Combining Social Media Data and Traditional Surveys to Nowcast Migration Stocks

Emilio Zagheni[1,2], Kivan Polimis[3], Monica Alexander[4], Ingmar Weber[5], and Francesco C. Billari[3]

[1]University of Washington, Seattle
[2]Max Planck Institute for Demographic Research
[3]Bocconi University, Milan
[4]University of California, Berkeley
[5]Qatar Computing Research Institute

April 7, 2018

Abstract

Social media and Web data offer new opportunities to improve demographic knowledge and to complement more traditional data sources. Facebook, for example, can be thought of as a large digital census that is constantly updated. However, its users are not representative of the underlying population. Contrastingly, the American Community Survey (ACS) relies on smaller samples that may be noisy and that are published with a substantial delay from data collection. Additionally, ACS samples are representative of the underlying population and have historical depth. We generate nowcasts, present and near-future predictions, of migration stocks that combine the best of the two complementary sources using a Bayesian hierarchical model. Facebook data, obtained via the Marketing API are timely, but lack demographic constraints on trends in age patterns that we extrapolate from ACS time series. Combining data sources and modeling strategies enables the researcher to weigh down inconsistencies and extract valuable insights without ignoring existing information. Although the focus of this article is on migration, our methods are general and contribute to the emerging literature on complementing social media and traditional data sources in various contexts.

1 Introduction

Social media and Web data offer new opportunities for demographic research. However, while these are typically *big data* with large samples that often provide information that is qualitatively different from what we can obtain from surveys, they are also not representative of the underlying population. A lack of understanding of the biases of these data may lead to incorrect predictions. For example, *Google Flu* [3], a tool that Google developed to track epidemics based on real time Web queries related to influenza symptoms, suffered from systematic biases that ultimately led to its demise [5].

The use of incomplete and deficient data sources to estimate demographic parameters when alternatives are not present is however not a novelty. Indeed, the whole set of indirect demographic estimation techniques has revolved around exploiting traditional data sources that suffer from biases or incompleteness to generate estimates of parameters of interest [7]. Three approaches have addressed the use of non-representative social media and Web data to generate population estimates. First, calibration and statistical corrections have been used to remove or model the bias in geo-located e-mail data and Twitter data [12, 10]. These approaches build on the long tradition of indirect demographic estimation techniques (e.g., [1, 13]). Second, statistical tools like difference-in-differences have been used to evaluate trends over time in Twitter data [11]. Third, survey research methods, like post-stratification, together with Bayesian multi-level models, have proven useful to analyze online surveys that rely on non-representative samples.[9]

In this article we want to go beyond correcting for bias in order to make estimates and predictions that are based on information extracted from both social media data and more traditional sources. We combine traditional survey data and social media data within a unified statistical framework using a Bayesian hierarchical model that exploits the complementarity of the two kind of data sources. Social media data are available almost in real time, and are particularly useful in detecting recent changes in trends. Traditional data typically offer longer time series and can be used to impose constraints, like regularities in age profiles.

More specifically, in order to nowcast stocks of migrants, we combine data from the "Facebook Adverts Manager", a source that offers a census of Facebook users, with time series from the American Community Survey (ACS from now onwards), a representative survey of the U.S. Census Bureau, designed to supplement the decennial Census in the United States. Facebook data offer a timely picture of stocks of immigrants in US states. The ACS is used to evaluate and

extrapolate regularities in age profiles of immigrants. The model that we develop is flexible enough that it could be used to incorporate additional data sources or in contexts that are not limited to migration.

2 Data

2.1 Facebook Adverts Manager

The main source of revenue for social media companies is online advertising. Facebook, LinkedIn and Twitter have dedicated substantial resources to the estimation of demographic characteristics, behavior and interests of their users in order to improve targeted advertisement. Facebook, in particular, has developed a platform, called Adverts Manager, that allows advertisers to select detailed characteristics of users to whom their ads should be shown.[1] The dimensions that can be targeted include both information explicitly reported by Facebook users, such as their age or gender, as well as information automatically inferred from their interaction on Facebook and affiliate websites, such as their interests.

