages of each population demographic group from the ACS estimates. We measured the bias by the median percentage error and median absolute percentage error for each demographic group at the county level; the results are shown in Table 1. As expected, Twitter users at ages 20–34 are over-represented while other age groups are under-represented. The biases for gender and race/ethnicity are difficult to explain. This is partly due to the fact that the percentage error measures are sensitive to distributional features in the data (e.g., skewed distributions) and do not provide simple and intuitive descriptions of the biases. Therefore, we will develop a representation index in the next step. We also use the Symmetric Mean Absolute Percentage Error (SMAPE) measures [94–96] to evaluate biases because of their two advantages: (1) the SMAPE scales the same for small proportions (e.g., percent Hispanic) and large proportions (e.g., percent Male), facilitating comparisons; and (2), it allows a computation of an average SMAPE for multiple criteria (e.g., a single summary measure that is the average error for gender, age, and race/ethnicity).

Table 1. Biases of Twitter User Estimates at the County Level in 2014

	Population Median	Twitter Users Median	Median Percentage Error	Median Absolute Percentage Error
Female	50.4%	52.3%	4.2%	8.3%
Male	49.6%	47.7%	–4.2%	8.3%
Age 20–24	10.1%	28.5%	173.7%	173.7%
Age 25–34	19.1%	49.3%	155.6%	155.6%
Age 35–44	18.3%	14.3%	–20.7%	26.0%
Age 45–54	19.8%	5.2%	–74.5%	74.5%
Age 55–64	18.7%	1.9%	–90.1%	90.1%
Age 65+	22.5%	1.4%	–93.7%	93.7%
Non-Hispanic White	89.2%	72.6%	–15.9%	17.6%
Non-Hispanic Black	2.3%	12.6%	369.4%	369.4%
Hispanics	4.1%	5.9%	31.6%	45.7%

4.2. Measures of Representativeness

We know that Twitter users are not representative of the population as a whole and that the use of Twitter data for population research could easily involve sampling biases. Before we seek insights to model and explain the determinants of biases, we need to get a better sense of the representativeness of the Twitter user population versus the population estimates from the census across counties. We will create measures to evaluate how representative Twitter users in each county are of the total population in each county. Our first measure is a representation (*r*) index, which is defined as:

$$r_j = (\frac{T_j}{T_{all}} / \frac{P_j}{P_{all}}) \qquad \text{(Eq. 1)}$$

where, T_j denotes the number of Twitter users in county j, P_j denotes the population in county j, T_{all} denotes the number of total Twitter users, and P_{all} denotes the total population of the U.S. A value of r_i equal to 1 indicates that the percentage of Twitter users in county j is equal to the national average of Twitter users. A value of r_j greater than 1 indicates an overrepresentation of Twitter users in county j. A value of r_j less than 1 indicates an underrepresentation of Twitter users in county j. Figure 2 shows the representation index for the continental U.S. in 2014. Twitter users are overrepresented in metropolitan areas and underrepresented in rural areas.

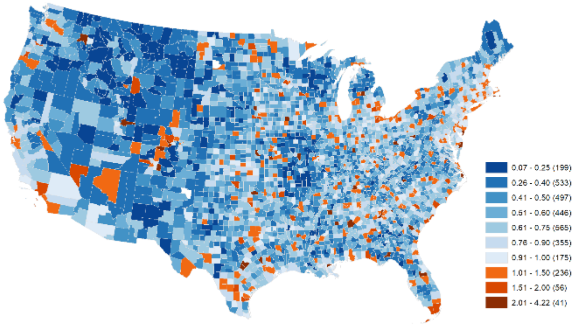


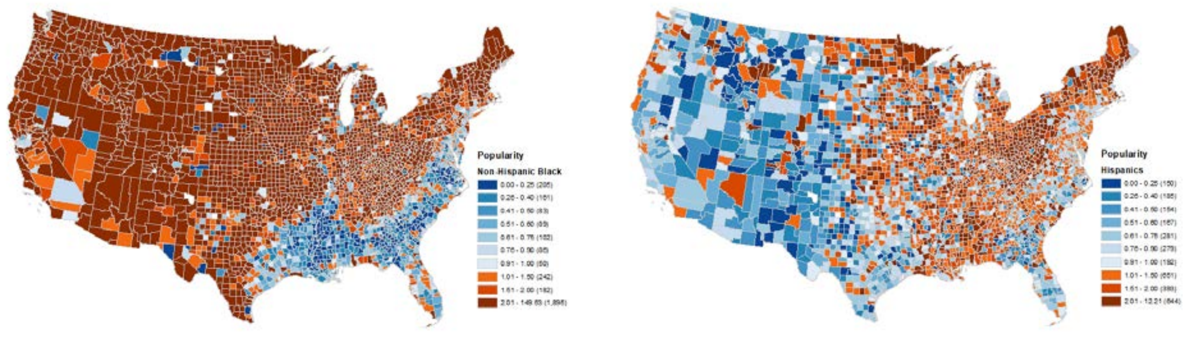
Figure 2. Representation Index for All Twitter Users in 2014

We can also apply Eq. (1) to each demographic group. We provide basic statistics of the representation index for each demographic group in Table 2. Twitter users are over-represented at ages 20–34 but under-represented at other age groups. The representation index decreases as the age group becomes older (Appendix 1).

Table 2. Representation Index of Twitter User Estimates at the County Level in 2014

	Mean	Median	Standard Deviation
Female	1.06	1.04	0.17
Male	0.95	0.96	0.16
Age 20–24	2.76	2.74	1.05
Age 25–34	2.49	2.56	0.82
Age 35–44	0.79	0.79	0.33
Age 45–54	0.27	0.25	0.14
Age 55–64	0.12	0.10	0.08
Age 65+	0.08	0.06	0.06
Non-Hispanic White	0.93	0.84	0.32
Non-Hispanic Black	10.52	4.69	16.33
Hispanics	1.65	1.32	1.22

Non-Hispanic Black and Hispanic Twitter users are over-represented (Table 2). Non-Hispanic Black Twitter users are under-represented in the south and southeast but over-represented in the rest of the country (Figure 3a). This is in contradictory to the distribution of non-Hispanic Black population. Also, although Hispanics have a smaller percentage of the total population in the northeast, Hispanics in this region use Twitter more than those in the other regions (Figure 3b). This contradiction might be due to the fact that the representation index (*r*) is skewed by the denominator.



