Estimating US Internal Emigration Flows in a timely way by complementing the American Community Survey with geolocated Twitter Data

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 necessary to guide policy decisions and to improve our understanding of migration processes. However, obtaining these estimates remains an elusive goal. We propose an approach to combine geo-located Twitter data for over 2 million users (2010-2016) with data from the American Community Survey (ACS) in order to estimate US internal emigration flows at state-level. We leverage the correlation structure for state-level biases in Twitter data by proposing a Bayesian hierarchical space-time model that structurally models the bias by space and time. We show that Twitter-based estimates can be combined with ACS estimates to improve short-term predictions of internal migration flows or predict years where there are no official statistics.

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 particular, migration data, especially measures of migration flows, are unavailable or unreliable in many contexts. For a number of practical and conceptual reasons, the movement of people across distance is difficult to measure. Cross-sectional survey estimates and estimates from administrative data are often limited in geographic and temporal scope and costly to produce in terms of time and money. Even where data do exist, different institutions use different definitions of migration to suit their specific needs which can make harmonization of estimates difficult (Nowok and Willekens, 2011).

Hampered by these issues, migration scholars have begun developing methods for using new, non-traditional sources of data, like social media 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). While noisy and biased, social media have a handful of potentially attractive characteristics, namely their real-time availability and their arguably globe-spanning coverage (Malik et al., 2015). With the right methods, these data can be used to provide more timely estimates of migration or even give estimates of migration where none currently exist.

A shortcoming of the growing methodological literature on migration and social media data, however, is a lack of sophistication with respect to the spatial and temporal structure of bias in social media data. Much of the work that has been done so far assumes that the relationship between social media estimates and survey estimates are consistent. In this paper, we take a more robust approach to determine whether a model that takes into account the variable spatial and temporally structure of bias in social media estimates improves overall model prediction. As such, this paper asks three interrelated questions:

- (Q1) how biased are estimates from social media data?
- (Q2) are there statistical models that stabilize the relationship between estimates from social media data and the true population rates?
- (Q3) can we combine estimates from social media data and official statistics to improve short-term predictions?

We draw on data from geo-located Twitter data for over 2 million users (2010-2016) with data from the American Community Survey (ACS) to estimate state-level internal migration in the United States. Given the research questions, this paper adopts the following analytical strategy:

- Utilize Twitter data in the US to obtain estimates of **state-level** estimates of **emigration rates**
- Assess bias by comparing with official statistics from the American Community Survey (ACS)

- Combine Twitter and ACS data using a Bayesian space-time to improve prediction

Our results suggest that raw estimates from Twitter are inconsistently biased across years and across states and are not useful for prediction. However, after we use a model that leverages the spatial and temporal component of the migration rates and bias structure, we are able to combine Twitter and ACS data to forecast short-term migration rates and achieve a higher prediction accuracy.

Although the paper utilizes Twitter data and the American Community Survey as the example, the approach is generic to accommodate other sources of digital data such as cell phone data or other social media with geolocation features, and other survey data such as estimates from OECD or Eurostat. We believe that by incorporating data from digital sources, we may be able to overcome obstacles that have impeded migration research in the past.

2 Migration estimation and approaches to incorporate digital data

Though often overlooked in favor of international migration, *internal* migration is central to a wide range of phenomena including urban growth and development (Greenwood, 1981), housing dynamics (Clark et al., 2000), and the market for labor (Moretti, 2013; Molloy et al., 2017, 2011). In the case of the U.S., data exist and are of reasonable quality; however, there is a considerable temporal lag to the release of migration statistics. The American Community Survey (ACS) reports interstate flows, but it often takes over a year for these estimates to be produced. As such, internal migration in the U.S. makes for a useful testing ground for developing methods for using social media data to measure migration. High quality, regularly released survey estimates exist and can be used to assess the bias of social media data estimates, but there still is a need for more timely data which social media estimates could potentially provide.

