DEMOGRAPHIC RESEARCH
A peer-reviewed, open-access journal of population sciences

DEMOGRAPHIC RESEARCH

VOLUME 37, ARTICLE 46, PAGES 1477–1514
PUBLISHED 22 NOVEMBER 2017
http://www.demographic-research.org/Volumes/Vol37/46/
DOI: 10.4054/DemRes.2017.37.46

Research Article

Using Twitter data for demographic research

Dilek Yildiz **Ramine Tinati**

Jo Munson **Jennifer A. Holland**

Agnese Vitali

This publication is part of the Special Collection on "Social Media and Demographic Research," organized by Guest Editor Emilio Zagheni.

Contents

Demographic Research: Volume 37, Article 46
Research Article

Using Twitter data for demographic research

Dilek Yildiz[1]

Jo Munson[2]

Agnese Vitali[2]

Ramine Tinati[2]

Jennifer A. Holland[3]

Abstract

BACKGROUND
Social media data is a promising source of social science data. However, deriving the demographic characteristics of users and dealing with the nonrandom, nonrepresentative populations from which they are drawn represent challenges for social scientists.

OBJECTIVE
Given the growing use of social media data in social science research, this paper asks two questions: 1) To what extent are findings obtained with social media data generalizable to broader populations, and 2) what is the best practice for estimating demographic information from Twitter data?

METHODS
Our analyses use information gathered from 979,992 geo-located Tweets sent by 22,356 unique users in South East England between 23 June and 4 July 2014. We estimate demographic characteristics of the Twitter users with the crowd-sourcing platform CrowdFlower and the image-recognition software Face++. To evaluate bias in the data, we run a series of log-linear models with offsets and calibrate the nonrepresentative sample of Twitter users with mid-year population estimates for South East England.

[1] Wittgenstein Centre for Demography and Global Human Capital (IIASA, VID/ÖAW, WU), Austria.
E-Mail: Dilek.Yildiz@oeaw.ac.at.
[2] University of Southampton, UK.
[3] Erasmus Universiteit Rotterdam, the Netherlands.

RESULTS
CrowdFlower proves to be more accurate than Face++ for the measurement of age, whereas both tools are highly reliable for measuring the sex of Twitter users. The calibration exercise allows bias correction in the age-, sex-, and location-specific population counts obtained from the Twitter population by augmenting Twitter data with mid-year population estimates.

CONTRIBUTION
The paper proposes best practices for estimating Twitter users' basic demographic characteristics and a calibration method to address the selection bias in the Twitter population, allowing researchers to generalize findings based on Twitter to the general population.

1. Introduction

Social media promises to become a rich source of social science data, offering insights into attitudes, behaviour, discourse, and the social linkages and interactions between individuals. The growing importance of social media data for use in the social sciences, including demography, arises from the availability and affordability of swiftly collected real time, large-scale data (Savage and Burrows 2007; Housley et al. 2013; McCormick et al. 2015). With this data, we can follow individuals and their networks over time and across space, offering a new source of demographic information to augment population estimates, particularly when data quality is poor or when data is unavailable (Zagheni and Weber 2015).

Despite its potential, social media data presents several challenges for social scientists. User demographic information is not always readily available. Increasingly scholars are developing methodologies to estimate these and other characteristics, however best practices have not yet been established. A second challenge is to evaluate the selection bias in these nonrandom, nonrepresentative samples (Zagheni and Weber 2015). This evaluation is essential for understanding whether it is possible to investigate social media data from a population perspective, mapping findings onto populations.

The paper asks to what extent findings that are obtained with social media data are generalizable to broader populations. We focus on data derived from user profiles and posts on the micro-blogging platform Twitter. As of 2016, an estimated 310 million people worldwide used Twitter, with over 100 million active daily users generating hundreds of millions of Tweets (Smith 2016). This data is increasingly used in social science research because it is readily available to researchers and provides rich

qualitative and social network information. However, since Twitter does not require users to provide their demographic characteristics (such as sex and age), often the demographics of Twitter users must be inferred or estimated.

With an interest in understanding populations and sampling on Twitter, we analysed a sample of geo-located Tweets posted in the days just prior to and following the UK Office for National Statistics' 2014 mid-year population estimates (Office for National Statistics 2015), i.e., the estimated counts of the usual resident population of England and Wales on 30 June 2014 (n_{Tweets} = 2,219,495; n_{users} = 132,991). Data was collected using DataSift's Twitter Firehose connection,[4] representing 100% of the Twitter public data stream. In this 'proof of concept' study, we restricted our analyses to the 67 local authorities constituting the South East region of England (n_{Tweets} = 979,992; n_{users} = 22,356).

To estimate demographic characteristics (age and sex) of the users, we tested two commonly employed methodologies: crowdsourcing via the CrowdFlower Crowdsourcing platform, and image recognition via Face++ (http://en.faceplusplus.com). Using these estimates, combined with location information from the geo-located Tweets, we extended a calibration methodology developed by Yildiz and Smith (2015) to the framework proposed by Zagheni and Weber (2015). We compared the Twitter population by age group, sex, and location to the usual resident population of South East England, using the 2014 mid-year population estimates. Employing this calibration methodology to combine Twitter data with auxiliary population estimates, we can reduce the bias in the Twitter data. Moreover, the methodology allows us to compare the accuracy of the age and sex estimates produced by the crowd-sourcing and image-recognition approaches. These results are useful for developing best practices for estimating demographic (and potentially other) characteristics of social media users.

In the following section, we discuss how we accessed and processed the Twitter data. We then outline the methodologies and results of the two approaches used for estimating the age and sex of users (Section 3). In Section 4, we develop a calibration methodology for matching the Twitter data with population estimates and a model to assess the bias in samples derived from our age and sex estimation approaches. Results of the calibration models and conclusions are presented in Sections 5 and 6, respectively.

[4] After our data was collected, Twitter revoked DataSift's access to the Twitter Firehose. One of the benefits of social media data is that it is dynamic, updating, and changing in real time, providing a current picture of the social world. However, this also presents challenges for researchers, with respect to replicability. This challenge is further complicated by the dynamic nature of the data and methodological platforms used to collect and analyse the data.

2. Data

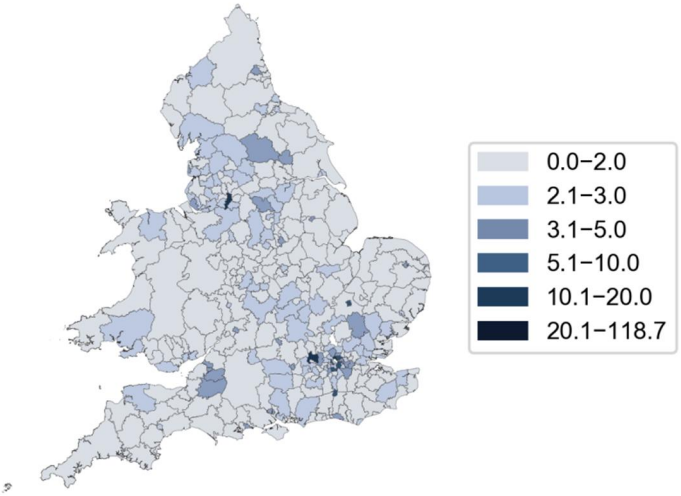
Using DataSift's Twitter Firehose connection, we collected all English-language (tweet_lang: en) Tweets geo-tagged within the latitude/longitude boundary box of the United Kingdom and posted between 23 June and 4 July 2014, a period straddling the mid-year population estimates for the usual resident population of England and Wales on 30 June 2014. From this data, we derived a sample of 'active' Twitter users, i.e., those users who tweeted at least once during our observation period. In addition, to be able to estimate the location of Twitter users, we restricted our sample to those users who had at least one geo-located tweet during the week of observation or who completed the location field in their Twitter profile. The final sample comprises over 132,000 unique users for the reference period in 2014, of which 22,356 are located in our region of interest, the South East. Table 1 provides an overview of the data collected. Figure 1 shows the density of Twitter users relative to the resident population in England and Wales. The density is highest in highly populated urban areas such as London. Our focal region, South East England, shows considerable heterogeneity in Twitter density, including local authorities with the highest and lowest densities in the sample.

Table 1: Description of the Twitter sample

Dataset/details	Twitter, 2014
Filter	Geo polygon: UK boundary (twitter.geo)
API feed	Firehose (via DataSift)
Time frame	23 June to 4 July 2014
Tweets*	2,219,495
Unique Twitter users*	132,991
Tweets*, South East England	979,992
Unique Twitter users*, South East England	22,356

Notes: *After assignment to the boundaries of a local authority within England and Wales.

Figure 1: Number of (geo-located) Twitter users per 1,000 population, England and Wales, 2014



3. Estimating demographic characteristics of Twitter users

3.1 Experimental setup

Twitter does not require users to provide demographic information; therefore, the demographics of its population need to be estimated. In this study, we focus on the location, age, and sex of the Twitter users. Following a precedent set in previous studies, we restrict our analyses to those Twitter users who geo-tagged at least one tweet during the week of observation or who completed the location field in their Twitter profile (Mislove et al. 2011; Zagheni et al. 2014). We note that self-reported information is not verified and may not accurately reflect the usual location of the user at the time of tweeting. To estimate age and sex, we utilize and compare two approaches: crowdsourcing and image recognition. These two approaches were selected for their efficiency and for the ability to compare human and nonhuman labelling techniques. While we do not consider them here, other approaches to estimate Twitter users' demographic characteristics include textual and language analysis of Tweet content (Sloan et al. 2013; Sloan et al. 2015) and analyses of self-reported names (Mislove et al. 2011; Sloan et al. 2013).

Our first approach employs the CrowdFlower crowdsourcing platform. CrowdFlower was initially developed in 2012 as part of a research project interested in crowdsourcing technologies to improve the cost, speed, and quality of work (Van Pelt and Sorokin 2012). The project was later turned into a commercial platform that enables (paying) customers to design, launch, and monitor crowdsourcing 'tasks,' and linking these customers to a large pool of (paid) contributors, i.e., 'crowd-workers,' who take part in these tasks.

