2. Using social media advertising data to estimate migration trends over time

Monica Alexander

2.1 INTRODUCTION

Measuring population movements, and understanding how they change over time and across space, is essential for understanding broader population change and has implications for urban policy and planning. However, compared to births and deaths, the other two components of population change, migration is more difficult to measure. Part of this is due to definition issues: migration involves both a geographic and temporal component, so the definition of who constitutes a migrant or what constitutes migration varies depending on context. For example, one may be interested in migration across national borders (international) or within countries (internal); in addition, the focus could either be on short-term or long-term migration.

Monitoring migration patterns over time and space is also challenging because of a lack of available data. In contrast to births and deaths, where data are often recorded through vital registration systems, detailed data on migration are less likely to be centrally recorded. For example, data on internal migration may be lacking because the migration events were not formally recorded. In terms of international migration, countries generally have a larger incentive to record and track in-migration (or immigration), rather than out-migration (or emigration). Even if official data on migration exists, it is often released after long delays, or without the granularity of information that is required for reliable population projection (for example, a breakdown by age or sex). Of all the components of population change, the need for timely and informative data is arguably the most important for migration, because flows can change dramatically in such a short period of time, for example, in response to war and conflict, a natural disaster such as a hurricane or flood, or global pandemic.

As a consequence of data sparsity and delays, migration researchers and policymakers have begun to investigate the potential for using other 'big data' sources of information. In particular, large amounts of demographic data are produced through the use of social media websites, such as Facebook, Instagram, LinkedIn, and Twitter. Indeed, from a demographic perspective, the users of social media websites can be thought of as their own population, with births occurring when users sign up to the platform, and deaths occurring when users delete their account. Demographic events of interest in a more traditional sense are potentially recorded through a user's lifetime, such as birthdays, births, and migration to other cities or countries. Data are produced in a number of ways; in terms of mobility information, updates can be made both explicitly through users updating their information (e.g., location of residence), and implicitly through 'digital traces', if the location of a user is inferred from their IP

address, mobile phone location, or photo backgrounds. Data on the population using a social media website can be thought of as a digital census, which is updated essentially in real time. Social media data therefore offers large potential for complementing traditional data sources of migration information.

There are, however, some notable issues with using social media data to estimate migration indicators. First, the population of users on a social media website is unlikely to be representative of the broader population of interest. The level of reliability of data will depend on a variety of factors, including how widespread the platform is used in a population. Facebook, for example, has a user base that ranges from less than 20 percent in most of Africa to around 90 percent in the United States and Canada (Internet World Stats 2020). In addition, the user base tends to vary substantially by demographic characteristics, such as age and education. Using social media data without appropriate adjustments would therefore likely lead to heavily biased results. A second issue is data accessibility. Unless there exists a direct agreement or connection with the company who runs the social media platform, it is very unlikely that an outside party would be able to access the 'raw' data produced from the website. And finally, even if access to the micro-level data were possible, there are confidentiality and ethical concerns with using an individual's data for a purpose other than what was originally intended. The dangers of such data use were highlighted by the Cambridge Analytica controversy, and are still pertinent today.

Ideally, when using social media data to produce estimates of migration, the strength of these data would be combined with the strength and reliability of traditional, 'gold-standard' demographic data sources. The challenge is to objectively and systematically combine these multiple sources of data, in order to fully take advantage of all information, while adjusting for known issues.

This chapter illustrates the use of demographic data sourced from social media advertising platforms, in combination with traditional demographic data sources to obtain up-to-date estimates of migration. The use of advertising data, which is publicly available and aggregated by population sub-group, alleviates issues regarding data access and confidentiality. I introduce a statistical framework for combining traditional data sources and the social media data. The framework is presented in a general way and emphasizes the importance of three main components: adjusting for non-representativeness in the social media data; incorporating historical information from reliable demographic data; and accounting for different errors in each data source. I will illustrate how data from Facebook's advertising platform can be used to estimate migrant stocks in North America. The contribution of this chapter to the existing literature is methodological in nature, outlining a broad statistical approach that can be used to robustly combine information from different data sources.

The remainder of the chapter is structured as follows. The next section introduces the characteristics of advertising data from social media and reviews previous research that uses such data in demographic research. The Data section then outlines the general approach to collecting social media advertising data and discusses potential gold-standard sources. The Model section outlines a general statistical framework that can be used to combine social media data with traditional data sources. An example of estimating Mexican migrant stocks by state in the United States is then illustrated. The final section concludes.

2.2 BACKGROUND

2.2.1 What is Social Media Advertising Data?

The majority of social media websites rely on advertising for a large portion of their revenue. Individuals or companies can advertise on platforms, for goods, services, or to take a survey. The advertisements appear to users embedded within their social media 'feed', and depending on the ad design, can be difficult to distinguish from the usual social media content. Advertisements are designed, managed and posted through a platform that is associated with the social media website. For example, Facebook's advertising platform (called the 'Ads Manager') can be accessed through www.facebook.com/business. Anyone with a Facebook account can post advertisements, for a fee that is charged as a rate per post.[1] The remainder of this section, and this chapter, will focus on Facebook as an example, although many other platforms have similar functionality.

