MAX PLANCK INSTITUTE
FOR DEMOGRAPHIC RESEARCH

Konrad-Zuse-Strasse 1 · D-18057 Rostock · Germany · Tel +49 (0) 3 81 20 81 - 0 · Fax +49 (0) 3 81 20 81 - 202 · www.demogr.mpg.de

MPIDR Working Paper WP 2020-019 I April 2020
https://doi.org/10.4054/MPIDR-WP-2020-019

Modeling the bias of digital data: an approach to combining digital and survey data to estimate and predict migration trends

Yuan Hsiao
Lee Fiorio
Jonathan Wakefield
Emilio Zagheni I sekzagheni@demogr.mpg.de

Modeling the bias of digital data: an approach to combining digital and survey data to estimate and predict migration trends

Yuan Hsiao
Department of Sociology, University of Washington
Department of Statistics, University of Washington

Lee Fiorio
Department of Geography, University of Washington

Jonathan Wakefield
Department of Statistics, University of Washington
Department of Biostatistics, University of Washington

Emilio Zagheni
Max Planck Institute for Demographic Research

Abstract

Reliable and timely estimates of migration flows are needed to guide our policy decisions and to improve our understanding of migration processes. However, obtaining timely and fine-grained estimates remains an elusive goal. Digital data provide granular information on time and space based on large sample sizes, but because these samples are often not representative of the general population, the estimates obtained by analyzing these data are biased. We propose a generic method for combining digital and survey data for the purposes of migration estimation by accounting for the bias structure of digital data. Specifically, we show that if the bias has a structure over time and space that can be statistically modeled, we can combine different sources of data for the purposes of prediction. We illustrate our approach by combining geo-located Twitter data for more than two million users (2010-2016) with data from the American Community Survey (ACS) to estimate state-level emigration in the United States. We propose a joint model that draws from both ACS and Twitter data by modeling the spatial and temporal correlation structure of Twitter biases. We show that while Twitter-based estimates are upwardly biased, when these estimates are combined with ACS estimates, the resulting predictions of internal migration flows are more accurate than predictions based on ACS data only. Our method can be used to forecast future migration flows or to fill in missing time periods for which survey estimates are not available. Finally, our model is flexible and can be extended to incorporate multiple sources of data, such as Twitter data, cellphone records, administrative reports, and survey estimates.

Contents

1 Introduction

Despite the fundamental role it plays in a wide range of social, political and economic processes, migration is difficult to study (Massey et al., 1993; Clark, 1983). In many contexts, migration data – and especially measures of migration flows – are unavailable, unreliable, or not sufficiently timely. The largest sources of migration data, survey estimates and administrative data, are often limited in geographic and temporal scope, and are costly to produce.

Because of the limitations of conventional data, migration scholars have begun developing methods for using innovative sources of digital data in the study of migration (Zagheni et al., 2014; Hawelka et al., 2014; Jurdak et al., 2015; Fiorio et al., 2017; Hughes et al., 2016). Digital data come in many forms, such as Twitter records, cellphone data, or email IP addresses. While noisy and biased, digital data have some attractive characteristics, including their real-time availability and their potentially globe-spanning coverage (Malik et al., 2015). As digital data can track where people are, often down to the second, they provide individual-level records of mobility at a very fine granular level of time and space. When the right methods are applied, these data can be used to provide estimates of migration that are more timely than previous estimates, or for contexts where no migration estimates currently exist.

It is, however, important to bear in mind that because digital data are not drawn from a representative sample of the population, estimates based on digital data are inherently biased. The looming question that must be resolved before digital data can be rendered useful for generating such estimates is how to deal with the bias. Yet a shortcoming of the growing literature on migration and digital data has been a lack of sophistication in addressing the structure of the bias inherent in digital data. Many of these studies have assumed that the relationship between social media estimates and survey estimates is constant. In this paper, we develop a more flexible approach to determining whether a model that takes into account the spatial and temporal structure of the bias in digital data estimates improves the model's overall predictive power. We argue that digital data are not a replacement for traditional survey data, but instead represent a complementary source of information. By combining digital and traditional data sources, we can investigate the structure of the bias, and can thus combine the advantages of the representativeness of conventional surveys with the timeliness of digital data.

As such, this paper asks three interrelated questions: (Q1) What is the size of the bias in digital data, and how does it change over space and time? (Q2) How can we formulate statistical models that capture the bias relationship between estimates from digital data and the true population processes? (Q3) How can we combine estimates from digital data and official statistics to improve the accuracy of predictions?

We approach these questions by developing a model that decomposes the spatial and temporal processes of the bias. This enables us to show that the biases can be modeled statistically and combined with survey data to improve the accuracy of our predictions. The model is generic, and can incorporate

multiple sources of data that meet some simple requirements. More specifically, the model incorporates two categories of information. (1) First, the model includes data from "unbiased" sources, such as representative surveys. While some of these surveys have relatively small samples, are not be very timely, or cover only certain points in time, we can apply standard tools of statistical inference to the data they provide. (2) Second, the model includes data from potentially biased sources, such as data from Twitter, geo-located website log-ins, or cellphone records. While these data are typically not representative, they are more timely than data from unbiased sources. In addition, these data often provide more information at finer levels of geographic and temporal granularity.

We illustrate our method using state-level emigration rates in the US for the years 2010-2016. We combine data from the geo-located tweets of more than two million Twitter users (2010-2016) with data from the American Community Survey (ACS). First, we use the Twitter data to obtain estimates of state-level emigration rates. We then inquire how the bias is the distributed across time and space. Third, we select the optimal model that captures the structure of the bias. Finally, we utilize this optimal model to combine Twitter and ACS data and improve the accuracy of our predictions.

The results suggest that the raw estimates from Twitter tend to overestimate emigration rates, and that, when taken alone, these estimates are not useful for making predictions. However, we also find that both the emigration process and the bias follow structured space-time interaction processes. We further show that we can use these estimates to improve our forecasts or predict emigration rates for years that are missing in the official statistics.

Our approach may be especially relevant in two scenarios. First, we can use this method to generate more accurate forecasts of future emigration rates. Whereas official statistics are often published with a time lag, the immediate availability of the digital data allows us to predict emigration rates without having to wait for the official statistics to appear. For instance, we can forecast emigration rates for the year 2019 before the official statistics become available.

Second, we can better predict emigration rates for years that are missing in the official statistics. Although the ACS provides annual estimates of emigration for the US, estimates for certain years are often missing for the many countries that do not conduct such annual surveys. Combining data from these traditional sources with data from social media can help researchers close such gaps, and thus improve their understanding of migration trends over time.

This paper contributes a methodological approach that we illustrate using examples that focus on migration research. However, the scope of the potential applications of this approach is broader. As our lives become increasingly digitalized, social scientists have access to more and and more information about population and social processes. A key challenge that researchers across the social science disciplines face is how scientifically rigorous inferences can be made using data drawn from a combination of representative and non-representative sources. This article contributes to this line of research by showing how our method can be used to combine unbiased and biased sources of data for the purposes of population science research. These sources can include cellphone

data, social media data, administrative records, and various kinds of geographical and temporal information.