Previous studies have shown that crowd-workers tend to be young, female, and located in the United States and India (Ipeirotis 2010; McCormick et al. 2015). Age and sex demographics were not available for the crowd-workers employed on our tasks, but 71% of the workers declared 'country of origin' (Table 2). In contrast to the findings of Ipeirotis (2010) and McCormick et al. (2015), the modal origin group was from Venezuela (13.4%), followed by India and the United States.

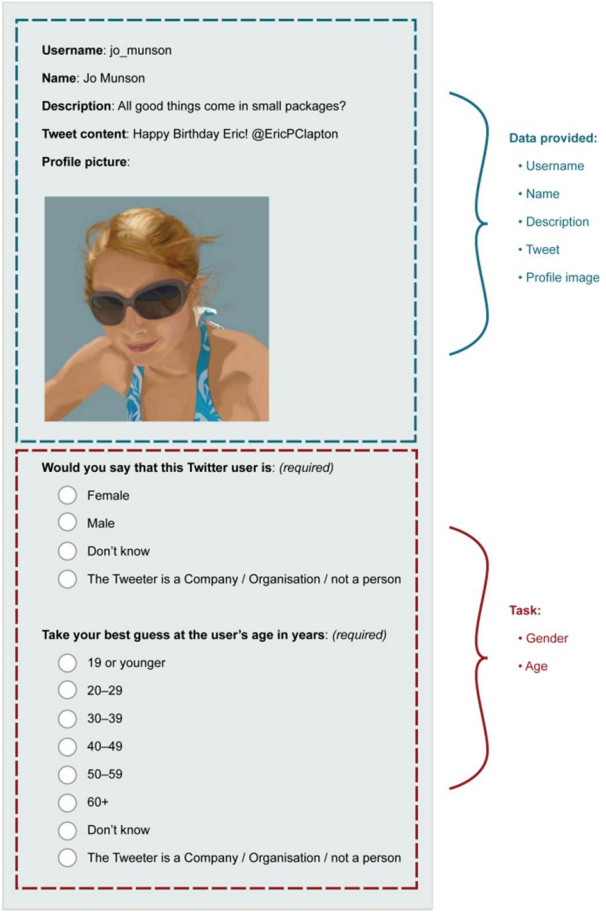
Crowd-workers' country of origin has implications for accuracy of age and sex classifications in our experiment because the tasks included the interpretation of free text written in English. Of the 71% of users with a declared country of origin, only 9% originated from countries with English as a primary or native language. A further 6% came from countries where 50% or more of the population speaks English, while the remaining 85% originated from countries where less than 50% of the population speaks English.

Table 2: Top countries of origin of CrowdFlower workers employed on this study

Country	Count	Percentage
Venezuela	294	14.20%
India	150	7.20%
Brazil	105	5.10%
United States	105	5.10%
Italy	102	4.90%
Serbia	91	4.40%
Spain	85	4.10%
Mexico	72	3.50%
Russia	67	3.20%
Turkey	60	2.90%
Bosnia-Herzegovina	53	2.60%
Canada	53	2.60%

Using the CrowdFlower Markup Language (CML), we designed a task that presented workers with specific fields from the Twitter data and asked them to answer two questions relating to sex (male or female) and age group (0–19, 20–29, 30–39, 40–49, 50–59, or ≥60). Figure 2 illustrates the layout of the task. Five tasks were listed per page, with a pricing of two cents a page for completing all tasks. While we are aware of the importance of a task's cost (Cheng, Teevan, and Bernstein 2015; Cheng et al. 2015; Kittur et al. 2013), we keep the price fixed for all experiments conducted.

Figure 2: Exemplar CrowdFlower task



The second approach uses Face++, an automated face-detection algorithm developed by Megvii Inc. (2013) that is frequently used for estimating the demographic characteristics of Twitter users (e.g., Zagheni et al. 2014; Fan et al. 2014; Zhou et al. 2013; Vikatos et al. 2017). Face++ takes links to image files as its input variable and outputs an age, sex, and ethnicity/race estimate (Figure 3). For this study, we use only the age and sex output fields. We applied the Face++ software to Twitter user profile pictures. Face++ requires one or more distinguishable faces in the image provided to return a valid result. Images that show nonhuman entities or where the algorithm is unable to identify a face return a null result. We acknowledge that an alternative software programme for image recognition is Microsoft Computer Vision API.

Figure 3: Exemplar Face++

Input

https://someimage.jpg

Output:

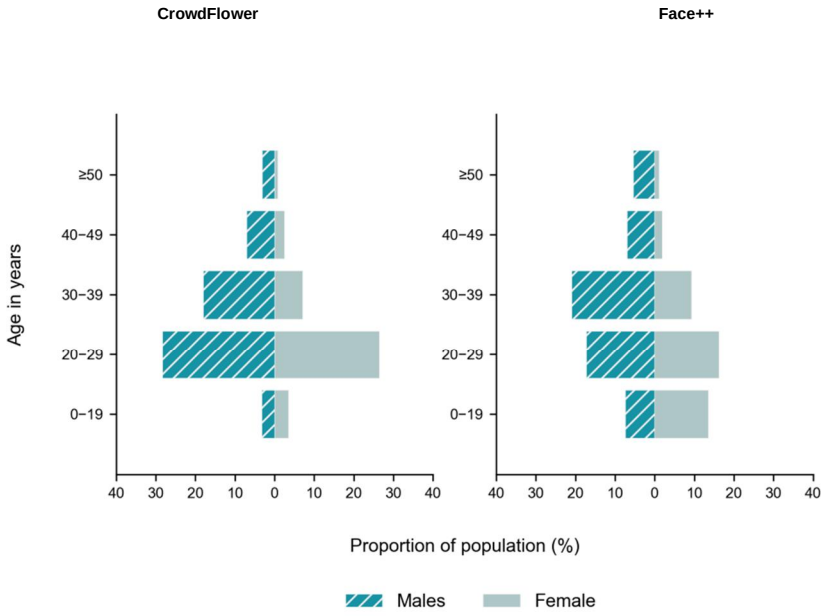

```
{'face' : [
        {'attribute' {
                'age' : {'value' : 18},
                'gender' : {
                        'confidence' : 94.6508,
                        'value' : 'Female'},
        }
]},
```

3.2 Demographic characteristics of Twitter users in South East England

For both CrowdFlower and Face++, males outnumber females in all age groups, except for the age group 0–19 in Face++. According to the 2014 mid-year population estimates, on average there are 96.8 males per 100 females in South East England. According to the sex estimates based on CrowdFlower, we find an average of 149 males per 100 females in the Twitter sample. According to Face++ estimates, on average there are 138.6 males per 100 females in the Twitter sample.

Figure 4 reports the population pyramids from the 2014 Twitter population with demographic information estimated via CrowdFlower and Face++. Due to small Ns, here we collapse the categories 50–59 and 60+. According to CrowdFlower, the age group 20–29 represents the modal age for both males and females, followed by the age group 30–39. The age groups 0–19 and 50+ are, as expected, the least represented age groups in the Twitter sample. For Face++, the most frequent group in the Twitter sample is men aged 30–39, followed by both men and women aged 20–29. According to image-recognition estimates, the youngest age group represents a higher proportion of the total Twitter population, as compared to the CrowdFlower estimates, particularly among females. This result could be partially explained by the fact that parents may use the picture of their child(ren) as their profile picture.

Figure 4: 2014 population pyramids based on Twitter data, demographic variables estimated with Crowdflower and Face++



3.3 Performance of demographic determination algorithms

The most common measure of performance for algorithms that attempt to assign data points to one of two or more categories is the total accuracy measure:

$$\text{Total Accuracy} = (TN + TP) / (FN + FP + TN + TP), \quad (1)$$

where:

TN: count of true negatives
TP: count of true positives
FN: count of false negatives
FP: count of false positives.

A 'gold standard' set was determined by taking a random sample of 250 geo-tagged Twitter users from the collected Tweets and using multiple search techniques to manually verify age and sex, including LinkedIn profiles, electoral roll listings, personal websites, Twitter descriptions, and Twitter profile images. Analysis of the gold standard set indicates a predominance of male users, with a sex ratio of 2.7 males for each female (Figure 5), and a positive skew in the age distribution of users (Figure 6). For sex, 5% of users were classified as 'Not a person/unknown.' For age, 28% of the gold standard set was classified as 'Not a person/unknown,' reflecting the relative difficulty of determining the age of users.

Figure 5: Gold standard sex distribution

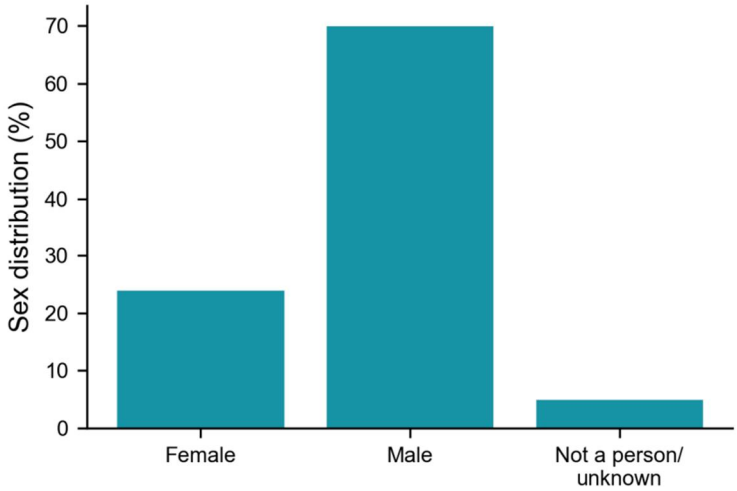


Figure 6: Gold standard age distribution

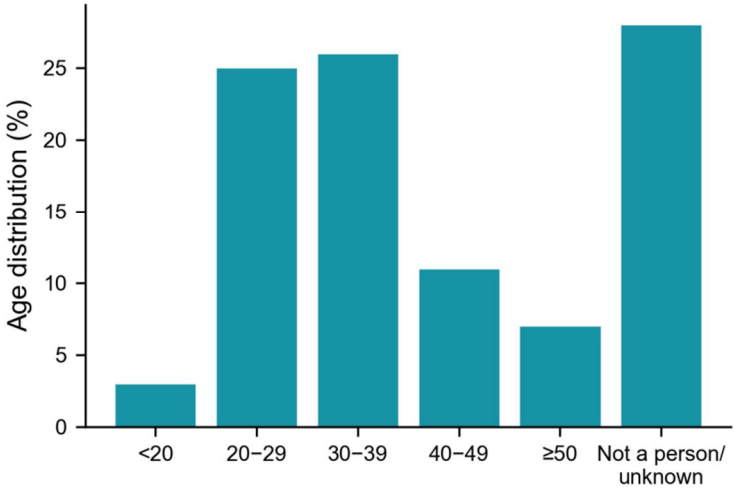


Table 3: CrowdFlower and Face++ performance

	Total accuracy (N=250)		Total accuracy, valid images (N=114)	
	Age	Sex	Age	Sex
CrowdFlower	60.8%	86.4%	56.1%	93.9%
Face++	40.4%	44.8%	32.5%	87.7%

