Armstrong et al. *EPJ Data Science* (2021) 10:1
https://doi.org/10.1140/epjds/s13688-020-00254-7

EPJ Data Science
a SpringerOpen Journal

REGULAR ARTICLE Open Access

Challenges when identifying migration from geo-located Twitter data

Caitrin Armstrong[1*], Ate Poorthuis[2], Matthew Zook[3], Derek Ruths[1] and Thomas Soehl[4]

*Correspondence:
caitrin.armstrong@mail.mcgill.ca
[1]School of Computer Science, McGill University, Montréal, Canada
Full list of author information is available at the end of the article

Abstract

Given the challenges in collecting up-to-date, comparable data on migrant populations the potential of digital trace data to study migration and migrants has sparked considerable interest among researchers and policy makers. In this paper we assess the reliability of one such data source that is heavily used within the research community: geolocated tweets. We assess strategies used in previous work to identify migrants based on their geolocation histories. We apply these approaches to infer the travel history of a set of Twitter users who regularly posted geolocated tweets between July 2012 and June 2015. In a second step we hand-code the entire tweet histories of a subset of the accounts identified as migrants by these methods. Upon close inspection very few of the accounts that are classified as migrants appear to be migrants in any conventional sense or international students. Rather we find these approaches identify other highly mobile populations such as frequent business or leisure travellers, or people who might best be described as "transnationals". For demographic research that draws on this kind of data to generate estimates of migration flows this high mis-classification rate implies that findings are likely sensitive to the adjustment model used. For most research trying to use these data to study migrant populations, the data will be of limited utility. We suspect that increasing the correct classification rate substantially will not be easy and may introduce other biases.

Keywords: Migration; Twitter; Global human mobility

1 Introduction: the promise of social media data to study international migration

For all the attention it attracts, international migration is a minority phenomenon; 97% of the world's population lives in their country of birth [1]. At the same time, in many developed countries of the global north a non-trivial share of the population is foreign-born ranging from about 10% in Europe, 15% in the US to 20% in Canada and almost 30% in Australia.[a] In most of these countries immigration will be an important aspect of population development for the foreseeable future. Similarly foreign workers, often with precarious legal status, make up a much larger share of the population in some of the Gulf states where in some cases more than three-quarters of the population is foreign born.

Being by definition a mobile population, and, in most countries, constituting a relatively small share of the population, migrants are a hard demographic to study. Although developed states generally have the capacity to collect statistics on their migrant populations, the definition of who is a migrant vary across countries making comparisons difficult, and even in the best of cases there is a substantial lag between migration flows and the availability of migration statistics which makes it especially hard to study and track the sudden population movements that occur in response to political crises. In the global south, where in fact a large majority of the migration occurs, the situation in terms of data availability, accuracy, and access is worse still.

Given these challenges, approaches that draw on digital trace data have sparked the interest of social scientists, governments, and international organizations alike promising to deliver timely estimates of migration flows measured in a consistent way across countries [2, 3]. Beyond studying population movements, some of these data might also provide new ways to study post-migration processes like assimilation and cultural change [4] or even track online hate speech against immigrants [5]. There is now a substantial body of work on this topic and a number methods have been proposed for detecting migrants and migration flows from digital traces [6–9]. Righi provides a comprehensive summary of work to date, referencing some 50 works in the area [8].

Given this work and excitement about the potential of this data it is surprising, that there seems little in terms of research that *directly* evaluates whether or not these methods work in the sense that they actually find migrants. When various data and data processing techniques are evaluated this is, with the exception of one paper that coded a small (N = 37) subset of Twitter users [10], done on the aggregate—that is by comparing estimates of migrant populations or migration trends obtained using digital trace data to estimates based on official statistics or surveys.

Filling this lacuna we use detailed hand-coding to test whether a method proposed in previous research as well as a new, theoretically optimal, method, actually find migrants. Our results are disconcerting: a detailed hand-coding of entire tweet histories shows that both methods we consider have a high false positive rate, classifying what appear to be highly mobile individuals as migrants. Although our results draw from data centered on one country only we have reason to think that the same issues arise in other contexts as well. In fact Canada, given its comparatively high rate of permanent migration, might be a relatively favourable context. We argue that one reason for this high error rate is the fact that migration is a very small share of global human mobility and that twitter users are very select, and skew towards being a highly mobile population.