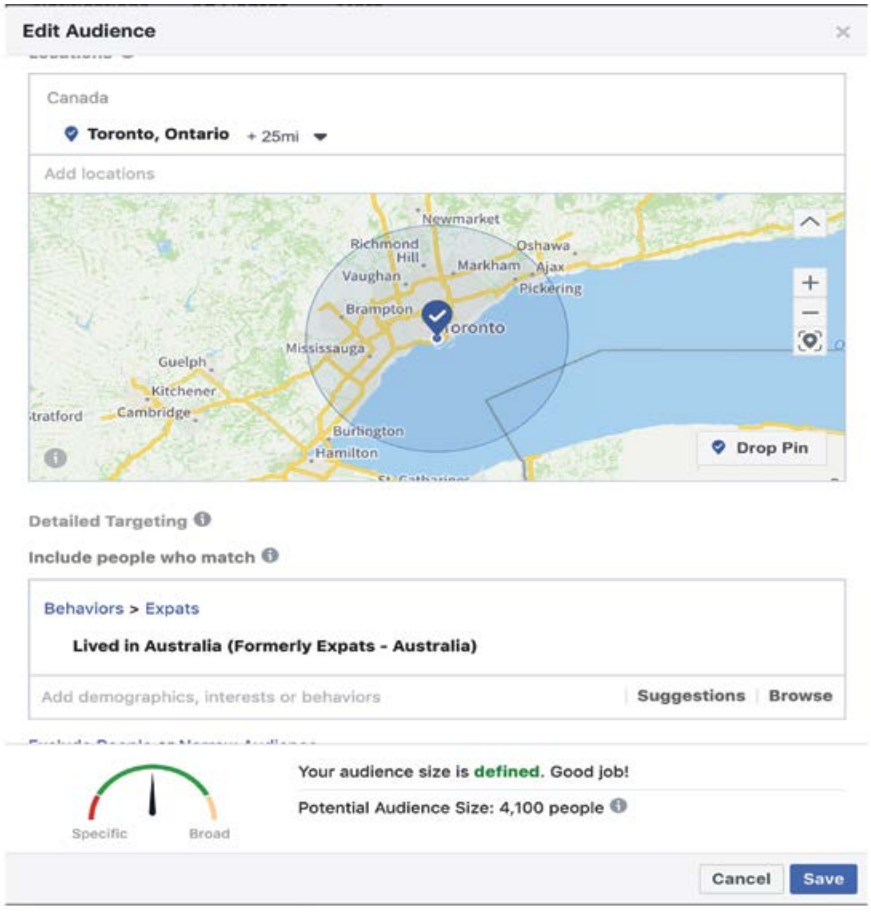

Figure 2.1 Facebook's advertising platform, as at 29 May 2020

Facebook's Ads Manager allows advertisers to target the audience of their ad based on demographic, geographic and socioeconomic characteristics of the users. An advertiser can select on characteristics such as age, sex, education, location of residence, whether the person is traveling, or whether they are an expat. For example, imagine an advertiser that would like to target Australians who live in Toronto. A screenshot of this selection is shown in Figure 2.1. Once the advertiser selects the relevant target groups, Facebook's advertising platform displays an estimate of the potential reach of the ad. Figure 2.1 shows that for Australian expats in Toronto, the potential reach was 4,100 people, as at 29 May 2020.

From an advertising perspective, the size of the target audience is useful to gauge how many people will likely see the ad content and whether this is too broad or narrow for the purposes of the business. However, there is a secondary utility of these data: as a demographic estimate.

In the example above, the potential reach of that ad is a data point: an observation of the number of Australian migrants in Toronto. We know that this data point is likely to be an imperfect observation of the true number of Australian migrants in Toronto, given that it is unlikely that all Australians use Facebook, and there are other likely errors based on how Facebook arrives at that number. However, it is an up-to-date observation that is somehow related to the true value. The estimates of potential reach are publicly available, and are available free of charge.[2] The estimates are provided at the aggregate level, in the form of counts by subgroup, so we never have access to any individual-level data.

Just as we obtained an observation of the number of Australians in Toronto, we could observe other migrant groups in Toronto, or any other city, state or country. In addition, it is possible to collect data from the advertising platform at various points over time, building up a time series of observations, to potentially study relative sizes of migrant groups, and changes in flows over time.

2.2.2 Previous Work

Over the past decade, the use of social media data in demographic research has gained an increasing amount of traction. Some of the earliest work in this area used geo-located data from emails (from service providers such as Yahoo!) and 'tweets' from Twitter to track small-scale mobility across cities (Ferrari et al. 2011; Noulas et al. 2011; Zagheni and Weber 2012). The use of geo-located data remains popular to study spatial variation in exposure to and reaction to major events such as natural disasters (MacEachren et al. 2011; Crooks et al. 2013; Martín et al. 2020). Methods of text analysis are commonly combined with spatial models to analyse Twitter and other text data to assess attitudes and beliefs, particularly in the political landscape (Barberá 2015; Halberstam and Knight 2016).

More specifically, social media advertising data has been used to study many aspects of demographic phenomena, including fertility, migration, and gender inequality. For example, Rampazzo et al. (2018) used Facebook advertising data to estimate the mean age at childbearing for both males and females. Garcia et al. (2018) used Facebook data to create an index of the internet gender divide in 217 countries, showing that this indicator encapsulated gender equality indices in education, health and economic opportunity. Ribeiro et al. (2020) develop a post-stratification framework for correcting demographic biases in Facebook data.

In the migration context, Facebook advertising data has been used to study both migrant stocks and flows over time at a subnational level. Zagheni et al. (2017) illustrated the strong

correlation between migrant stocks as reported in Facebook and the American Community Survey (ACS), a large and nationally-representative annual survey in the United States. Alexander et al. (2019) showed how these data could be used to derive reasonable estimates of the extent of out-migration from Puerto Rico following Hurricane Maria in 2017. Finally, Alexander et al. (2020) present a statistical forecasting method to 'nowcast' migrant stocks in the United States, combining information from both the ACS and Facebook. The method presented in that paper is a special case of the more general framework presented here.

2.3 DATA SOURCES

This section discusses practically how to collect demographic information from an advertising platform, with a particular focus on using Facebook's Marketing Application Programming Interface (API). A general discussion of potential 'gold-standard' data sources on migration statistics is also presented.

2.3.1 Obtaining Facebook Data Using an API

Figure 2.1 illustrated how estimates of potential reach appear within the Ads Manager. Theoretically, one could manually select the options within the platform to display the potential reach for each of the population sub-group of interest and note down the result in a data file. However, this would quickly become time consuming and prone to data entry error. For example, collecting information on migrant stocks by province in Canada by age and sex would lead to over 16,000 distinct subgroups.