(a) Non-Hispanic Black

(b) Hispanics

Figure 3. Representation Index for Non-Hispanic Blacks (a) and Hispanics (b) in 2014

Therefore, our second measure of representativeness is a relative representation difference (*rrd*) that compares each demographic group's Twitter representation and all Twitter users' representation. That is:

$$rrd_{jh} = r_{jh} - r_j \qquad \text{(Eq. 2)}$$

where h denotes a demographic group h, r_{jh} denotes the representation of demographic group h in county j, and r_j denotes the representation of all Twitters users in county j. A positive *rrd* indicates that the corresponding Twitter demographic group over-represents the Twitter user population in a county. A negative *rrd* indicates that the corresponding Twitter demographic group underrepresents the Twitter user population in a county. Figure 4 shows the relative representation difference for Hispanics in 2014. In general, we see that Hispanics are more likely to use Twitter than other race/ethnic groups in areas where Hispanics are less concentrated. Specifically, Hispanics are more likely to use Twitter than other racial/ethnic groups in the less-urbanized areas of the Northeast and Midwest and are less likely in the Southwest, Florida, Chicago and urban areas along the eastern seaboard, which are places where Hispanics are concentrated. We do not know the reason for this spatial pattern (which is exactly why we propose to understand the biases of Twitter demographics versus population demographics) but one possibility is that it reflects a heterogeneous process of Hispanic migration. Regardless, this underscores that we cannot use Twitter data in its raw form—not only are the biases large, but also the size of the biases are different in different places and for different groups.

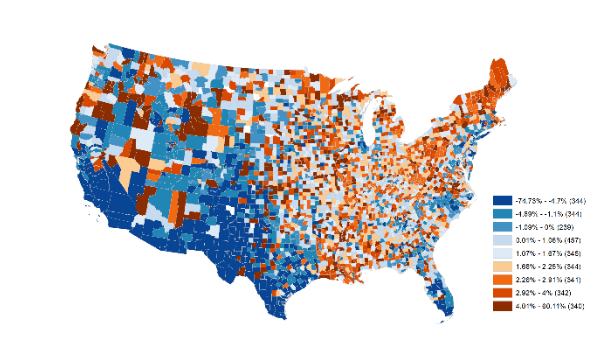


Figure 4. Relative Representation Difference for Hispanics in 2014

To summarize, the representation index is a global measure that considers Twitter users and population across all the counties. The relative representation difference serves as a local measure that compares a demographic group to the population. Because Twitter users and the population are heterogeneously distributed across the geographical space, to understand whether there are spatial patterns of the biases regarding the representativeness of the Twitter user population, we perform Local Indicators of Spatial Association (LISA) to understand the geographical distributions of the two representativeness measures [97]. The results from the preliminary studies show the existence of statistically significant spatial clusters of the biases (Appendix 2), possibly contributed by a set of structural factors across the areal units [78].

5. Determinants of Under- or Over-representation

We next model the under- or over-representation of demographic groups among Twitter users. Related studies have suggested several potential determinants that can contribute to the biases. As Twitter data are essentially Internet data, multiple factors affect the adoption, access, and usage patterns of the new technology, which lead to a digital division of the population. For example, Internet access is strongly correlated with various sociodemographic dimensions such as income, age, education, and gender [98]. People with lower incomes in the U.S. are less likely to have smartphones; this is particularly true of older residents, whose smartphone ownership can be as low as 16% [98]. Some research found that the Twitter user population is skewed towards younger adults, people with higher education level are more likely to use Twitter, and that the ratio between Twitter users and the whole population is significantly higher in urban areas than in rural areas [68]. We model the determinants of biases in linear functions using the following covariates: (1) rural or urban setting, (2) density of mobile-phone tower and Wi-Fi coverage, (3) income and education levels, and (4) the proportion of corresponding demographic groups. Their descriptive statistics are provided in Table 3.

Table 3. Variables and Descriptive Statistics

Variable	Description	Mean	Median	Standard deviation	Min	Max
Rural	Rural or metro areas (1 = rural; 2 = metro)	0.35	0	0.48	0	1
High school	% having high school degree	35%	35%	7%	9%	64%
College	% having college degree	30%	30%	5%	13%	47%
Income	Median household income ($)	46,358	44,731	11,940	19,146	123,966
Internet	# high-speed Internet providers	2.16	2	0.98	0	5

Notes: data for Internet speed are from the Federal Communications Commission (https://www.fcc.gov/internet-access-services-reports); data for all other variables are from the US Department of Agriculture Economic Research Service (https://www.ers.usda.gov/data-products/county-level-data-sets/download-data; https://www.ers.usda.gov/data-products/rural-urban-continuum-codes).

In our preliminary analysis, we fit ordinary least squares (OLS) regression models to examine the determinants of under- or over-representation of demographic groups among Twitter users. The dependent variable is the representation index of each demographic group as calculated in the previous section. The independent variables are rural status, high school, college, income, Internet, and the percentage of the corresponding demographic group in a county. If an independent variable has a statistically significant (*p* <=0.05) association with the representation index, the sign (positive or negative) of the coefficient is reported in Table 4. Female and male Twitter users are under-represented in rural counties. The representation of Twitter users in most demographic groups has a positive association with the income level of a county, but a negative association with the educational attainment. Existing studies suggest that people are more likely to use Twitter if they are well educated and have higher income. The conflicting finding may be due to the possible collinearity among the income and education variables. In our next step we will create a community disadvantage index (CDI) to integrate income, education, and unemployment variables. The representation index increases as the Internet coverage increases for females, males, age group 20–54, and non-Hispanic white.

Table 4. The Sign of Coefficients of the Ordinary Least Squares Regression Models (Dependent Variables = Representation Index)

Dependent variable	Rural	High school	College	Income	Internet	% demographic group
Female	–	–	–	+	+	+
Male	–	–	–	+	+	+
Age 20–24				+	+	
Age 25–34		–	–	+	+	+
Age 35–44		–	–	+	+	+
Age 45–54		–	–		+	+
Age 55–64		–	–		–	+
Age 65+	–	–	–	+	–	+
Non-Hispanic White		+		+	+	–
Non-Hispanic Black	+		+		–	–
Hispanics		+	–	–		–

Note: only statistically significant (*p* <=0.05) coefficients are reported. % demographic group refers to the percentage of the corresponding demographic group population.