2 Related work and challenges

Researchers have experimented with a number of ways to leverage digital traces to create new sources of data for migrants with each approach presenting unique opportunities but also significant challenges. Principally one can distinguish data that provide information on populations or "groups" but do not allow individual level analysis and data that provide information on individuals. An example of the former is research that leverages the audience estimates generated through Facebook's advertising platform as a "digital census" of sorts to create estimates of migrant stocks [11], or even more specifically to track changes in interests in migrant populations [4]. This work uses the fact that Facebook classifies expatriate populations by national origin as one of the many advertising audiences. The

exact classification however remains a "black box" since both the algorithm and the underlying data are proprietary. And although impressive in its reach and ability to reproduce, with some adjustment, figures from census data, a key limitation of this data is that it does not allow researchers to obtain individual level information constraining the analysis to national-origin or ethnic groups. For example, research that uses these data to study immigrant acculturation [4] has to fall back on group-level analysis, which stands in tension with the state of the art in the social sciences which try to move away from analysis that take groups as the basic constituents of social life [12].

In order to obtain individual level samples of migrants researchers have used a number data sources each representing it's own challenges. One approach that builds on individual level data uses geolocation of IP addresses. Tracking users through repeated logins on accounts such as web-based e-mail accounts one can then estimate migration flows [13] [14]. Accurate at the country level and in the case of e-mail data with high frequency the data provide a way to track population movements almost in real time. However, the kind of data that contain IP address information are generally proprietary and, as in the case of e-mail, contain sensitive private information. Academic researchers normally do not have access to these data limiting research and especially transparency and reproducibility. Location data gathered from cell-phone records similarly provides high frequency and accuracy but does not provide much in terms of demographic information on the users and comes with similar access challenges and is generally not available to researchers. Since cell-phone networks are generally national these data are, if available, used to examine within-country mobility patterns [15].

Data from social networks are an attractive alternative when they can be gleaned from public profiles and researchers have used the now defunct Google+ social network [16] and LinkedIn [17] to study migration patterns. The most popular source in that regard is geo-located twitter data. The data is publicly available and samples of tweets can be accessed relatively easily through an API. A variety of location data is embedded in tweets and user profiles [18] some of which, as we discuss below, is more reliable than others. Although the Twitter population is clearly a non-random sample of the larger migrant population and even more so when we restrict ourselves to those users who enable the geo-tagging feature in Twitter [19], the unique features of the data one could collect from streams of tweets may well worth be the trade-offs. For population-level research if we can find reasonable adjustment factors we could still infer migration streams at the aggregate level. Even for research that has the aim to study individual migrants the selection bias may be offset by the unique characteristics this data could afford. For example, if it was possible to identify migrants on Twitter and pinpoint their time-window of migration with reasonable certainty one could then analyze Twitter stream before and after migration as an indicator of social involvement in different spheres or even track the network of followers over time.

There is a wealth of literature regarding online user geolocation—usually on Twitter. The task is to identify a user's location—usually implicitly defined as the user's home location. There are two general approaches to a solution—social network based [20] and text-based [21]. While current state of the art on this task has advanced to over 98% prediction capability for large-region prediction [22], the task described is not the same as the *migration* task: neither accommodates changes in location nor integrates the concept of a permanent change in home location as seen in the migration task. Both tasks take

a users' declaration of location, either through the Twitter location field or through geo-tagged content as a potential solution. To identify the location of the remaining majority of users who do not share such information, social network solutions work by focusing on a user's friends and followers who have provided this information, making the assumption that a user's network will have enough co-located users to provide a reasonable inference as to their location. To our knowledge these approaches have so far only been used to identify a single location per user, rather than an explicit *permanent or long term change* in home location. But, if these data are to form the base for research on migrants or post-migration processes these approaches have liabilities as they include features that are likely to change with the migration and integration process to identify the study population. For example using social networks may systematically miss those migrants that use Twitter to stay engaged with social circles 'back home'.

3 Why identifying migrants is hard

Before proceeding it is worth stepping back and consider the larger process by which geo-located twitter data is generated, observed and how that maps onto the challenge of studying migration.

Surprisingly, scholars who study migration rarely define what distinguishes migration from other forms of human mobility. While some make distinctions between short-term (one year or less) and long-term migration, there is generally agreement that migration involves a change of the usual place of residence—at the same time there is broad agreement that legal status and the precise impetus for movement or even the length of stay can note be used as defining characteristics. For example "The UN Migration Agency (IOM) defines a migrant as any person who is moving or has moved across an international border or within a State away from his/her habitual place of residence, regardless of (1) the person's legal status; (2) whether the movement is voluntary or involuntary; (3) what the causes for the movement are; or (4) what the length of the stay is." [23]. Yet clearly not all human mobility is migration and at the very minimum when talking about migration we probably want to exclude types of human mobility that do not involve a change of residence such as vacations, business travel, or other short-term mobility.

