Using Geolocated Twitter Data to Study Recent Patterns of International and Internal Migration in OECD Countries

Ingmar Weber[1], **Venkata Rama Kiran Garimella**[1], **Emilio Zagheni**[2], **and Bogdan State**[3]

[1]Qatar Computing Research Institute
[2]Queens College - CUNY & Wittgenstein Center
[1]Stanford University

November 15, 2013

Abstract

Data about migration flows are largely inconsistent across countries, typically outdated, and often inexistent. Despite the importance of migration as a driver of demographic change, there is limited availability of migration statistics. Generally, researchers rely on census data to indirectly estimate flows. However, little can be inferred for specific years between censuses and for recent trends. The increasing availability of geolocated data from online sources has opened up new opportunities to track recent trends in migration patterns and to improve our understanding of the relationships between internal and international migration. In this paper, we use geolocated data for about 500,000 users of the social network website "Twitter", during the period May 2011- April 2013, for OECD countries. We evaluated, for the subsample of users who have posted geolocated tweets regularly, the movements within and between countries for independent periods of four months, respectively. Since Twitter users are not representative of the OECD population, we cannot infer migration rates at a single point in time. However, using a difference-in-differences approach, we could evaluate trends in out-migration rates for single countries, and the heterogeneity in mobility patterns of migrants and non-migrants. We obtained estimates of the age and gender of users using a face recognition software (Face++) with the profile pictures of users. Preliminary results indicate that the approach may be useful to predict turning points in migration trends which are particularly relevant for migration forecasting. We observed quite a bit of heterogeneity in the relationship between within- and across-countries mobility for OECD countries. Our analysis relies uniquely on publicly available data that could be potentially available in real time and that could be used to monitor migration trends.

1 Introduction

Migration is one of the major sources of demographic change [7]. Projecting migration rates and unveiling relationships between internal and international migrations are thus important tasks for demographers and social scientists. Limited data availability has been one of the main bottlenecks for empirical analyses and for theoretical advances in the study of migrations. In particular, data about international migration flows are largely inconsistent across countries, typically outdated, and often inexistent [4, 5].

Demographers and international organizations are very interested in improving migration statistics. For instance, EUROSTAT has worked in partnership with researchers at the University of Southampton and NIDI to generate consistent estimates of migration flows between European countries [6]. Abel [1] has developed statistical techniques to estimate flows of migrants from census data, in order to generate historical time series of migration flows. Among others, Abel's work is intended to inform IIASA's population projections. The Population Division of the United Nations (UN) has recently moved towards offering probabilistic population projections [8]. Forecasting migrations remains one of the most difficult tasks for the UN. Currently, there is a continuing collaboration between the UN and the University of Washington to develop statistical models to forecast net migration rates for all countries [2].

Most of the existing work in the literature relies on time series of historical data to produce forecasts for the future. One of the main limitations is that data for the "present" are typically unavailable. In other words, there is a substantial lag between data collection and production of migration statistics. Even in those fortunate cases where it is possible to resolve inconsistencies between sources from different countries, it may take a few years before data become available, especially when the main source is a census.

The lack of timely data may strongly affect migration projections. Forecasts are particularly sensitive to recent trends. Therefore, extrapolations that fail to include information about the recent past may lead to much larger errors in the medium- and long-term. In this paper, we use geolocated data from the social network website "Twitter" to evaluate recent trends in migrations in OECD countries. The main goals of our work are to complement existing migration statistics, and to develop methods for harnessing publicly available online data in order to improve migration forecasts and our understanding of populations of migrants.

The increasing availability of geolocated data from online sources has opened up new opportunities to identify migrants and to follow them, in an anonymous way, over time. Two main lines of literature have emerged in the the last few years. Recent trends in international flows of migrants have been estimated by tracking the locations of users who repeatedly login into a Web service [10, 9]. For these types of analyses, geographic locations are inferred from IP addresses, which are very accurate at the country level, but are not accurate enough to pinpoint exact locations within a region of a country. In order to evaluate internal migration, use of geolocated cellphone data has been suggested [3]. Cellphone data offer fairly accurate information about the location of cellphone users. However, the geographic scope is typically limited to single countries. Therefore cellphone data are suitable for analyses of internal migration, but not to track cross-border mobility. Geolocated Twitter data is unique in the sense that

the information about locations is very accurate and the geographic scope is potentially global. Our paper is a first analysis of the possibilities offered by Twitter data to estimate recent migration trends in a timely fashion, and to improve our understanding of the relationships between internal and international migration.