The representations of non-Hispanic black and Hispanic Twitter users are very different from the other demographic groups. Non-Hispanic black Twitter users are over-represented in rural counties than in metro counties, and more represented in lower Internet coverage areas than in higher ones. Hispanic Twitter users are more represented in lower income counties than higher ones. These puzzling results need further investigation, by (1) creating a CDI index, (2) examining the bivariate relationship and possible nonlinear relationships between the dependent variable and each independent variable, and (3) adding the relative representation indices as dependent variables.

6. Next Steps

This draft presents preliminary results of the generalizability of Twitter users for population. By using a combination of multiple existing methods for demographic estimates of Twitter users, we were able to estimate demographics for a higher percentage of Twitter users. We also created two indices to measure the representation of Twitter users. While our findings echo those from existing studies showing higher usages of Twitter by younger population and minorities, our findings also suggest a strong spatial heterogeneity of Twitter usages.

Between now and the 2019 annual meeting of the Population Association of America (PAA), we will focus on investigating the determinants of under- or over-representation of Twitter users as discussed in the previous section. We will also use data from 2014–2017 to re-conduct the analyses. In addition, we will compare the 100% tweets with geotagged tweets in terms of the representation of Twitter data. Only a small percentage of tweets are geotagged and thus our geotagged tweets are not representative of all tweets. To deal with this issue, in future research we will purchase 100% tweets in the first week of October, 2014–2017 from Twitter Inc., which include both geotagged and non-geotagged ones, to test the difference in representativeness of the two types of Twitter data.

A further step beyond this PAA study would be to develop and validate weights to generalize Twitter data for population research. We will develop and compare five types of weighting procedures to generalize Twitter data for population research; we will assess whether weighted Twitter-based estimates come in line with population parameters and assess the replicability of these procedures for out-of-sample predictions. The five weighting procedures are: simple ratio weights, raking extension, propensity scores, matching methods, and multilevel regression with post-stratification [99,100]. If one of these methods achieves high levels of validity, it will be a breakthrough for using Twitter data for population research. We will adapt five different rebalancing methods to the particular features of Twitter data and then produce five corresponding sets of weighted population estimates at the county level. The assessment of these different methods will be conducted in a split-sample design in two ways. One, we will use a random selection of counties from 2014–2017 as "training" data and the remaining counties in the same years as "test" data. We will use the training data to assess the representativeness of Twitter data relative to census data for all U.S. counties, model the bias as a function of the differences between Twitter derived estimates and Census counts. Two, we will train the 2014–2015 data and test the 2016–2017 data. Here the question is whether the weighting procedures using older Census benchmarks, can be used to accurately track subsequent trends in population composition within counties. If successful, this endeavor will produce the necessary *tool* for social scientists to use Twitter data to study population characteristics and behaviors and social problems that cannot be measured well by traditional surveys.

Acknowledgements

This research was supported in part by the National Science Foundation (Awards # 1541136 and #1823633), the Eunice Kennedy Shriver National Institute of Child Health and Human Development (Award # P2C HD041025), and the Social Science Research Institute, Population Research Institute, and Institute for CyberScience of the Pennsylvania State University.

References

[1] Boyd, Danah and Kate Crawford. 2012. "Critical Questions for Big Data: Provocations for a Cultural, Technological, and Scholarly Phenomenon." *Information, Communication & Society* 15(5):662–679.

[2] Metaxas, Panagiotis T. and Eni Mustafaraj. 2012. "Social Media and the Elections." *Science* 338(6106):472–473.

[3] Beauchamp, Nicholas. 2017. "Predicting and Interpolating State-Level Polls Using Twitter Textual

Data." *American Journal of Political Science* 61(2):490–503.

[4] Frias-Martinez, Vanessa and Enrique Frias-Martinez. 2014. "Spectral Clustering for Sensing Urban Land Use Using Twitter Activity." *Engineering Applications of Artificial Intelligence* 35:237–245.

[5] Paul, Michael J. and Mark Dredze. 2011. "You Are What You Tweet: Analyzing Twitter for Public Health." Pp. 265–272 in *Proceedings of the Fifth International AAAI Conference on Weblogs and Social Media*, vol. 20. Barcelona, Spain, July 17-21, 2011.

[6] Signorini, Alessio, Alberto Maria Segre, and Philip M. Polgreen. 2011. "The Use of Twitter to Track Levels of Disease Activity and Public Concern in the US during the Influenza A H1N1 Pandemic." *PLOS One* 6(5):e19467 (DOI: 10.1371/journal.pone.0019467).

[7] Ruths, Derek and Jürgen Pfeffer. 2014. "Social Media for Large Studies of Behavior." *Science* 346(6213):1063–1064.

[8] Zagheni, Emilio, Venkata Rama Kiran Vrk Garimella, Ingmar Weber, and Bogdan State. 2014. "Inferring International and Internal Migration Patterns from Twitter Data." Pp. 439–444 in *Proceedings of the 23rd International Conference on World Wide Web - WWW '14 Companion*. Seoul, Korea, April 7-11, 2014.

[9] Zagheni, Emilio and Ingmar Weber. 2015. "Demographic Research with Non-Representative Internet Data." *International Journal of Manpower* 36(1):13–25.

[10] Cook, Fay Lomax. 2016. "Dear Colleague Letter: Robust and Reliable Research in the Social, Behavioral, and Economic Sciences, (NSF 16-137)." Retrieved January 3, 2018 (https://www.nsf.gov/pubs/2016/nsf16137/nsf16137.jsp).

[11] Gabel, Tim J. and Cathy Tokarski. 2014. "Big Data and Organizational Design: Key Challenges Await the Survey Research Firm." *Journal of Organization Design* 3(1):37–45.

[12] Smith, Michael A. and Brant Leigh. 1997. "Virtual Subjects: Using the Internet as an Alternative Source of Subjects and Research Environment." *Behavior Research Methods, Instruments, & Computers* 29(4):496–505.