Face++ is not an effective tool for the measurement of age. The CrowdFlower age assignment is significantly better at 60.8%. The counterintuitive worsening of the Total Accuracy when the sample is restricted to valid images reflects the fact that our sample included Twitter accounts associated with a company or organisation. The Face++ algorithm matched some of these corporate accounts to individuals due to the appearance of one or more people in the profile image.

Although a good baseline measure, Total Accuracy does not expose where the errors of matching occur, for example, whether an approach produces more false negatives or false positives, or how effectively the approach identifies true positives and/or true negatives. Confusion Matrices are a useful tool for identifying the sources of error in classification algorithms (Figure 7; Table 4).

Figure 7: Format of a Confusion Matrix

		Estimate from social media	
		Positive	*Negative*
Gold standard	*Positive*	True positive	False negative
	Negative	False positive	True negative

There is a general overweighting of false negatives over false positives, representing a tendency for the Face++ algorithm and the crowd-workers to assign a value to a given age category only if they were fairly sure and otherwise to allocate to 'unknown.' This is also evident in the relatively high number of false negatives in the 'Not a person/unknown' category. One notable exception is the CrowdFlower 20–29 category, where there are nearly twice as many false positives as false negatives. This may represent a source of bias wherein the difficulty of differentiating teenagers from people in their twenties or thirties resulted in the crowd defaulting to the middle, 20–29, category if unsure. The categories with the most users, namely the 20–29 and 30–39 groups are also the least accurately predicted by both Face++ and CrowdFlower. Again, this may reflect the relative difficulty in differentiating age visually during this life stage.

Table 4: Confusion matrices for age prediction, CrowdFlower and Face++

		CrowdFlower			Face++	
Age group		Positive	Negative		Positive	Negative
<20	Positive	2	6	Positive	3	5
	Negative	6	236	Negative	15	227
20–29	Positive	46	17	Positive	15	48
	Negative	34	153	Negative	10	177
30–39	Positive	33	33	Positive	16	50
	Negative	26	158	Negative	25	159
40–49	Positive	9	18	Positive	2	25
	Negative	8	215	Negative	17	206
50–59	Positive	4	10	Positive	0	14
	Negative	3	233	Negative	7	229
≥ 60	Positive	0	3	Positive	0	3
	Negative	2	245	Negative	3	244
Not a person/ unknown	Positive	58	11	Positive	65	4
	Negative	19	162	Negative	72	109

Table 5 provides further analysis of the efficacy of the two methods by determining precision (the percentage of users identified as belonging to a particular age group that are indeed in that age group) and recall (the percentage of the total number of users in a given age group captured by the algorithm). For Face++, precision is highest for the 20–29 age group, where 60% of the individuals identified by the algorithm as aged 20–29 were indeed in that category. This was the only category for which Face++ (marginally) outperformed the CrowdFlower precision (57.5%). This provides some evidence that Face++ may be effective in correctly discerning age from physical characteristics for this age band. For CrowdFlower, precision across age categories was generally around 50–60%, but significantly poorer for the youngest (<20) and oldest (60+) age categories. This result is in part attributable to the low frequency of users in these categories in the gold standard set, such that any error in assignment is amplified. Recall was generally poor for both algorithms, although the CrowdFlower workers could correctly identify 73% of the users aged 20–29 and half of those aged 30–39.

If we consider precision and recall to be equally important, then the weighted harmonic means (F1) indicate that both methods are effective for identifying users that are not a person/unknown and that the CrowdFlower workers are fairly effective at identifying users aged 20–29 and 30–39, with F1 scores of 0.64 and 0.53 respectively.

Table 5: Total accuracy, recall, and precision for CrowdFlower and Face++

	CrowdFlower				Face++			
Age group	Total acc.	Recall	Precision	F1	Total acc.	Recall	Precision	F1
<20	95.2%	25.0%	25.0%	0.25	92.0%	37.5%	16.7%	0.23
20–29	79.6%	73.0%	57.5%	0.64	76.8%	23.8%	60.0%	0.34
30–39	76.4%	50.0%	55.9%	0.53	70.0%	24.2%	39.0%	0.30
40–49	89.6%	33.3%	52.9%	0.41	83.2%	7.4%	10.5%	0.09
50–59	94.8%	28.6%	57.1%	0.38	91.6%	0.0%	0.0%	N/A
≥ 60	98.0%	0.0%	0.0%	N/A	97.6%	0.0%	0.0%	N/A
Not a person/ unknown	88.0%	84.1%	75.3%	0.79	69.6%	94.2%	47.4%	0.63

Since the age categories are sequential, it could be argued that a false positive allocation to a category adjacent to the true value should not be penalised as severely as an allocation to an age category significantly higher or lower than the true value. For this reason, we present hierarchical performance measures for the two models, as proposed by Wang et al. (2001) and refined in Costa et al. (2007), which reweight accuracy based on a tolerance level of one adjacent category (Table 6). This adjustment

does improve the Total Accuracy of both approaches but does not significantly improve the Face++ accuracy enough to make it a viable alternative to CrowdFlower for the assignment of age.

Table 6: CrowdFlower and Face++ adjoining category adjusted total accuracy for age prediction

	Total acc.	Adjusted total acc.
CrowdFlower age	60.8%	82.8%
Face++ age	40.4%	63.2%

In summary, CrowdFlower proved the most robust sex assignment approach overall, with sex assignment accuracy of 86.4% and age assignment accuracy of 60.8%. Age assignment was most accurate for the 20–29 and 30–39 age groups. Face++ accuracy was poor overall for both age and sex, although when the matching was restricted to only those users with a clear human in the profile image, the sex matching accuracy was 87.7%. Our results confirm previous analyses by An and Weber (2016), who compared the demographics estimated with Face++ with those obtained by looking at the profile description of Twitter users.

4. Calibration methodology and modelling approach

Zagheni and Weber (2015) refer to the representativeness problem of the Twitter sample as a calibration problem. They propose a framework for evaluating the measurement error incurred when using nonrepresentative internet data to estimate quantities of interest. If an auxiliary data source exists that can be assumed to measure the 'true' population, it can be combined with the dataset containing the counts from the Twitter population. The source of auxiliary data might be a representative survey or a census. The combination of the Twitter data with the auxiliary data source allows for a calibration exercise whose result will help to reduce the bias in the Twitter data or, in other words, to improve the quality of the Twitter data. This approach proposes to compare the 'true' counts of specific population subgroups defined over certain demographic characteristics in each geographical location obtained from the 'ground truth' data source with those obtained from the nonrepresentative sample, in our case the Twitter population that was active in the South East region of England between 23 June and 4 July 2014. In our application, the auxiliary dataset is the England and Wales Mid-Year Population Estimates for 2014 (MYEs). MYEs refer to the usual resident

population[5] on 30 June and are a combination of registration, survey, and administrative data from sources including the General Register Office, the International Passenger Survey, the Higher Education Statistics Agency, the National Health Service Central Register, and the Ministry of Justice, which are used to estimate the different components of population change (Office for National Statistics 2017). The Office for National Statistics publishes them every year. The MYE is assumed to represent the 'true' population count in each of the 67 local authorities in South East England, by sex and age group. We acknowledge that the MYEs will contain measurement error; the number of individuals in each local authority is an estimate and not the true population count. Even so, the MYEs, estimated using the cohort component method, are the official population estimates for England and Wales (Office for National Statistics 2017). Additionally, the definition of the population estimated by MYEs and Twitter sample are not exactly the same. MYEs reflect the usual residence population whereas Twitter data reflects the presence of a user at a specific place during a specified period. It is possible that people are more mobile during the summer (in our case 23 June to 4 July 2014) than during other times of the year when they are more likely to Tweet from their usual residences. Nevertheless, this time period was selected because it was the best match to the MYEs, corresponding to the population on 30 June 2014.

There are two sources of bias in Twitter data. The first source is the demographic information estimated by Face++ and CrowdFlower. Although they both provide good estimates of sex and age group (as discussed in Section 3.3), they are not free of error. The second source of bias is due to selection arising from the fact that the demographic structure (i.e., age, sex, and location) of the Twitter user population is different from the usual resident population of the South East region. The calibration methodology proposed in this section aims to reduce the selection bias in the nonrepresentative Twitter data. This methodology can be applied more broadly for demographic research using other nonrepresentative data sources thought to be characterized by measurement error. For example, Yildiz and Smith (2015) use a similar approach to reduce the bias in the England and Wales Patient Register data source, using marginal information from population census estimates as the auxiliary data source. Further applications could also include other digital populations, such as Facebook and LinkedIn, and nonrandom or convenience survey samples.