As an illustrative example, Facebook supports showing ads exclusively to expats from India, aged 13 and over, who live in the state of California. Before actually launching an ad, which then incurs a cost to the advertiser, Facebook provides an estimate of the selected audiences' sizes. In the example above, Facebook reports a "potential reach" of 390,000 users[2]. This reach estimate refers to the number of monthly active users on Facebook that are classified as matching the described criteria [3]. For our analyses, we obtain a number of these reach estimates in a programmatic way via Facebook's Marketing API[4] (the Marketing API is the back-end data supplier to the Adverts Manager platform). As we did not proceed to actually launch an ad, these data were collected free of charge. Our data collection started in January 2017. Since then, we have been collecting data at regular periods of time, obtaining an updated snapshot every 1-2 months.

More specifically, we make use of the category "Expats (*)" that Facebook provides. Expat groups were found by querying "Ex-pat" and "Expat" in the Graph and Marketing APIs. Currently, these APIs support 90 countries or territories of origin when targeting expats of a particular origin, such as "Expats (Mex-

[1] This platform can be accessed for free at https://www.facebook.com/ads/manager/creation/creation/.
[2] As of April 2nd, 2018.
[3] https://www.facebook.com/business/help/1438142206453359
[4] https://developers.facebook.com/docs/marketing-apis

ico)"[5]. Additionally, one can target "Expats (All)", which also includes users of other countries of origin. Facebook does not provide a very detailed definition of "Expats". In the Adverts Manager's documentation, expats are defined as "people whose original country of residence is different from the current country". Despite the lack of documentation, we can infer, based on the literature produced by researchers who work internally at Facebook [4], that there are two factors that play a key role in the estimation of expats at Facebook. The first one is the self-reported "current city" and "hometown" in the list of "places you have lived" that people fill in for their Facebook profile. The second one is the network structure of friendships (e.g., having at least two Facebook friends in the home country and two Facebook friends in the destination country).

State-level estimates of expat-origin groups were collected from the Facebook Marketing API using the Python module the pySocialWatcher [2]. pySocialWatcher supports data extraction from multiple categories available in the Facebook Marketing API such as education, household composition, life transitions and events, interests, and behaviors. The expat-origin groups are categorized as behaviors with the Marketing API. From January 2017 to March 2018, six waves of data were collected using Amazon Web Services (AWS) Elastic Computing (EC2) servers [6]. At each wave, 3 types of data were collected. First demographically disaggregated data on all Facebook users were collected by state for the following demographic groups: gender (female, male, and total population) as well as 10 standard age-groups (15-19, 20-24, 25-29, 30-34, 35-39, 40-44, 45-49, 50-54, 55-59, 60-65; an 11$^{\text{th}}$ age group from 13-65, the entire available age range on Facebook, was also added). Secondly, expat populations for all 90 expat groups disaggregated by state, gender, and age group were gathered.

[5]Here is the list of the 90 supported countries and territories of origin that we considered for our analysis: Algeria, Argentina, Australia, Austria, Bangladesh, Belgium, Brazil, Cameroon, Canada, Chile, China, Colombia, Congo DRC, Cuba, Cyprus, Czech Republic, Denmark, Dominican Republic, El Salvador, Estonia, Ethiopia, Finland, France, Germany, Ghana, Greece, Guatemala, Haiti, Honduras, Hong Kong, Hungary, India, Indonesia, Ireland, Israel, Italy, Ivory Coast, Jamaica, Japan, Jordan, Kingdom of Saudi Arabia, Kenya, Kuwait, Latvia, Lebanon, Lithuania, Luxembourg, Malaysia, Malta, Mexico, Monaco, Morocco, Nepal, Netherlands, New Zealand, Nicaragua, Nigeria, Norway, Peru, Philippines, Poland, Portugal, Puerto Rico, Qatar, Romania, Russia, Rwanda, Senegal, Serbia, Sierra Leone, Singapore, Slovakia, Slovenia, South Africa, South Korea, Spain, Sri Lanka, Sweden, Switzerland, Tanzania, Thailand, UAE, Uganda, United Kingdom, United States, Venezuela, Vietnam, Zambia, Zimbabwe

[6]Wave 1: 2017/01/25 to 2017/02/11. Wave 2: 2017/04/20 to 2017/05/19. Wave 3: 2017/06/03 to 2017/06/30. Wave 4: 2017/01/17 to 2017/10/29. Wave 5: 2018/01/26 to 2018/02/24. Wave 6: 2018/03/15 to 2018/04/02.

