Using Twitter data for Population Estimates

Usare dati Twitter per Stime di Popolazione

Dilek Yildiz, Jo Munson, Agnese Vitali, Ramine Tinati and Jennifer Holland

Abstract Twitter is increasingly being used as a source of data for the Social Sciences. However, deriving the demographic characteristics of users and dealing with the non-random non-representative populations from which they are drawn represent challenges for social scientists. This paper has two objectives: first, it compares different methods for estimating demographic information from Twitter data based on the crowd-sourcing platform CrowdFlower and the image-recognition software Face++. Second, it proposes a method for calibrating the non-representative sample of Twitter users with auxiliary information from official statistics, hence allowing to generalize findings based on Twitter to the general population.

Abstract *Twitter è sempre più usato come fonte di dati per la ricerca sociale. Derivare le caratteristiche demografiche degli utenti di Twitter e la natura non-random e non rappresentativa del campione, però, rappresentano una sfida. Questo lavoro si propone due obiettivi: il primo è di confrontare due diversi metodi per derivare le caratteristiche demografiche degli utenti di Twitter, uno basato sulla piattaforma di crowd-sourcing CrowdFlower, l'altro sul software di riconoscimento di immagini Face++. Il secondo obiettivo propone un metodo per calibrare il campione non rappresentativo di Twitter con informazioni sulla popolazione ottenute da fonti di statistica ufficiale, in modo da poter fare inferenza sulla popolazione di interesse, partendo dal campione non rappresentativo di Twitter.*

Key words: Calibration, Population, Social Media, Twitter

[1] Dilek Yildiz, Wittgenstein Centre (IIASA, VID/ÖAW, WU), VID/ÖAW; email: Dilek.Yildiz@oeaw.ac.at

Jo Munson, University of Southampton; email: J.Munson@soton.ac.uk

Agnese Vitali, University of Southampton; email: A.Vitali@soton.ac.uk

Ramine Tinati, University of Southampton; email: R.Tinati@soton.ac.uk

Jennifer Holland, University of Rotterdam; email: j.a.holland@fsw.eur.nl

Alessandra Petrucci, Rosanna Verde (edited by), *SIS 2017. Statistics and Data Science: new challenges, new generations. 28-30 June 2017 Florence (Italy). Proceedings of the Conference of the Italian Statistical Society*
ISBN (online) 978-88-6453-521-0 (online), CC BY 4.0, 2017 Firenze University Press

1 Introduction

Twitter data, just like data from other social media, are increasingly being used for Social Science research. However, such data are not representative of the total population: the inference made using social media is hence invalid. Also, the basic demographic characteristics of the Twitter users are not readily available and hence need to be estimated. This paper proposes a method based on calibration for reducing the existing bias between the Twitter population and the total population and it compares the results obtained using two different methods for estimating the demographic characteristics of Twitter users with the aim of establishing best practices which were used in previous research (e.g. McCormick et a. 2015; Zagheni et al. 2014).

2 Data

We collected Twitter data between 23 June and 4 July 2014 using DataSift's Twitter Firehose connection. This one-week period straddles the mid-year population estimates (MYE) for the usual resident population of England and Wales on 30 June 2014 which are produced annually by the Office for National Statistics (2015). Our Twitter sample consists of users who tweeted at least once during the reference week. In addition, we restrict our sample to those Twitter users who have at least one geo-located tweet in South-East England during the week of observation. The final sample comprises 22,356 unique users.

3 Estimating Age and Sex of Twitter Users

We estimate age and gender of the Twitter users using two distinct methodologies: crowdsourcing, via the CrowdFlower Crowdsourcing platform, and the image-recognition software Face++. By restricting our sample to all geo-located tweets, we further have information on the location of the users.