2.1 Concepts in the Measurement of Migration

Migration data are often thought of as coming in one of two forms: information about stocks and information about flows. In general, stock data are easier to observe. With information about where an individual resides and where that individual was born, lifetime migration can be inferred. Someone living in a different U.S. state from the one in which they were born, for example, is an interstate migrant. The problem with stock data, however, is that they do not provide very timely estimates of current migration processes. Moreover, people may move more than once, either onward or return, resulting in the underestimate of the overall level of migration.

Bilateral flow data, i.e. estimates of flows from all O origins to D is a much richer form of migration data. These data are essential for measuring and understanding migration systems. When people migrate, they form a link between origin and destination, transforming both places in the process. That being said, flow data come with considerably more complexity than stock data. Flow data require either repeatedly observing a panel of individuals multiple times as they relocate (or not) across space or asking individuals to retrospectively report the locations they have been. The latter method is further complicated by the issue of time interval. Different surveys may use different temporal intervals for retrospective reporting (e.g. ”where did you live one year ago?” versus ”where did you live five years ago?”), reducing the comparability of different measures (Rogers et al., 2003).

What is more, migration remains a relatively rare event, even within a highly mobile society like the United States. On average, only 1.5 to 2% of the population changes states each year (Molloy et al., 2017). This makes accurately estimating the full bilateral flow matrix difficult, especially when the population is unevenly distributed across states. Accurately estimating the flow of people from California to Texas each year is easier than accurately estimating the flow of people from Connecticut to Idaho.

Certain kinds of social media data, like location information associated with Twitter posts, lend themselves naturally to the study of migration flows. Individuals are observed repeated in time and space, allowing for researchers to convert their movements into estimates of flows. Moreover, given the relative size of the population of active Twitter users, Twitter data potentially provide a robust signal of all bilateral flows large (e.g. California to Texas) and small (e.g. Connecticut to Idaho). But more research is needed to understand how flows of Twitter users might be biased with respect to the space of these flows and their change over time.

Nevertheless, just knowing that social media data are biased does not render them useful for prediction, as the critical question is *whether we can statistically model the bias*. If we can provide a probabilistic approach that systematically captures the bias in relation to the true migration process, we would be able to incorporate information from social media data to improve the estimation and prediction of migration flows. In this paper, we provide a Bayesian space-time model on state-level emigration rates and show that the bias of social media data can be modeled statistically and combined with estimates from survey data.

3 Data

The data is the paper come from two sources:

1. Twitter data from the 1% historical archive of Twitter
2. Offical statistics from the American Community Survey (ACS)

3.1 Twitter data

Twitter data come from the 1% historical archive of Twitter. In this paper, we include only geo-located tweets within the US, resulting in **2,226,467** users and

554,229,541 tweets. Each geo-located tweet includes information on the **userid** and the **latitude** and longitude of the **tweet**. The time-frame for analysis in this paper is from **Jan 1, 2010** to **Dec 31, 2016**

3.2 ACS statistics

The ACS inquires respondents on the current state/country they live in currently, and the state/country the resided in one year ago. From the information on current and previous residency, the ACS produces estimates of state-to-state migration flows on a yearly basis. In this paper, we draw from ACS estimates for 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 from the replicate weights in ACS.

4 Twitter estimates and assessing bias

4.1 Obtaining estimates from Twitter data

The goal of the paper would be to estimate λ_{st}, the population (out-)migration rate for state s, year t. We use Twitter data to obtain estimates of this population rate:

In this paper, we follow (Zagheni et al., 2014) and use the following method:

1. For each tweet, from the latitude and longitude construct the state of the tweet

2. For each user, for each year, calculate the number of tweets in each state

3. For each user, for each year, calculate the modal state and second modal state

4. For every two years (e.g., year 2010 and year 2011), discard users where the number of tweets for the modal state is less than 3 in at least one of the two years, or the ratio between the number of tweets in the modal state and the number of tweets in the second modal state is less than 3.

   For example, if a user has 15 tweets in Washington state and 8 tweets in Ohio state in year 2010, and 20 tweets in Washington state and 3 tweets in Ohio state in year 2011. The user would be discarded because the ratio in year 2010 is less than 3.