2.2 American Community Survey

The American Community Survey (ACS) is an annual survey of the U.S. Census Bureau, designed to supplement the decennial census. Based on the long-form version of the census, the ACS collects information on topics including population, housing, employment and education from a nationally representative sample. The survey is sent to around 3.5 million households annually[7].

Data on migrant stocks can be readily obtained from the ACS. In particular, in every year of the ACS, the survey has contained a question asking the birthplace of the person; if it is inside the United States, the state is recorded, and if it is outside the United States, the country is recorded. This birthplace variable is recorded as a three digit code to indicate the US state or country of birth. In addition to the birthplace variable, the ACS has information on current state of residence. Thus, we can tabulate the number of migrants from a particular country living in a particular state by looking at the combination of these two variables.

Tabulations of migrants by age and state for every year in the period 2000-2016 can be obtained through the IPUMS online data analysis system[8].

3 Methods

In this section, we describe the methods that we developed in order to model data from Facebook Adverts Manager and the American Community Survey. In particular we present, step by step, how we combine all the available information within a unified Bayesian hierarchical model.

The intuition for our approach is simple. We start by using the most recent data from the ACS in order to calibrate Facebook estimates for a corresponding time interval. A simple regression model has been effective for evaluating the extent of bias across subgroups of the population. Without additional information, we could use this simple model to nowcast stocks of migrants. However, the ACS also provides historical time series for our quantity of interest and for the subgroups of the population that we are interested in, namely age and sex groups. We reduce the dimensionality of schedules of migrants in US states by country of origin using Singular Value Decomposition (SVD). In particular, the approach that we use is intimately connected with classic methods for modeling life tables and

[7] `https://www.census.gov/content/dam/Census/programs-surveys/acs/about/ACS_Information_Guide.pdf`

[8] https://usa.ipums.org/usa/sda/

demographic forecasting. Finally, we combine the regression model and the time series approach within a unified Bayesian hierarchical model. The resulting forecasts can be interpreted in a loose way as weighted averages of the two models. They account for information contained in the most recent Facebook data, while also avoiding values that are inconsistent with historical trends and age patterns. The following sections provide more details about our modeling strategy. Section 3.1 describes the approach for evaluating the relationship between Facebook data and the ACS; section 3.2 offers the intuition behind the time series model for the ACS data; finally, section 3.3 provides the details for our main contribution: a Bayesian hierarchical model that combines the ideas described in sections 3.1 and 3.2.

3.1 Modeling Facebook data

It has been shown that a simple regression model can be used to evaluate patterns of bias in Facebook data by age and sex for stocks of migrants [14]. Our modeling approach therefore builds on the work of Zagheni et al. (2017). Here is the simple regression model that we consider:

$$
\begin{aligned}
logit(\text{ACS fraction of foreign born pop}_{ij}^{z}) = {} & \beta_0 + \beta_1 logit(\text{Facebook fraction of expats}_{ij}^{z}) + \\
& + \beta_2 \mathbb{1}(\text{Origin 1}) + \cdots + \beta_{30} \mathbb{1}(\text{Origin 29}) + \\
& + \beta_{31} \mathbb{1}(\text{Age group 20-24}) + \cdots + \\
& + \beta_{38} \mathbb{1}(\text{Age group 55-59}) + \\
& + \epsilon_{ij}^{z}
\end{aligned}
\tag{1}
$$

where '*ACS fraction of foreign born pop*$_{ij}^{z}$' is the fraction of people in the age-sex group z born in country i and living in US state j. '*Fraction of Facebook expats*$_{ij}^{z}$' is the fraction of expats in the age-sex group z from country i who live in US state j. $\mathbb{1}(\text{Origin 1})$ is an indicator variable for country of origin 1. $\mathbb{1}(\text{Age group 20-24})$ is an indicator variable for age group 20-24 years old. In this model, each country of origin has its own coefficient, which serves as a 'level' parameter. Each age group also has its own coefficient that serves as a 'shape' parameter.