This heterogeneity in forms human mobility and among the part that we consider international migration overlaps with processes that generate the geo-located data at that is used to make inferences about migration. The probability of a person tweeting and appending geo-location information is clearly not random but highly structured along a number of dimensions including age, socio-economic status, education but also the legal and citizenship status of an individual. Even if just looking at mobile populations it is likely that those fleeing violence, moving undocumented will rarely, if ever, post geo-located tweets. These populations will be systematically missing. On the other hand for those for whom cross-country mobility is a "lifestyle", leisure travellers and jet setting business executives may be dis-proportionally represented.

Thus two related issues need to be considered when using twitter data to study international migration. First, these data will almost certainly under-represent (or completely lack) large aspects of migration. Second, even if data gleaned from twitter can to be used to study those migrants that generate geo-located twitter data—the second problem, correctly identifying them remains. That is we need to distinguish between migration and

other-non migration forms of human mobility (e.g. vacation, business travel) among mobile Twitter users with a reasonable degree of certainty. As we show this is not a simple task and poses both conceptual and data/information issues.

Migration statistics based on administrative data generally draw on data on residence permits or population registries so vacationers in a hotel or visitors to a conference are not included. Most survey-based research on migration similarly sidesteps the issue. Since in almost all cases survey research samples residents only (whether based on registry data, or other sampling frames, random digit telephone dialing etc...) the residence question is already 'solved'. Conditional on being surveyed, a respondent who was born outside of the survey is simply considered an immigrant.

The aspiration of providing more real-time migration data creates the challenge of making inferences about residence based on information about a persons location history. Several challenges arise: First, since migration is defined as a change in residence this means each person has to have one (and only one) residence for any period of time. Conceptually (and operationally) this makes it hard to classify people whose location history does not suggest one single residence over a period of time. For example, individuals may split their time relatively evenly between two places/countries or they might live "on the road" moving from country to country over extended periods of time. Second, it requires differentiating between mobility that represents a change in residence from mobility that does not. For any one instance of mobility this is a daunting task as fundamentally it goes to the intention behind the move. Any one instance of mobility could be a vacation (or other short-term trip) or a permanent relocation. Ascertaining the intention behind a move from location data alone is generally not possible. In fact assessing intention to change residence or not is a critical aspect of migration control and one that states invest considerable resources in. Modern states have sophisticated bureaucratic processes and documentary regimes [24], and collect large amounts of information in visa applications and visa interviews.

Looking retrospectively at the location histories of individuals makes this problem tractable at least in cases where we have location histories with a high frequency—that is where we can pinpoint the location of an individual at narrow time-intervals (say daily) over long periods of time and where one location is clearly dominant over longer stretches of time (relative to the observation frequency). The sparser the geolocated information and the less clearly one location dominates the location history, the less clearly we can distinguish between mobility that involves a change in residence and mobility that does not.

Geolocated Twitter data is unfortunately not ideal in that regard. Overall, only a small percentage of tweets have embedded geolocation information and even for Twitter users who enable geo-tagging on their Twitter application (volunteered geographic information), geo-located tweets tend to come in "bursts" with often extended time periods where we can't locate users. Figure 1 illustrates how there are only very few users who do not have large gaps between consecutive tweets—most have at least one large gap. Indeed, 93% of users in our dataset have gaps in their tweet histories of over 30 days, with 99% having gaps of over 5 days, and 63% having gaps of over 4 months. In addition we suspect that many people use the geo-location feature of Twitter not in the same way all the time but are more likely to post their location when traveling than when in their places of residence.

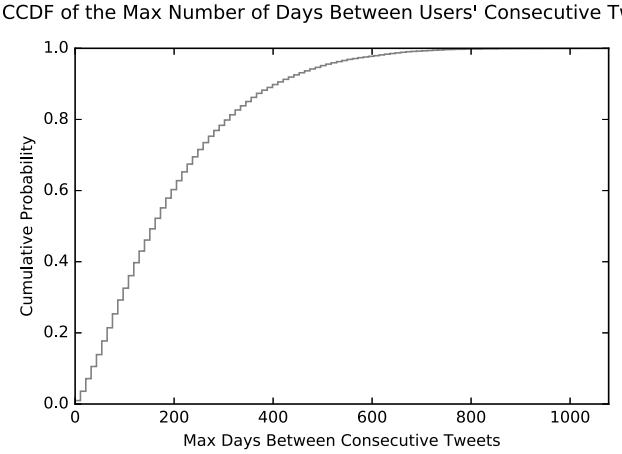

Figure 1 A CDF of the max number of days between consecutive tweets for each user. Only a small share of users have consistent tweets without at least one large gap during out time window—a criterion that is necessary for the migration algorithms developed thus far

A final challenge for identifying migration is that migration that involves long-term changes in residence is a very small part of global human mobility. If three percent of the worlds population are migrants a much smaller share of the world's population will migrate in any given year. Current estimates are that in any five year period about 0.5% migrate [25]. In contrast a substantial share of the world's population will cross international borders as tourists or business travellers. The United Nations World Tourist Organization estimates that there were 1.4 billion international tourists in 2018[b] almost 20% of the worlds population. The ratio will likely be worse still in the global north where tighter border control means migration is relatively rare and international travel and tourism is much more prevalent than in the global south.