However, Facebook's Advertising Manager has an associated API, called the Facebook Marketing API (Facebook 2020). An API is essentially a system that allows data from a website to be retrieved in an automatic, programmatic way. Thus, instead of manually selecting each subgroup of interest, we can make 'calls' to the API from a programming script (in a language such as R or Python) to retrieve the information of interest. This information can then be stored in a data frame and loaded at a later date for analysis. For more information and guidance on how to set up code to query Facebook's Marketing API in the context of collecting demographic data, see Gil-Clavel (2019).

There are several settings that relate to migration measurement in the Ads Manager. One of the main sources is from the 'Expats' variable, which indicates whether a person lived in a particular country in the past (but is currently residing somewhere else). There is currently information available for 89 different expat origin groups.[3] This variable can be used to get an estimate of migrant stocks in a particular place. If multiple waves of Facebook data are collected over time, this variable can also be used to infer flows (Alexander et al. 2019). Information about shorter-term migration behaviors can be collected through the 'Travel' variables (which include settings such as 'returned from travel 1 week ago'), or by selecting people who have recently traveled in a particular location.

There are also many demographic and socioeconomic variables of interest. Sex and ages between 13 and 65+ are available. There is highest education level, whether or not a person is currently studying and information on income and occupation. Data can be collected at a wide range of geographic levels, including country, state/province, or city level. In practice, subgroups that have small sizes will have potential reach estimates that are heavily rounded,

so the optimal granularity on which to collect information depends on the specific populations of interest.

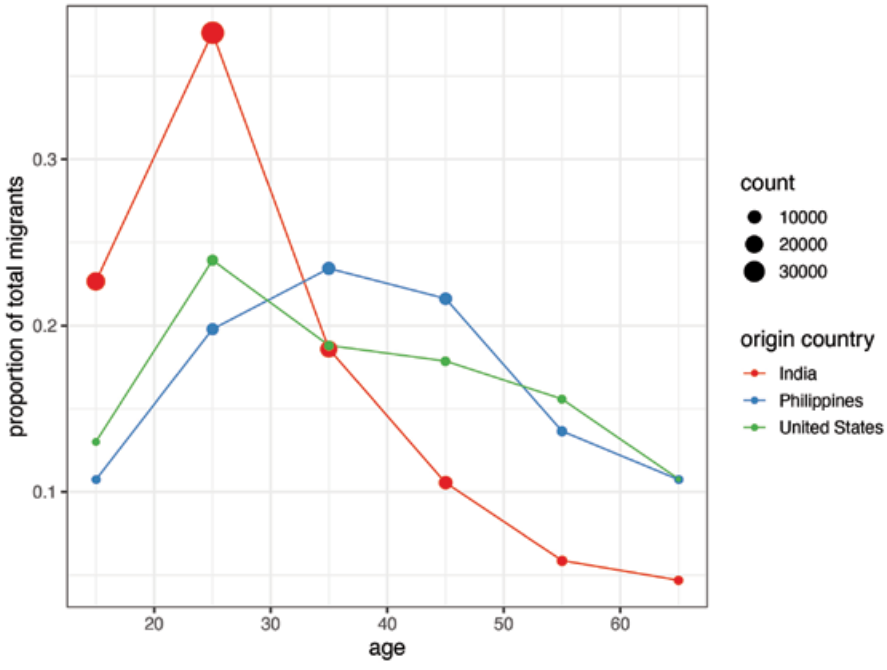

Figure 2.2 *Migrant stock estimates from Facebook's Advertising Platform in Ontario, Canada, 2019*

Figure 2.2 illustrates data collected on migrants from India, the Philippines and USA in Ontario in August, 2019. The proportion of total migrants in each ten-year age group is plotted, and the size of the dot indicates the size of the population reported. The three origin countries differ in both size and age distribution, particularly for India, which has a much higher peak at younger ages. The key for using these data to produce migrant estimates, however, is to relate these distributions to data from a more representative data source.

2.3.2 Representative 'Gold-Standard' Data Sources

In order to effectively use social media data to obtain representative estimates, it is important to be able to compare such data to high-quality representative data that is collected for the same or similar population. In essence, we would like to learn from good-quality data that is available for past time periods in order to use social media data for the current and future periods.

There are three main sources of good-quality data on migration statistics: censuses, nationally-representative surveys and migration data from government agencies. These data sources have complementary strengths and weaknesses. For example, censuses are

particularly good sources for obtaining a relatively complete picture of migrant stocks (that is, the number of migrants living in a location at a particular point in time). This information is derived from questions about a person's birthplace. Migrant flows (that is, the number of people moving in to or out of an area over a time period) may be derived from questions that have the form: where was your place of residence X years ago? Migrant flows could also be inferred by differences in stocks across adjacent censuses; however, given censuses are usually only run every five or ten years, these may be of little utility. Nationally-representative surveys can be useful to obtain estimates of either stocks or flows, depending on the design, frequency, and objectives of the survey. For example, surveys designed to supplement information from censuses may be good for estimating stocks, whereas labour force surveys may be useful to estimate flows. Migration data from central government agencies is particularly useful to get data for the size of migrant flows into a particular area over a certain time period. However, these types of data often lack the demographic breakdown of migrants (e.g., by age or sex) required for population composition analysis.

Note that for a particular country or population of interest, some or all of these data sources may be available. In general, most high-income countries run censuses every five or ten years, have government data on in-migration readily available, and will most likely have at least one large-scale nationally-representative survey that captures some information on the level of migration. For example, in Canada, broad-scale trends in migration can be gleaned from a combination of the Census, which is run every five years, and data from Immigration, Refugees and Citizenship Canada. In the United States, the decennial Census is supplemented by the annually-run ACS, and broad-level immigration statistics are available every year through Homeland Security. In contrast, many low-income countries lack readily available data on nationally-representative migration statistics, although some information is usually available through censuses.

2.4 A STATISTICAL FRAMEWORK TO COMBINE DATA SOURCES TO ESTIMATE MIGRATION OVER TIME

This section describes a general methodological framework to combine social media and traditional data sources on migration. It is assumed that observations of the migration indicator of interest are available from traditional data sources (censuses, surveys or government data) for time periods in the past, and that the goal is to produce up-to-date estimates, or 'nowcasts', of migrants for the current period and the short-term future.