3.2 Modeling data from the American Community Survey

We consider a time series model in combination with an SVD approach that reduces the dimensionality of the matrix of age-time schedules of stocks of migrants. The modeling framework is inspired by the Lee-Carter mortality model [6], which was originally formulated to forecast US life expectancy. More broadly, the method is related to the general approach that demographers use when dealing with model life tables or other instances in which regularities in age patterns of demographic rates can be leveraged to reduce the dimensionality of the data problem. The forecasted indexes in lower dimensions are then used to infer entire future schedules of demographic rates. In sum, the model is based on the idea that age-specific patterns of stocks of migrants are fairly stable and change in a relatively regular way over time.

Consider migrants from a particular country living in a given US state. Define p_{xt} to be the fraction of these migrants in age group x and year t. These p_{xt} are modeled as

$$logit(p_{xt}) = a_x + \kappa_t \cdot b_x + \varepsilon_{x,t} \tag{2}$$

where a_x is the age-specific mean of the logit-transformed age schedules in each year. The b_x term represents the contribution of each age group to the total amount of change. The b_x can be obtained via Singular Value Decomposition of the logit-transformed, demeaned matrix of age-specific migrant populations in each year. In particular, b_x is equal to the first right singular vector. For our modeling purposes, the logit transformation is particularly convenient; however, other transformations could be used in various contexts. κ_t is typically modeled using a time series approach. Forecasts of κ_t can be plugged into equation 2 in order to generate forecasts of p_{xt}.

3.3 Combining Data from Facebook and the ACS within a Bayesian Hierarchical Model

The previous two sections provided the intuition for separate modeling strategies related to the use of Facebook data only or ACS data only. Here we summarize the Bayesian hierarchical model that we propose in order to combine the two separate approaches within a unified framework. The general idea is that we have qualitatively different data sources that provide different perspectives on the underlying phenomenon. Combining data sources and modeling strategies enables the researcher to weigh down inconsistencies and extract valuable insights without ignoring existing information.

First we include a formal presentation of the model. Then we present, in loose terms, the interpretation for each component. The following model is for one migrant group from one specific country of origin living in one particular US state:

$$
\begin{aligned}
\operatorname{logit} p_{xt} &\sim N(\mu_{xt}, \sigma^2) && (3)\\
\mu_{xt} &= a_x + \kappa_t b_x + \varepsilon_{xt} && (4)\\
\Delta^2 \kappa_t &\sim N(0, \sigma^2_\kappa) && (5)\\
\varepsilon_{xt} &\sim N(\rho_x \varepsilon_{x,t-1}, \sigma^2_\varepsilon) && (6)\\
\rho_x &\sim U(-1, 1) && (7)\\
\sigma^2 &= \begin{cases} \sigma^2_p, & \text{if } 2001 \le t \le 2016 \\ \sigma^2_p + \sigma^2_{bias} + \sigma^2_{ns}, & \text{if } t = 2017 \end{cases} && (8)\\
p_{xt} &= \begin{cases} \text{from ACS}, & \text{if } 2001 \le t \le 2016 \\ p^*_{xt} \text{ (estimated proportion)}, & \text{if } t = 2017 \end{cases} && (9)\\
p^*_{xt} &= \overline{p^*_{xtj}} && (10)\\
p^*_{xtj} &= \hat{\beta}_0 + \hat{\beta}_1 \cdot p^{\text{facebook}}_{xtj} + X\hat{\beta} && (11)
\end{aligned}
$$

where

- p_{xt} is the observed proportion of migrants in age group x at time t;

- μ_{xt} is the 'true' unobserved proportion;

- κ_t is a time-varying coefficient that serves a role similar to the one in the Lee-Carter mortality model. Here large second-order differences are penalized;

- ε_{xt} is an age-time random effect. It can be thought of as introducing distortions around the mean, and it is modeled as an autoregressive process of order 1;

- σ^2 is the variance of proportions around the truth on the logit scale;

- a_x is the mean age schedule (on logit scale) of ACS proportions from 2001-2015.