5. For every two years, 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.

6. For every two years, calculate the estimated migration rate for the first year (define as $\hat{\lambda}$) as $N_{Migrants}/N_{Users}$.

4.2 Assessing bias of Twitter estimates using the ACS

4.2.1 Bias ratios across space

The above provides raw estimates from Twitter data. However, it would be helpful to understand how estimates from Twitter compare to official statistics from the ACS. We assess bias using a simple bias ratio formula:

Define BR as the bias ratio. Let $\hat{\lambda}_{st}$ be the estimate for state s, year t. Let $\hat{\lambda}_{Twitter-st}$ be the raw estimates from Twitter and $\hat{\lambda}_{ACS-st}$ be the official estimates from ACS.

Then:

$$BR_{st} = \hat{\lambda}_{Twitter-st}/\hat{\lambda}_{ACS-st}$$

Since the ACS uses a representative sample, we assume that the point estimates from ACS are equal to λ_{st} (incorporating standard errors of the ACS would be the next step). Throughout this paper, we would use the bias ratio to assess the degree of discrepancy between estimates from Twitter and the estimates from ACS.

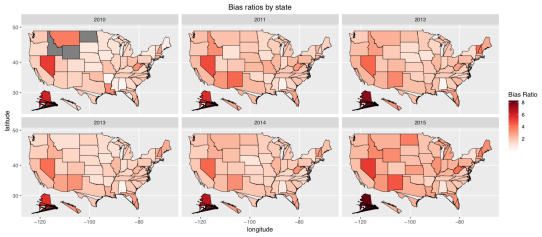

Figure 1: Bias ratios across states

Figure 1 plots the average bias ratios across states for each year. Compared to the ACS estimates, perfect estimation yields a bias ratio of 1, and numbers large than 1 indicate that the raw Twitter estimates **overestimate** the migration rate (e.g., a bias ratio of 1.15 indicates an over-estimation of 15%), while numbers smaller than 1 indicate **underestimation** (e.g., a bias ratio of 0.78 indicates an under-estimation of 22%). As seen, the bias ratios are not only larger than 1.00 (i.e., red colors in the maps), but also fluctuate a lot across years.

4.2.2 Bias ratios across years

We then plot the bias ratios over the years in Figure 2. We produce separate plots for each census division so that we can also informally assess spatial correlation. In general, the bias ratios are lower in year 2010 and higher in year 2015, with small dips and bumps in between.

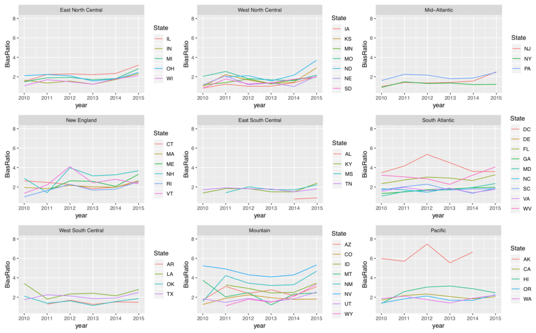

Figure 2: Bias ratios across years

4.2.3 Short summary

In short, we see that Twitter estimates are in general larger than the ACS estimates, indicating over-estimation bias. However, by utilizing models that examine correlations of biases by space or by time, we might able to combine the Twitter and ACS estimates to predict emigration rates.

5 A statistical approach to combine ACS and Twitter data for forecasting/prediction

5.1 Intuition of the model

In this paper, we use a **joint-modeling approach** to understand the data-generating mechanisms for both ACS and 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 estimates this process with measurement error, while the Twitter data estimates this process with measurement error and bias. In other words, the ACS and the Twitter estimates for a particular state-year would be bivariate normal with a partially shared mean component (the part without the bias). Over the states and years, the ACS and Twitter estimates would be multivariate normal with a partially shared mean component.

The key modeling choices would be:

- How do we model the true process?
- How do we model the bias?