4 Data and methods

To directly investigate the prevalence and ease of identification of migrants on Twitter, we use human annotators to evaluate results generated by two migration detection algorithms operating at the level of the individual user. We filter our dataset to maximize our chances of finding migrants in a programmatic manner, and then evaluate the potential migration events in a rigorous high-agreement annotation procedure.

To keep the coding manageable and consistent we focus on one case—Canada. In many ways Canada presents a best-case scenario for using twitter data to find migrants. Since it is a country in the global north and due to its selection of highly skilled migrants likely has a relatively larger share of migrants who use twitter. Given that among OECD countries Canada has, relative to its population, one of the largest inflows of permanent migration, the ratios are unlikely to be more favourable in other countries. Still the odds are daunting. A generous accounting would give us just under 850,000 persons who came to Canada on a reasonably long-term basis in 2016. This includes just under 300,000 permanent residents and a roughly equal number of temporary foreign workers (286,000) and counts everyone who holds a foreign student permit (266,000) even though not all students will migrate in one year.[c] This number is a small share of the overall number of individuals crossing in and out of Canada in one year. Comprehensive statistics are hard to come by but excluding

any trips by immigrants there were about 83 million trips to Canada in 2016.[d] So in the most generous accounting, migration represents about 1 percent of this mobility.

4.1 Data collection and pre-processing

We obtained data from the Digital Online Life and You (DOLLY) database which ingested all geo-tagged tweets during the study period. The system achieves this by defining seven parallel bounding boxes with each covering a continent. By narrowing the stream of geo-tagged tweets to these smaller geographical areas, the system can use the 1-percent API interface to capture the universe of geotagged tweets. The system also makes use of an elevated 10% streaming API access ('gardenhose') to get an estimate of the total amount of messages (including non-geotagged tweets) sent each day, as well as the relative share of geotagged tweets within the overall corpus of messages. Using this approach the DOLLY project began collecting approximately 10 million geotagged tweets every day since June 2012 [26].

As discussed, we center our analysis on Canada and define our study period as the 36 months from July 2012 to June 2015. From this data we selected all Twitter users who have tweeted 10 or more times from Canada during the last 12 months of the study period. This results in approximately 171,000 unique users. From within this group, we then selected all users who made 10 or more tweets from outside of Canada during the entire 36 month period. This results in a group of approximately 52,000 unique users. For each of these users, we extract their entire Twitter activity during the study period.

As found in previous research using Twitter data to study migration, there is a trade-off between larger sample that has sparse geolocation information and a smaller sample that has more information [6] In light of the issues discussed in the previous section and in order to make our analysis comparable to previous work we decided to use a sample for whom we have consistent information and select those users for whom we have at least one geolocated tweet per month (30 day period) for their entire Twitter history. This dramatically reduces the sample to 3413 accounts. We further remove 61 accounts that were located in 100 countries or more as well as all accounts that were since deleted bringing the final sample to 2961 users. Figure 2 provides a density estimation (distribution) of the average days between tweets for our sample, illustrating that while those that were eventually identified as migrant or non-migrant were similar, those users that we excluded had longer average gaps between tweets.

4.2 Parsing location histories to identify migrants

Working with this sample we then applied two representative approaches (modal and ResIn) to infer residence histories from the location data. While there are many different methods that have been devised to infer user residence patterns from geolocated data, we chose these two methods because (1) the modal method is among the most commonly-used methods in the literature and (2) ResIn was recently proven to infer the most accurate residence history, given an assumption of parsimony (which effectively all existing methods use).

4.2.1 Modal tweet approach

The first approach which we will call the "modal tweet" method follows work by Zagheni and colleagues [6]. We divide the location history into three months intervals. The location

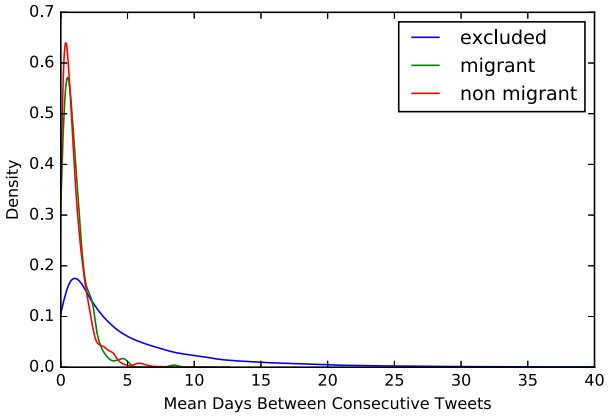

Figure 2 A KDE of the average number of days between tweets for those users that were excluded through our 30 day criteria, those that were eventually classified as migrants through either of the two methods, and those that were not classified as migrants. The x axis has been constrained to 40 tweets maximum to reduce graph scale, the max value extends to 140 along the same horizontal line

where the modal tweet is sent during that three months interval is assumed to be the residence of the individual during that period. If a user is assigned to different countries in two consecutive time periods this is classified as a change in residence.[e]