- b_x is the first singular vector from the Singular Value Decomposition (SVD) of the demeaned age by time matrix of ACS proportions from 2001-2015;

- σ_p^2 is the variance in model estimates associated to the sampling error and time variability in the ACS;

- σ_{bias}^2 is the variance associated to bias-adjustment predictions for the Facebook regression model;

- σ_{ns} is the variance associated to other non-sampling error in Facebook data. For example, we have different waves of Facebook data, each of them including numbers that could be rounded in different ways;

- p_{xt}^* is the proportion of migrants in age group x at time t estimated from observed Facebook data;

- p_{xtj}^* is the observed proportion of migrants observed in Facebook wave j, as we collected data several times over the course of the year;

- $\hat{\beta}_0, \hat{\beta}_1, \hat{\beta}$ are estimated coefficients based on the regression model run using 2016 ACS data as dependent variable, and the first Facebook wave in January 2017 as covariate data.

- Non-informative priors were selected for all variance parameters,and the βs.

Equation 4 is essentially the principal component model described in section 3.2. Equation 5 sets constraints on the time-varying parameter κ_t: we expect second differences over time to be small on average, meaning that trends are expected to be linear on average, with temporal autocorrelated distortions described by equation 6. Equation 8 distinguishes the types of variability that are modeled. For the period between 2001 and 2016, the variance in the model estimates is mainly related to the time series variability and the sampling error in the ACS. For the nowcast for 2017, in addition to variability in temporal trends, we also have variance related to the predictive regression model for bias in Facebook, as well as variance associated to other non-sampling errors in Facebook data (e.g., we collected several waves of Facebook data for each year and each of these waves includes values that could have been rounded in different ways). Equation 9 indicates that for the years for which we have data from the ACS, the data points come from the ACS. For the most recent year, for which we do not have data from the ACS, the estimate from the Facebook model serves as a provisional data point.

4 Results

In this section, we show some illustrative results from the Bayesian hierarchical model that we propose. Figure 1 shows nowcasts of fractions of people living in California who were born in Mexico, by age and time. The figure includes both estimates from the Bayesian hierarchical model and the nowcasts that we would have obtained based on the Facebook model only, without accounting for trends in time series from the ACS. For most age groups, the two values for the nowcasts are fairly close to each other, but that is not always the case. For people who are 35-39 years old, recent data from Facebook indicate a reversal in trends that is also predicted by the hierarchical model. For people who are 45-49 years old, the Facebook-only estimate is lower than the prediction from the hierarchical model. For the 45-49 age group, the Facebook-only predictions have less leverage as the outcome would be more unlikely given the historical age-time patterns. To elaborate on this point, figure 2 shows age schedules for estimates and nowcasts over time. Patterns that deviate from historical regularities are weighed down by the model.

Figure 3 offers some insights into the leverage that Facebook data points have on the estimates obtained from the hierarchical model. More specifically figure 3 shows nowcasts of fractions of Mexican-born people in California for 2017, including a fictitious scenario in which Facebook values for 2017 were higher than what they actually were. For the age group 40-44, if Facebook suddenly showed an increase in migrant stocks of Mexicans by 10%, the estimates from the Bayesian hierarchical model would be on average 2.6% higher across all age groups for Mexicans in California. The effect of the new information from Facebook would range from 0.9% to 5.6% across all age groups. The latest figures from Facebook would have a higher impact on age groups with historical trends that are less linear and have higher variance. In those cases, deviations from historical trends are less 'unusual' and more likely to be accommodated. Nonetheless, an extreme 10% increase in Facebook values would not result in a 10% increase in the final estimates. This is an important feature of our model, which accounts for new information, but also weighs down the effect of new data that are inconsistent with historical age schedules or unlikely given temporal trends.