[13] Alberti, Marina, John M. Marzluff, Eric Shulenberger, Gordon Bradley, Clare Ryan, and Craig Zumbrunnen. 2003. "Integrating Humans into Ecology: Opportunities and Challenges for Studying Urban Ecosystems." *BioScience* 53(12):1169–1179.

[14] Watts, Duncan J. 2007. "A Twenty-First Century Science." *Nature* 445(7127):489–489.

[15] Deville, Pierre, Catherine Linardc, Samuel Martin, Marius Gilbert, Forrest R. Stevens, Andrea E. Gaughan, Vincent D. Blondel, and Andrew J. Tatem. 2014. "Dynamic Population Mapping Using Mobile Phone Data." *Proceedings of the National Academy of Sciences* 111(45):15888–15893.

[16] Stevenson, Amanda Jean. 2014. "Finding the Twitter Users Who Stood with Wendy." *Contraception* 90(5):502–507.

[17] Willekens, Frans, Douglas Massey, James Raymer, and Cris Beauchemin. 2016. "International Migration under the Microscope." *Science* 352(6288):897–899.

[18] Berry, Brent. 2006. "Friends for Better or for Worse: Interracial Friendship in the United States as Seen through Wedding Party Photos." *Demography* 43(3):491–510.

[19] Palmer, John R. B., Thomas J. Espenshade, Frederic Bartumeus, Chang Y. Chung, Necati Ercan Ozgencil, and Kathleen Li. 2013. "New Approaches to Human Mobility: Using Mobile Phones for Demographic Research." *Demography* 50(3):1105–1128.

[20] Blumenstock, Joshua, Gabriel Cadamuro, and Robert On. 2015. "Predicting Poverty and Wealth from Mobile Phone Metadata." *Science* 350(6264):1073–1076.

[21] Lienemann, Brianna A., Jennifer B. Unger, Tess Boley Cruz, and Kar Hai Chu. 2017. "Methods for Coding Tobacco-Related Twitter Data: A Systematic Review." *Journal of Medical Internet Research* 19(3):e91 (DOI: 10.2196/jmir.7022).

[22] McCormick, T. H., H. Lee, N. Cesare, A. Shojaie, and E. S. Spiro. 2017. "Using Twitter for Demographic and Social Science Research: Tools for Data Collection and Processing." *Sociological Methods & Research Science* 46(3):390–421.

[23] Social Science Research. 2016. "Special Issue on Big Data in the Social Sciences." Retrieved January 12, 2018 (http://www.sciencedirect.com/journal/social-science-research/vol/59/suppl/C).

[24] Blumenstock, Joshua E. and Ott Toomet. 2014. "Segregation and 'Silent Separation': Using Large-Scale Network Data to Model the Determinants of Ethnic Segregation." Presented at *Annual Meeting of the Population Association of America*. Boston, MA, May 1-3, 2014.

[25] Cesare, Nina, Emma Spiro, and Hedwig Lee. 2015. "Self-Presentation and Information Disclosure on Twitter: Understanding Patterns and Mechanisms Along Demographic Lines." Presented at *Annual*

Meeting of the Population Association of America. San Diego, CA, April 30-May 2, 2015.

[26] Massey, D. 2016. “Measuring Racial Prejudice Using Google Trends.” Presented at *Annual Meeting of the Population Association of America*. Washington, DC, March 31-April 2, 2016.

[27] Mateos, P. and J. Durand. 2014. “Netography and Demography: Mining Internet Forums on Migration and Citizenship.” Presented at *Annual Meeting of the Population Association of America*. Boston, MA, May 1-3, 2014.

[28] Zagheni, Emilio, Ingmar Weber, and Krishna Gummadi. 2017. “Leveraging Facebook’s Advertising Platform to Monitor Stocks of Migrants.” *Population and Development Review* 43(4):721–734.

[29] Williams, Matthew L., Pete Burnap, and Luke Sloan. 2017. “Crime Sensing with Big Data: The Affordances and Limitations of Using Open-Source Communications to Estimate Crime Patterns.” *British Journal of Criminology* 57(2):320–340.

[30] Wilson, Robin et al. 2016. “Rapid and near Real-Time Assessments of Population Displacement Using Mobile Phone Data Following Disasters: The 2015 Nepal Earthquake.” *PLOS Currents* 8:1–25.

[31] Ginsberg, Jeremy, Matthew H. Mohebbi, Rajan S. Patel, Lynnette Brammer, Mark S. Smolinski, and Larry Brilliant. 2009. “Detecting Influenza Epidemics Using Search Engine Query Data.” *Nature* 457(7232):1012–1014.

[32] Bollen, Johan, Huina Mao, and Xiaojun Zeng. 2011. “Twitter Mood Predicts the Stock Market.” *Journal of Computational Science* 2(1):1–8.

[33] Chang, Ray M., Robert J. Kauffman, and Youngok Kwon. 2014. “Understanding the Paradigm Shift to Computational Social Science in the Presence of Big Data.” *Decision Support Systems* 63:67–80.

[34] McFarland, Daniel A., Kevin Lewis, and Amir Goldberg. 2016. “Sociology in the Era of Big Data: The Ascent of Forensic Social Science.” *The American Sociologist* 47(1):12–35.

[35] Taylor, Linnet. 2016. “No Place to Hide? The Ethics and Analytics of Tracking Mobility Using Mobile Phone Data.” *Environment and Planning D: Society and Space* 34(2):319–336.

[36] Morstatter, Fred, Shamanth Kumar, Huan Liu, and Ross Maciejewski. 2013. “Understanding Twitter Data with TweetXplorer.” Pp. 1482–1485 in *Proceedings of the 19th ACM SIGKDD international conference on Knowledge discovery and data mining - KDD ’13*. Chicago, IL, August 11-14, 2013.

[37] Twitter Inc. 2018. “Twitter Docs.” Retrieved January 10, 2018 (https://developer.twitter.com/en/docs).

[38] Twitter Inc. 2018. “Twitter Enterprise API.” Retrieved January 3, 2018 (https://developer.twitter.com/en/enterprise).