The modelling approach proposed by Zagheni and Weber (2015) aims to correct the selection bias in a nonrepresentative internet data source. Their proposed model takes the following form:

[5] Usual residence population "includes people who reside, or intend to reside, in the country for at least 12 months, whatever their nationality. Visitors and short-term migrants (who enter or leave the UK for less than 12 months) are not included" (Office for National Statistics 2016). We refer to the Office for National Statistics (2016) quality and methodology information paper for more information on the estimation of usual resident population.

$$y_{ij}^{t} = \left(k \times \mu_{ij}^{t}\right) + \alpha_{ij}^{t} + \epsilon_{ij}^{t}, \quad (2)$$

where y_{ij}^{t} is any quantity of interest coming from the ground truth source for location i and demographic group j (in our case, the population counts by age, sex, and local authority), μ_{ij}^{t} is the respective count extracted from the Twitter data source, k is a constant used to rescale the quantities, α_{ij}^{t} is the bias for location i and demographic group j, and finally ϵ_{ij}^{t} is the normally distributed error term. Zagheni and Weber (2015) propose to model the bias term α_{ij}^{t} in a regression framework, as a function of other variables that are thought to be relevant, such as the Twitter penetration rate.

In this paper, we propose a different approach for correcting the selection bias. As in Zagheni and Weber (2015), we rely on a regression framework for calibrating the nonrepresentative sample of Twitter users with the auxiliary marginal information from the ground truth data source, but our modelling approach makes use of log-linear models and does not rely on the collection of additional relevant information, such as internet penetration rate. We evaluate a series of log-linear models with offsets, measuring the degree to which the models can calibrate the Twitter users' data set. These models are in two broad groups. The first group includes four models. The Total Model reweights the Twitter sample so that estimates will match the MYEs total population size. The Age (A) Model reweights the Twitter sample to match the MYEs age group totals (e.g., the 30–39 population). The Age-Sex (AS) Model reweights the sample to match the MYEs age-group/sex totals (e.g., the 30–39 female population). And finally, the Age-Location (AL) Model reweights the sample to match the MYEs age-group/location totals (e.g., 30–39 population in Local Authority of Southampton).

Models in the second group are the AS,L Model, AS,SL Model, and AS,AL Model. The estimates of these models match two sets of marginal total population sizes of the MYEs separated by a comma. For example, the estimates of AS,L Model will match both the age-group/sex population marginal totals and location totals of the MYEs, e.g., the 30–39 female population and the total population of Southampton separately. However, they will not match the three-way age-group, sex, and location population, e.g., 30–39 female population in Southampton. The sets of log-linear models with offsets are estimated by an iterative process and are similar to multiplicative weighting, raking, or raking ratio estimation (Bethlehem et al. 2011). It is possible to solve the likelihood equations of such models by the Newton-Raphson method or the iterative proportional fitting algorithm (IPF) (Agresti 2013). In this paper, we employ the IPF algorithm, which is both easier and more transparent than the Newton-Raphson method. See Bishop, Fienberg, and Holland (1975) and Agresti (2013) for more information on IPF algorithm, and Willekens (1983, 1999) for examples of fitting log-linear models with offsets using the IPF algorithm.

We denote the MYEs by C_{asl}, where a denotes age groups 0–19, 20–29, 30–39, 40–49, and 50+,[6] s is sex (males and females), and l corresponds to the 67 local authorities in South East England. Assume that C_{asl} comes from a super population model and has Poisson distribution with mean μ_{asl}. T_{asl} is the 'offset' term and denotes the count of Twitter users in local authority l who are estimated to be in age group a and sex s. Finally, P_{asl}^{M} denotes the corresponding population count estimated by the M[th] Model for age group a, sex s, and local authority l, and M denotes one of the models in the model space, M = {A; AL; AS; AS,L; AS,SL, AS,AL}.

The models fitted in this paper are only a part of all possible models. There are three reasons for our choice of models. First, comparing the estimates produced by all possible models with both the ground truth and each of the other models will unnecessarily expand and complicate the paper, obscuring the fundamental purpose and aim of the proposed methodology. Second, in Sections 3.2 and 3.3 we demonstrate that the sex ratio and the age distribution of Twitter users are significantly different from the usual resident population (see Figure 1). However, deviations in the sex ratio and age distributions of the Twitter population did not differ greatly across local authorities in South East region. As such, we start with a relatively basic model and then expand it, focusing on differences by age-group and sex. Finally, we aim to maximize the reduction of bias using the most parsimonious model possible, in order to keep the information required from the auxiliary data source to a minimum. By reducing data demands, this methodology will be more flexible for future applications.

It is useful to present the modelling approach starting from the Total Model, which calibrates the Twitter sample with a constant weight. In this case the weight is equal to C_{+++}/T_{+++} where '+' denotes summation over indexes a, s, and l, respectively. Hence, the multiplicative Total Model combining the Twitter sample with auxiliary total population counts from the MYE can be expressed as:

$$\mu_{asl} = e^{\lambda} T_{asl}. \tag{3}$$

The term e^{λ} in Equation 3 is equal to the weight C_{+++}/T_{+++}. The log-linear model representation of the Total Model is:

$$\log \mu_{asl} = \lambda + \log(T_{asl}). \tag{4}$$

This model relies on the implausible assumption that the Twitter sample has the same age, sex, and local-authority association structures as observed in the ground truth

[6] In the estimation of the age group from CrowdFlower we have distinguished between the age groups 50–59 and 60+, however, due to their small sample size, in the modelling approach we combine these two age groups.

data source. It assumes that, after rescaling the Twitter data with the constant e^{λ} (Equation 3), the population count estimates for age group *a*, sex *s* and location *l* will be the same as the population counts in the ground truth data source (μ_{asl}).

Our aim is to identify an improved version of the Total Model, which takes into account the fact that the Twitter sample differs from the ground truth data in terms of association structures between age, sex, and location. In total, we run seven models (including the Total Model). We describe their characteristics in the remaining part of the section.

The A Model calibrates and reweights the Twitter sample by using the total population size and age distribution from the MYE for 2014. It can be written as follows:

$$\log \mu_{asl} = \lambda + \lambda_a^A + \log(T_{asl}), \tag{5}$$

where the factor λ calibrates the Twitter sample to match the South East total population count and λ_a^A calibrates its age distribution, irrespective of sex and location. The A Model corrects the age distribution in the Twitter data so that, in the final estimates, the total population in age groups summed over all local authorities and sex, P_{a++}^A, will equal the total population in age groups as measured in the ground truth data source, C_{a++}. In other words, the age distribution in the estimates will be the same as the ground truth data source. However, this model does not imply that the resulting estimates would provide the same local-authority or sex distributions as the ground truth data source.

The AS Model (Equation 6) combines the Twitter sample with the auxiliary age-group/sex-association structure from the ground truth data source, irrespective of location. The model estimates will have the same age-group/sex distribution as in the ground truth data source. This means that, in the final estimates, the total (fe)male population in a given age group *a* summed over all local authorities, P_{as+}^{AS}, will equal the (fe)male population in age group *a* in the ground truth data source, C_{as+}.

$$\log \mu_{asl} = \lambda + \lambda_a^A + \lambda_s^S + \log(T_{asl}) \tag{6}$$

Similarly, the AL Model (Equation 7) combines the Twitter sample with the auxiliary age-group/location association structure from the ground truth data source, irrespective of sex. In the final estimates, the total population of a local authority *l* in a given age group *a* summed over males and females will equal the population of local authority *l* in age group *a* in the ground truth data source, i.e., $P_{a+l}^{AL} = C_{a+l}$.

$$\log \mu_{asl} = \lambda + \lambda_a^A + \lambda_l^L + \log(T_{asl}). \quad (7)$$

The A, AS, and AL Models can be considered as a linear weighting or post-stratification exercise; they weight the Twitter sample to match the auxiliary marginal totals from the ground truth data source (Bethlehem et al. 2011). The second group models improve upon the AS Model. They require iterative estimation processes to produce estimates that match with both the age-group/sex distribution from the ground truth source and the local-authority margins, the sex/local-authority distributions, or the age-group/local-authority distributions, respectively.

In the AS,L Model (Equation 8) both the age-group/sex and the local-authority totals are corrected separately according to the ground truth population counts. This means that the sum over local authorities of the total male population in a given age group a in the final estimates is equal to the sum of the total male population in age group a in the auxiliary data, $P_{as+}^{AS,L} = C_{as+}$. Additionally, the local authorities will have the same sum as the local authority sums in the auxiliary data, $P_{++l}^{AS,L} = C_{++l}$. For example, the total population for the local authority of Southampton in the final estimates will be equal to the ground truth population of Southampton, and the population of 30–39-year-old males in the South East region in the estimates will be equal to the 30–39-year-old males in the South East region in the MYE. However, the model does not impose the three-way association structure (age group/sex/local authority) as observed in the MYE, i.e., $P_{asl}^{AS,L} \neq C_{asl}$. Hence, the population of 30–39-year-old males in Southampton in the final estimates will not be equal to the population of 30–39-year-old males in Southampton in the ground truth MYE.

$$\log \mu_{asl} = \lambda + \lambda_a^A + \lambda_s^S + \lambda_l^L + \lambda_{as}^{AS} + \log(T_{asl}). \quad (8)$$

The AS,SL Model produces estimates for which the marginal age-group/sex and sex/local-authority marginal totals are equal to the ground truth marginal totals ($P_{as+}^{AS,SL} = C_{as+}$ and $P_{+sl}^{AS,SL} = C_{+sl}$), but the three-way age-group/sex/local-authority association structure is different from the ground truth data source. The AS,SL Model can be written as follows:

$$\log \mu_{asl} = \lambda + \lambda_a^A + \lambda_s^S + \lambda_l^L + \lambda_{as}^{AS} + \lambda_{sl}^{SL} + \log(T_{asl}). \quad (9)$$

Similarly, the AS,AL Model produces estimates for which the marginal age-group/sex and age-group/local-authority marginal totals are equal to the MYE marginal totals ($P_{as+}^{AS,AL} = C_{as+}$ and $P_{a+l}^{AS,AL} = C_{a+l}$). The AS,AL Model can be written as follows:

$$\log \mu_{asl} = \lambda + \lambda_a^A + \lambda_s^S + \lambda_l^L + \lambda_{as}^{AS} + \lambda_{al}^{AL} + \log(T_{asl}). \quad (10)$$

To ease the interpretation of results, the models are evaluated by using percentage differences between the Twitter population and the population estimates in the ground truth data source. The percentage differences are the ratio of the difference between the model estimate and the 'true' population counts divided by the 'true' population counts of the respective cell (*asl*), defined as follows:

$$D_{asl}^M = 100 \times \frac{P_{asl}^M - C_{asl}}{C_{asl}}, \quad (11)$$

where C_{asl} denotes the population counts estimated by the MYE for age group *a*, sex *s*, and local authority *l*; P_{asl}^M denotes the corresponding population counts estimated by the M[th] Model for age group *a*, sex *s*, and local authority *l*; and M denotes one of the models in the model space M = {A; AL; AS; AS,L; AS,SL, AS,AL}. If the model correctly estimates the population count in a given cell (i.e., P_{asl}^M approaches C_{asl}), the percentage difference D_{asl}^M will approach 0. The quantity D_{asl}^M will be positive if the model overestimates the population count, and it will be negative if the model underestimates the population count. The extreme negative case is when the model yields a very small estimate for a given cell (e.g., P_{asl}^M approaches 0), in which case the percentage difference will approach –100. However, in theory, there is no upper limit. The D_{asl}^M will continue to increase as $P_{asl}^M - C_{asl}$ increases.