4.2.2 Residence inference algorithm (*ResIn*)

The second approach is the ResIn method [27]. This method uses a dynamic programming approach to find the optimal residence history for a given sequence of location data. Here "optimal" refers to finding a residence history in which the user spends the most time in her home locations (as defined by the residence history). As a result, ResIn is able to efficiently assess all possible residence histories and identify the one which maximizes the user's time in their home location. At a high-level, the critical distinction between ResIn and the modal tweet method is that ResIn returns the best residence history whereas the modal tweet method is a heuristic and returns a good, but rarely optimal residence history.

4.2.3 Results

Starting from our sample of 2961 users, 2151 were identified as not having changed residence by both approaches and 249 accounts where identified by both algorithms as having changed residence. In addition, the modal tweet method classified 510 users as having changed residence that were not classified as migrants using the ResIn approach. In turn, the ResIn approach identified 51 accounts as migrants that had not been classified as such by the modal tweet approach.

For those users that were classified as having changed residence, we construct a time-window in which the move might have occurred. For ResIn the estimated moving date is 30 days before they were identified as last being in their former and 30 days after they last entered the new country, recognizing that users may travel back and forth for some time. For the modal algorithm the estimated range was within the 120 day range where a move was estimated to occur.

4.3 Verification—manual coding of user accounts

4.3.1 Process

In the final step we set out to verify whether or not the two approaches in fact did identify migration (changes in residence). To do that we hired a team of undergraduate students and trained them to verify whether the changes in location identified by the algorithms migration actually represent instances of migration or not. Wherever possible coders were asked to review the entire tweeting history of the accounts paying special attention to the time-frame where the migration should have occurred according to the classification by the algorithms. In a few cases where users were tweeting extremely frequently we reviewed only parts of the twitter history. The coding was done retro-actively—during the academic year 2018-2019.

Initially we double coded each profile and switched to single-coding accounts once we had established high inter-coder reliability (more than 90 percent of codes identical). We covered a number of languages including French, English, Spanish, Italian, Turkish, Arabic, Japanese, and Portuguese leaving us with only 6 accounts we could not code for language reasons. We coded a total of 164 unique accounts in this fashion.

Since both algorithms identified a good share of individuals as moving several times and make these assignments (including the time-window of the move) slightly differently this corresponds to a total 431 putative changes in residence.

For a subset of 48 accounts coders were also asked examine the tweets in detail and to give their best assessment of the type of mobility, if any, based on a close reading Twitter timeline. We distinguish vacation/leisure travel and work business trips from migration. A fourth category captures those where coders could not find any evidence for mobility in the tweet history. In addition, we coded accounts that were private and tweeters that were too prolific as to assess the tweeting history with reasonable effort. In addition we also asked the coders to provide a brief characterization of the account user drawing on their reading of the tweet history as well as the account profile.

Our annotation procedure involved considerable time and effort into reading and coding the Twitter activity of each account, and care was take to establish high agreement. Nevertheless, determining whether or not a Twitter user migrated based on a tweet timeline inevitably involves judgement calls. However, as we show, the results of this hand-coding exercise are clear enough that we are confident that the substantive conclusions we draw are robust.

4.3.2 Findings

In all these accounts we found a total of seven individuals who our coders classified as possibly having relocated permanently. Whether or not students who spend a semester abroad should be counted as migration is debatable, but especially longer term exchanges or years abroad would fit the definition of residence change both the ResIn and the Modal Tweet approach were designed to detect. To address this we provide a second definition that includes student exchange as migration. As summarized in Table 1, this corresponds to a correct classification rate between 14% including students and 6% when only taking those accounts where we did find signs of permanent migration. The rates are remarkably consistent across both approaches. In sum, the large majority of mobility that is detected by these methods is almost certainly not migration.