2.4.1 Overview

Conceptually, estimates and projections of migration can be made by either projecting forward the historical time series of data, or adjusting recent social media data to be more representative. Ideally, we would combine both sources of information, objectively weighting the two options. A statistical framework to achieve such a goal has three main features:

1. A model or method for adjusting the social media data.
2. A time series model that allows historical data to be projected forward.
3. A 'data model' that allows observations from social media and gold-standard data to have different amounts and sources of error.

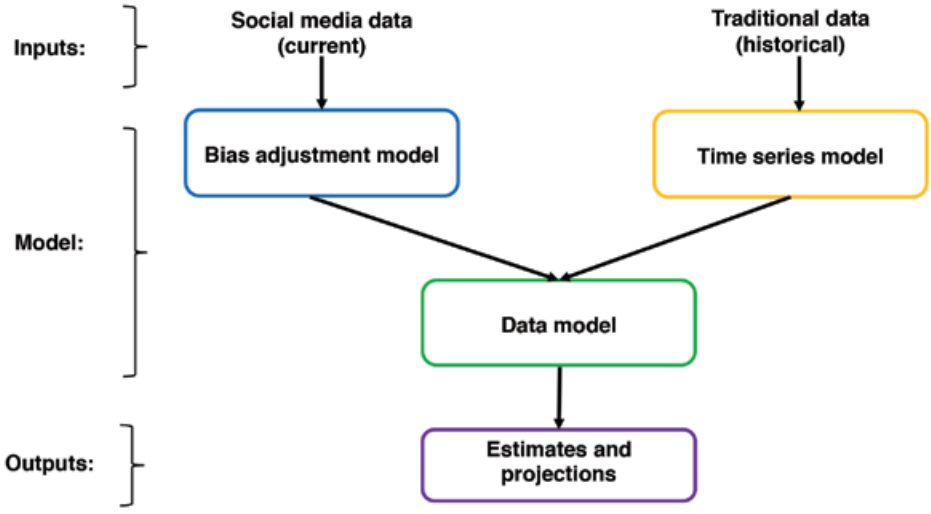

Figure 2.3 Framework to combine social media and traditional data sources

These three features are illustrated in Figure 2.3. The bias-adjustment model is the key to adjusting the social media data to be more representative. The time series model allows for patterns in migration in the past to be captured and projected forward. The data model is the key to bringing the two aspects together.

It is worth noting that while we collect data on migration *counts*, from a modelling perspective it is usually more practical to model migration *proportions*, that is, the proportion of each subgroup of interest that are migrants. Modelling proportions naturally constrains the upper bound of the count (to be at most equal to the total population). Counts can be easily obtained post estimation by multiplying the estimated proportions by population counts. In general, the logarithm of the proportion is modelled and then estimates are transformed back to the natural scale, to ensure positive values.

Let $p_{og}(s,t)$ be the observed proportion of migrants from origin country *o* in group *g*, from data source *s* and at time *t*. The group *g* may represent a subgroup of the population stratified by age and sex, for example. Denote $s = 1$ if the proportion is observed from social media and $s = 2$ if it is observed from a traditional data source. Assume we are interested in obtaining estimates for the expected value of the observed proportion, $\mu_{og}(t)$.

2.4.2 Bias-Adjustment Model

The goal of the bias-adjustment model is to account for the non-representativeness of the social media data. That is, we would like to adjust the $p_{og}(1,t)$ to account for biases. In practice, we expect the users of social media to be, on average, younger than the population of interest, and potentially may have a higher level of education and income than the general population. Previous research has also highlighted gender disparities in Facebook users worldwide (Gil-Clavel and Zagheni 2019). For example, Figure 2.4 shows the age distributions for

Mexican migrant stocks by state in the United States in 2016, as observed in Facebook advertising data and ACS data. The migrant stocks are plotted in terms of proportion of the total population in each age group. In general, the proportions are much lower at older ages in the Facebook data.

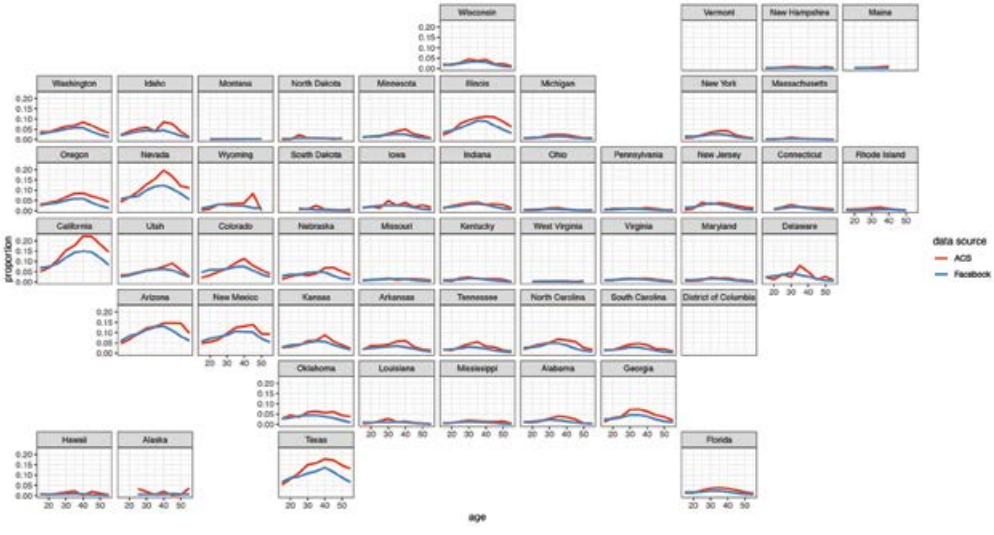

Figure 2.4 Age patterns in Facebook compared with ACS data by US state, 2016

While biases like those shown in Figure 2.4 are substantial, they are also fairly systematic across time and space. This means that biases lend themselves well to being modelled. There are several potential methods of bias adjustment. Perhaps the most common and straightforward approach is to use a form of post-stratification (for example, see Ribeiro et al. 2020). This involves calculating the proportion of migrants in each subgroup of interest and then comparing those to the same proportions from a representative data source (for example, census counts) to obtain a set of correction factors. For example, a correction factor for migrants from origin country *o* in group *g* would be