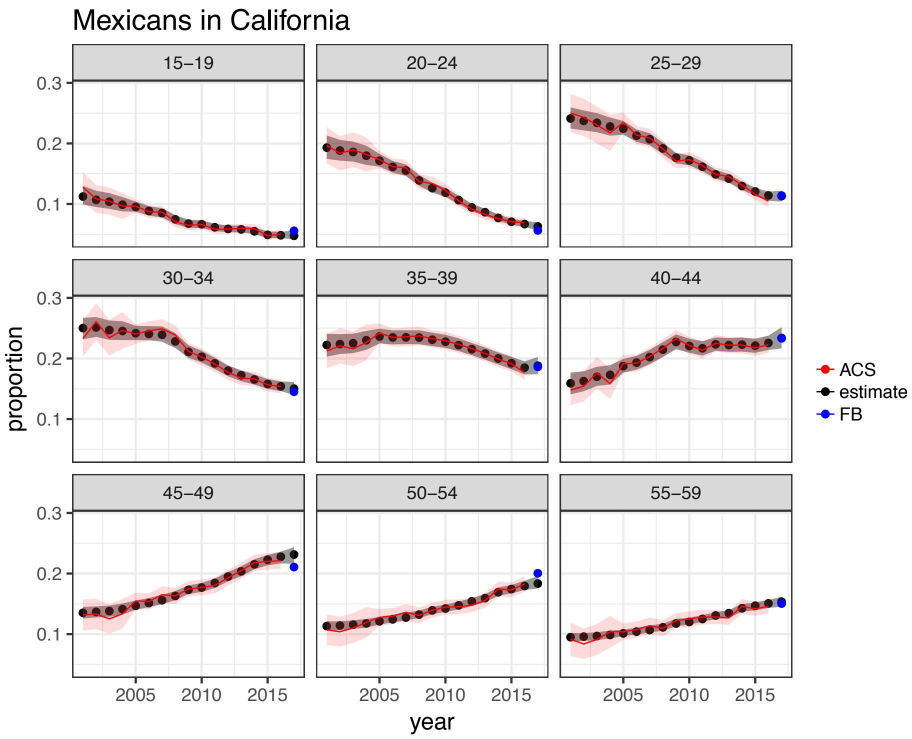

Figure 1: Illustrative example of nowcasts of fractions of people living in California who were born in Mexico obtained from the Bayesian hierarchical model. The solid red line represents the estimates from the American Community Survey, with the red shaded area indicating confidence intervals related to sampling variability. The black dots are the estimates and nowcasts of the Bayesian hierarchical model, with the grey shaded area indicating the associated credibility intervals. The blue dots are the nowcasts that we would have obtained based on the Facebook model only, without accounting for trends in time series from the ACS.