We present the percentage differences between the estimates obtained from the calibration models and those from the ground truth auxiliary data by means of two types of graphical representations. First, we plot the mean percentage differences by age group and sex for each of the M models (with the exception of the Total Model) across all local authorities (L=67), calculated as follows:

$$D_{as+}^M = \frac{\sum_{l=1}^{L} D_{asl}^M}{L}. \quad (12)$$

Second, we use choropleth maps to plot the mean percentage differences for each local authority *l*, calculated as follows:

$$D_{++l}^M = \frac{\sum_{a=1}^{A}\sum_{s=1}^{S} D_{asl}^M}{A \times S}, \quad (13)$$

where A (=5) and S (=2) denote the number of age groups and sexes respectively.
Because the demographic characteristics of the Twitter users are estimated using two alternative methods based on Crowd Flower and Face++, we estimate the log-linear models separately for the two Twitter samples.

5. Results

In the following section, we present the results of our modelling exercise comparing the Twitter sample and the usual resident population, as measured in the mid-year population estimates (MYEs) for the South East region of England in 2014. We present results for a total of six models and for two alternative sources of demographic information for the Twitter population based on CrowdFlower and on Face++. We do not present the results from the Total Model since it does not calibrate the association structures in the Twitter sample. We start by commenting on the model results based on the Twitter sample whose demographic characteristics were estimated by CrowdFlower, and then we present the Face++ results. The range of mean percentage differences changes according to different models. Hence, figures are produced using different scales.

The A Model aims to calibrate the Twitter sample by combining it with the ground truth age distribution from the MYE. The mean percentage differences for the A Model across all local authorities are reported in the upper-left panel of Figure 8. The mean percentage differences at ages 30+ are, in absolute value, higher than 50%. This means that the difference between the Twitter estimates and the MYEs are, on average, half of the size of MYEs for a given cell for the particular age group and sex when summed over local authorities. This suggests that the Twitter sample is not only biased by age, but potentially also by sex and location, and it is an indication that the calibration could be improved by adding the auxiliary age-group/sex association from the MYEs (i.e., the AS Model). The upper-left panel of Figure 9 shows how the mean percentage differences estimated by the A Model vary by local authorities when summed over age group and sex. The map shows that the A Model underestimates the population of some local authorities while it overestimates the population of other local authorities, supporting the idea that the local-authority distribution also needs to be calibrated in the Twitter sample.

Figure 8: Mean percentage difference between the MYEs and calibrated models based on the Twitter users' population according to age groups and sex, 2014 CrowdFlower

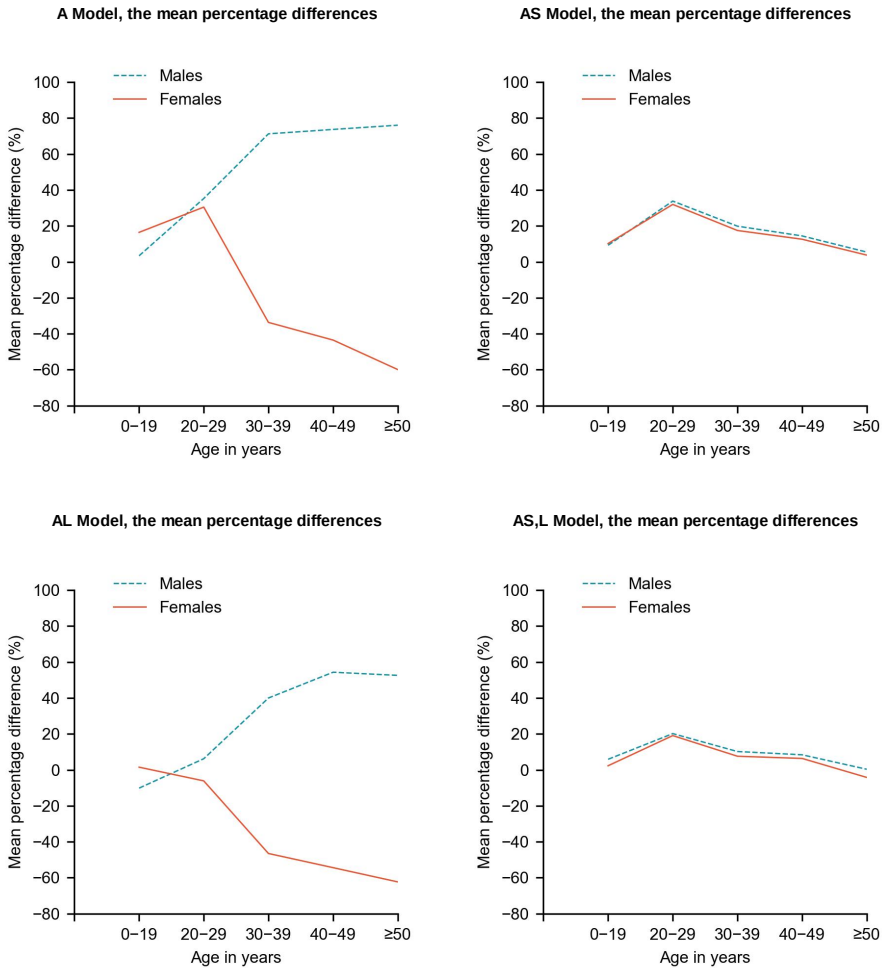


Figure 8: (Continued)

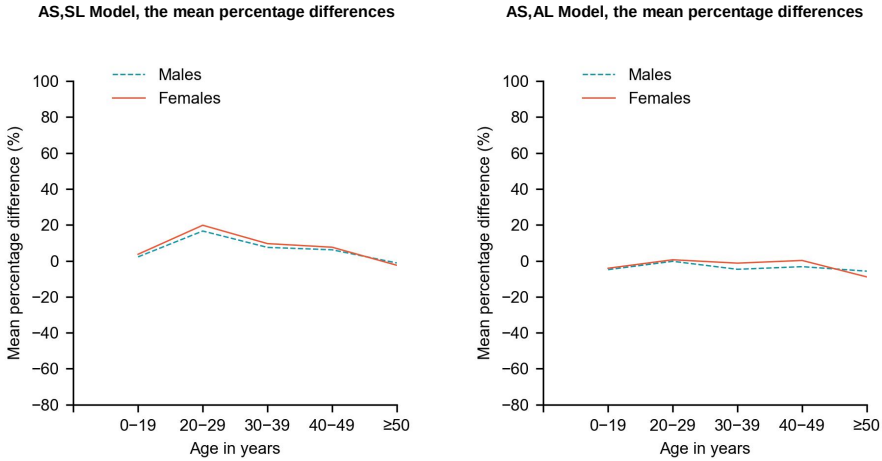


Figure 9: Mean percentage difference between the MYEs and calibrated models based on the Twitter users' population according to age groups and local authority, 2014 CrowdFlower

