Assessing Biases of Facebook Data to Nowcast Migrant Stocks in the United States

Esther Denecke*[a,b], Monica Alexander[c], Emanuele Del Fava[a], Emilio Zagheni[a]

[a] *Max Planck Institute for Demographic Research*
[b] *University of Rostock*
[c] *University of Toronto*

November 23, 2021

Work in progress. Please do not cite without author's permission.

Short Abstract

Data on migration are often insufficient or released with a large delay. While the delayed release of migration statistics is a general problem, the Covid-19 pandemic further emphasizes the need for timely data to assess the impact of discontinuities on migration. New and innovative data sources, such as social media data, could help monitoring and estimating migrant stocks and flows in real time. However, these data are highly biased. Only if this bias is properly accounted for, social media data may supplement existing representative data sources such as surveys. Often, it is assumed that this bias is constant over time which may not be tenable as both platforms and user behavior underlie frequent changes. We revisit the assumption of constant biases using data from the Facebook Ads Manager assessing the variability of the Facebook data over time.

*Corresponding author: denecke@demogr.mpg.de

Introduction

Data on migration are often insufficient or released with a large delay. While the delayed release of migration data is a general problem, the ongoing Covid-19 pandemic emphasizes the need for more timely data once more. The pandemic and especially its aftermath could lead to the discontinuation of past trends (Guadagno, 2020), and as such, it is of interest to obtain real-time estimates, so called "nowcasts". Researchers have explored the feasibility of using social media data to aid in this task (Zagheni et al., 2017; Alexander et al., 2019; Alexander et al., 2020). While social media data can be collected in real-time it is well known that these data are biased (e.g. Zagheni et al. (2017)). Previous work on monitoring migrant stocks with social media data has attempted to account for these biases but adjustments have mostly been relatively simple, often relying on the strong assumption that biases are constant over time (Zagheni et al., 2017; Alexander et al., 2020). In this work in progress, we revisit this simplistic assumption using Facebook data collected over the years 2017 and 2018.

Zagheni et al. (2017) explore the use of Facebook (FB) data for estimating migrant stocks and study its biases by means of a linear regression model; regressing representative data from the American Community Survey (ACS) on the Facebook data. Alexander et al. (2020) combine Facebook data and data from the American Community Survey to nowcast migrant stocks in the US. They (i) use a regression model to adjust the biased Facebook data ('bias-adjustment model') and (ii) a time series model to make use of all available information in the historic ACS data. Finally, the two approaches are combined in a larger framework by means of a Bayesian hierarchical model. In the bias-adjustment step, the representative data from the American Community Survey (ACS), used as a gold-standard, are regressed on the nonrepresentative Facebook data. Subsequent waves of Facebook data are then adjusted using the estimated regression coefficients. Any changes in the Facebook data are attributed to a change in migrant stocks (though the model attributes more uncertainty to the final estimates of migrant stocks). Thus, biases and measurement errors in this one wave are propagated through subsequent adjustments. This is known to be problematic as online platforms are subject to constant change (e.g. Zagheni et al. (2017, p. 730), Lazer et al. (2014), and Salganik (2019, chapter 2)). In fact, it has been noted that there may be variation in the Facebook data due to reasons other than migration (Alexander et al., 2019).

In addition to statically propagating biases and measurement errors through subsequent adjustments, it is unclear which wave of Facebook data should be used as a predictor in a regression with one year of ACS data as outcome. As a proof-of-concept, Zagheni et al. (2017) use Facebook data collected in 2016 with the ACS estimates of 2014, and Alexander et al. (2020) use Facebook data collected in January 2017 with ACS data of 2016. Additionally, for fitting the model, both studies rely on one wave of Facebook data only. We recognize that this problem of matching the ACS data with the Facebook data, and the assumption of constant biases, are inherently intertwined. To illustrate, it is unclear how the Facebook data evolve and vary over the course of one year and as a consequence, it is unclear how using, for example, the latest wave of one year or all waves collected in one year to adjust another wave of Facebook data influence the adjustment.