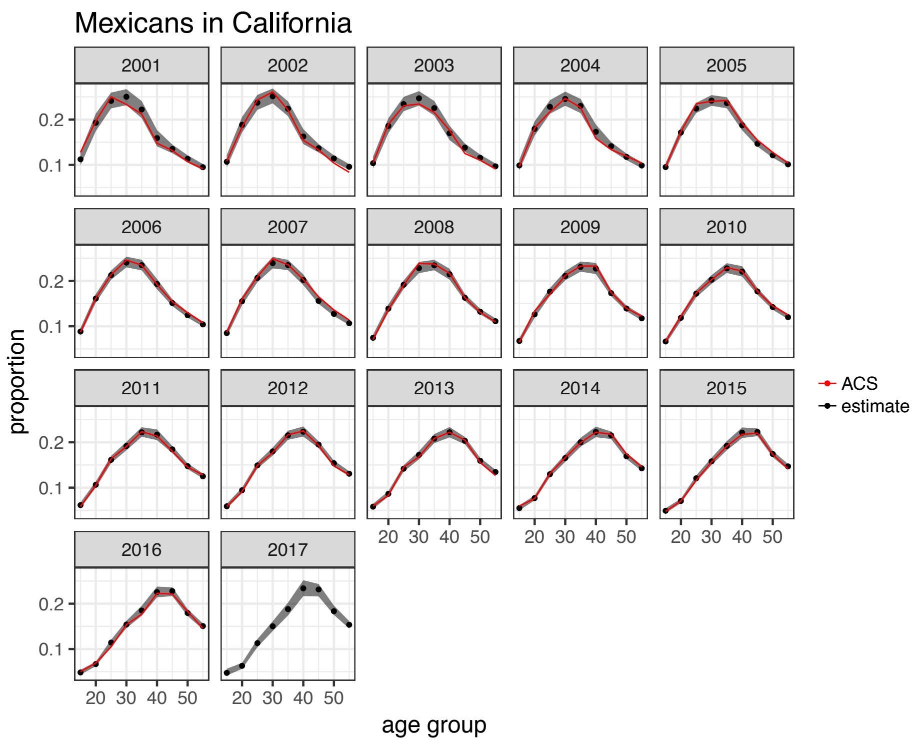

Figure 2: Age schedules for estimates and nowcasts of fraction of Californians born in Mexico from 2001-2017. The red lines indicate estimates from the ACS. The black dots are the estimates from the Bayesian hierarchical model. The grey shaded areas indicate the credibility intervals for the estimates from the Bayesian hierarchical model.

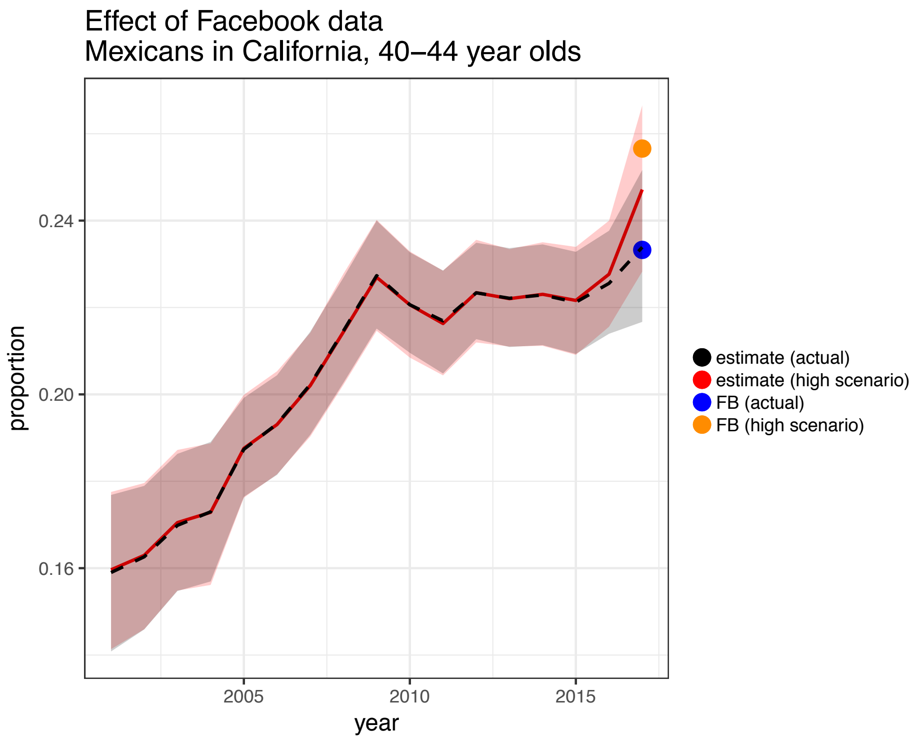

Figure 3: Nowcasts of fractions of Mexican-born people aged 40-44 in California for 2017. The black dashed line represents estimates from the hierarchical model. The red solid line shows estimates from the hierarchical model under a fictitious scenario in which Facebook values for 2017 were 10% higher than what they actually were.

5 Discussion and Next Steps

In this article we developed a Bayesian hierarchical model to generate nowcasts of migration stocks in US states by age, sex and country of origin. The approach combines traditional survey data and social media data within a unified framework.