Of the randomly selected set of just under 50 accounts we coded in more detail, we could determine the type of mobility for about two thirds of the Modal Tweet classified

Table 1 Percent of accounts and instances of mobility identified as migration by coders by approach

	Modal tweet		ResIn	
	Accounts	Moves	Accounts	Moves
Permanent	6%	3%	7%	1%
+ Students	14%	8%	14%	2%
N	114	212	100	219

Table 2 Detailed coding of type of mobility for 48 accounts identified as migrants by Modal Tweet and ResIn approach

	All accounts ($N = 48$)		Exclude not able to code ($N = 39$)	
	Modal Tweet	ResIn	Modal Tweet	ResIn
Leisure	36%	22%	46%	27%
Business	21%	28%	26%	33%
Study abroad	5%	2%	7%	2%
Migration	5%	2%	7%	2%
No mobility	12%	31%	15%	37%
Not able to code	21%	16%		
N	58	58	46	49

accounts and for about half of the ResIn accounts. For 15 accounts we could not make a determination as either tweets were deleted, private, or there were too many tweets to read through the timeline with reasonable effort. These not-coded accounts correspond to 21% and 16% of the identified moves for the Modal tweet and ResIn accounts respectively. For another 12% of the Modal tweet and 31% of the ResIn identified moves our coders could not see any sign of mobility in the tweet timeline. Due to the small sample size in this step one should not over-interpret these differences but one reason might be that the time-window that the ResIn algorithm identifies for the move is narrower than the time-window we can derive from the modal tweet algorithm. While we cannot pin down the type of mobility we observe in these time windows we are fairly certain that it is not migration or even long-term student exchanges but rather short term trips that left little trace on the Twitter timeline.

While the exact proportions are not meaningful due to small sample size, the overall pattern is clear: only a small minority (14% by the most expansive definition) of moves represent migration with the large majority being business or leisure trips. We suspect that this is at least in part due to the fact that people tweet more frequently during such trips as compared to their "normal" life. This would, for example, create modal tweets in places where people spend a relatively short amount of time and similarly suggest residences to the ResIn algorithm in places that are in fact simply vacation locations or destinations of business trips.

5 What kinds of mobility can we capture with geolocated Twitter data?

As this analysis has shown using a variety of coding approaches only a small share of moves classified as changes in residence appear to actually be migration as commonly understood but rather other types of mobility. So, if not migrants, what populations do these approaches capture? Given the relatively small size of the sample we coded in greater detail, our findings here are clearly exploratory. Nevertheless beyond simply mis-classifying vacations and business trips as migrations, some patterns appear in the data.

Given the public nature of Twitter (or at least the unprotected Twitter accounts we were able to code) and the fact that we filter the data for those who have at least one geolocated tweet per month, which in general implies a high frequency of tweeting, it is not surprising that the population we capture tends extrovert. For some a combination of frequent travel and an public facing persona appears to be part lifestyle and part professional freelancers or consultants in creative/cultural industries. Some travel mis-classified as migration was indeed migrants but the mobility picked up was not migration but return trips to visit friends and family.

We also found several examples of people who do not have to appear a single residence at least during the period of our study. One set of examples were people living on the road in the literal sense for example working for musicians on tours. Other examples resembled the ideal-type of the "transnational global citizen" seemingly living in two (or more) countries at the same time and seamlessly moving across borders. Future work could break down these categories similar to the typology identified by Liao et al. [28]. Figure 3 provides an illustration of several such accounts.

6 Conclusion

The existing work using geo-located Twitter data to study migration has focused on estimating migrant flows between countries for example using geo-located Twitter data as a signal to infer estimates of migration flows using demographic and statistical adjustment models [6]. For this type of work, having a fairly imprecise classification might be acceptable as long as the processed data does provide some signal. That said, if the rate of false positives really is in the range we find in our analysis (between 85 and 95%) then the signal would be quite noisy and a lot of a study's outcome rides on the method chosen to make inferences.[f]

On the one hand, for studying migrants or post-migration processes, our study suggests that these approaches are likely not useful unless the accurate detection of migrant Twitter users can be improved dramatically. Doing so will prove challenging. If using geotagging approaches, one might require more regular geotagged content—but this will dramatically reduce sample size (by up to 98% according to our study). Imposing longer residence periods will create similar problems—yielding potentially very few accounts classified as potential migrants in the first place.

Further, even if classification issues can be solved, an issue of biased sampling remains. Geotagging based techniques don't allow us to see those who are too marginalized and vulnerable to leave digital traces. Perhaps more concerning is that they don't provide ways of estimating the prevalence of those omitted populations.

On the other hand, even if the utility of Twitter data for studying international migration is limited, it might prove useful for studying the types of mobility we saw more frequently in the data and which are, in fact, a much larger part of global mobility than migration, but that received scant attention from social science researchers so far [29]. Studying the life-worlds of these mobile populations could be an important counter-weight in the very migration-focused study of global human mobility and, more generally, provide insights into types of social life that transcend the nation-state [30]. The data that can be gleaned from Twitter accounts might be a good start for such an endeavor especially when combined with other social science approaches such as surveys or targeted in-depth interviews.