$$C_{og} = \frac{p_{og}(1,t)}{p_{og}^{GS}(t)}$$

where *GS* is the gold standard. Another alternative is to first model the proportion of migrants in each group using a multi-level regression model before calculating correction factors (Park et al. 2006). The multi-level regression step may produce more robust estimates particularly if subgroup counts are small. In this context, the $p_{og}(1,t)$ above would be replaced by its estimate, $\hat{p}_{og}(1,t)$. Alternatively, the relationship between the proportion of migrants in the social media and the gold-standard data can be modelled directly, accounting for the relationship to vary by demographic groups and geography. The resulting regression coefficients then act as correction factors in a similar way to post-stratification (Alexander et al. 2020). One of

the advantages of using regression methods over standard post-stratification correction factors is that there is a resulting estimate in the error in the bias-adjustment process, which can be input into the data model (see below). Regardless, the key commonality between these approaches is that we rely on there being an overlap of observations from social media data and a gold-standard data source, such that the extent of biases can be formally assessed.

The bias-adjustment model gives a method to produce adjusted proportions, from social media observations, of migrants from the origin country of interest in each group of interest, $p^*_{og}(1,t)$. If the social media data was collected today, then $p^*_{og}(1,t)$ could be thought of our 'best guess' of the current proportion of migrants from *o* in group *g*.

2.4.3 Time Series Model

The second component of the model is a time series model that captures temporal patterns in historical data and forms the basis for the data to be projected forward. The goal of the time series model is to express the expected value $\mu_{og}(t)$ as a function of what has occurred in the past, i.e. $\mu_{og}(t-1)$ and potentially previous time periods $(t-2), (t-3), \ldots, 1$. By doing so, we have a mechanism to project forward $\mu_{og}(t)$ in time.

There are many different options to model time series data, and the right choice is contextually dependent. A simple moving average approach may be sufficient, or Box-Jenkins approaches to ARIMA models may be suitable (Makridakis et al. 2008). For example, a first-order auto-regressive model would have the form

$$log\,\mu_{og}(t) = \phi log\,\mu_{og}(t-1) + \epsilon(t)$$

with $\phi \epsilon [-1, 1]$ and $\epsilon(t) \sim N(0, \sigma^2)$. Other temporal smoothing models that could be used include P-Spline models (Currie and Durban 2002) and Gaussian Processes (Wu and Wang 2018). Lee–Carter-based time series approaches are also particularly common in the demographic projection of age patterns (Lee and Carter 1992). Indeed, the case study described in this chapter uses a hierarchical Lee–Carter approach to model patterns over time. Regardless of the choice it allows the data we do have from the past to be projected forward.

2.4.4 Data Model

A data model is the key feature that allows the observations from social media and observations from historical data sources to be combined together in a systematic way. We observe proportions of interest at time *t* and from data source *s*, $p_{og}(s, t)$. These proportions are either observed from social media ($s = 1$), or from projecting the historical time series forward ($s = 2$). For the social media data source, we consider the bias-adjusted observations, $p^*_{og}(1,t)$. A data model assumes that the observed proportions are a draw from a distribution that is centered at the expected value of the proportion of interest, $\mu_{og}(t)$, with some variance that depends on the data source, i.e.

$$\tilde{p}_{og}(s,t) \sim N\left(\log \mu_{og}(t), \sigma_s^2\right)$$

where $s = 1$ if observed from social media and $s = 2$ if observed from historical data or projections, $\tilde{p}_{og}(1,t) = p^*_{og}(1,t)$ and $\tilde{p}_{og}(2,t) = p^*_{og}(2,t)$. This is equivalent to assuming that the observed proportions are equal to the expected value of the proportion of interest, plus some error, and the error depends on the data source, i.e.

$$\log \tilde{p}_{og}(s,t) = \log \mu_{og}(t) + \epsilon(s,t)$$

where $\epsilon \sim N\left(0, \sigma_s^2\right)$. The underlying μ_{og} is the quantity we are interested in estimating. If we have overlapping estimates from both data sources, that is $\tilde{p}_{og}(1,t)$ and $\tilde{p}_{og}(2,t)$ then we essentially have two observations of the same underlying quantity. However, the relative size of the variance term σ_s^2 allows the observations to be weighted in different ways. The bigger σ_s^2, the less weight the model places on $\tilde{p}_{og}(s,t)$.

In practice, there is usually at least some information about what σ_s^2 should be. For example, in terms of traditional data sources, if $\tilde{p}_{og}(2,t)$ comes from a survey, the size of σ_2^2 could be inferred from sampling error. In terms of social media, it seems reasonable to assume that σ_1^2 would be larger than σ_2^2. Potential sources of error include: sampling error, non-sampling error, and the error as a consequence of the bias-adjustment model. The use of a data model allows potentially overlapping observations from different sources to be both taken into consideration, and the relative weight of each source is dependent on the size of the error or variance assumed from that data source.

2.4.5 Summary of Model Framework

The three components discussed above can be summarized in a hierarchical framework as follows:

$$\log \tilde{p}_{og}(s,t) \sim N\left(\log \mu_{og}(t), \sigma_s^2\right)$$

$$\log \tilde{p}_{og}(s,t) = \begin{cases} p^*_{og}(s,t) \text{ if } s = 1 \\ P_{og}(s,t) \text{ if } s = 2 \end{cases}$$

$$\sigma_s^2 = \begin{cases} \sigma_1^2 + \ldots \text{ if } s = 1 \\ \sigma_2^2 + \ldots + \epsilon_{bias} \text{ if } s = 2 \end{cases}$$

$$\log \mu_{og}(t) = f\left(\log \mu_{og}(t-1), \ldots\right) (\text{time series model})$$

$$\log p^*_{og}(1,t) = f\left(logP^{GS}_{og}\right), P_{og}(1,t)) + \ldots + \epsilon_{bias} \left(\text{bias adjustment model}\right)$$

In terms of implementation, while it would be potentially possible to estimate each component separately in a sequential manner, this approach would fail to adequately propagate the model uncertainty associated with each step. However, models of this structure can be implemented in software that estimates Bayesian hierarchical models using Markov Chain Monte Carlo

(MCMC) algorithms, such as JAGS (Plummer 2003) or Stan (Carpenter et al. 2017). Example data and code to implement the model used in the case study below can be found at: https://github.com/MJAlexander/fb-migration-bayes.