Thus, revisiting the assumption of constant biases, the present work in progress assesses the variability of Facebook data over time. In the next section, the data are introduced. Then, we explore the variability of the Facebook data and show an illustrative example of how to deal with this variability. We conclude by giving an outlook on potential areas of development.

Data

The project utilizes representative data from the American Community Survey (ACS) (Ruggles et al., 2020) as well as nonrepresentative data collected via Facebook's advertising platform. The ACS contains information on respondent's birthplace, which allows us to calculate the number of migrants from a particular origin in each state for different age and sex groups. An international migrant is defined as a person that was born outside of the US.[1]

The second data source consists of information extracted from Facebook's advertising platform. The number of monthly active users that could potentially be reached by advertisers can be extracted through this platform. To allow for targeting of advertisements, several characteristics of the potential audience can be chosen. This enables the collection of data with a specified age group, sex, and origin in a particular state. It is, however, not clear how this number is calculated and, additionally, these calculations may change over time. The data collection took place via Facebook's Marketing API[2] with the Python module pySocialWatcher (Araujo et al., 2017). Four waves of such data were collected in each 2017 and 2018, giving a total of eight waves.[3] In addition to being biased, Facebook can change its algorithms and the API at any time. This happened in the midst of wave 5: the lower bound of the number of expats was raised from 20 to 1000 – adding censoring to our data.

Here, we focus on male Mexican migrants. For both data sources, we have the number of migrants in nine different age groups, $15-19, 20-24, \ldots, 55-59$ in the different US states. Note, however, that there are missing values in the ACS data and left-censoring in the Facebook data. Additionally, there are some implausible values in wave 8. We pre-process the Facebook data by removing left-censored and implausible observations. After this, we remove all those observations which have a missing value in the ACS data. This leaves us with a total of 39 states (which are not necessarily complete with regard to the age groups) and a total of 285 observations per wave; with the exception of wave 8 which has 270 observations after pre-processing.[4]

Descriptives & Illustrative Example

All results in this section were produced using the software `R` (R Core Team, 2021), data wrangling was done with the package `data.table` (Dowle and Srinivasan, 2021) and plots were produced with `ggplot2` (Wickham, 2016).

Figure 1 shows the number of male Mexican migrants in California according to both data sources in 2017 and 2018. There is systematic bias in the Facebook data: Both in 2017 and 2018, the number of migrants in the older age groups is underestimated by the Facebook data whereas in some of the younger age groups the number of migrants is overestimated. Additionally, there is some variability over the different waves of the Facebook data. In 2017 this variability is larger for

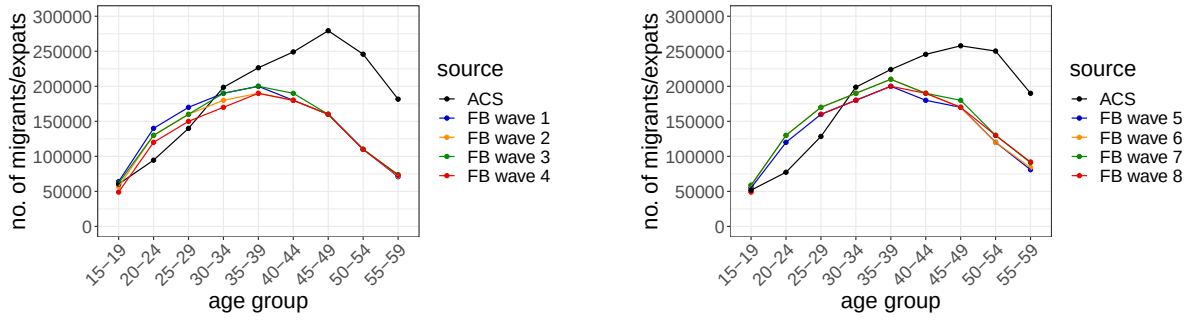
[1] This follows the definition of international migrants used by the United Nations Department of Economic and Social Affairs, Population Division (United Nations Department of Economic and Social Affairs, Population Division, 2020, p. 5).