CrowdFlower provides access to a large pool of crowd-workers who will execute a specific task in exchange of a monetary reward. We designed a task which presented crowd-workers with a user's profile description and picture (if available) and random tweet, and asked them two questions: "Would you say this Twitter user is female; male; don't know; the Tweeter is a company/organization/not a person" and "Take the best guess at the user's age in years: 0-19; 20-29; 30-39; 40-49; 50+". Given the cost of such experiment, we restrict the sample to be analysed by the crowd-workers to Twitter users in the South-East England.

Face++ is an automated face-detection algorithm developed by Megvii Inc. (2013). Face++ takes links to image files as its input variable and outputs an age and gender estimate. Face++ demands that there are one or more distinguishable faces in the image provided in order to return a valid result, hence images showing non-human entities, or where the algorithm is unable to identify a face return a null result.

Figure 1 reports the population pyramids from the 2014 Twitter population with demographic information estimated via CrowdFlower and Face++. For both Crowdflower and Face++, males outnumber females in all age groups, with the exception of ages 0-19 in Face++. According to the gender estimates based on CrowdFlower and Face++, we find that the average number of males per 100 females in the Twitter sample to be equal to 149 and 138.6 males, respectively, whereas there are 96.8 males per 100 females according to the 2014 MYE.

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 the males aged 30-39, followed by both males and females aged 20-29. The youngest age group represents a higher proportion of the total Twitter population compared to the CrowdFlower estimates, especially among females.

In order to compare CrowdFlower and Face++, we compute a measure of performance for algorithms which attempt to assign data points to one of two or more categories, i.e. the Total Accuracy, as follows:

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

where T and F stands for True and False and N and P stands for Negative and Positive, respectively. In order to compute the Total Accuracy, we refer to a gold standard set of 123 randomly selected users with a valid profile picture, for whom we know the true age and sex as these were manually verified using LinkedIn profiles, Electoral Roll listings, personal websites. As Table 1 shows, the accuracy is higher with CrowdFlower. The gender matching when there is a valid profile picture is nearly 92% accurate for Face++ and 97% for CrowdFlower, but the age matching is only 35% accurate for Face++ vs. 79% for CrowdFlower.

Figure 1: Population pyramids based on Twitter data, demographic variables estimated with Crowdflower and Face++