2 Background on a demographic challenge: Quantifying migration

Migration data come in one of two forms: stock data and flow data. Stock data, or counts of migrants and non-migrants at a particular time, are more intuitive and easier to quantify than flow data. If we have information about where an individual currently resides and where that individual was born, whether s/he has migrated over his/her lifetime can be inferred. If, for example, someone is living in a US state that is not the state where s/he was born, the person is considered an interstate migrant. A major problem with these stock data is, however, that they do not provide direct or timely information on current migration processes. Stocks will change over time as people migrate. However, an individual may move to or out of a given place more than once. As these patterns can be difficult to track, the overall level of migration may be underestimated. Thus, information on migration stocks is not sufficient when the goal is to infer migration trends over time or to capture mobility patterns with different time spans, such as repeat migration or short-term migration.

Bilateral flow data – i.e., estimates of flows from all origins to all destinations – represent a much richer form of migration data, as they are better able to capture migration patterns over time than stock data. Having access to such data is crucial for measuring and understanding migration systems. When people migrate, they form a link between their origin and their destination, transforming both places in the process. Migration stocks are changed by migration flows. However, estimating flows is challenging, because tracking these flows requires researchers to either observe a panel of individuals multiple times as they relocate (or remain in place), or to ask individuals to retrospectively report the places where they have lived. In both cases, survey data or administrative data must be collected and analyzed.

As this discussion of stock data and flow data makes clear, producing timely estimates of migration patterns is difficult. This lack of timely information can hinder the implementation of important research agendas and policy initiatives. For example, while the American Community Survey (ACS) produces high-quality annual estimates of interstate migration trends in the United States, these estimates are often publicly released with a time lag of more than a year. Given that detailed knowledge of interstate migration patterns is essential for understanding a wide range of phenomena, including urban growth and development (Greenwood, 1981), housing dynamics (Clark et al., 2000), and the labor market (Moretti, 2013; Molloy et al., 2017, 2011), this delay is problematic. In addition, as climate-related disasters become more numerous and more severe, being able to produce immediate estimates of climate-related migration response becomes increasingly important.

Given their high degree of granularity of time and space, digital data may prove useful for generating more timely migration estimates. A growing number of population scientists are seizing this opportunity to understand human mobility (Blumenstock, 2012; Zagheni et al., 2014; Jurdak et al., 2015; Fiorio et al., 2017; Hughes et al., 2016; Zagheni and Weber, 2012). The main reason why digital data represent a rich source for studying migration is that users share their location (explicitly or implicitly) each time they interact with a digital platform. By tracking the locations of users over time, the mobility of individuals can be estimated. These individual estimates can, in turn, be aggregated and converted into estimates of migration flows. For instance, when Twitter users post their tweets with "geo-location," each tweet shows the location where the tweet was posted. The advantage of these digital data is that they offer a high level of granularity of time and space, as the information on each tweet is often provided by the second (for time) and the precise geographic coordinates of the user (for space) when it was posted. Thus, scholars are able to obtain much richer information on users from these highly granular data than they are from survey data. For instance, Blumenstock (2012) drew on cellphone records to study internal migration patterns in Rwanda. As each cellphone call is associated with a cellphone tower, it is possible to infer the area where a given caller was located from the location of the tower that transmitted the call. The temporal and spatial granularity of cellphone records therefore enable researchers to identify patterns of temporary and circular migration that are not easily captured by government surveys. Similarly, Zagheni and Weber (2012) used email data to estimate international migration rates. As each email is associated with an IP address, the authors were able to use the IP addresses to identify the location of each email sender and the time when each email was sent. This information was, in turn, analyzed to infer migration flows between countries. Moreover, Fiorio et al. (2017) used Twitter data to investigate the relationship between migration patterns at different levels of granularity. Many Twitter posts have "geo-tags" that reveal when and where the tweet was sent. These data were aggregated to investigate migration patterns with different time spans (i.e., short-term and long-term migration).

However, a key problem researchers face when basing their estimates on digital data is that because users of digital platforms are not random samples of the population, there is inherent bias in the estimates. For instance, Twitter users tend to be younger than the general population (Mislove et al., 2011), and younger people tend to be more mobile. If we took estimates of mobility based on data from Twitter at face value, we would likely overestimate the migration rates.

Fortunately, digital data are not the only sources of data. In addition to digital data, governments and large organizations collect representative survey data at larger time intervals. Although survey data are not as timely and or as geographically granular as digital data, they provide relatively unbiased estimates of migration flows. Thus, one reasonable approach to producing timely estimates might be to combine social media data with official statistics. By combining different sources of data, we may be able to take advantage of both

the representativeness of official statistics and the timeliness of digital data.

Based on these considerations, we can reframe the question of how the relationship between these biased and unbiased forms of data should be modeled. If the relationship between estimates from digital data and estimates from survey data follows statistical patterns, we can draw information from both sources to aid in estimation. As a simplified example, if estimates from digital data are always higher than the population rates by a constant factor, then we can rescale and obtain reasonable estimates from social media data. In other words, the problem is not whether digital data are biased, but rather how the structure of the bias should be modeled. By accounting for this bias, digital data can nonetheless be used to compensate for the deficiencies of survey estimates, and thus to produce predictions of higher quality.

3 A joint model that combines survey and digital data

We contend that digital data are not a replacement for traditional survey data, but should instead be seen as a complementary source. To combine one type of data with the other, we introduce a "joint" model that incorporates both unbiased and biased data sources.

3.1 Intuition of the model

We illustrate the method using the empirical case of annual state-level emigration rates in the US. The biased data we draw on are Twitter data, which provide highly granular spatial and temporal information. We also draw on ACS data, which are representative, but are not as timely or as granular as the Twitter data. We propose using a "joint-modeling" approach in order to gain a better understanding of the data-generating mechanisms for both the ACS and the Twitter data.

The intuition of the model is that there is a common process for the "true emigration rates" in the population, and we have two sources of data that measure this process. The ACS data estimate this process with measurement error and little bias, while the Twitter data estimate this process with both measurement error and bias. In other words, the ACS and the Twitter estimates have a shared true emigration process that is unbiased. However, for the Twitter estimates, there is an additional bias term that needs to be modeled.

This decomposition of a true process and a bias term leads us to ask two key questions: (1) How should we model the true process? (2) How should we model the bias?

To answer these questions, we draw from a well-established literature on space-time models in population estimates (Knorr-Held, 2000; Mercer et al., 2015; Waller and Gotway, 2004; Wakefield et al., 2019). As the name suggests, this collection of models incorporate information on time and space. Specifically, these models take into account that observations within the same space (or

time) are likely to be similar to one another. For instance, if the demographic composition of the state of Connecticut is more or less stable, we would expect migration rates for Connecticut to be correlated over time. These models also account for the possibility that observations adjacent in space (or in time) are correlated. For instance, we might expect observations in adjacent years to be correlated with one another. The space-time models decompose population processes into spatial and temporal processes with the aim of leveraging these effects to improve the accuracy of our predictions.