To illustrate the features of the model, we showed estimates and nowcasts for people born in Mexico and living in California. One of the next steps is to provide estimates for other states and for other countries of origin. We are in the process of testing whether each state and country of origin should be modeled separately or useful information can be borrowed across space and time. More broadly, future steps will build on the method that we presented. Improvements to this model could include refinements of the Facebook regression model, or of the time series approach or of the way in which different types of information are combined and regularities in patterns leveraged.

Although we applied our methods on two data sources (Facebook and ACS), the approach could include additional data sets. For example, other sampling surveys or information from other social media sites could be incorporated. As more data will be included, the approach will be more resilient to any potential sudden change in the availability of each of the data sources used. The type of social media data that we used, aggregate-level counts of users, are becoming increasingly available as a result of advertisers' needs. Given the nature of the data (aggregate-level information that is not available for very small groups of users and that preserve the privacy of users), we expect that more data sources like the one that we used from Facebook will be available in the future.

As social media and Web data become more and more normalized in demography, it is important to develop tools to model biases, and to combine complementary sources of information. Although the focus of this paper is on migration, the methods that we develop are general and could be used for a number of applications in different substantive areas. We see this as an opportunity to increase the value for researchers of both traditional surveys and digital trace data.

Our paper shows that it is possible to exploit data that are the by-product of our digital life to gain a better understanding of the world, including for 'official' demographic estimation. We hope that by presenting a concrete application, our article contributes also to the general debate[8] about the value of digital trace data and the importance of protecting these data, of making them available to researchers while preserving privacy, as well as of developing methods for understanding them.

References

[1] L. Alkema, A. E. Raftery, P. Gerland, S. J. Clark, and F. Pelletier. Estimating trends in the total fertility rate with uncertainty using imperfect data: Examples from west africa. *Demographic research*, 26(15), 2012.

[2] M. Araujo, Y. Mejova, I. Weber, and F. Benevenuto. Using facebook ads audiences for global lifestyle disease surveillance: Promises and limitations. WebSci '17, New York, NY, USA, 2017. ACM.

[3] J. Ginsberg, M. H. Mohebbi, R. S. Patel, L. Brammer, M. S. Smolinski, and L. Brilliant. Detecting influenza epidemics using search engine query data. *Nature*, 457(7232):1012–1014, 2009.

[4] A. Herdağdelen, L. Adamic, W. Mason, et al. The social ties of immigrant communities in the united states. In *Proceedings of the 8th ACM Conference on Web Science*, pages 78–84. ACM, 2016.

[5] D. Lazer, R. Kennedy, G. King, and A. Vespignani. The parable of google flu: traps in big data analysis. *Science*, 343(6176):1203–1205, 2014.

[6] R. D. Lee and L. R. Carter. Modeling and forecasting us mortality. *Journal of the American statistical association*, 87(419):659–671, 1992.

[7] T. A. Moultrie, R. Dorrington, A. G. Hill, K. Hill, I. Timæus, and B. Zaba. *Tools for demographic estimation*. International Union for the Scientific Study of Population, 2013.

[8] H. Shah. Use our personal data for the common good. *Nature*, 556(7), 2018.

[9] W. Wang, D. Rothschild, S. Goel, and A. Gelman. Forecasting elections with non-representative polls. *International Journal of Forecasting*, 31(3):980–991, 2015.

[10] D. Yildiz, J. Munson, A. Vitali, R. Tinati, and J. Holland. Using twitter data for population estimates. 2017.

[11] E. Zagheni, V. R. K. Garimella, I. Weber, and B. State. Inferring international and internal migration patterns from twitter data. In *International World Wide Web Conference*, pages 439–444, 2014.

[12] E. Zagheni and I. Weber. You are where you e-mail: using e-mail data to estimate international migration rates. In *Web Science*, pages 348–351, 2012.

[13] E. Zagheni and I. Weber. Demographic research with non-representative internet data. *International Journal of Manpower*, 36(1):13–25, 2015.

[14] E. Zagheni, I. Weber, and K. Gummadi. Leveraging facebook's advertising platform to monitor stocks of migrants. *Population and Development Review*, 43(4):721–734, 2017.