Regarding how to model the true process and the bias, 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., 2018). The advantage of space-time models is that they incorporate information within space and across space, as well as information within time and across time. For instance, if the demographic composition of the Connecticut state is more or less stable, we would expect migration rates *within* Connecticut to be correlated across different years. However, we may also wish to incorporate information that Connecticut is adjacent to Massachusetts, and if their demographics are similar there would be spatial correlation *between* the two states. Similarly, we might expect that states *within* year 2015 to be potentially correlated in general, but also adjacent years to be correlated with one another. The space-time models decompose population processes into spatial and temporal processes with within and between effects, which have benefited estimation.

Following (Mercer et al., 2015), we could 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.

Similarly, we could conceptualize 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. In the paper, we would test these possibilities and select the appropriate model.

5.2 Mathematical formulation of the model

Since emigration rates or probabilities that lie between zero and one, we model the **logit** of the emigration rates.

Formally, denote:

- Y_{ACS-st} as the logit of the migration estimates from ACS for state s, year t
- Y_{TW-st} as the logit of the migration estimates from Twitter for state s, year t

Then:

- $Y_{ACS-st} \sim N(\mu_{st}, V_{ACS-st})$
- $Y_{TW-st} \sim N(\mu_{st} + B_{st}, V_{TW-st})$

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

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

Notice B_{st}, which represents this bias term for Twitter estimates (we assume that ACS estimates are unbiased). Also, we assume that the covariance terms are 0 as the measurement errors of ACS and Twitter are independent because the data are drawn from independent samples, and measurement errors should be unrelated.

The choice would then becomes how to model μ_{st} and B_{st}. For the true process μ_{st}, we could model μ_{st} as:

- An ICAR spatial process: $\mu_{st} = \mu + \theta_s + \phi_s$
- A Random Walk 2 temporal process: $\mu_{st} = \mu + \alpha_t + \gamma_t$
- A space-time independent process: $\mu_{st} = \mu + \theta_s + \phi_s + \alpha_t + \gamma_t$
- A space-time interaction process: $\mu_{st} = \mu + \theta_s + \phi_s + \alpha_t + \gamma_t + \delta_{st}$

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, δ_{st} is a structured interaction between the ICAR process and the RW2 process.

Similarly, we could model B_{st} as:

- An ICAR spatial process: $B_{st} = \mu + \theta_s + \phi_s$
- A Random Walk 2 temporal process: $B_{st} = \mu + \alpha_t + \gamma_t$
- A space-time independent process: $B_{st} = \mu + \theta_s + \phi_s + \alpha_t + \gamma_t$
- A space-time interaction process: $B_{st} = \mu + \theta_s + \phi_s + \alpha_t + \gamma_t + \delta_{st}$

We fit these models using the *INLA* package (CITATION) in the statistical software *R*. Since the ACS data is representative, our analytical strategy would be to use only the ACS data first to select the best model for μ_{st}. We use three model selection criteria: log-cpo (higher indicates better fit), DIC (lower indicates better fit) and WAIC (lower indicates better fit) [CITATIONS AND EXPLAIN WHAT THESE ARE].

After we select the optimal model for the true process, we estimate the joint model that utilizes both ACS and Twitter data. We then use the same fit indices (log-cpo, DIC, WAIC) to select the optimal joint model.

To validate whether the joint model improves prediction accuracy, we use a cross-validation approach. We remove one year of the ACS data as the target and estimate the joint model on the rest of ACS data and all the Twitter data. We compare this model with a "ACS only" model that only uses the rest of ACS data. The comparison mimics the scenario where we have Twitter data but not official statistics for a particular year. For instance, one application would be to forecast timely estimates in the future when Twitter data is available but official statistics are not yet produced. Another application would be when we wish to understand migration trends but there are years where official statistics are not available. Assuming that official statistics such as the ACS is the "truth", we wish to test whether the joint model that incorporates Twitter data can better uncover the "truth" compared to a model that does not incorporate Twitter data.