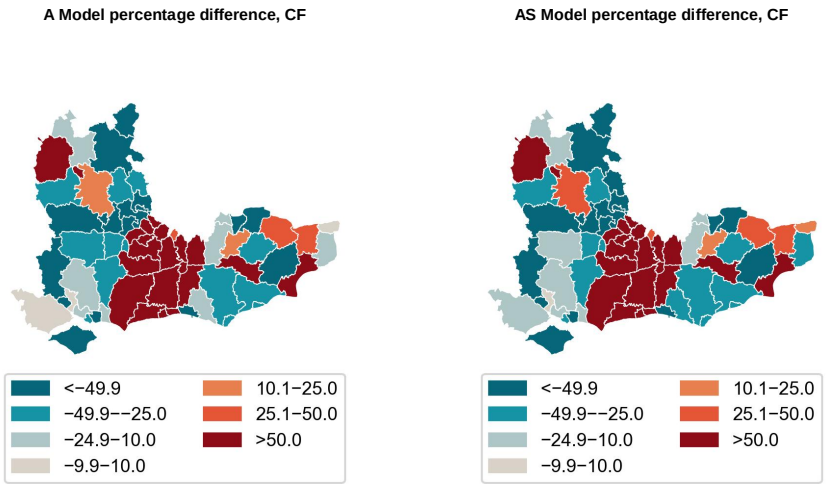
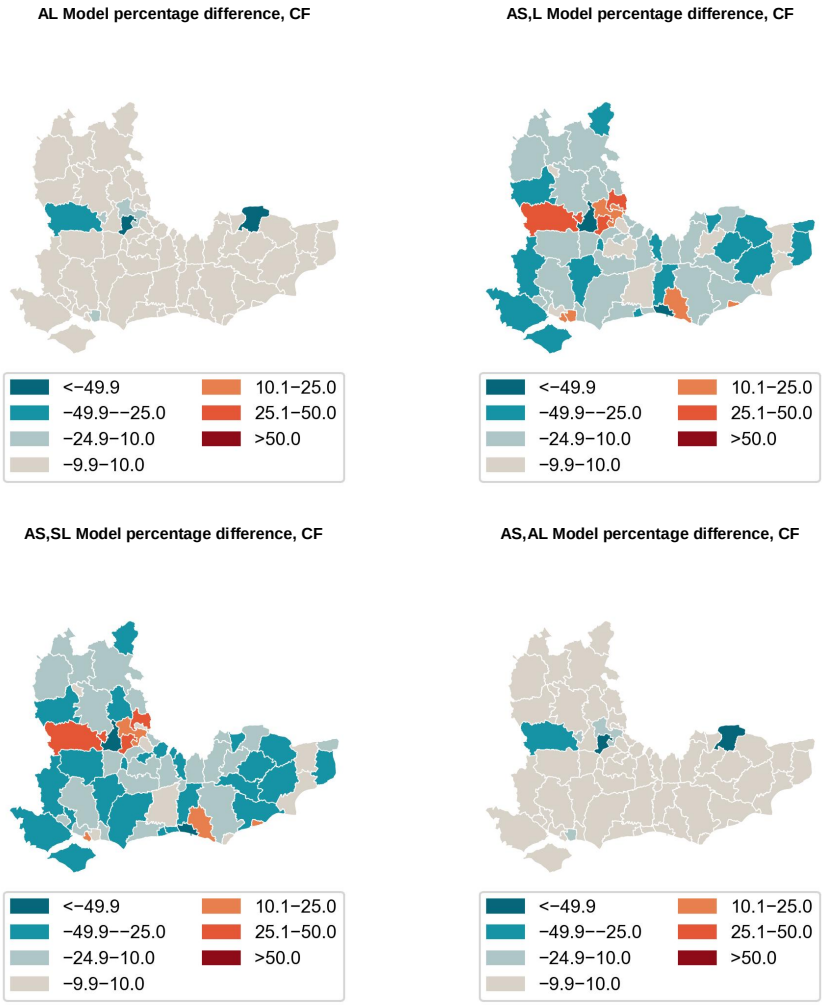


Figure 9: (continued)



When incorporating additional information about sex distributions from the MYEs, the bias in the Twitter sample decreases, i.e., the percentage differences are smaller, particularly for older age groups (Figure 8, upper-right panel). However, we find little change in the variation by local authority (Figure 9, upper-right panel), since the local-authority parameter has not been included in the AS Model.

We continue with the AL Model, which calibrates the Twitter sample by auxiliary age-group and local-authority association structure from the MYEs. The middle-left panel in Figure 8 shows that the age-group/sex mean percentage differences of the AL Model are smaller than those obtained from the A Model, but only for the first two age groups, and are fairly similar for the other ages. However, the map (Figure 9, middle-left panel) shows that, as expected, the mean percentage differences for local authorities decrease substantially as compared to previous models.

We continue with more complicated models that use additional information from the MYEs. The AS,L Model represents an improvement compared to the AS Model in terms of both local-authority estimates and age-group/sex estimates because it combines the Twitter sample with the local-authority distribution in addition to age-group/sex-association structure (Figures 8 and 9, middle right panel).

Similar to the AS,L Model, the AS,SL Model also represents an improvement from the AS Model because it adds the sex/local-authority association structure to the model. Figure 8 bottom left panel shows that when the Twitter sample is calibrated using auxiliary age-group/sex and age-group/local-authority association structures, the mean percentage differences by age groups and sex decrease. In other words, the bias in the Twitter sample not only stems from the age-group/sex association but also from the sex/local-authority association structure. After controlling for both association structures, the mean percentage differences decrease to reach the 0%–20% range. However, the mean percentage differences do not improve as much as in the AS,L Models. In fact, the map in Figure 9 bottom-left panel shows that adding the sex/local-authority structure distorts the local-authority estimates.

We continue with investigating the AS,AL Model. We expect the AS,AL Model to combine the strengths of the AS Model and the AL Model in terms of decreasing the mean percentage differences both by sex-specific age groups and local authorities. The bottom right panel of Figure 8 shows that combining Twitter sample with auxiliary age-group/sex and age-group/local-authority association structures indeed decreases the bias in the Twitter sample substantially: the mean percentage differences decrease to reach the 0%–5% range in the final model. In addition, the bottom right map in Figure 9 shows that this model yields low mean percentage differences for local-authority estimates, as also observed in the AL Model. The AS,AL Model proves to be the best model, reducing the bias the most. Adjusting the Twitter sample by both the age-group/sex association and the age-group/location association is needed to minimize the mean percentage differences with the ground truth data source.

All models except the AS,AL Model yield an overestimation of the male population for all ages (with the exception of the AL Model, which yields an underestimation of the male population aged 0–19, and the AS,L Model which yields an underestimation of the male population aged 50+). The female population is more

often underestimated. This is the case in the A Model for females aged 30+, in the AL Model for those aged 20+, and in the AS,L Model for those aged 50+. When both the male and the female population are overestimated, the mean percentage difference tends to be greater for the male population, except in the AL,SL Model.

The AS,AL Model yields an overall (slight) underestimation of both the male and the female populations with the exception of females aged 20–29 and 40–49, which are overestimated by 0.6% and 0.22% respectively. For both sexes, the 50+ age category is most often underestimated (–5.72 for males and –8.92 for females).

Unsurprisingly, Figure 9 shows that the mean percentage differences by local authorities (i.e., the bias in the final population estimates) decreases substantially when the Twitter sample is combined with the auxiliary local-authority distribution. The decrease in the mean percentage differences is higher when the age-group/local-authority association structure is included in the model (AL and AS,AL Models) than when only the local-authority distribution (AS,L) or the sex/local-authority association structure (AS,SL) is included. This suggests that the age-group/local-authority association structure in the Twitter sample differs vastly from the one observed in the MYEs.

In the final step of our analyses, we replicate the same models but using the demographic information for age group and sex obtained via Face++. Results from this exercise (Figure 10) show that the performance of the models is similar across the two Twitter samples (CrowdFlower and Face++). When using Face++, the model that yields the best fit (i.e., the lowest bias and the lowest mean percentage differences between the Twitter sample and the MYEs) is again the AS,AL Model. The maps reporting the mean percentage differences by local authority are very similar to those obtained for CrowdFlower (Figure A-1 in the Appendix). However, the overall performance of the model is better when the demographic information is obtained via CrowdFlower. For example, the mean percentage differences decrease to reach the 0–5.72% range in the best model (i.e., the AS,AL for CrowdFlower), whereas the range equals to 0–8.92% for Face++.

When the AS,AL Model is run with the Face++ demographic enhancement, the mean percentage differences are lower than with CrowdFlower for the younger age group 0–19 for both sexes, but especially for females. The mean percentage differences are very similar for the 20–29 age group: After calibration, CrowdFlower and Face++ underestimate the male population aged 20–29 by a factor of 0.25 and 0.28, respectively, and they both overestimate the female population by a factor of 0.6. The mean percentage differences for the population of females aged 50+ are considerably lower when using Face++ rather than CrowdFlower (–3% vs. –8.92%,), whereas they are considerably higher for the population of males aged 50+ (–8.96 vs. –5.72) and the population of females aged 40–49 (–7.42 vs. –3.22). The mean percentage differences

differ substantially between the two samples in two other cases: The population of males aged 40–49 and the population of females aged 30–39 are overestimated when using Face++, while they are underestimated when using CrowdFlower.

Figure 10: Mean percentage difference between the MYEs and calibrated models based on the Twitter users' population according to age groups and sex, 2014 Face++

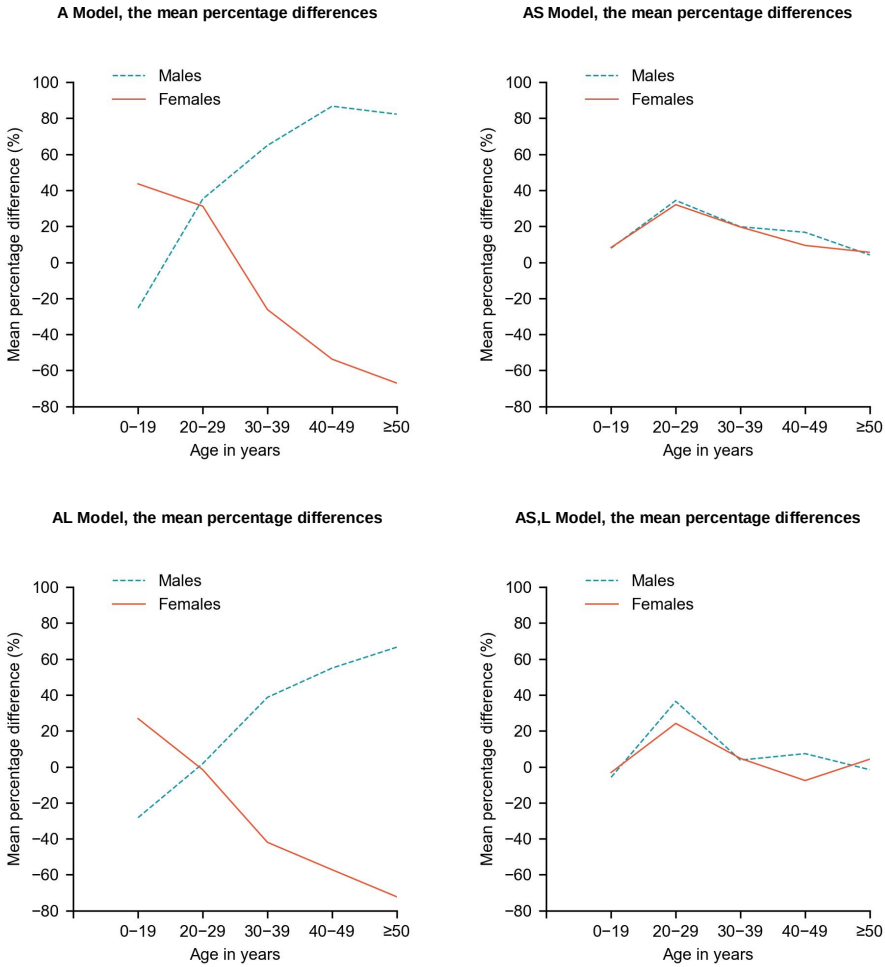
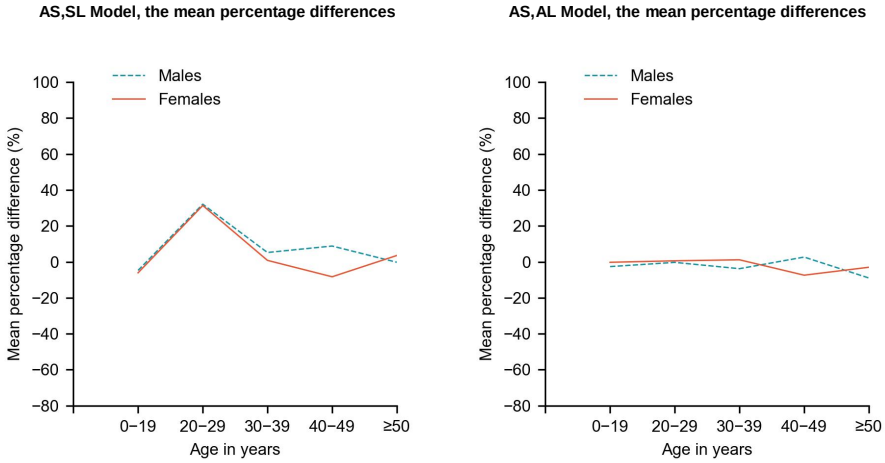