2 Data and Methods

2.1 Data collection and pre-processing

For this project we downloaded geolocated Twitter 'tweets' for about 500,000 users who have posted at least one geolocated tweet. We have 'full coverage' of the tweets for these users for the period from May 2011 to April 2013. Since the Twitter API limits the number of Tweets that can be downloaded for each user to 3,200, only partial coverage for periods before May 2011 is available. In other words, we have information prior to May 2011 only for those users who have not posted a large number of tweets.

We sampled tweets for users who have posted at least one geolocated tweet in OECD countries from an initial 'seed' of users. We sampled users with the goal of generating a balanced sample. In other words, for countries where out-migration is a relatively rare phenomenon, we oversampled. For countries with higher rates of mobility we obtained relatively smaller samples.

We started with about 500,000 users with at least one geolocated tweet. Of these, about 345,000 had at least 10 geolocated tweets. About 150,000 posted at least 100 geolocated tweets. On average, users in the sample posted 142 geolocated tweets. The distribution is fairly skewed, with a median number of geolocated tweets equal to 34.

The average number of days between the first and the last tweet is 225. The average number of days between tweets is about 12. The average number of days between tweets reduces to about 6 for users who have posted at least 10 geolocated tweets.

There is a trade-off between a large sample of users for whom we may have sparse information over time, and a smaller sample of users for whom we have detailed and consistent information over time. We decided to select a sample of users for whom we have detailed and consistent geographic information since the early 2011. This decision is motivated by the fact that users for whom we have information over a longer period of time are more likely to provide reliable information. Moreover, they are more likely to continue to post on Twitter and therefore we can follow them in the future. We splitted the dataset into separate periods of four months, from May 2011 to April 2013, and we considered only those users for whom we have at least 3 geolocated tweets for each period. The final sample size reduces to about 15,000 users in OECD countries for whom we have detailed information over time.

For each user we have, among others, a unique identifier, the text of his or her tweets, the date of the posts, and the geographic coordinates for the location from where the user 'tweeted'. We are in the process of complementing the dataset with information about the age and sex of users. We are using the face recognition software Face++[1] to obtain estimates of the gender and age of users based on their profile picture. We are testing the validity of our estimates of age and sex, and the uncertainty

[1] `http://en.faceplusplus.com`

associated with them. We expect to be able to provide fairly precise statements as we make progress with this project.

2.2 Estimation of trends in out-migration rates with a difference-in-differences approach

For each user for whom we have at least 3 geolocated tweets for each period of four months, we estimated the country of residence for the given period as the country from where most of the tweets were posted (the 'modal' country). If the uncertainty is large, i.e. if the number of tweets in the modal country is not at least three times as large as the number of tweets for the second most frequent country, then we discarded the information for that user for the specific period.

For a given user, if the modal country for two consecutive periods is the same, then we estimate that the user did not move over the period of eight months. If, for the first period of four months the modal country is A and then, for the second period, the modal country is B, we estimate that the user moved from country A to country B over the eight months considered.

The migration rates that we estimated cannot be considered representative of OECD countries. They represent the experience of migration and mobility of the subset of Twitter users who regularly post geolocated tweets. This population of Twitter users could be of great interest in itself. Describing and studying this population may be very relevant for instance for anthropological studies. At the very least we are describing the experience of a fairly large and significant population. However, we would like to be able to use the information in our dataset of Twitter users to make some generalizations for the whole population.

We propose a difference-in-differences approach to estimate recent trends in mobility rates. Although we cannot make statistical inference about mobility rates at single points in time, under certain assumptions we can extract information about trends. Let m_c^t be the out-migration rate from country c to all other countries, at time t. Consider then the average of this quantity across all countries, m_{oecd}^t, i.e., the average of the migration rates at time t for all the OECD countries considered. If the population of Twitter users changes in similar ways across all the OECD countries over time, for instance due to Twitter's expanding user base, then we can use a difference-in-differences estimator to evaluate relative changes in trends:

$$\hat{\delta} = (m_c^t - m_{oecd}^t) - (m_c^{t-1} - m_{oecd}^{t-1}) \tag{1}$$

In other words, selection bias prevents us from making statistical inference for single points in time, because the population of Twitter users is not representative of the whole population. In addition, changes in the composition of the Twitter population over time prevent us from using time series of estimated out-migration rates to make statistical inference about changes over time. However, if changes in the composition of Twitter users are consistent across countries, then the comparison of relative changes for a single country with relative changes for the group of reference can be used to provide information about trends. For example, if the proportion of Twitter users who are 25 years old is larger than the fraction of people who are 25 years old in the population,

then estimates of migration rates based on Twitter data would tend to overestimate flows (because people in their 20s are more mobile than people in other age groups). Analogously, if the proportion of Twitter users who are 25 years old changes from one period to the next one, then we cannot compare the two estimates from Twitter. However, if the proportion of Twitter users who are 25 years old changes in similar ways across all countries, then we can expect that for those countries where we observe more rapid increases in out-migration rates, the population-level migration rates have been increasing, relative to other countries.

These difference-in-differences estimates have limitations. There is a certain amount of uncertainty around numeric values. However, there is generally less uncertainty about the signs of the estimates. Twitter data are virtually the only source of information that is publicly available in almost real-time. Geolocated Twitter data are very important because they can complement existing statistics and be used to predict turning points in migration rates. These types of predictions are extremely relevant to forecast international migration rates in the short- and medium-term.

2.3 Measuring the relationship between internal and international mobility

Geolocated Twitter data include the geographic coordinates from where individuals post their tweets. The level of detail for the geographic information is very high and thus allows us to compute measures of mobility for users, and to evaluate patterns of mobility. In particular, we can distinguish between trajectories of mobility for those users that we classify as migrants and for those who continue to reside in the same country. The unique features of the data set allow us to tackle one dimension of the important issue of the relationship between international and internal migration.

We use the radius of gyration as a measure of the distance covered by users over a certain period of time [3]. The radius of gyration is the average distance of the geolocated tweets from the baricenter. In particular, we evaluate this quantity for subsets of the population (e.g., migrants vs non-migrants in their home countries) to assess similarities and differences in the experiences across countries.

3 Preliminary Results

Figure 1 shows out-migration rates for selected countries, and the average of out-migration rates for OECD countries. These estimates were obtained using the methods described in the previous session. The time series shown in Figure 1 are the starting point to generate difference-in-differences estimates of recent trends.

Figure 2 shows estimates of $\hat{\delta}$s for out-migration rates for OECD countries (for which we consistently have a sample of at least 100 Twitter users for each period of four months). The results shown are the average of $\hat{\delta}$s evaluated for the periods May-Aug and Sept-Dec of 2011 against the estimates for the respective periods in 2012. Positive values indicate a relative increase in out-migration rates.

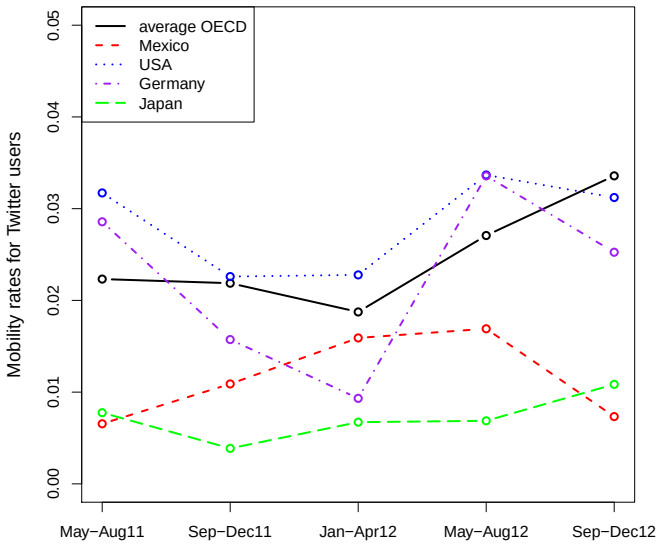

Figure 1: Out-migration rates for selected countries, and the average of out-migration rates for OECD countries. The rates are computed by evaluating the most frequent location of a user over the course of consecutive periods of four months.