2.5 CASE STUDY: 'NOWCASTS' OF MEXICAN MIGRANT STOCKS BY STATE IN THE US

This section discusses a specific case study, with chosen forms of the bias-adjustment model, time series model, and data model. The goal of the study is to produce 'nowcasts', that is estimates for the current year, for age-specific rates of migrants from Mexico currently living in each state in the United States. Mexican immigrants represent by far the largest migrant group in the United States, with relative proportions of the population exhibiting substantial variation across state (Borjas 2007). Data and methods for this study are explained in detail elsewhere (Alexander et al. 2020), but are outlined briefly below.

2.5.1 Data

The social media data used for this project were collected through Facebook's Marketing API, with collection waves beginning in December 2016. For each wave of data, state-level estimates of all Facebook users (by age, sex, and gender) were collected, as well as state-level estimates of the 89 expat groups mentioned above, including Mexico.

Historical data on migrant stocks were obtained from the ACS for each year between 2001 and 2017 using micro-data available through the Integrated Public Use Microdata (IPUMS) US project (Ruggles et al. 2000).

2.5.2 Model

2.5.2.1 Bias-adjustment model

As shown in Figure 2.4, proportions of Mexican migrants by age and state observed in the Facebook data appear to be biased in systematic ways. The bias-adjustment model has the following form:

$$\log p_{xts}^{GS} = \alpha_0 + \alpha_1 \log p_{xts}^{FB} + \beta \mathrm{X} + \varepsilon_{bias} \tag{2.1}$$

where p_{xts} is the proportion of Mexican migrants in age group x, time t and state s, $\mathbf{X}$ is a covariate matrix containing an indicator variable for each age group (15–19, 20–24, . . . , 50–54) and each of the 50 states plus Washington DC.

In order to obtain estimates of the coefficients α_0, α_1 and the vector of β's, the first wave of the Facebook data and the ACS data from 2016 were used as inputs for p_{xts}^{FB} and p_{xts}^{GS}. Once obtained, these coefficient estimates are then used to adjust subsequent waves of Facebook data.

2.5.2.2 Time series model

The time series model chosen for this context can be described as hierarchical Lee–Carter model. In essence, age-specific migration proportions are modelled as a combination of two 'principal components' plus an auto-correlated error term:

$$\log p_{xts} = \beta_{ts,1} Z_{x,1} + \beta_{ts,2} Z_{x,2} + \varepsilon_{xts} \tag{2.2}$$

where $Z_{x,1}$ and $Z_{x,2}$ are principal components that capture the main patterns of variation in migration proportions across age, and ε_{xts} is the auto-correlated error term. In addition, the coefficients $\beta_{ts,1}$ and $\beta_{ts,2}$ are modelled hierarchically such that information about migrant trends is shared across geography. For specific details refer to Alexander et al. (2020). This form of time series model is appropriate in this context because: (1) age-specific rates tend to show strong regularities which are well-captured by principal component methods, and (2) the data has a natural hierarchy in that we are modelling states within the United States.

2.5.2.3 Data model

The data model assumes

$$\log p_{xts} \sim N(\log \mu_{xts}, \sigma_p^2)$$

where σ_p^2 depends on the data source:

$$\sigma_p^2 = \begin{cases} \sigma_s^2, \text{if ACS} \\ \sigma_s^2 + \sigma_{bias}^2 + \sigma_{ns}^2, \text{if Facebook} \end{cases}$$

Here, σ_s^2 refers to sampling error, and is assumed to be present in both ACS and Facebook data. For the ACS data, sampling errors are calculated based on guidelines from the US Census Bureau (2020). For Facebook data, the sampling error is calculated assuming the binomial approximation to the Normal distribution.

For the Facebook data there are two additional error terms. σ_{bias}^2 refers to the error associated with our bias-adjustment model (equation 2.1). Additionally, a non-sampling error term σ_{ns}^2 is estimated within the model, which aims to capture additional uncertainty like variation in the way potential reach is estimated across waves.

Data and code to implement this model in R using JAGS software is available at: github.com/MJAlexander/fb-migration-bayes.

2.5.3 Results

Figure 2.5 shows the resulting estimated age distributions of Mexican migrants by US state in 2008 and 2018. In general, Mexican populations across the US are ageing, which most likely suggests a slowdown of immigration, as the existing migrant stocks age over time. The shift to the right of the age distributions is particularly noticeable on the west coast.

To illustrate the mechanics of the model more closely, Figure 2.6 shows the estimates and projections of migration proportions for 20–24 and 45–49-year-olds in California. The black