[39] Leetaru, Kalev, Shaowen Wang, Anand Padmanabhan, and Eric Shook. 2013. “Mapping the Global Twitter Heartbeat: The Geography of Twitter.” *First Monday* 18(5):5–6.

[40] Tufekci, Zeynep. 2014. “Engineering the Public: Big Data, Surveillance and Computational Politics.” *First Monday* 19(7):(DOI: 10.5210/fm.v19i7.4901).

[41] Jurgens, David, Tyler Finnethy, James McCorriston, Yi Tian Xu, and Derek Ruths. 2015. “Geolocation Prediction in Twitter Using Social Networks: A Critical Analysis and Review of Current Practice.” Pp. 1-10 in *Proceedings of the 9th International Conference on Weblogs and Social Media (ICWSM)*. Oxford, UK, May 26-29, 2015.

[42] Davis, Clodoveu A., Gisele L. Pappa, Diogo Rennó Rocha de Oliveira, and Filipe de L. Arcanjo. 2011. “Inferring the Location of Twitter Messages Based on User Relationships.” *Transactions in GIS* 15(6):735–751.

[43] Li, Rui, Shengjie Wang, Hongbo Deng, Rui Wang, and Kevin Chen-Chuan Chang. 2012. “Towards Social User Profiling: Unified and Discriminative Influence Model for Inferring Home Locations.” Pp. 1023–1031 in *Proceedings of the 18th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. Beijing, China, August 12-16, 2012.

[44] Li, Rui, Shengjie Wang, and Kevin Chen-Chuan Chang. 2012. “Multiple Location Profiling for Users and Relationships from Social Network and Content.” *Proceedings of the VLDB Endowment* 5(11):1603–1614.

[45] Rout, Dominic, Kalina Bontcheva, Daniel Preoţiuc-Pietro, and Trevor Cohn. 2013. “Where’s@ Wally?: A Classification Approach to Geolocating Users Based on Their Social Ties.” Pp. 11–20 in *Proceedings of the 24th ACM Conference on Hypertext and Social Media*. Paris, France, May 1-3, 2013.

[46] McGee, Jeffrey, James Caverlee, and Zhiyuan Cheng. 2013. “Location Prediction in Social Media Based on Tie Strength.” Pp. 459–468 in *Proceedings of the 22nd ACM International Conference on Information and Knowledge Management*. San Francisco, CA, October 27-Novemeber 1, 2013.

[47] Kong, Longbo, Zhi Liu, and Yan Huang. 2014. “SPOT: Locating Social Media Users Based on Social

Network Context." Pp. 1681–1684 in *Proceedings of the VLDB Endowment*, vol. 7. Hangzhou, China, September 1-5, 2014.

[48] Backstrom, Lars, Eric Sun, and Cameron Marlow. 2010. "Find Me If You Can: Improving Geographical Prediction with Social and Spatial Proximity." Pp. 61–70 in *Proceedings of the 19th international conference on World wide web - WWW '10*. Raleigh, NC, April 26-30, 2010.

[49] Jurgens, David. 2013. "That's What Friends Are for: Inferring Location in Online Social Media Platforms Based on Social Relationships." Pp. 273–282 in *Proceedings of the 7th International AAAI Conference on Weblogs and Social Media*. Cambridge, MA, July 8-11, 2013.

[50] Compton, Ryan, David Jurgens, and David Allen. 2014. "Geotagging One Hundred Million Twitter Accounts with Total Variation Minimization." Pp. 393–401 in *Proceedings of 2014 IEEE International Conference on Big Data*. Washington, DC, October 27-30, 2014.

[51] 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):260–271.

[52] Jurdak, Raja, Kun Zhao, Jiajun Liu, Maurice AbouJaoude, Mark Cameron, and David Newth. 2015. "Understanding Human Mobility from Twitter." *PLOS One* 10(7):e0131469 (DOI: 10.1371/journal.pone.0131469).

[53] Yin, Junjun, Yizhao Gao, Zhenhong Du, and Shaowen Wang. 2016. "Exploring Multi-Scale Spatiotemporal Twitter User Mobility Patterns with a Visual-Analytics Approach." *ISPRS International Journal of Geo-Information* 5(12):187 (DOI: 10.3390/ijgi5100187).

[54] DiGrazia, Joseph, Karissa McKelvey, Johan Bollen, and Fabio Rojas. 2013. "More Tweets, More Votes: Social Media as a Quantitative Indicator of Political Behavior" edited by L. M. Martinez. *PLOS One* 8(11):e79449 (DOI: 10.1371/journal.pone.0079449).

[55] West, Joshua Heber. 2012. "Temporal Variability of Problem Drinking on Twitter." *Open Journal of Preventive Medicine* 2(1):43–48.

[56] Hanson, Carl L., Scott H Burton, Christophe Giraud-Carrier, Josh H West, Michael D Barnes, and Bret Hansen. 2013. "Tweaking and Tweeting: Exploring Twitter for Nonmedical Use of a Psychostimulant Drug (Adderall) among College Students." *Journal of Medical Internet Research* 15(4).

[57] Daniulaityte, Raminta et al. 2015. "'Time for Dabs': Analyzing Twitter Data on Marijuana Concentrates across the U.S." *Drug and Alcohol Dependence* 155:307–311.

[58] Weber, Ingmar, Venkata Rama Kiran Garimella, Emilio Zagheni, and Bogdan State. 2014. "Using Geo-Located Twitter Data to Study Recent Patterns of International and Internal Migration in OECD Countries." Pp. 1–9 in *Proceedings of 2014 European Population Conference*. Budapest, Hungary, 25-28 June, 2014.

[59] Hübl, Franziska, Sreten Cvetojevic, Hartwig Hochmair, and Gernot Paulus. 2017. "Analyzing Refugee Migration Patterns Using Geo-Tagged Tweets." *ISPRS International Journal of Geo-Information* 6(10):302 (DOI: 10.3390/ijgi6100302).

[60] 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* 103–110. Troy, NY, June 25-28, 2017.

[61] Flores, René D. 2017. "Do Anti-Immigrant Laws Shape Public Sentiment? A Study of Arizona's SB 1070 Using Twitter Data." *American Journal of Sociology* 123(2):333–384.