Following Mercer et al. (2015), we model the true process as a spatial ICAR process, a temporal Random Walk process, a combination of independent spatial and temporal processes, or a process with space-time interactions. These classes of statistical models capture different potential dependencies between space and time (see the online appendix for detailed explanations). ICAR models assume that areas that are adjacent to one another are correlated and have similar levels of emigration probabilities. Random Walk models assume that observations that are adjacent in time are correlated with one another. Depending on the specification, these classes of models capture space only, time only, or a combination of the two. In this paper, we test these possibilities, and select the model that has the best fit.

Similarly, we specify the bias as a spatial ICAR process, a temporal Random Walk process, a combination of independent spatial and temporal processes, or a process with space-time interactions. Again, we test these possibilities and select the appropriate model.

After selecting the appropriate model (for both the true process and the bias process), we can leverage this joint model to generate predictions when we do not have official statistics, but we do have digital data.

3.2 Mathematical formulation of the model

We define the emigration rate for a particular state in a particular year as the probability that a person who is living in the state will out-migrate in that year. Thus, for the purposes of this paper, we use the terms migration rate and migration probability interchangeably.

Since emigration rates are probabilities that lie between zero and one, we model the *logit* of the emigration rates. Define $p_{s,t}$ as the emigration probability for state s, time t, then $Y_{s,t}$ is the logit of $\hat{p}_{s,t}$ (the estimate of $p_{s,t}$), defined as $Y_{s,t} = \log(\frac{p_{s,t}}{(1-p_{s,t})})$

Formally, denote:

$Y_{s,t}^{ACS}$ as the logit of the migration estimates from ACS for state s, year t

$Y_{s,t}^{TWIT}$ as the logit of the migration estimates from Twitter for state s, year t

Then, since logits are continuous variables bounded by $(-\infty, \infty)$, the true process for both the ACS data and the Twitter data follows a normal distribution. We denote this true process as $\mu_{s,t}$ For Twitter data, on top of $\mu_{s,t}$ we

include $B_{s,t}$ as the bias term for the Twitter estimates (we assume that ACS estimates are the gold standard for comparison):

$$Y_{s,t}^{ACS} \sim N(\mu_{s,t}, V_{s,t}^{ACS})$$
$$Y_{s,t}^{TWIT} \sim N(\mu_{s,t} + B_{s,t}, V_{s,t}^{TWIT})$$

where $V_{s,t}^{ACS}$ and $V_{s,t}^{TWIT}$ are the estimated variances that acknowledge the study design.

Notice the common mean component μ_{st} in both the ACS and the Twitter estimates. Because of this common mean component, we model the two processes *jointly*. That is:

$$\begin{matrix} Y_{s,t}^{ACS} \\ Y_{s,t}^{TWIT} \end{matrix} \sim N\left[\begin{pmatrix} \mu_{s,t} \\ \mu_{s,t} + B_{st} \end{pmatrix}, \begin{pmatrix} V_{s,t}^{ACS} & 0 \\ 0 & V_{s,t}^{TWIT} \end{pmatrix}\right]$$

We assume that the covariance terms are zero, as the measurement errors of the ACS and the Twitter data are independent because they are drawn from independent samples. Thus, the measurement errors should be unrelated.

The question then becomes how $\mu_{s,t}$ and $B_{s,t}$ should be modeled. For the true process $\mu_{s,t}$, we use extensions of the Fay-Herriot model (Fay and Herriot, 1979) (see also (Mercer et al., 2015)). These models assume that each area (e.g., a state) or each time point (e.g., a particular year) have area/time-specific random effects. Following this approach, there are multiple ways to specify $\mu_{s,t}$ as including such random effects:

An ICAR (BYM2) spatial process: $\mu_{s,t} = \mu + \theta_s + \phi_s$

A Random Walk 2 and IID temporal process: $\mu_{s,t} = \mu + \alpha_t + \gamma_t$

A space-time main effect only process: $\mu_{s,t} = \mu + \theta_s + \phi_s + \alpha_t + \gamma_t$

A space-time interaction process: $\mu_{s,t} = \mu + \theta_s + \phi_s + \alpha_t + \gamma_t + \delta_{st}$

Where μ is an overall mean, θ_s is a spatial intrinsic conditional autoregressive process (ICAR), ϕ_s is a random IID intercept for each state, α_t is a random walk of order 2 process (RW2), γ_t is a random IID intercept for each year, and δ_{st} is a structured interaction between the ICAR process and the RW2 process. In short, θ_s and ϕ_x capture the spatial random effects, α_t and γ_t capture the temporal random effects, and δ_{st} captures the dependency between the spatial and temporal random effects. In other words, these terms capture the within and adjacent effects of space and time (see the online appendix for a detailed explanation of each of the model components).

This statistical model shares information between contiguous neighbors and close time periods. The information is shared by specifying probability distributions that penalize contributions to the mean of $Y_{s,t}$ that are very different in areas that are geographically and/or temporally close. Hence, similarity in estimates is encouraged.

In line with the approach we used to model the common mean process $\mu_{s,t}$, we can model $B_{s,t}$ as:

An ICAR spatial process: $B_{s,t} = \mu + \theta_s + \phi_s$

A Random Walk 2 temporal process: $B_{s,t} = \mu + \alpha_t + \gamma_t$

A space-time independent process: $B_{s,t} = \mu + \theta_s + \phi_s + \alpha_t + \gamma_t$

A space-time interaction process: $B_{s,t} = \mu + \theta_s + \phi_s + \alpha_t + \gamma_t + \delta_{st}$

In other words, the true process may include space and time effects. However, the bias may also vary over space and time, as the demographic composition of different states and years can vary considerably.

3.3 Model selection and model evaluation

We fit these models using the *INLA* package in the statistical software *R* (Rue et al., 2009). Since the ACS data are representative, our analytical strategy is to first use only the ACS data to select the best model for s,t. We use three model selection criteria: log-CPO (higher indicates a better fit), DIC (lower indicates a better fit), and WAIC (lower indicates a better fit) (also see Krainski et al. (2018)). In brief, the CPO captures the sum of the probability density of predicting each data point from the rest of the data points via the model, which is, in spirit, similar to a leave-one-out cross validation. The DIC and WAIC are generalizations of the Akaike information criterion (AIC), which rewards a better fit to the data, but penalizes model complexity. See the online appendix for a detailed explanation of each of these evaluation criteria.

From these criteria we select a "best" model for the common mean process ($\mu_{s,t}$) from a model that uses only ACS data. Then following the specification of the common mean process, we estimate the bias structure ($B_{s,t}$) with a joint model that uses both ACS and Twitter data.