[2] https://developers.facebook.com/docs/marketing-apis

[3] Some of the data have been used in Alexander et al. (2019) and Alexander et al. (2020). Also see the descriptions therein.

[4] For the sake of brevity, we are not giving a complete description here.

the younger age groups and virtually non-existent for the older age groups. In 2018, there is some variability in all age groups but the maximum difference in the number of migrants between two waves is smaller. It is important to note that the source of the variability is unclear.



(a) ACS 2017, Facebook waves 1–4

(b) ACS 2018, Facebook waves 5–8,

Figure 1: Number of male Mexican migrants in California according to both data sources. Missing value in wave 8 for age group 20–24. Figures inspired by Zagheni et al. (2017), Alexander et al. (2019), and Alexander et al. (2020).

It is an open question whether using several waves or only the latest wave of Facebook data in a bias-adjustment model is the best choice. In the spirit of the models in Zagheni et al. (2017) and Alexander et al. (2020) we fit four simple regression models. In our case, with two years of overlapping Facebook and ACS data, we can match the years, i.e. ACS data from 2017 are regressed on FB data from the same year.

Let $\text{pop}_{a,s}^{\text{ACS},2017}$ be the number of male Mexican migrants in the 2017 ACS data in a particular age group a in state s, $a = 1, \ldots, 9, s = 1, \ldots, 39$, and $\text{pop}_{a,s}^{\text{FB}}$ the number of expats in the Facebook data. We explain the four models we fit with the following regression model:

$$\begin{aligned}\log(\text{pop}_{a,s}^{\text{ACS},2017}) = &\beta_0 + \beta_1 \log(\text{pop}_{a,s}^{\text{FB}}) \\ &+ \beta_2 \mathbb{1}(\text{State } 1) + \cdots + \beta_{39} \mathbb{1}(\text{State } 38) \\ &+ \beta_{40} \mathbb{1}(\text{Age Group } 1) + \cdots + \beta_{47} \mathbb{1}(\text{Age Group } 8) \\ &+ \epsilon_{a,s},\end{aligned}$$

where $\mathbb{1}(\cdot)$ denotes the indicator function and $\epsilon_{a,s}$ the error. The four models are as follows:

1. Baseline model: The Facebook data do not enter the model, i.e. the blue part is removed.
2. Wave 4 only: Only wave 4 enters the model.
3. Stacked: All waves from 2017 (waves 1 – 4) enter the model. For this purpose we create a stacked data set where the ACS values are repeated for the different waves.
4. Averaged: For each age group, state, and origin, the values of all waves from 2017 (waves 1 – 4) are averaged.

Figure 2 shows the adjusted number of migrants from wave 6 in 2018 based on the four different models as well as the ACS data in California. In this example, the Facebook data add value. Overall, the adjusted number of migrants from the models including the Facebook data are closer

to the ACS data than those of the baseline model – with exceptions in age groups $35-39$ and $40-44$.

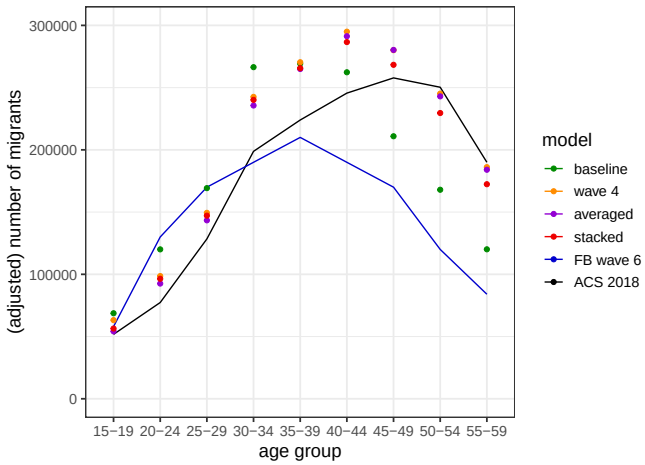


Figure 2: Estimated number of migrants in 2018 based on the four different models; adjustment of wave 6 collected in 2018. Also shown are wave 6 of the Facebook data (blue line) as well as the 2018 ACS data (black line).