[62] Seabold, Skipper, Alex Rutherford, Olivia De Backer, and Andrea Coppola. 2015. *The Pulse of Public Opinion: Using Twitter Data to Analyze Public Perception of Reform in El Salvador*. Retrieved (https://openknowledge.worldbank.org/bitstream/handle/10986/22656/The0pulse0of0p0eform0in0El0Salvador.pdf?sequence=1&isAllowed=y).

[63] Kim, Annice E., Heather M. Hansen, Joe Murphy, Ashley K. Richards, Jennifer Duke, and Jane A. Allen. 2013. "Methodological Considerations in Analyzing Twitter Data." *Journal of the National Cancer Institute - Monographs* 47:140–146.

[64] Mislove, Alan, Sune Lehmann, Yong-yeol Ahn, Jukka-pekka Onnela, and J.Niels Rosenquist. 2011. "Understanding the Demographics of Twitter Users." Pp. 554–557 in *Proceedings of the Fifth International Conference on Weblogs and Social Media*. Barcelona, Catalonia, Spain, July 17-21, 2011.

[65] Nguyen, Quynh C. et al. 2016. "Leveraging Geotagged Twitter Data to Examine Neighborhood

Happiness, Diet, and Physical Activity." *Applied Geography* 73:77–88.

[66] Jones, Nickolas M., Sean P. Wojcik, Josiah Sweeting, and Roxane Cohen Silver. 2016. "Tweeting Negative Emotion: An Investigation of Twitter Data in the Aftermath of Violence on College Campuses." *Psychological Methods* 21(4):526–541.

[67] Duggan, Maeve and Aaron Smith. 2014. "Social Media Update 2013." *Pew Research Center*. Retrieved January 12, 2018 (http://pewinternet.org/Reports/2013/Social-Media-Update.aspx).

[68] Yildiz, Dilek, Jo Munson, Agnese Vitali, Ramine Tinati, and Jennifer A. Holland. 2017. "Using Twitter Data for Demographic Research." *Demographic Research* 37:1477–1514.

[69] Gayo-Avello, Daniel, Panagiotis Takis Metaxas, and Eni Mustafaraj. 2011. "Limits of Electoral Predictions Using Twitter." Pp. 490–493 in *Proceedings of the Fifth International AAAI Conference on Weblogs and Social Media*. Barcelona, Catalonia, Spain, July 17-21, 2011.

[70] Longley, Paul A., Muhammad Adnan, and Guy Lansley. 2015. "The Geotemporal Demographics of Twitter Usage." *Environment and Planning A* 47(2):465–484.

[71] Longley, Paul A. and Muhammad Adnan. 2016. "Geo-Temporal Twitter Demographics." *International Journal of Geographical Information Science* 30(2):369–389.

[72] Luo, Feixiong, Guofeng Cao, Kevin Mulligan, and Xiang Li. 2016. "Explore Spatiotemporal and Demographic Characteristics of Human Mobility via Twitter: A Case Study of Chicago." *Applied Geography* 70:11–25.

[73] Vidal, Leticia, Gastón Ares, Leandro Machín, and Sara R. Jaeger. 2015. "Using Twitter Data for Food-Related Consumer Research: A Case Study On 'what People Say When Tweeting about Different Eating Situations.'" *Food Quality and Preference* 45:58–69.

[74] 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):1–9.

[75] Tang, C., K. Ross, N. Saxena, and R. Chen. 2011. "What's in a Name: A Study of Names, Gender Inference, and Gender Behavior in Facebook." *Database Systems for Advanced Applications* 344–356. Retrieved (http://link.springer.com/chapter/10.1007/978-3-642-20244-5_33).

[76] Megvii Inc. 2018. "Face++ Face Detection API." Retrieved January 4, 2018 (https://www.faceplusplus.com/).

[77] Kairos. 2018. "Kairos Face Recognition API." Retrieved January 4, 2018 (https://www.kairos.com/face-recognition-api).

[78] Chi, Guangqing, Junjun Yin, and Jennifer Van Hook. 2017. "Predicting Twitter User Demographics as a First Step in Big Data for Population Research." Presented at the 28th International Population Conference of the International Union for the Scientific Study of Population, Cape Town, South Africa, October 29-November 4, 2017.

[79] Mahmud, Jalal, Jeffrey Nichols, and Clemens Drews. 2014. "Home Location Identification of Twitter Users." *ACM Transactions on Intelligent Systems and Technology* 5(3):1–21.

[80] Twitter Inc. 2018. "Twitter Streaming API." Retrieved January 11, 2018 (https://developer.twitter.com/en/docs).

[81] Pittsburg Supercomputing Center. 2018. "Bridges Supercomputer." Retrieved January 11, 2018 (https://www.psc.edu/bridges).

[82] Dean, Jeffrey and Sanjay Ghemawat. 2010. "MapReduce: A Flexible Data Processing Tool" *Communications of the ACM* 53(1):72–77.

[83] Zhang, Chao Michael and Vern Paxson. 2011. "Detecting and Analyzing Automated Activity on Twitter." Pp. 102–111 in *Lecture Notes in Computer Science*, vol. 6579. Springer, Berlin, Heidelberg.

[84] Wang, Alex Hai. 2010. "Detecting Spam Bots in Online Social Networking Sites: A Machine Learning Approach." Pp. 335–342 in *Data and Applications Security and Privacy XXIV*, edited by S. Foresti and S. Jajodia. Springer, Berlin, Heidelberg.

[85] Stringhini, Gianluca, Christopher Kruegel, and Giovanni Vigna. 2010. "Detecting Spammers on Social Networks." Pp. 1–9 in *Proceedings of the 26th Annual Computer Security Applications Conference*. Austin, TX, December 6-10, 2010.

[86] Yin, Junjun, Aiman Soliman, Dandong Yin, and Shaowen Wang. 2017. "Depicting Urban Boundaries from a Mobility Network of Spatial Interactions: A Case Study of Great Britain with Geo-Located Twitter Data." *International Journal of Geographical Information Science* 31(7):1293–1313.

[87] Mateos, Pablo, Paul A. Longley, and David O'Sullivan. 2011. "Ethnicity and Population Structure in Personal Naming Networks". *PLOS One* 6(9):e22943 (DOI: 10.1371/journal.pone.0022943).