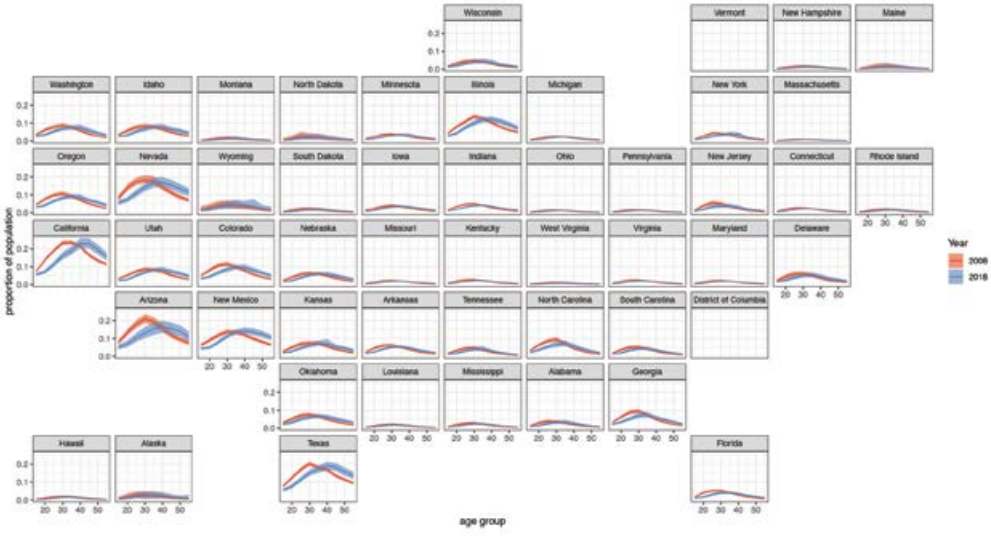

Figure 2.5 *Age distributions of Mexican migrants in 2008 and 2018*

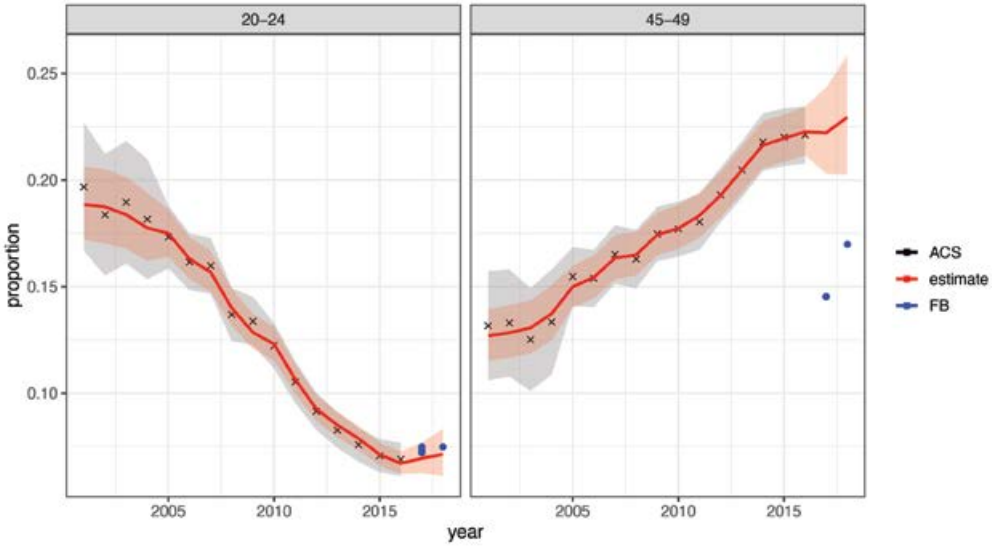

Figure 2.6 *Estimated and projected proportion of Mexican migrants in California, for 20–24 and 45–49 age groups*

crosses show the historical ACS data, and the black shaded area represents the ACS sampling error. The red line and associated shaded are the model estimates and 95 percent credible intervals, and the blue dots represent observations from Facebook data. For both age groups, the uncertainty around the estimates in the projected periods (2017 and 2018) increases – this

is a consequence of the data model, and the fact that the error around the Facebook estimates is larger than for the ACS. For the younger age group, the Facebook data are more closely in line with the ACS data, and projections are influenced by both historical trends and Facebook. In contrast, for the older age group, the Facebook data are less accurate, and the projections are more influenced by past trends in the ACS.

2.6 DISCUSSION

The large amount of data being produced online through social media websites presents exciting opportunities to leverage this information to improve estimates of human mobility and migration. Data sourced from the advertising platforms of websites such as Facebook, Twitter, and LinkedIn offers a rich set of demographic data that is freely available and updated essentially in real time. However, despite the clear strengths of such data, there are obvious drawbacks, most notably with issues of bias and non-representativeness.

In this chapter, the nature of advertising data from social media websites was discussed, with a particular focus on data from Facebook's Advertising Platform. A general statistical modelling framework to estimate migration patterns over time that incorporates social media advertising data was presented. The model has three main components: a bias-adjustment model; a time series model; and a data model. In particular, the data model allows the social media data to be combined with historical data from representative sources in a probabilistic way, by explicitly modelling different sources of potential error in the observations from each source. A specific example of this modelling framework was illustrated to estimate stocks of Mexican migrants by state in the US.

Future work in this area will focus on better understanding biases in the Facebook advertising data and how they change over time. In addition, there is potential for incorporating more sources of information, including from other social media platforms, such as Twitter. The case study presented here focuses on annual projections of migrants, but these types of data could also be used to track short-term mobility within cities and in response to exogenous shocks.

Social media data are most powerful when viewed as complementary data sources to existing, more traditional sources of demographic information, such as censuses and large-scale surveys. Combining the two types of information into the one modelling framework allows the strengths of both to be realized while also accounting and adjusting for known issues. While the application in this paper was to model migration over time, the general approach of recognizing the strength and weaknesses of different sources of data, and building analyses that incorporate both, has broader implications in methodological areas of demography and geography.

ACKNOWLEDGEMENTS

I would like to thank Michael Chong for his invaluable research assistance. Thank you also to Rohan Alexander and Edward Morgan for helpful feedback. A large part of my research in this area to date has been in collaboration with Kivan Polimis and Emilio Zagheni.

NOTES

1. Advertisements are screened before they are posted to ensure they meet Facebook's standards. As scrutiny into the spread of false information online has increased, the amount of vetting on Facebook advertisements has also increased.
2. Although it costs money to post an ad on Facebook, at the time of writing, information on the size of the potential audience is available before the ad is actually posted, thus can be extracted free of charge.
3. Algeria, Argentina, Australia, Austria, Bangladesh, Belgium, Brazil, Cameroon, Canada, Chile, China, Colombia, Congo, Democratic Republic of the, Côte d'Ivoire, 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, Jamaica, Japan, Jordan, Kenya, Korea, South, 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, Saudi Arabia, Senegal, Serbia, Sierra Leone, Singapore, Slovakia, Slovenia, South Africa, Spain, Sri Lanka, Sweden, Switzerland, Tanzania, Thailand, UAE, Uganda, United Kingdom, United States, Venezuela, Viet Nam, Zambia, and Zimbabwe.