We evaluate the performance of the joint model over an "ACS-only" model using a "Leave-One-Year-Out" cross validation. For example, suppose that estimates from the ACS on year t are the prediction target. We remove the observations from the ACS for that year and compare the model predictions from the joint model with those from the "ACS-only model." Our goal is to recover the target estimates from the model predictions, with the difference being the prediction error of the model. If the prediction error is lower for the joint model than for the "ACS-only model," we can conclude that the Twitter information improves the accuracy of our emigration predictions. Our validation strategy is motivated by our empirical concerns. Because survey estimates are often not timely enough or have missing years, predictions are frequently made for all observations in a given year (or multiple years). Thus, our model is particularly useful when the goal is to forecast the future before survey data are available, or if we are seeking to better understand migration trends by filling in years for which survey estimates are missing.

4 Data

The data in this paper come from two main sources: (1) Twitter data from the 1% historical archive of Twitter; and (2) offical statistics from the American Community Survey (ACS).

4.1 Twitter data

The Twitter data used in this paper were assembled from a historical archive of a 1% sample of the Twitter stream (Archive.org, 2016) between January 1, 2010 and December 31, 2016. From this sample, we selected all tweets containing a geo-tag (i.e., information on latitude and longitude) that occurred within the United States, and use this information to place each tweet in one of 50 US states. We define migration as a change in residency between US states. Our initial selection results in a sample of 554,229,541 tweets from 2,226,467 users.

While we have already discussed some of the reasons why Twitter data may be biased, we should also point out that these biases are not randomly distributed across time and space. First, the user base is not consistent over the period. While the mean number of geo-tagged tweets associated with each user is 267, these tweets do not necessarily occur evenly over time. On average, a user appears in a 24-week spread over a period of about a year. This means, for example, that the 2011-2012 interstate migration flows are estimated from a different set of users than the 2012-2013 flows. Thus, including time-specific effects when capturing the bias becomes imperative. Second, geo-tagged tweets will pick up all kinds of movement – e.g., holiday travel, travel for business, and short-term mobility for education or family-related reasons – and not just the semi-permanent relocation associated with migration. Thus, different people will migrate for different reasons to and from different places. We can therefore expect to observe spatial patterns in the degree of bias in the Twitter data. For example, the popularity of destinations like Las Vegas and Miami might result in the overestimation of migration to Nevada and Florida. Third, in early 2015, Twitter made a top-down change to the kind of geographic information captured in geo-tags. Instead of precise latitude and longitude information being collected more or less passively with each tweet, users had to opt to share a specific location tag associated with a place. This change resulted in fewer geo-tagged tweets overall, and likely increased the amount of travel-related information captured in the latter years.

For these reasons, we expect migration estimates derived from Twitter data to be biased with respect to time and space. We argue, however, that we can model the structure of these biases to better isolate the migration signal from the Twitter data.

4.2 American Community Survey (ACS) data

The ACS asks respondents to name the state where they currently live and the state/country where they were living one year before the interview. From the

information on current and previous residence, the ACS produces estimates of state-to-state migration flows on an annual basis. In this paper, we draw from ACS estimates for the years 2010-2016.

By aggregating the migration flows, we can obtain a point estimate for the emigration rate for each state (e.g., the number of migrants from Arizona is the sum of the migrants from Arizona → Florida, Arizona → Kentucky, etc.). We also compute the standard errors that acknowledge the survey design using the replicate weights in the ACS.

5 A diagnosis on the bias of Twitter estimates

5.1 Obtaining estimates from Twitter data

Before starting the analysis, we need to transform the raw Twitter data into emigration estimates for each state-year. To do this, we follow Zagheni et al. (2014) and use the following procedure:

1. For each geo-tweet, we use the latitude and longitude to identify the US state from where the tweet was posted.

2. For each user and for each year, we calculate the number of tweets posted in each US state.

3. For every two-year period (e.g., 2010 and 2011), we discard users for whom the number of tweets posted in the modal state is less than three in at least one of the two years, or for whom the ratio between the number of tweets posted in the modal state and the number of tweets posted in the second modal state is less than three. For example, if a user posted 15 tweets in Washington state and eight tweets in Ohio in 2010, and 20 tweets in Washington state and three tweets in Ohio in 2011, the user would be discarded because the ratio in 2010 is less than three. This threshold is somewhat arbitrary, but was chosen based on the underlying goal of achieving a compromise between ensuring that the state of residence is identified accurately, while also maintaining a large sample of tweets.

4. For every two-year period, and for users who meet the threshold criteria described above, if the modal state in the first year is different from the modal state in the second year, the user is classified as a migrant. If the modal state is the same, the user is classified as a non-migrant. For the purposes of this paper, we consider internal migration only. However, this approach could also be applied to international migration.

5. For every two-year period, t and $(t+1)$, we calculate the estimated migration probability for the first year (defined as $\hat{p_t}$) in the state as $N_{Migrants}/N_{Users}$.

5.2 Assessing bias of Twitter estimates using the ACS

5.2.1 Overview of bias diagnostics

After obtaining the raw estimates from Twitter, it would be helpful to compare these estimates, which are drawn from a non-representative sample, to estimates from the ACS, which are based on a sample that is representative of the US population.

We assess the degree of bias using a simple bias ratio formula. More specifically, we define BR as the bias ratio. Let $\hat{p}_{s,t}$ be the estimate of the emigration probability for state s, year t. Let $\hat{p}_{s,t}^{TWIT}$ be the raw estimates from Twitter and $\hat{p}_{s,t}^{ACS}$ be the official estimates from ACS.

Then:

$$BR_{s,t} = \frac{\hat{p}_{s,t}^{TWIT}}{\hat{p}_{s,t}^{ACS}}$$

The bias ratio is used to assess the relative discrepancy between the Twitter estimates and the ACS estimates. If the Twitter estimates were perfectly in line with the the ACS estimates, the bias ratio would be one. Values larger than one indicate that the raw Twitter estimates overestimate the emigration rate (e.g., a bias ratio of 1.15 indicates an over-estimation of 15%), whereas values smaller than one indicate that the Twitter estimates underestimate the emigration rate (e.g., a bias ratio of 0.78 indicates an underestimation of 22%). The goal of this section is to assess the bias structure of the Twitter estimates. Simply knowing that Twitter estimates are biased does not help us improve the accuracy of our estimates and predictions of emigration trends; instead, we need to leverage the potential spatial/temporal variation and spatial/temporal correlation of the bias in our statistical estimation model.

5.2.2 A visual diagnosis of bias over space and time

Our first step is to visually diagnose the bias for each state over time. Figure 1 plots the bias ratios across states for each year. Each subplot shows the map of the bias ratios for a given year. The colors indicate the size of the bias. Red colors are associated with bias ratios that are larger than one; the darker the red color, the larger the bias. Conversely, blue colors indicate bias ratios that are smaller than one. Gray colors indicate that data are missing in the state for the year, which occurs in 2010 only.

As Figure 1 shows, there are red colors, but no blue colors. This pattern indicates that the Twitter estimates of internal migration rates are always higher the the ACS estimates. This is to be expected, given that, on average, Twitter users are younger than the general population (Mislove et al., 2011) and have higher mobility rates.

Furthermore, although the Twitter estimates tend to be upwardly biased, this bias is not randomly distributed. There appears to be spatial variation in the distribution of bias, as states vary consistently in their degree of bias. For

instance, it appears that states like Alaska and Nevada have much larger biases over time (i.e., darker reds for each subplot), while states like Alabama have smaller biases over the period considered.