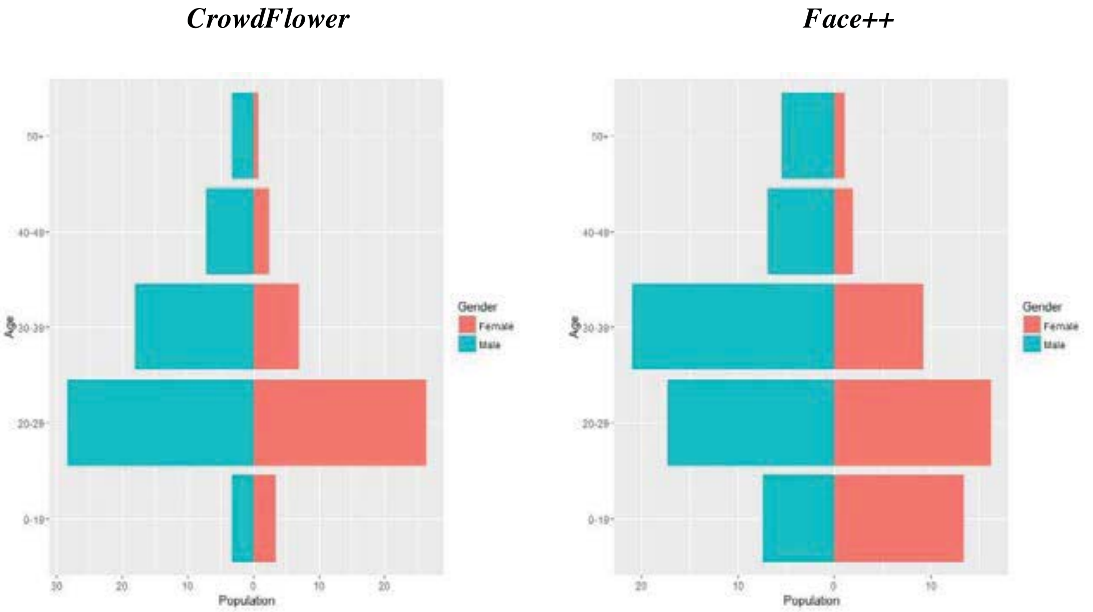


Table 1: Accuracy of Face++ and CrowdFlower

	Total Accuracy, valid images (N.=123)	
	Age	***Gender***
Face++	35.8%	91.9%
CrowdFlower	73.2%	97.6%

4 Calibration Methodology

We propose a calibration approach for correcting the selection bias in a non-representative internet population. This approach relies on a regression framework for calibrating the non-representative sample of Twitter users with the auxiliary marginal information from the ‘ground truth’ data source, using log-linear models with offsets. We extend a calibration methodology developed by Yildiz and Smith (2015) to the framework proposed by Zagheni and Weber (2015).

If an auxiliary data source exists which can be assumed to measure the ‘true’ population, it can be combined with the dataset containing the counts from the Twitter population. This approach proposes to compare the ‘true’ counts of specific population subgroups by age and sex in each geographical location obtained from the ‘ground truth’ data source, with those obtained from the non-representative sample. In this example, we compare the Twitter population to the usual resident population of South-East England using the 2014 mid-year population estimates.

This source is assumed to represent the ‘true’ population count in each of the 67 local authorities in South East England, by gender and age group.

We fit a sets of log-linear models with offsets which takes into account the fact that the Twitter sample differs from the ‘ground truth’ data in terms of association structures between age and/or gender and/or location. Such models are estimated by an iterative process are similar to multiplicative weighting, raking or raking ratio estimation. We employ the IPF algorithm to fit the log-linear models with offsets and produce maximum likelihood estimates. We evaluate the capability of each model of calibrating the Twitter users’ data.

The best model, i.e. the model which reduces the bias between the Twitter sample and the total population the most, is the AS, AL model (the best model was chosen according to the mean percentage differences –see below–; results for other models are not shown). This model calibrates the Twitter population counts so that the marginal age-sex and sex-local authority marginal totals are equal to the ‘ground truth’ marginal totals. Instead, the three-way age-sex-local authority association structure is different from the ‘ground truth’ data source. The AS,SL Model can be written as follows:

$$\log(\mu_{asl})= \lambda+ \lambda_a^A+ \lambda_s^S +\lambda_l^L+\lambda_{as}^{AS}+\lambda_{sl}^{SL}+\log(T_{asl}) \qquad (2)$$

We denote the Census estimates and the MYE for age group a, sex s, and local authority l by C_{asl} where a denotes age groups “0-19”, “20-29”, “30-39”, “40-49” and “50+”; and s =1, 2 for males and females respectively. We assume that C_{asl} comes from a super population model and has Poisson distribution with mean μ_{asl}. In this application we focus on the South-East region of England which consists of 67 local authorities, i.e. l =1, 2,…, 67. 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. The factor λ calibrates the Twitter sample to match the South-East total population count; λ_a^A calibrates its age distribution, irrespectively of sex and location; λ_{as}^{AS} calibrates its age-sex distribution, irrespectively of location; etc.