REFERENCES

Alexander, Monica, Kivan Polimis, and Emilio Zagheni. 2020. Combining social media and survey data to nowcast migrant stocks in the United States. *arXiv Preprint arXiv:2003.02895*.

Alexander, Monica, Emilio Zagheni, and Kivan Polimis. 2019. The impact of Hurricane Maria on out-migration from Puerto Rico: Evidence from Facebook data. *Population and Development Review* 45 (3): 617–30.

Barberá, Pablo. 2015. Birds of the same feather tweet together: Bayesian ideal point estimation using Twitter data. *Political Analysis* 23 (1): 76–91.

Borjas, George J. 2007. *Mexican Immigration to the United States*. Chicago: University of Chicago Press.

Carpenter, Bob, Andrew Gelman, Matthew D. Hoffman, Daniel Lee, Ben Goodrich, Michael Betancourt, Marcus Brubaker, Jiqiang Guo, Peter Li, and Allen Riddell. 2017. Stan: A probabilistic programming language. *Journal of Statistical Software* 76 (1), 1–32.

Crooks, Andrew, Arie Croitoru, Anthony Stefanidis, and Jacek Radzikowski. 2013. Earthquake: Twitter as a distributed sensor system. *Transactions in GIS* 17 (1): 124–47.

Currie, Iain David, and M. Durban. 2002. Flexible smoothing with P-Splines: A unified approach. *Statistical Modelling* 2 (4): 333–49.

Facebook. 2020. Marketing API. https://developers.facebook.com/docs/marketing-apis.

Ferrari, Laura, Alberto Rosi, Marco Mamei, and Franco Zambonelli. 2011. Extracting urban patterns from location-based social networks. In *Proceedings of the 3rd Acm Sigspatial International Workshop on Location-Based Social Networks*, 9–16.

Garcia, David, Yonas Mitike Kassa, Angel Cuevas, Manuel Cebrian, Esteban Moro, Iyad Rahwan, and Ruben Cuevas. 2018. Analyzing gender inequality through large-scale Facebook advertising data. *Proceedings of the National Academy of Sciences* 115 (27): 6958–63.

Gil-Clavel, Sofia. 2019. Using the Facebook Marketing API. https://github.com/sofiag11/using_facebook_api.

Gil-Clavel, Sofia, and Emilio Zagheni. 2019. Demographic differentials in Facebook usage around the World. In *Proceedings of the International AAAI Conference on Web and Social Media*, 13 (01): 647–50.

Halberstam, Yosh, and Brian Knight. 2016. Homophily, group size, and the diffusion of political information in social networks: Evidence from Twitter. *Journal of Public Economics* 143: 73–88.

Internet World Stats. 2020. https://www.internetworldstats.com/facebook.htm.

Lee, Ronald D., and Lawrence R. Carter. 1992. Modeling and forecasting US mortality. *Journal of the American Statistical Association* 87 (419): 659–71.

MacEachren, Alan M., Anthony C. Robinson, Anuj Jaiswal, Scott Pezanowski, Alexander Savelyev, Justine Blanford, and Prasenjit Mitra. 2011. Geo-Twitter analytics: Applications in crisis management. In *25th International Cartographic Conference*, 3–8.

Makridakis, Spyros, Steven C. Wheelwright, and Rob J. Hyndman. 2008. *Forecasting Methods and Applications*. New York: John Wiley & Sons.

Martín, Yago, Susan L. Cutter, Zhenlong Li, Christopher T. Emrich, and Jerry T. Mitchell. 2020. Using geotagged tweets to track population movements to and from Puerto Rico after Hurricane Maria. *Population and Environment* 42: 1–24.

Noulas, Anastasios, Salvatore Scellato, Cecilia Mascolo, and Massimiliano Pontil. 2011. An empirical study of geographic user activity patterns in Foursquare. In *Fifth International AAAI Conference on Weblogs and Social Media*.

Park, David K., Andrew Gelman, and Joseph Bafumi. 2006. State-level opinions from national surveys: Poststratification using multilevel logistic regression. In J.E.Cohen (ed.) *Public Opinion in State Politics*, Stanford University Press, 209–28.

Plummer, Martyn. 2003. JAGS: A program for analysis of Bayesian graphical models using Gibbs sampling. In *Proceedings of the 3rd International Workshop on Distributed Statistical Computing*, 124 (125.10):1–10. Vienna, Austria.

Rampazzo, Francesco, Emilio Zagheni, Ingmar Weber, Maria Rita Testa, and Francesco Billari. 2018. Mater Certa Est, Pater Numquam: What can Facebook advertising data tell us about male fertility rates? In *Twelfth International AAAI Conference on Web and Social Media*.

Ribeiro, Filipe N., Fabrício Benevenuto, and Emilio Zagheni. 2020. How biased is the population of Facebook users? Comparing the demographics of Facebook users with census data to generate correction factors. *arXiv Preprint arXiv:2005.08065*.

Ruggles, Steven, Catherine A. Fitch, Patricia Kelly Hall, and Matthew Sobek. 2000. IPUMS–USA: Integrated public use microdata series for the United States. *Handbook of International Historical Microdata for Population Research.* Minneapolis: Minnesota Population Center, pp. 259–84.

US Census Bureau. 2020. PUMS Accuracy. Https://Www2.census.gov/Programs- Surveys/Acs/Tech_docs/Pums/Accuracy/.https://www2.census.gov/programs-surveys/ acs/tech_docs/pums/accuracy/.

Wu, Ruhao, and Bo Wang. 2018. Gaussian process regression method for forecasting of mortality rates. *Neurocomputing* 316: 232–39.

Zagheni, Emilio, and Ingmar Weber. 2012. You are where you e-mail: Using e-mail data to estimate international migration rates. In *Proceedings of the 4th Annual ACM Web Science Conference*, 348–51.

Zagheni, Emilio, Ingmar Weber, and Krishna Gummadi. 2017. Leveraging Facebook's advertising platform to monitor stocks of migrants. *Population and Development Review* 43 (4): 721–34.