A potential diagnosis of the spatial correlation is that there are similar degrees of bias in the New England area and similar degrees of bias in the West Coast region. While these patterns are not obvious, we can see that these areas tend to have clusters of red in states that share neighbors.

[Figure 1 about here]

An alternative visual diagnosis is based on an examination of how these biases vary across states. We plot the bias ratios over the years in Figure 2. Each subplot represents a state, which is geographically mapped to its relative position in the United States. Within each subplot, the lines represent the bias ratios over the years for each state.

We can see that the relative positions of the lines are generally consistent over the years. This finding again suggests that there is spatial variation in the degree of bias, as states with higher overall levels of bias tend to have higher levels of bias across the years. For instance, Alaska and the District of Columbia have higher bias ratios, while Alabama has a very low bias ratio.

Additionally, we find evidence of temporal variation. The uptick that can be observed in many subplots leads us to conclude that, in general, the level of bias is higher in 2015 and is lower in 2010.

Finally, we see that most of the lines are relatively smooth with only a few bumps, which suggests that there is a temporal correlation, whereby the bias of the current year is related to the bias in the previous year.

[Figure 2 about here]

We observed that the Twitter estimates are always higher than the ACS estimates, which indicates that the level of bias is overestimated if we use the raw estimates from Twitter. Nonetheless, the diagnostics also show that there is spatial and temporal dependency in the biases, which suggests that by capturing this spatio-temporal structure of the bias, we may be able to combine the Twitter and ACS estimates to predict emigration rates. Our visual diagnoses give us some initial confidence that the use of a space-time model can help us capture the bias statistically, and, in turn, improve the accuracy of our predictions.

5.3 Selecting the best model for the true process and the bias process

Recall that our modeling strategy is to model the bias structure in addition to the true emigration process. Thus, before we model the structure of the bias, we first need to model the structure of the true emigration process. To accurately estimate this true emigration process, we construct a set of models

from the unbiased ACS data to prevent a contamination of biases (i.e., "ACS-only models"). We consider different modeling options, including a spatial-only model, a temporal-only model, a space-time independent model, and a space-time interaction model.

We compare the fit statistics for these "ACS-only models" in order to select the best model for the true emigration process (i.e., $\mu_{s,t}$). As Table 1 shows, the space-time interaction model has the highest log-CPO, the lowest DIC, and the lowest WAIC; which suggests that it is the model with the best fit for the true emigration process.

[Table 1 about here]

[Table 2 about here]

Next, we take on the task of statistically modeling the bias. After specifying the space-time interaction structure for the true emigration process, we explore different models that incorporate the bias structure (i.e, $B_{s,t}$). Again we consider a space-only model, a time-only model, a space-time independent model, and a space-time interaction model.

As Table 2 shows, the space-time interaction joint model best captures the bias structure, as it has the highest log-CPO, the lowest DIC, and the lowest WAIC. From the comparison statistics, we select the joint model that specifies the true process as a space-time interaction process, with the bias structure also having a space-time interaction process. In the next section, we evaluate whether this joint model outperforms the best "ACS-only model" in terms of forecasting and prediction.

5.4 Results on forecasting/predicting emigration rates

5.4.1 Overview

To determine whether a joint model that utilizes Twitter data outperforms an ACS-only model, we need to compare the prediction error from both models. As we mentioned in Section 3.3, we use a "Leave-One-Year-Out" validation to test our ability to recover missing years from the model predictions. This means that we have one year for which official statistics are not available, and can only be predicted from existing models.

To assess the prediction error, we use the Root Mean Squared Error (RMSE, a measure of absolute prediction error) and the Mean Absolute Prediction Error (MAPE, a measure of relative prediction error) to evaluate the performance of the models.

Regarding RMSE, let $p_{s,t}^{ACS}$ be the emigration rate for the ACS target year (i.e., the year with ACS data removed), and $\hat{p}_{s,t}$ be the predicted emigration rate from the model. Then the RMSE for year t would be:

$RMSE_t = \sqrt{\sum_{s=1}^{51}(\hat{p}_{s,t} - p_{s,t}^{ACS})^2}$

A lower RMSE indicates a more accurate prediction. We calculate the RMSE for each year to see how well the models perform.

Conversely, since emigration rates tend to be low (often around 5%), the MAPE measures the percentage of the error compared to the target. Specifically, the MAPE for year t would be:

$MAPE_t = \sum_{s=1}^{51} |(\hat{p}_{s,t} - p_{s,t}^{ACS})/p_{s,t}^{ACS}|$

A lower MAPE indicates a more accurate prediction. We calculate the MAPE for each year to see how well the models perform.

5.4.2 Results on prediction error

We compare the RMSEs for both models for each year in Figure 3. In the plot, the horizontal axis represents the year for which we calculate the RMSE, and the vertical axis is the value of the RMSE. The red line represents the RMSE for the ACS-only model, while the blue line represents the RMSE for the joint model that utilizes the Twitter data. It is clear that for every year, the joint model has a lower RMSE than the ACS-only model.

[Figure 3 about here]

We then compare the MAPEs for the two models in Figure 4. Again, we see that for every year, the joint model outperforms the "ACS-only model."

[Figure 4 about here]

These results are encouraging, as they show that for both an absolute measure of error and a relative measure of error, the joint model outperforms the "ACS-only model." The findings therefore indicate that using Twitter data can improve the accuracy of predictions more than using ACS data only. Regardless of whether the goal is forecasting/nowcasting or filling in missing years for which there are no official statistics, it appears that models that include both digital data and traditional survey data produce better outcomes than models that rely on only one type of data.

6 Discussion and Conclusion

Our social lives have been increasingly digitalized. As a result of these technological developments, the amount of information on population and social processes that is available to social scientists is growing rapidly. While studies that use digital sources are proliferating (Zagheni et al., 2014; Hawelka et al., 2014; Jurdak et al., 2015; Fiorio et al., 2017; Hughes et al., 2016), a key problem that social scientists of all disciplines face is that these data come from specific groups in the general population. While they overlap to some extent, the users of cellphones are not the same as the users of Twitter or the users of email. Although each digital source leaves traces of human activity with a high degree

of spatial and temporal granularity, none can provide a representative sample of the general population.

Concerns about the non-representativeness and the biased inferences of digital data have plagued scientists who wish use these data to conduct rigorous research. Our view on this issue is that to utilize digital data effectively, we need to conceptualize digital data as complementing, rather than replacing, traditional sources. We contend that recent advances in the statistical sciences allow us to make scientifically rigorous inferences based on a combination of representative and non-representative sources.

We showed in this paper that Twitter estimates are always more upwardly biased than ACS estimates. If we analyzed Twitter data alone, the results we would produce would be highly misleading. Nevertheless, we also showed that by combining Twitter and ACS data, this bias can be modeled statistically. By decomposing the bias into spatial and temporal processes, we were able to estimate the bias structure, and incorporate information drawn from both Twitter and the ACS into a joint model that enhances prediction accuracy.