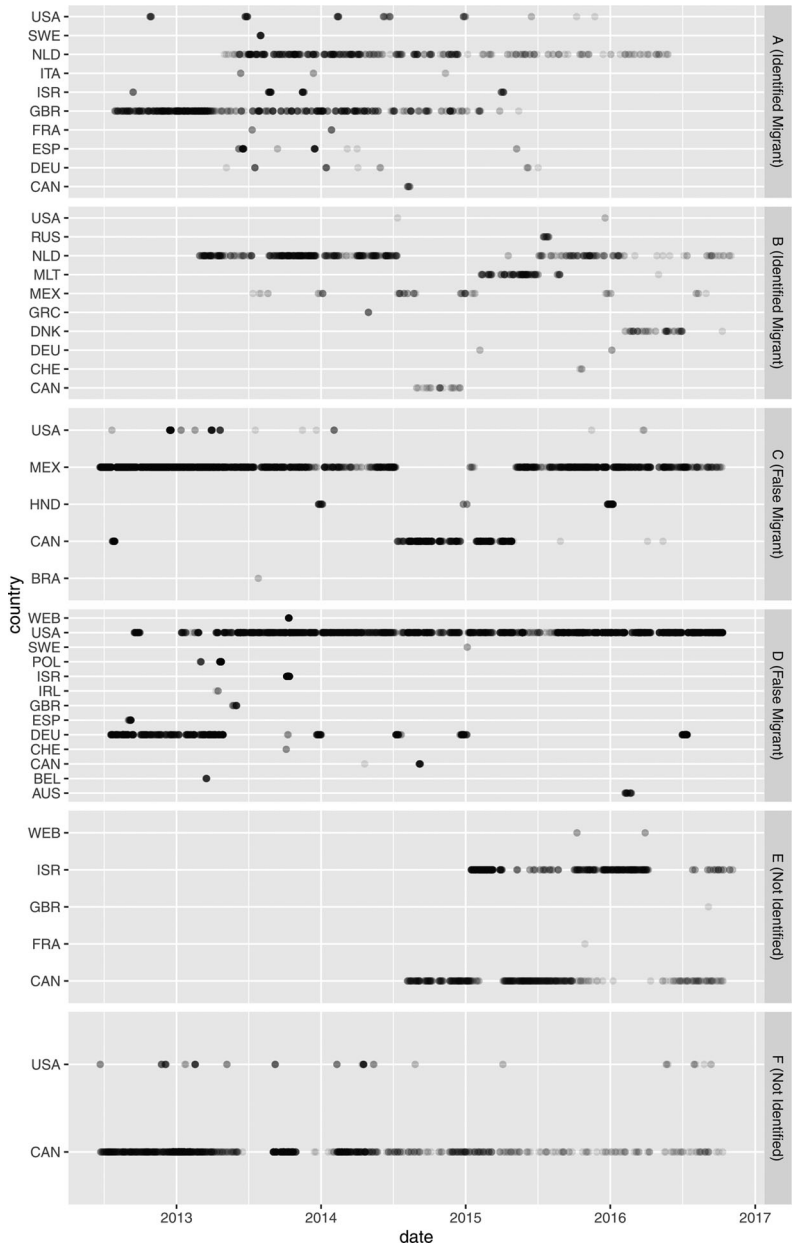

Figure 3 We present the geolocated Twitter timelines of 6 representative Twitter users. We observe that tweet behaviour is bursty, and that users take many short trips to other locations. Users A–D were classified by both algorithms as migrants, however only users A and B were confirmed by our annotators to display an actual migration event. *User A* mentions directly moving to Amsterdam from Great Britain, while travelling for business. At one point they mention (paraphrased), "maybe one day I'll be at my new home for more than 2 days!". This user is a migrant, but seemingly geo-located in both the Netherlands and Britain over an extended period. *User B* was identified as a student migrant. While both algorithms identified several moves, only the initial move from the Netherlands to Mexico was identified by our annotators as a true migration event. In both cases, it is possible that a user was using a VPN to access restricted content from another country. *User C* is a resident of Canada originally from Mexico. Despite the geolocation of the tweets, there was no mention of the user actually being in Mexico, or moving back to Canada. *User D* likely moved from Germany to the United States, however due to the overlap in time periods, neither algorithm identified the correct move window: during the time window flagged as the time of move the user had not yet made the move to the United States, but had only returned for a brief family visit. The algorithm was, however, correct in principle. Finally, users E and F were not identified as migrants by either of the algorithms. *User E* is an artist who often travels to Israel to paint, but mentions living in Montreal, Canada. *User F* is a business professional who takes occasional trips to the United States for work

This raises a final point, which is that one notable opportunity for future work is the automated disentangling of migrants and otherwise-mobile individuals. Certainly those who migrate and those who are hyper-mobile travel with different intentions—which may well be encoded in the timing, content, and other metadata of tweets. One potentially productive formulation of this exercise would be as a graphical model in which tweets are observations, the latent variable being the type of individual (migrant vs. mobile individual). Such a method, built on top of the methods considered here, could well contribute to both lines of research on global human mobility.