[88] U.S. Social Security Administration. 2014. “Names from US Social Security Card Application.” Retrieved January 11, 2018 (https://www.ssa.gov/ssnumber/).

[89] Tang, Cong, Keith Ross, Nitesh Saxena, and Ruichuan Chen. 2011. “What’s in a Name: A Study of Names, Gender Inference, and Gender Behavior in Facebook.” Pp. 344–356 in *Database Systems for Advanced Applications*.

[90] Edmonds, Rick, Emily Guskin, Amy Mitchell, and Mark Jurkowitz. 2013. *The State of the News Media 2013*. Retrieved (http://www.pewresearch.org/topics/state-of-the-news-media/).

[91] Microsoft Azure. 2018. “Face Verification.” Retrieved January 4, 2018 (https://azure.microsoft.com/en-us/services/cognitive-services/face/).

[92] Sightsound Inc. 2018. “Benchmarks, Age, Gender and Emotion.” Retrieved January 3, 2018 (https://www.sighthound.com/technology/age-gender-emotion/benchmarks).

[93] United States Census Bureau. 2016. “Surnames from US 2010 Census.” Retrieved January 3, 2018 (https://www.census.gov/topics/population/genealogy/data/2010_surnames.html).

[94] Armstrong, J.Scott. 1985. *Long-Range Forecasting: From Crystal Ball to Computer*. 2nd. Wiley-Interscience.

[95] Flores, Benito E. 1986. “A Pragmatic View of Accuracy Measurement in Forecasting.” *Omega* 14(2):93–98.

[96] Tofallis, Chris. 2015. “A Better Measure of Relative Prediction Accuracy for Model Selection and Model Estimation.” *Journal of the Operational Research Society* 66(8):1352–1362.

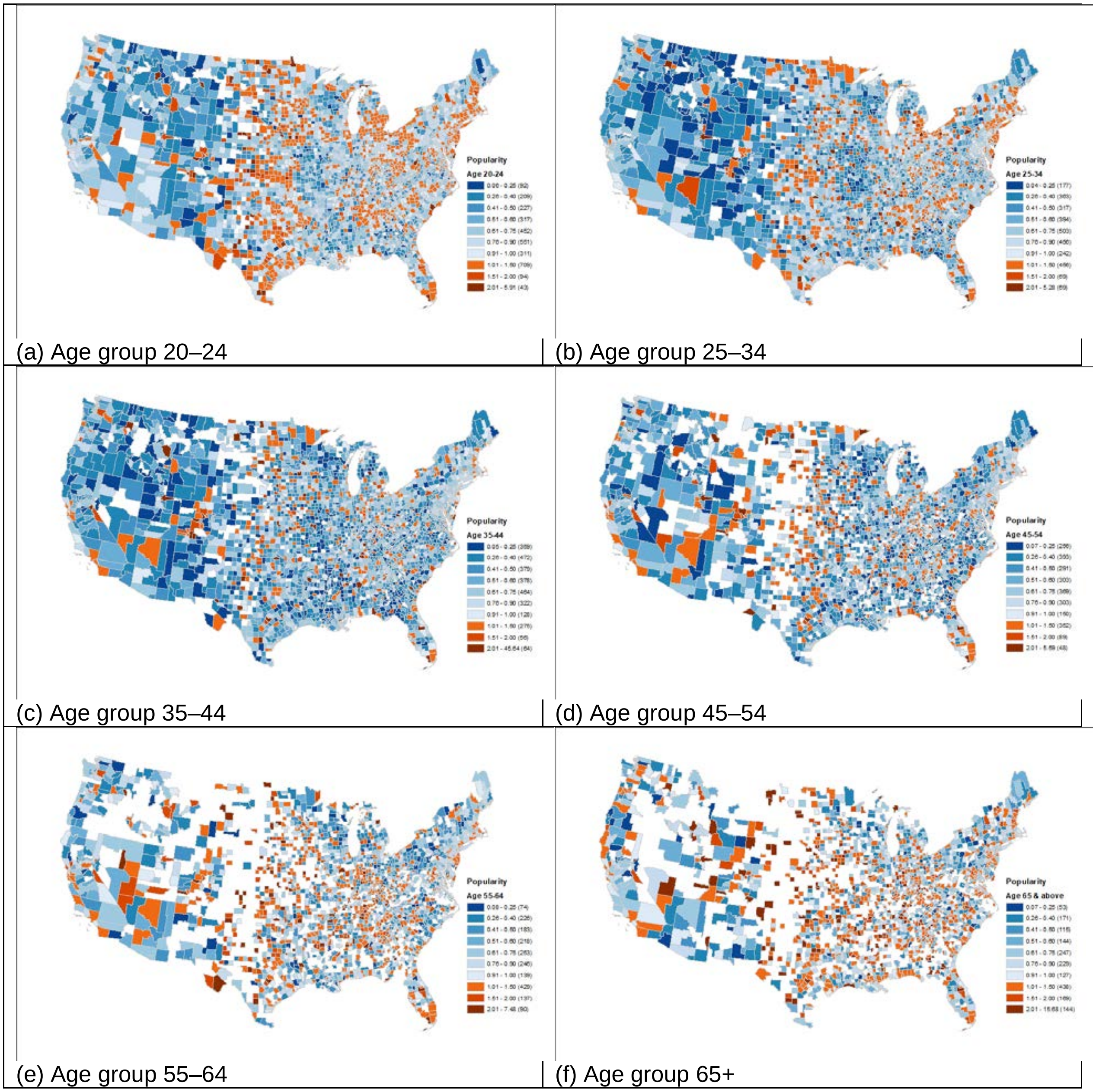
[97] Anselin, Luc. 1995. “Local Indicators of Spatial association—LISA.” *Geographical Analysis* 27(2):93–115.

[98] Friemel, Thomas N. N. 2016. “The Digital Divide Has Grown Old: Determinants of a Digital Divide among Seniors.” *New Media and Society* 18(2):313–331.

[99] Schnell, Rainer, Marcel Noack, and Sabrina Torregroza. 2017. “Differences in General Health of Internet Users and Non-Users and Implications for the Use of Web Surveys.” *Survey Research Methods* 11(2):105–123.