Although we illustrated how the method might be applied by using Twitter and ACS data to measure emigration rates, the method can be used to study other issues while drawing from other data sources. The requirements for our generic method would be (1) to establish a "gold standard estimate" drawn from official statistics with representative estimates, such as survey data; (2) while also taking into account (often biased) data that provide fine-grained information on the locations of individuals over time.

Thus, the method has many potential applications. First, although we used Twitter data in our example, we also showed that the method can combine a wide range of unbiased and biased sources of data, such as cellphone data, social media data, administrative records, and various sources of geographical and temporal information.

Second, our generic method can be applied to cases beyond than that of emigration rates in the US. The model is a generic approach that merely specifies two processes: one that is unbiased and one that is structurally biased. Future applications of the model might include measuring immigration rates or migration stocks, or replicating the results in other countries.

Finally, our generic method is not limited to including only two sources of data. Although we illustrated the method using Twitter and ACS data, the method could be easily extended to incorporate multiple sources of data. The method could, for example, take into account census data, large survey estimates from organizations, Twitter data, cellphone records, or email histories; while using different bias terms for each source of data.

We believe that in an age when humans are routinely leaving behind digital traces, combining new and traditional sources of data and methods will enable population scientists to investigate social processes in innovative ways. We look forward to future studies that adopt a perspective similar to our own.

References

Archive.org (2016). Archive.org of twitter 1% sample. https://archive.org/details/twitterstream. Accessed: 2016-09-30.

Blumenstock, J. E. (2012). Inferring patterns of internal migration from mobile phone call records: evidence from rwanda. *Information Technology for Development*, 18(2):107–125.

Clark, G. L. (1983). *Interregional migration national policy and social justice*. Totowa NJ: Rowman and Allanheld.

Clark, W. A., Deurloo, M. C., and Dieleman, F. M. (2000). Housing consumption and residential crowding in us housing markets. *Journal of Urban Affairs*, 22(1):49–63.

Fay, R. and Herriot, R. (1979). Estimates of income for small places: an application of James–Stein procedure to census data. *Journal of the American Statistical Association*, 74:269–277.

Fiorio, L., Abel, G., Cai, J., Zagheni, E., Weber, I., and Vinué, G. (2017). Using twitter data to estimate the relationship between short-term mobility and long-term migration. In *Proceedings of the 2017 ACM on Web Science Conference*, pages 103–110. ACM.

Greenwood, M. (1981). *Migration and Economic Growth in the United States*. Academic Press.

Hawelka, B., Sitko, I., Beinat, E., Sobolevsky, S., Kazakopoulos, P., and Ratti, C. (2014). Geo-located twitter as proxy for global mobility patterns. *Cartography and Geographic Information Science*, 41(3):260–271.

Hughes, C., Zagheni, E., Abel, G. J., Sorichetta, A., Wi'sniowski, A., Weber, I., and Tatem, A. J. (2016). Inferring migrations: Traditional methods and new approaches based on mobile phone, social media, and other big data: Feasibility study on inferring (labour) mobility and migration in the european union from big data and social media data.

Jurdak, R., Zhao, K., Liu, J., AbouJaoude, M., Cameron, M., and Newth, D. (2015). Understanding human mobility from twitter. *PloS one*, 10(7):e0131469.

Knorr-Held, L. (2000). Bayesian modelling of inseparable space-time variation in disease risk. *Statistics in Medicine*, 19:2555–2567.

Krainski, E. T., Gómez-Rubio, V., Bakka, H., Lenzi, A., Castro-Camilo, D., Simpson, D., Lindgren, F., and Rue, H. (2018). *Advanced Spatial Modeling with Stochastic Partial Differential Equations Using R and INLA*. Chapman and Hall/CRC.

Malik, M., Lamba, H., Nakos, C., and Pfeffer, J. (2015). Population bias in geotagged tweets. In *ICWSM Workshop on Standards and Practices in Large-Scale Social Media Research*, pages 18–27.

Massey, D. S., Arango, J., Hugo, G., Kouaouci, A., Pellegrino, A., and Taylor, J. E. (1993). Theories of international migration: A review and appraisal. *Population and development review*, 19(3):431–466.

Mercer, L. D., Wakefield, J., Pantazis, A., Lutambi, A. M., Masanja, H., and Clark, S. (2015). Space-Time smoothing of complex survey data: Small area estimation for child mortality. *Ann. Appl. Stat.*, 9(4):1889–1905.

Mislove, A., Lehmann, S., Ahn, Y.-Y., Onnela, J.-P., and Rosenquist, J. N. (2011). Understanding the demographics of twitter users. In *Fifth international AAAI conference on weblogs and social media*.

Molloy, R., Smith, C. L., and Wozniak, A. (2011). Internal migration in the united states. *Journal of Economic perspectives*, 25(3):173–96.

Molloy, R., Smith, C. L., and Wozniak, A. (2017). Job changing and the decline in long-distance migration in the united states. *Demography*, 54(2):631–653.

Moretti, E. (2013). Real wage inequality. *American Economic Journal: Applied Economics*, 5(1):65–103.

Rue, H., Martino, S., and Chopin, N. (2009). Approximate Bayesian inference for latent Gaussian models by using integrated nested Laplace approximations. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 71:319–392.

Wakefield, J., Fuglstad, G.-A., Riebler, A., Godwin, J., Wilson, K., and Clark, S. J. (2019). Estimating under five mortality in space and time in a developing world context. *Statistical Methods in Medical Research*, 28(9):2614–2634.

Waller, L. and Gotway, C. (2004). *Applied Spatial Statistics for Public Health Data*. John Wiley and Sons.

Zagheni, E., Garimella, V. R. K., Weber, I., and State, B. (2014). Inferring international and internal migration patterns from twitter data. In *Proceedings of the 23rd International Conference on World Wide Web*, WWW '14 Companion, pages 439–444, New York, NY, USA. ACM.

Zagheni, E. and Weber, I. (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*, pages 348–351. ACM.