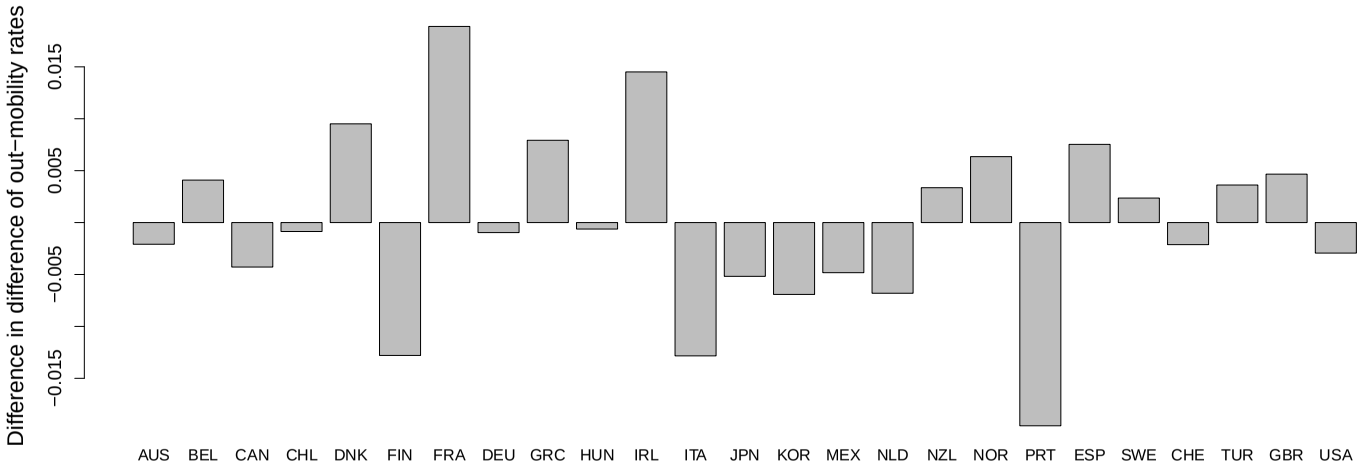

Figure 2: Values of the difference-in-differences estimator $\hat{\delta}$ for out-migration rates for OECD countries for which we consistently have a sample of at least 100 Twitter users for each period of four months.

The results provide interesting insights. For instance, they indicate a decline in out-migration rates from Mexico to other countries. This type of information would show up in official statistics only with a considerable delay. As a result, projections of migrations from Mexico to the US may tend to overestimate flows unless information about recent trends is incorporated.

Preliminary estimates of the radius of gyration for non-migrants and migrants, in their countries of origin, indicate that the distance from the baricenter is larger for larger countries, as expected. There is some heterogeneity. For most countries, international migrants, when in their home countries tend to travel shorter distances than people who did not migrate internationally. For some countries like the US, Greece and New Zealand, the opposite is true.

As we develop our project, we expect to be able to provide more detailed analyses and statements about trends and about the relationships between mobility within and across countries. We also expect to be able to provide more information about the demography of our sample, age-specific trends, and comparisons with existing statistics. Information about gender and age of users will be obtained with the use of the face recognition software Face++ applied to profile pictures of users.

References

[1] G. J. Abel. Estimating global migration flow tables using place of birth data. *Demographic Research*, 28(18):505–546, 2013.

[2] J. J. Azose and A. E. Raftery. Bayesian probabilistic projection of international migration rates. *arXiv preprint arXiv:1310.7148*, 2013.

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

[4] J. E. Cohen, M. Roig, D. C. Reuman, and C. GoGwilt. International migration beyond gravity: A statistical model for use in population projections. *Proceedings of the National Academy of Sciences*, 105(40):15269–15274, 2008.

[5] J. De Beer, J. Raymer, R. Van der Erf, and L. Van Wissen. Overcoming the problems of inconsistent international migration data: A new method applied to flows in europe. *European Journal of Population*, 26(4):459–481, 2010.

[6] R. v. d. Erf. Analysis of final results. In *Paper prepared for the IMEM progress meeting on 22-24 February in Asker, Norway*, 2012.

[7] R. Lee. The outlook for population growth. *Science*, 333(6042):569–573, 2011.

[8] A. E. Raftery, N. Li, H. Ševčíková, P. Gerland, and G. K. Heilig. Bayesian probabilistic population projections for all countries. *Proceedings of the National Academy of Sciences*, 109(35):13915–13921, 2012.

[9] B. State, I. Weber, and E. Zagheni. Studying inter-national mobility through ip geolocation. In *WSDM*, pages 265–274, 2013.

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