Figure 10: (Continued)



6. Conclusion and discussion

The Twitter population, as with other social media populations, represents a nonrandom, nonrepresentative sample of the total population. Consistent with previous research based on the United States (e.g., Mislove et al. 2011; Perrin 2015; Sloan et al. 2015), we find that Twitter users tend to be young adults, that male Twitter users outnumber female users, and that Twitter penetration is highest in urban areas. This implies that inference based on social media data, without any calibration, will be biased and invalid. The question then arises as to whether it is possible to use Twitter (and other social media data) to take a population perspective on social science questions. This paper proposed a modelling approach for reducing the selection bias in the Twitter population. The population estimates derived from our model allow for considerable improvement in the correction of the bias between the Twitter population and the real population, aiding researchers in making inferences based on this data.

To solve the problem of the selection bias in the Twitter sample, we used a series of log-linear models with offsets. The aim of this modelling exercise was to calibrate the Twitter population using population counts for each age group, sex, and local authority as measured in what can be assumed to be the ground truth population, in our case, the mid-year population estimates of the usual resident population of South East England. This calibration allows for the correction of the bias in the age-, sex-, and

location-specific population counts obtained from the Twitter population, on the basis of available information on the 'true' population. According to our results, the model that calibrates for both age-group/sex and age-group/location (AS,AL Model) provides the best population estimates for all age groups and sexes. The required age-group/sex and age-group/local-authority association structures are typically available from population censuses or from large-scale survey data.

The calibration modelling that we used in this paper could be improved (i.e., the percentage difference between the Twitter data and auxiliary data source could be reduced) if additional sociodemographic variables, such as the ethnicity of the Twitter users (Mislove et al. 2011; Messias, Vikatos, and Benevenuto 2017) or their work occupation (Sloan et al. 2005), were estimated and included in the model. It is also possible to extend the offset term to make inferences about variables that are present in the Twitter sample but are either not collected or not up to date in the auxiliary data source (Raymer, Abel, and Smith 2007). For example, Smith, Raymer, and Guilietti (2010) used log-linear modelling to study economic activity flows over time. All three data sources they combined were missing a component of the desired estimates or were not in the correct geographical detail. Similarly, if the Twitter sample included a variable of interest *x* that we would like to make inferences about, then the AS,AL Model could have been extended as follows:

$$\log \mu_{aslx} = \lambda + \lambda_a^A + \lambda_s^S + \lambda_l^L + \lambda_{as}^{AS} + \lambda_{al}^{AL} + \log(T_{aslx}). \tag{14}$$

Another extension of our approach for population inference would use the proposed log-linear models with offsets to calculate sample weights for individual Twitter users or Tweets. Such an approach would allow for inference from the Twitter sample to the reference population when treating the Twitter sample as an individual-level dataset for micro-level analyses (e.g., for measuring associations between variables obtained from Twitter data). For example, the associated weights for the AS,AL Model for age group *a*, sex *s*, and location *l* could be calculated as follows:

$$w_{asl} = \frac{e^{\log \mu_{asl}}}{C_{+++}} = \frac{e^{\lambda+\lambda_a^A+\lambda_s^S+\lambda_l^L+\lambda_{as}^{AS}+\lambda_{al}^{AL}}}{C_{+++}}. \tag{15}$$

However, the use of this calibration approach for individual-level weighting must be tested and applied with caution since variables of interest may or may not be correlated with the parameters in the auxiliary data source.

Unlike some social media data sources (e.g., Facebook), Twitter presents an additional challenge for social science researchers: The basic sociodemographic information of Twitter users is not collected and, therefore, must be estimated. This

estimation creates an additional source of bias. This paper measured and compared two different methods for estimating two basic demographic characteristics of Twitter users (age group and sex): CrowdFlower, a platform where crowd-workers were paid to guess the demographics of the Twitter users on the basis of their profile picture, profile description, and one random tweet, and the Face++ image-recognition software.

Our results suggest that CrowdFlower and Face++ perform similarly when the aim of the analysis is to obtain population estimates from Twitter. The calibration exercise yields only a slightly lower bias when estimated using the Twitter sample enhanced with demographic information obtained via CrowdFlower than when obtained with Face++. When using CrowdFlower, the AS,AL Model predicts the population estimates drawn from Twitter to lie within 5.7% of the mid-year population estimates, whereas when using Face++ the mean percentage error is only slightly higher (8.9%). When we compare the model estimates by sex and age group, the model based on Face++ actually performs better, on average, than the model based on CrowdFlower. In seven out of ten comparisons (five age groups x two sexes), the model based on Face++ yields a lower mean percentage difference. Face++ also provided lower mean percentage differences (when the average of absolute mean percentage differences is calculated) in five out of seven models for local authorities. However, these differences between the Face++ and CrowdFlower estimates are only minor.

Face++ offers the opportunity to analyse information at no cost. In this application, for example, we restricted our sample to the active Twitter population of South East England because of the costs incurred in performing the CrowdFlower task. Our experiment indicates that Face++ is a robust algorithm for the assignment of sex, where a clear profile image is available and could be combined with lexical analysis or partial crowd sourcing for the remaining user base. The results of this experiment indicate that age is more problematic to automate, although one potential compromise would be to assess the hierarchical accuracy of Face++ with smaller age categories.

Our modelling approach does not come without limitations. In the models, we use auxiliary information (age distribution, age-group/sex association, age/local-authority association) from the MYEs and compare our calibrated estimates with the same MYEs. A more robust check of the calibration method would use another source as auxiliary information in the model, such as a survey or census sample. We would expect that the use of an alternative source of auxiliary information would further reduce the bias, but will probably increase the variance of the estimates.

We also acknowledge that a given individual can have multiple Twitter accounts and hence we may be counting the same person more than once. So too does the Twitter population include nonperson accounts and so-called ‘bots.’ McCorriston, Jurgens, and Ruths (2015) documented that organizations represent about 10% of Twitter accounts, and they tend to be more connected than individual accounts. We

have not specifically dealt with nonperson accounts in this paper. Despite these shortcomings, this analysis provides an essential assessment framework for establishing best practices for the estimation of social media user characteristics and an innovative, yet straightforward to implement, calibration technique for statistical inference based on social media data.

7. Acknowledgements

This project received funding from the Web Science Institute and the Faculty of Social, Human and Mathematical Sciences at the University of Southampton via the Strategic Interdisciplinary Research and Development Fund. We are grateful to two anonymous reviewers and to participants at the workshop Social Media and Demographic Research: Applications and Implications (Cologne, 17 May 2016), the Stat-talk 2017, and the Intermediate Scientific Meeting of the Italian Statistical Society 2017 for useful comments and suggestions.

References

Agresti, A. (2013). *Categorical data analysis*. New Jersey: John Wiley and Sons, Inc.

An, J. and Weber, I. (2016). *#greysanatomy vs.#yankees: Demographics and hashtag use on Twitter*. Paper presented at the Tenth International AAAI Conference on Web and Social Media, Cologne, Germany, May 17–20, 2016.

Bishop, Y.M.M., Fienberg, S.E., and Holland, P.W. (1975). *Discrete multivariate analysis: Theory and practice*. Cambridge: MIT Press.

Cheng, J., Teevan, J., and Bernstein, M.S. (2015). Measuring crowdsourcing effort with error-time curves. *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems, Seoul, Korea, April 18–23, 2015*. New York: ACM: 1365–1374. doi:10.1145/2702123.2702145.

Cheng, J., Teevan, J., Iqbal, S.T., and Bernstein, M.S. (2015). Break it down: A comparison of macro- and microtasks. *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems, Seoul, Korea, April 18–23, 2015*. New York: ACM: 4061–4064. doi:10.1145/2702123.2702146.

Costa, E.P., Lorena, A.C., Carvalho, A.C.P.L.F., and Freitas, A.A. (2007). A review of performance evaluation measures for hierarchical classifiers. In: Drummond, C., Elazmeh, W., Japkowicz N., and Macskassy, S.A. (eds.). *Evaluation methods for machine learning II: Papers from the AAAI-2007 Workshop*. Palo Alto: AAAI Press: 1–6.

Fan, H., Yang, M., Cao, Z., Jiang, Y., and Yin, Q. (2014). Learning compact face representation. *Proceedings of the 22nd ACM international conference on Multimedia, Orlando, Florida, November 3–7, 2014*. New York: ACM: 933–936. doi:10.1145/2647868.2654960.