[100] Rivers, Douglas and Delia Bailey. 2009. “Inference from Matched Samples in the 2008 US National Elections.” Pp. 627–639 in *Joint Statistical Meetings Proceedings*, vol. 1. Washington, DC, August 1-6, 2009.

Appendix 1. Representation Index by Age Groups in 2014



(a) Age group 20–24

(b) Age group 25–34

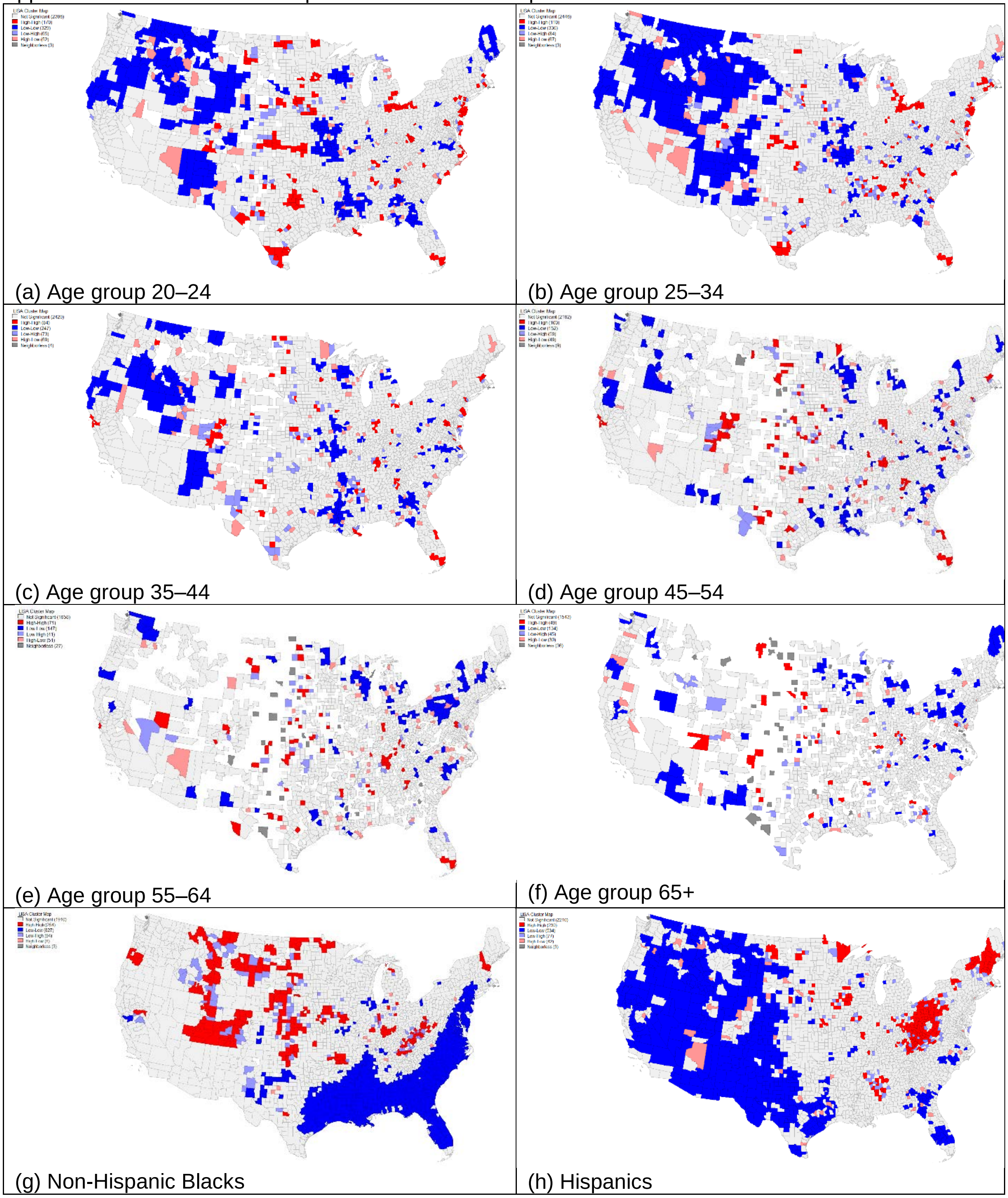
(c) Age group 35–44

(d) Age group 45–54

(e) Age group 55–64

(f) Age group 65+

Appendix 2. Local Indicator of Spatial Association of Representation Indices in 2014



(a) Age group 20–24

(b) Age group 25–34

(c) Age group 35–44

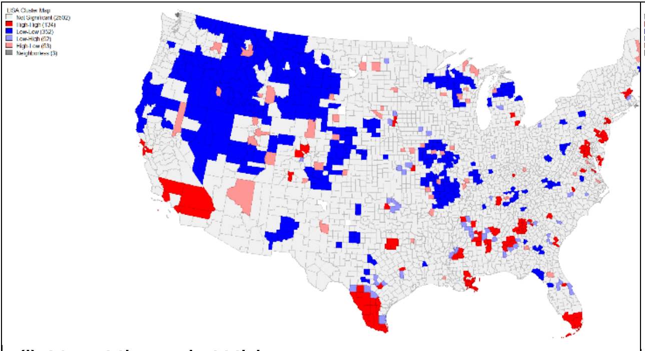
(d) Age group 45–54

(e) Age group 55–64

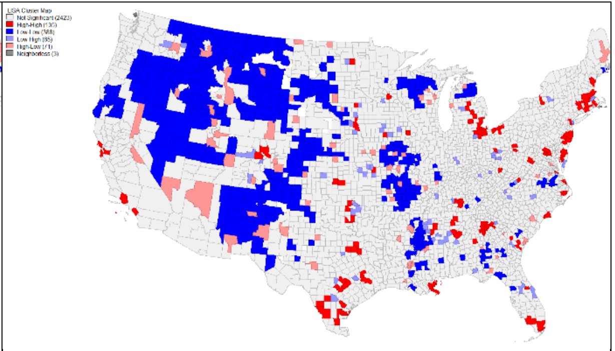
(f) Age group 65+

(g) Non-Hispanic Blacks

(h) Hispanics



(i) Non-Hispanic Whites



(j) Females