Potential Areas of Development

In this work in progress, we looked at the variability of the Facebook data in 2017 and 2018. As such, there is systematic bias in the Facebook data while there is also some variability. The source of the variability, however, is unknown. Additionally, we have proposed two ideas of integrating more than one wave of Facebook data into a model for bias adjustments; and shown these in an illustrative example together with a baseline model and one that uses one wave of Facebook data only. Several potential areas of development remain.

The models shown above have some issues that should be resolved: (i) Figure 2 does not show any uncertainty. It is, however, crucial to include the different sources of uncertainty. This would also help in understanding how meaningful differences between the different waves are. (ii) The model suffers from heteroscedasticity. (iii) We have considered the number of migrants only but we have not taken into account the total population in the different states. In the future, this should be taken into account by working with the proportion of migrants or a count model. (iv) We have removed all observations with left-censoring in the Facebook data. Lastly, we have shown results for male Mexican migrants in California only. In the future, both females as well as other states should be studied in more detail.

While our use case is the estimation of migrant stocks with Facebook data, we would like to point out that the variability of digital traces is a challenge that is neither unique to migration research nor to Facebook data.

References

Alexander, Monica, Kivan Polimis, and Emilio Zagheni (2019). "The Impact of Hurricane Maria on Out-migration from Puerto Rico: Evidence from Facebook Data". In: *Population and Development Review* 45.3, pp. 617–630.

Alexander, Monica, Kivan Polimis, and Emilio Zagheni (2020). "Combining Social Media and Survey Data to Nowcast Migrant Stocks in the United States". In: *Population Research and Policy Review*.

Araujo, Matheus, Yelena Mejova, Ingmar Weber, and Fabricio Benevenuto (2017). "Using Facebook Ads Audiences for Global Lifestyle Disease Surveillance: Promises and Limitations". In: WebSci '17. New York, NY, USA: ACM.

Dowle, Matt and Arun Srinivasan (2021). *data.table: Extension of 'data.frame'*. R package version 1.14.2. URL: `https://CRAN.R-project.org/package=data.table`.

Guadagno, Lorenzo (2020). *Migrants and the COVID-19 pandemic: An initial analysis*. Tech. rep. Geneva: Research Series N° 60.

Lazer, D., R. Kennedy, G. King, and A. Vespignani (Mar. 2014). "The Parable of Google Flu: Traps in Big Data Analysis". In: *Science* 343.6176, pp. 1203–1205.

R Core Team (2021). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing. Vienna, Austria. URL: `https://www.R-project.org/`.

Ruggles, Steven, Sarah Flood, Ronald Goeken, Josiah Grover, Erin Meyer, Jose Pacas, and Matthew Sobek (2020). *IPUMS USA: Version 10.0*. [dataset]. Minneapolis, MN: IPUMS.

Salganik, Matthew (2019). *Bit by Bit*. Princeton, NJ: Princeton University Press.

United Nations Department of Economic and Social Affairs, Population Division (2020). *International Migration 2020 Highlights*. (ST/ESA/SER.A/452).

Wickham, Hadley (2016). *ggplot2: Elegant Graphics for Data Analysis*. Springer-Verlag New York. URL: `https://ggplot2.tidyverse.org`.

Zagheni, Emilio, Ingmar Weber, and Krishna Gummadi (2017). "Leveraging Facebook's Advertising Platform to Monitor Stocks of Migrants". In: *Population and Development Review* 43.4, pp. 721–734.