Table 1: Comparison of fit statistics for ACS only models

	Spacial	Temporal	Space-time independent	Space-time interaction
log-cpo	234.1555	-134.1714	233.0876	323.7873
DIC	-479.3320	265.9899	-479.0940	-633.7412
WAIC	-470.9372	268.4685	-470.6455	-641.1218

Table 2: Comparison of fit statistics for joint models

	Constant	Spatial	Temporal	Space-Time independent	Space-Time interaction
log-cpo	356.6269	360.9662	382.1864	410.8564	424.3998
DIC	-473.3560	-485.3548	-549.0363	-620.6123	-722.5505
WAIC	-471.0102	-479.5892	-545.7481	-613.4820	-724.1266

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

We compare the fit statistics for "ACS-only" model to select the best model for the true emigration process (i.e., μ_{st}) in Table 1. As seen, the Space-Time interaction model has the highest log-cpo and the lowest DIC and WAIC, suggesting that it is the model with the best fit for the true emigration process.

On top of this space-time model for the true emigration process, we explore different joint models that incorporate the bias structure (i.e, B_{st}). As seen in Table 2, the space-time interaction joint model best captures the bias structure with the highest log-cpo and the lowest DIC and WAIC. From the fit statistics, we select as the optimal model a joint model that specifies the true process as a space-time interaction process, and also the bias structure as a space-time interaction process. In the next section, we compare whether this joint model outperforms the best "ACS-only" model in forecasting and prediction.

5.4 Results on forecasting/predicting emigration rates

We use the Root Mean Squared Error (RMSE) to evaluate model performances. Let $\hat{\lambda_{ACS-st}}$ be the emigration rate for the ACS target year (i.e., the removed ACS year), and $\hat{\lambda_{st}}$ be the predicted emigration rate from the model. Then the RMSE would be:

$$RMSE = \sqrt{\sum_{s=1}^{51}(\hat{\lambda_{st}} - \hat{\lambda_{ACS-st}})^2}$$

A lower RMSE indicates better prediction.

We compare the RMSEs for both models for each year in Figure 3. As seen, the joint model performs the ACS-only model for every year, especially the later years (i.e., year 2015 and year 2016). The results are encouraging as it shows that Twitter data can not only improve predictions in general, but may be particularly useful when forecasting the future.

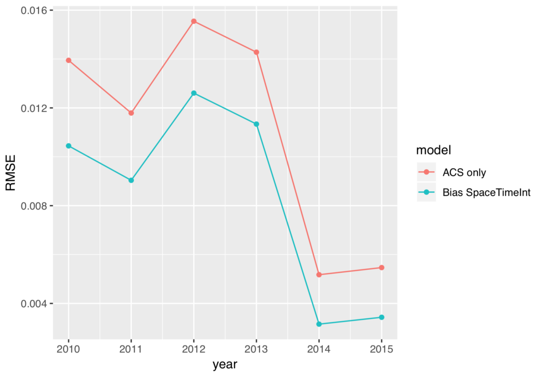

Figure 3: Comparison of RMSE of years

6 Discussion and Conclusion

We have shown that digital data can be helpful in predicting migration rates when combined with official statistics. This may be particularly useful official statistics are available but one wishes to increase prediction accuracy or produce estimates in smaller time intervals than the official statistics. However, raw estimates from Twitter data tends to be inconsistent in bias, and in turn are not appropriate for prediction. Nevertheless, we show a space-time model can statistically model the bias by accounting for the spatial and temporal structure and substantially increase prediction accuracy.

Note that the methods proposed in this paper are quite general and not limited to Twitter data. The requirement would be granular data on the locations of individuals over time, which could come from social media data, cellphone records, or administrative collections. With increasing availability of data sources, we believe we can gain better estimates of migration rates.

References

Clark, G. L. (1983). Interregional migration national policy and social justice.

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.

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.

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*, pages 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.

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.

Nowok, B. and Willekens, F. (2011). A probabilistic framework for harmonisation of migration statistics. *Population, Space and Place*, 17(5):521–533.

Rogers, A., Raymer, J., and Newbold, K. B. (2003). Reconciling and translating migration data collected over time intervals of differing widths. *The Annals of Regional Science*, 37(4):581–601.

Wakefield, J., Fuglstad, G.-A., Riebler, A., Godwin, J., Wilson, K., and Clark, S. J. (2018). Estimating under five mortality in space and time in a developing world context. *Statistical Methods in Medical Research.* To Appear.

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.