Housley, W., Williams, M., Williams, M., and Edwards, A. (2013). Special issue Computational Social Science: Research Strategies, Design, and Methods: Introduction. *International Journal of Social Research Methodology* 16(3): 173–175. doi:10.1080/13645579.2013.774164.

Ipeirotis, P.G. (2010). Demographics of mechanical turk. New York: New York University (NYU working paper No. CEDER-10–01). https://ssrn.com/abstract=1585030.

Kittur, A., Nickerson, J.V., Bernstein, M., Gerber, E., Shaw, A., Zimmerman, J., Lease, M., and Horton, J. (2013). The future of crowd work. *Proceedings of the 2013 conference on computer-supported cooperative work, San Antonio, Texas,*

February 23–27, 2013. New York: ACM: 1301–1318. doi:10.1145/2441776.2441923.

McCormick, T.H., Lee, H., Cesare, N., Shojaie, A., and Spiro, E.S. (2015). Using Twitter for demographic and social science research: Tools for data collection and processing. *Sociological Methods and Research* 46(3): 390–421. doi:10.1177/0049124115605339.

McCorriston, J., Jurgens, D., and Ruths, D. (2015). Organizations are users too: Characterizing and detecting the presence of organizations on Twitter. *Proceedings of the Ninth International AAAI Conference on Web and Social Media, Oxford, UK, May 26–29, 2015*. Palo Alto: AAAI Press: 650–653.

Megvii Inc. (2013). Face++ research toolkit [electronic resource]. Beijing: Face++ Cognitive Services. www.faceplusplus.com .

Messias, J., Vikatos, P., and Benevenuto, F. (2017). White, man, and highly followed: Gender and race inequalities in Twitter. *Proceedings of the IEEE/WIC/ACM International Conference on Web Intelligence, Leipzig, Germany, August 23–26, 2017*. New York: ACM: 266–274. doi:10.1145/3106426.3106472.

Mislove, A., Lehmann, S., Ahn, Y.Y., Onnela, J.P., and Rosenquist, J.N. (2011). Understanding the demographics of Twitter users. *Proceedings of the Fifth International AAAI Conference on Web and Social Media, Barcelona, Spain, July 17–21, 2011*. Palo Alto: AAAI Press: 554–557.

Office for National Statistics (2015). Annual mid-year population estimates: 2014 [electronic resource]. Newport: Office for National Statistics. https://www.ons.gov.uk/peoplepopulationandcommunity/populationandmigration/populationestimates/bulletins/annualmidyearpopulationestimates/2015-06-25.

Office for National Statistics (2016). Information paper: Annual mid-year population estimates: 2016 [electronic resource]. Newport: Office for National Statistics. https://www.ons.gov.uk/file?uri=/peoplepopulationandcommunity/populationandmigration/populationestimates/qmis/annualmidyearpopulationestimatesqmi/qmimyejune16finalforpub.pdf .

Office for National Statistics (2017). Annual mid-year population estimates QMI [electronic resource]. Newport: Office for National Statistics. https://www.ons.gov.uk/peoplepopulationandcommunity/populationandmigration/populationestimates/qmis/annualmidyearpopulationestimatesqmi .

Perrin, A. (2015). Social networking usage 2005–2015 [electronic resource]. Washington, D.C.: Pew Research Center. http://www.pewinternet.org/2015/10/08/2015/Social-Networking-Usage-2005-2015/.

Raymer, J., Abel, G.J., and Smith, P.W.F. (2007). Combining census and registration data to estimate detailed elderly migration flows in England and Wales. *Journal of the Royal Statistical Society. Series A* 170(4): 891–908. doi:10.1111/j.1467-985X.2007.00490.x.

Savage, M. and Burrows, R. (2007). The coming crisis of empirical sociology. *Sociology* 41(5): 885–899. doi:10.1177/0038038507080443.

Sloan, L., Morgan, J., Burnap, P., and Williams, M. (2015). Who tweets? Deriving the demographic characteristics of age, occupation and social class from Twitter user meta-data. *PloS ONE* 10(3): e0115545. doi:10.1371/journal.pone.0115545.

Sloan, L., Morgan, J., Housley, W., Williams, M., Edwards, A., Burnap, P., and Rana, O. (2013). Knowing the tweeters: Deriving sociologically relevant demographics from Twitter. *Sociological Research Online* 18(3): 7. doi:10.5153/sro.3001.

Smith, C. (2016). How many people use the top social media? Digital market ramblings [electronic resource]. Raymond: DMR. http://expandedramblings.com/index.php/resource-howmany-.

Smith, P.W.F., Raymer, J., and Guilietti, C. (2010). Combining available migration data in England to study economic activity flows over time. *Journal of the Royal Statistical Society, Series A* 173(4): 733–753. doi:10.1111/j.1467-985X.2009.00630.x.

Van Pelt, C. and Sorokin, A. (2012). Designing a scalable crowdsourcing platform. *Proceedings of the 2012 ACM SIGMOD International Conference on Management of Data*. New York: ACM: 765–766. doi:10.1145/2213836.2213951.

Vikatos, P., Messias, J., Miranda, M., and Benevenuto, F. (2017). Linguistic diversities of demographic groups in Twitter. *Proceedings of the 28th ACM Conference on Hypertext and Social Media, Prague, Czech Republic, July 4–7, 2017*. New York: ACM: 275–284. doi:10.1145/3078714.3078742.

Wang, K., Zhou, S., and He, Y. (2001). Hierarchical classification of real life documents. *Proceedings of the 2001 SIAM International Conference on Data Mining. Chicago, Illinois, April 5–7, 2001*: 1–16. doi:10.1137/1.9781611972719.22.

Willekens, F. (1983). Log-linear modelling of spatial interaction. *Papers of the Regional Science Association* 52(1): 187–205. doi:10.1007/BF01944102.

Willekens, F. (1999). Modelling approaches to the indirect estimation of migration flows: From entropy to EM. *Mathematical Population Studies: An International Journal of Mathematical Demography* 7(3): 239–278. doi:10.1080/08898489909525459.

Yildiz, D. and Smith, P.W.F. (2015). Models for combining aggregate-level administrative data in the absence of a traditional census. *Journal of Official Statistics* 31(3): 431–451. doi:10.1515/jos-2015-0026.

Zagheni, E., Garimella, V.R.K., Weber, I., and State, B. (2014). Inferring international and internal migration patterns from twitter data. *WWW '14 Companion Proceedings of the 23rd International World Wide Web Conference, Seoul, Korea, April 7–11, 2014*. New York: ACM: 439–444. doi:10.1145/2567948.2576930.

Zagheni, E. and Weber, I. (2015). Demographic research with nonrepresentative internet data. *International Journal of Manpower* 36(1): 13–25. doi:10.1108/IJM-12-2014-0261.

Zhou, E., Fan, H., Cao, Z., Jiang, Y., and Yin, Q. (2013). Extensive facial landmark localization with coarse-to-fine convolutional network cascade. *Proceedings of the IEEE International Conference on Computer Vision Workshops, Sydney, Australia, December 2–8, 2013*. New York: IEEE: 386–391. doi:10.1109/ICCVW.2013.58.

Appendix

Figure A-1: Mean percentage difference between the MYEs and calibrated models based on the Twitter users' population according to age groups and local authority, 2014 Face++

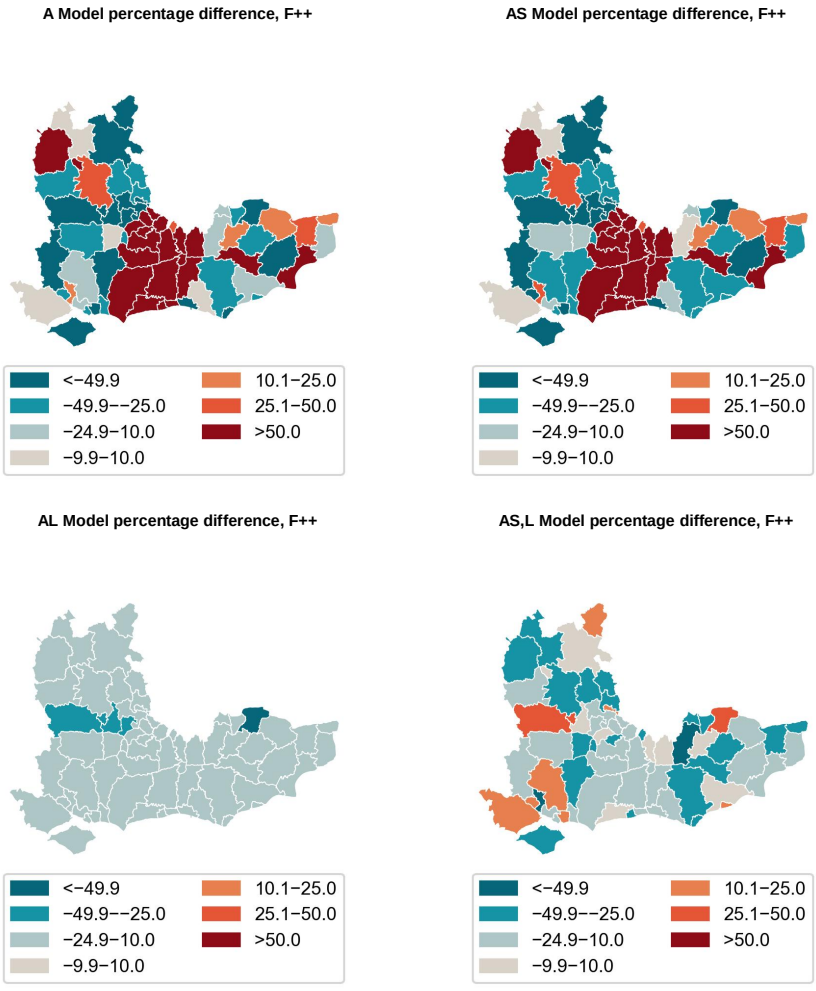


Figure A-1: (Continued)