7 Tables

Table 1: Comparison statistics for ACS-only models

	Spatial	Temporal	Space-time main effects	Space-time interaction
log-CPO	234	-134	233	324
DIC	-479	266	-479	-634
WAIC	-471	268	-471	-641

Table 2: Comparison statistics for joint models

	Spatial	Temporal	Space-Time main effects	Space-Time interaction
log-CPO	361	382	411	424
DIC	-485	-549	-621	-723
WAIC	-480	-546	-613	-724

8 Figures

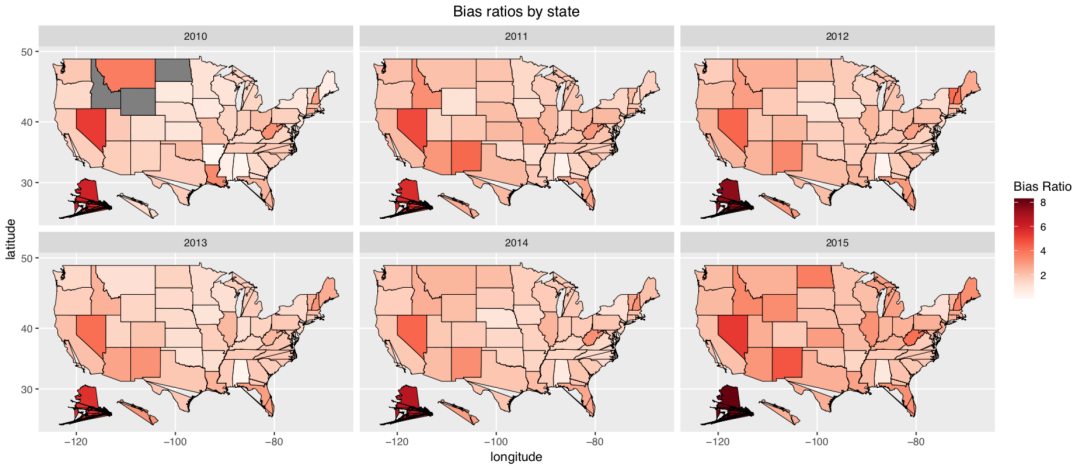

Figure 1: Map of bias ratios for each state across years (darker red colors indicate larger upward bias)

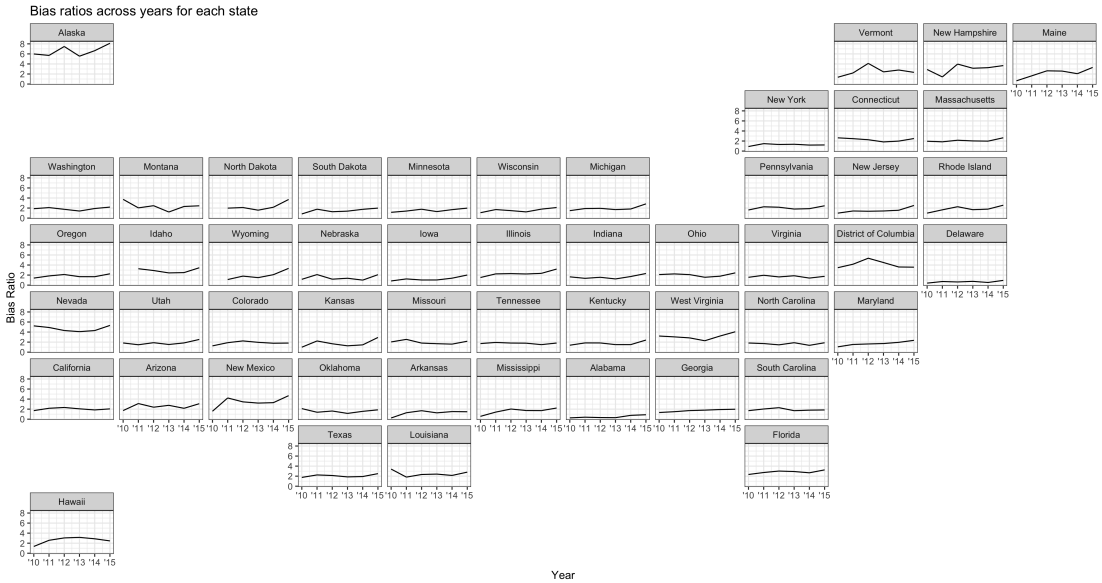

Figure 2: Lineplots of bias ratios within each state across years

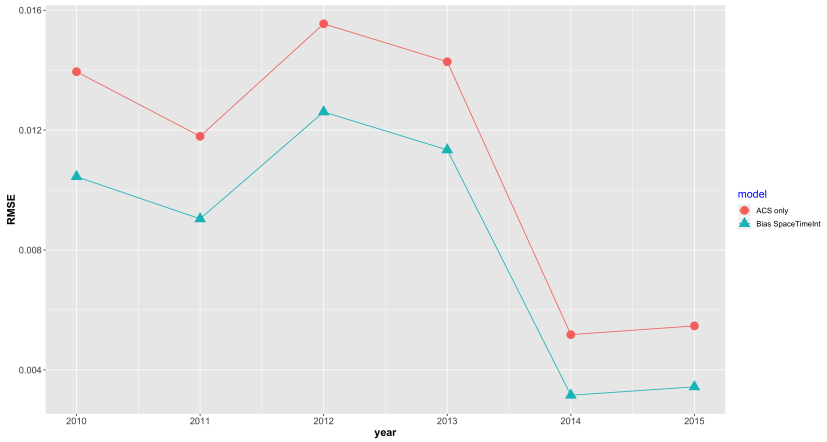

Figure 3: Comparison of RMSE of the joint model and the ACS-only model for each year

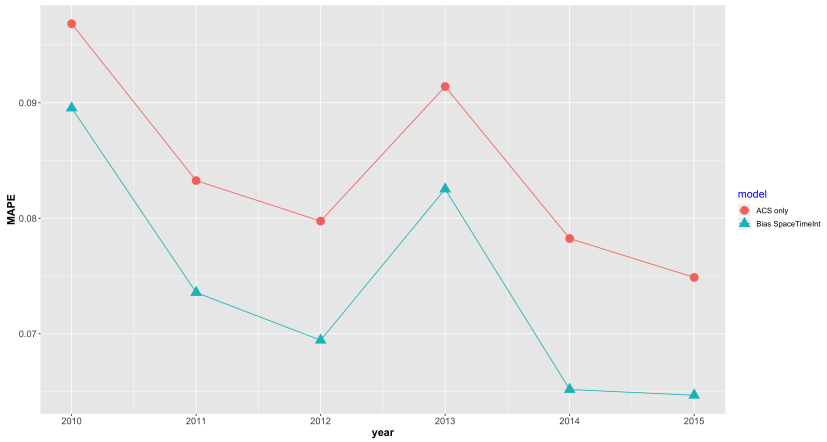

Figure 4: Comparison of MAPE of the joint model and the ACS-only model for each year

Online appendix for "Modeling the bias of digital data: an approach to combining digital and survey data to estimate and predict migration trends"

Contents

1 Explanation of model terms

For the full model $\mu_{s,t} = \mu + \theta_s + \phi_s + \alpha_t + \gamma_t + \delta_{st}$, we explain each of the parameters θ_s, ϕ_s, α_t, γ_t, δ_{st} in the following (alse see Rue and Held (2005)):

- ϕ_x and γ_t are independent random effects (i.e., effects with no spatial or temporal structure). These random effects have a generic $N(0, \sigma^2)$ form. This variance determines the amount of smoothing with small/large values favoring large/small amounts of smoothing. These capture state or year specific random effects. For instance, we might expect the emigration rate for Nevada state or year 2014 to be generally higher.