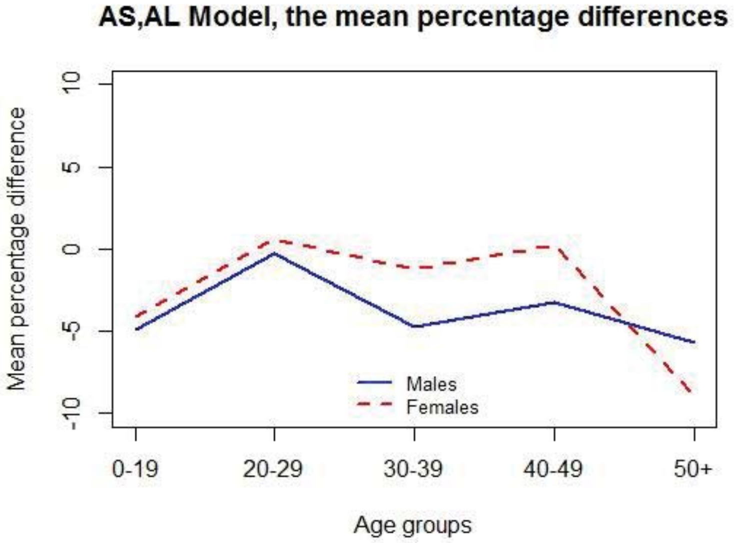
In order to ease the interpretation of results, the models are evaluated using percentage differences between the Twitter population and the population estimates in the ‘ground truth’ data source, defined as follows:

$$D_{asl} = 100 \times (P_{asl} - C^{asl}) / C^{asl} \qquad (3)$$

where C^{asl} denotes the population counts estimated by the MYE for age group a, sex s and local authority l and P^{asl} the corresponding population counts. Figure 2 plots the mean percentage differences by age group and sex for the AS,AL model, using the CrowdFower estimates of age and sex. This figure shows that combining the Twitter sample with auxiliary age-sex and age-location association structures indeed decreases the bias in the Twitter sample substantially: the mean percentage differences decrease to reach the 0-5% range. We conclude that adjusting the Twitter sample by both the age group-gender association and the age group-region association is needed in order to minimize the mean percentage differences with the 'ground truth' data source.

Figure 2 also shows that overall our model slightly underestimates the populations of both sexes. The age category which is underestimated the most is the 50+ for both sexes.

Figure 2: Mean percentage difference between the MYE and calibrated models based on the Twitter users' population according to age groups, 2014 CrowdFlower



5 Conclusions

This paper proposed a modelling approach based on log-linear models with offsets for reducing the selection bias in the Twitter population. The population estimates derived from the model allows a considerable improvement towards the correction of the bias between the Twitter population and the real population, allowing researchers to make inference from the non-representative Twitter sample to the population of interest.

Moreover, this contribution has compared the accuracy of the age and gender estimates produced by the crowd-sourcing and image-recognition approaches. One of the major drawbacks of the Face++ approach is that it takes only an image as its input variable. If there is no image available for a user, or if the image does not clearly display a human user, the Face++ algorithm fails. In contrast, CrowdFlower users are able to utilise the username, tweet content and description as well as the image to guess the demographics of the user. Whilst the CrowdFlower results are clearly the most accurate, Crowd-sourcing assignment is not free and can be time consuming. Face++ is free and comparatively quick and could thus be considered the best approach for gender matching where there is an identifiable user in the profile image, whereas Face++ is not an effective tool for the measurement of age.

References

1. Megvii Inc. (2013) Face++ Research Toolkit. Available at: http:// www.faceplusplus.com
2. McCormick, T. H., Lee, H., Cesare, N., Shojaie, A., and Spiro, E. S. Using Twitter for Demographic and Social Science Research Tools for Data Collection and Processing. Sociological Methods & Research, doi: 0049124115605339 (2015).
3. Office For National Statistics (2015). Annual Mid-year Population Estimates, 2014
4. Yildiz, D., and Smith, P.W.F. Models for Combining Aggregate-Level Administrative Data in the Absence of a Traditional Census. Journal of Official Statistics, 31(3):431-451 (2015).
5. Zagheni, E. and Weber, I. Demographic research with non-representative internet data. International Journal of Manpower, 36(1): 13-25 (2015)
6. Zagheni, E., Garimella, V.R.K., Weber, I. and State, B. Inferring international and internal migration patterns from twitter data. Proceedings of the Companion Publication of the 23rd International Conference on World Wide Web (WWW): 439-444 (2014)