- α_t is a random walk of order 2 process (RW2) on a yearly scale. RW2 are part of a larger family of Intrinsic Gaussian Markov Random Fields (IGMRF). IGMRFs are "improper", as they have precision matrices not of full rank, and take into account the limiting case of how a data point is fully dependent on neighboring observations (e.g., in space or time). For a RW2 process, the conditional mean of a data point is dependent only on the last two data points with a precision parameter τ_x to estimate. That is:

  - If we define $\Delta^2 x_i = \Delta(\Delta x_i)$, and $\Delta^2 x_i \sim_{iid} N(0, \tau_x^{-1})$. Then:
  - $E[x_{i+k}|x_1, \ldots, x_i, \tau_x] = (1+k)x_i - kx_{i-1}$
  - $Prec(x_{i+k}|x_1, \ldots, x_i, \tau_x) = \tau_x/(1 + 2^2 + \ldots + k^2)$

- Similarly, θ_s is a local spatial smoothing model from intrinsic conditional autoregressive (ICAR) model. An ICAR model again represents an IGMRF but the dependency of a data point is now on the adjacent neighbors of a state (which is similar to a generalization of a RW1 process to a lattice). Under an ICAR model, the distribution of a data point x_i is: $x_i|x_i, \tau_x \sim N(\frac{1}{m_i}\sum_{j:j\sim i} x_j, \frac{1}{m_i \tau_x})$. Again τ_x is the precision parameter to be estimated in the model.

- δ_{st} is the type IV interaction described by Knorr-Held (2000). The interaction term that assumes the spatially ICAR effect and and temporal RW2 effect interact at the yearly level. That is, the precision matrix Q_δ is the kronecker product of the precision matrix of the ICAR process and the precision matrix of the RW2 process: $Q_\delta = Q_{alpha} \otimes Q_\theta$.

2 Explanation of fit indices

We explain each of the model fit indices below:

- Log-cpo represents the "leave-one-out" predictive measures of fit. The CPO value $P(y_i|y_{-i})$, which is the probability density of an observed response based on the model fit to the rest of the data.

- The deviance information criterion (DIC) is a hierarchical modeling generalization of the Akaike information criterion (AIC). It draw from information from the deviance $D(\theta) = -2log(P(y|\theta)) + C$, where C is a constant that cancels out between models. Because of the negative term, lower values of DIC indicate better fit.

- The Watanabe-Akaike information criterion (WAIC) similarly draws from the deviance, but with modifications in averaging of the posteriors. See Gelman et al. (2014) for a detailed explanation.

3 Robustness checks with PC priors

In the main paper, we used the default priors in INLA (Rue et al., 2009) to information the spatial and temporal parameters in the model. However, an alternative in the recent literature on space-time models would be to use Penalised Complexity priors, or PC priors (Simpson et al., 2017). PC priors are invariant to reparameterisations, have a natural connection to Jeffreys' priors, are designed to support Occam's razor, and possess robustness properties (Simpson et al., 2017).

In PC priors, for each precision parameter, the user provides an U value and an α value. Generally, for precision parameter τ, the user specifies (U, α) so that $P(1/\sqrt{\tau} > U) = \alpha$. In this robustness check, we set $\alpha = 0.05$ and experiment with different values of U: $U = \{0.5, 1, 5, 10, 20\}$. We then rerun the validation procedure and recompute the RMSE to test if different priors affect the prediction error of the models.

Figure 1 shows the RMSE's under different PC priors. The results suggest that the prediction errors, whether from the default priors or the PC priors, are all better than the ACS-only model (i.e., the red solid line). In fact, in low values of U, the RMSE from models using PC priors are almost identical to results from the default priors, as on th plot the RMSE's overlap. Only under very large values of U (e.g., 5 or 10) do the results change.

Priors Robust.png

Figure 1: Comparison of priors

4 Validity check on information added by Twitter data

Our model presupposes that Twitter data aids prediction when official statistics are not available. On the other hand, when official statistics are available, Twitter data should not affect the estimation. We check this model assumption by comparing the estimates from the model that uses only ACS data and the joint model that utilizes both ACS and Twitter data.

The results indicate that the average difference in estimated probability across states and across years between the "ACS-only" model and the joint model is 0.000005010695, which is probably due to divergences in the optimization procedure. The results provide confidence that the model draws primarily from official statistics when available, and the estimation of the bias does not contaminate the estimation of the true migration process.

5 Testing model performance under random noise

In the main analysis we showed that a model with a space-time bias structure showed the best performance. However, one may wonder whether this model is robust to situations where there is **no bias**. In other words, since in many scenarios we cannot determine whether there is bias in certain sources of data a priori, if the space-time model assumes a bias structure we would be interested in whether this assumption affects the results.

To examine the issue, we simulate synthetic data that includes two series that follow the same space-time processes (μ_{st}), but with unequal random noise ϵ_1&ϵ_2. In other words, both series are unbiased of the true process, but with different variations of error:

$$Y^1_{s,t} = \mu_{s,t} + \epsilon^1_{s,t} = \mu + \theta_s + \phi_s + \alpha_t + \gamma_t + \delta_{s,t} + \epsilon^1_{s,t}$$
$$Y^2_{s,t} = \mu_{s,t} + \epsilon^1_{s,t} = \mu + \theta_s + \phi_s + \alpha_t + \gamma_t + \delta_{s,t} + \epsilon^2_{s,t}$$

In our synthetic data, without loss of generality we set our precision parameters for each of the terms as follows:

- $\tau_\theta = \tau_\phi = \tau_\alpha = \tau = 1$
- $\tau_{\epsilon^1} = 16$
- $\tau_{\epsilon^2} = 0.25$

The goal of the exercise is to compare the performance if we estimated the model with the "Baseline Model" just data from series 1 (which is analogous to the "ACS-only" model in the main analyses) or the "Joint Model" which specifies a space-time interaction structure for the bias in series 2 (which is analogous to the Joint model in the main analyses that uses both ACS and Twitter data). We use the log-CPO, which is the probability density of an

observed response based on the model fit to the rest of the data. A higher log-CPO indicates a better fit. We run the simulations 500 times and take the average sum of the log-CPO.

The results show that the average log-CPO for the "Baseline model" is -448, while the average log-CPO for the "Joint model" is -438. Even if we impose a model with a more complicated bias structure than the true process, the "Joint model" that uses data from both series still performs better the the "Baseline model" that only uses one series. This is probably because the "Joint model" has more information as it draws more information. In short, the "Joint model" that assumes a space-time interaction bias structure appears to be robust even when there is no bias.

References

Gelman, A., Hwang, J., and Vehtari, A. (2014). Understanding predictive information criteria for bayesian models. *Statistics and computing*, 24(6):997–1016.

Knorr-Held, L. (2000). Bayesian modelling of inseparable space-time variation in disease risk. *Statistics in Medicine*, 19:2555–2567.

Rue, H. and Held, L. (2005). *Gaussian Markov random fields: theory and application*. Chapman and Hall/CRC Press, Boca Raton.

Rue, H., Martino, S., and Chopin, N. (2009). Approximate Bayesian inference for latent Gaussian models by using integrated nested Laplace approximations. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 71:319–392.

Simpson, D., Rue, H., Riebler, A., Martins, T. G., Sørbye, S. H., et al. (2017). Penalising model component complexity: A principled, practical approach to constructing priors. *Statistical Science*, 32(1):1–28.