Acknowledgements
For diligent coding of twitter accounts we thank Tiphaine Le Corre, Philippe Chasse, Nour Daoud, Lucy Martinez-Gesualdi, Akiho Nitahara, Sarin Akpulat, and Lisiane Freitas Leal.

Funding
Financial support was provided by an Social Sciences and Humanities Emerging Scholars Accelerator Grant awarded by McGill University to Thomas Soehl.

Availability of data and materials
Due to privacy concerns our annotated dataset is not openly available.

Competing interests
The authors declare that they have no competing interests.

Authors' contributions
DR, CA and TA designed the research; PA and MZ provided data and pre-processed data; CA analyzed the data; CA and TS developed coding protocol and supervised coding; CA, DR and TS wrote the manuscript; all authors read and approved the final manuscript.

Author details
[1]School of Computer Science, McGill University, Montréal, Canada. [2]Department of Geography, Singapore University of Technology and Design, Singapore, Singapore. [3]Department of Geography, University of Kentucky, Lexington, United States. [4]Department of Sociology, McGill University, Montréal, Canada.

Endnotes
[a] https://data.worldbank.org/
[b] https://ourworldindata.org/tourism
[c] https://www.canada.ca/en/immigration-refugees-citizenship/corporate/publications-manuals/annual-report-parliament-immigration-2017.html
[d] https://www150.statcan.gc.ca/n1/pub/66-001-p/2017012/tbl/tbl-1-eng.htm
[e] The exact window over which to aggregate here is somewhat arbitrary—we chose three months as this corresponds to the commonly used distinction between visitors and longer terms stays as for example codified in visa-waiver programs where for stays less than 90 days visa applications are not necessary but still required for stays longer than 90 days.
[f] The fact that migration rates calculated from these approaches correlate highly with migration rates derived from administrative data or other official statistics is not necessarily solid evidence in this case as other types of mobility such as leisure or business travel will often correlate with migration links as well as predicted for example by world-systems approaches to international migration [31].

Publisher's Note

Received: 16 April 2020 Accepted: 18 November 2020 Published online: 07 January 2021

References

1. United Nations (2017) International migration report 2017: highlights
2. Weber I, State B (2017) Digital demography. In: Proceedings of the 26th international conference on world wide web companion. WWW'17 companion. International world wide web conferences steering committee, Republic and Canton of Geneva, pp 935–939. Switzerland
3. Hughes C, Zagheni E, Abel GJ, Sorichetta A, Wi'sniowski A, Weber I, Tatem AJ (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
4. Dubois A, Zagheni E, Garimella K, Weber I (2018) Studying migrant assimilation through facebook interests. ArXiv preprint. arXiv:1801.09430

5. Bosco C, Patti V, Bogetti M, Conoscenti M, Ruffo GF, Schifanella R, Stranisci M (2017) Tools and resources for detecting hate and prejudice against immigrants in social media. In: Social interactions in complex intelligent systems, pp 79–84. AISB
6. Zagheni E, Garimella VRK, Weber I, State B (2014) Inferring international and internal migration patterns from Twitter data. In: Proceedings of the 23rd international conference on world wide web. ACM, New York, pp 439–444
7. Fiorio L, Abel G, Cai J, Zagheni E, Weber I, 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. ACM, New York, pp 103–110
8. Righi A (2019) Assessing migration through social media: a review. Math Popul Stud 26:80–91
9. Nguyen H, Garimella K (2017) Understanding international migration using tensor factorization. 1702.04996
10. Hübl F, Cvetojevic S, Hochmair H, Paulus G (2017) Analyzing refugee migration patterns using geo-tagged tweets. ISPRS Intl J Geo-Inf 6(10):302
11. Zagheni E, Weber I, Gummadi K (2017) Leveraging Facebook's advertising platform to monitor stocks of migrants. Popul Dev Rev 43(4):721–734
12. Brubaker R (2004) Ethnicity without groups. Harvard University Press, Cambridge
13. Zagheni E, 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. ACM, New York, pp 348–351
14. Weber I, Zagheni E et al (2013) Studying inter-national mobility through ip geolocation. In: Proceedings of the sixth ACM international conference on web search and data mining. ACM, New York, pp 265–274
15. Hong L, Wu J, Frias-Martinez E, Villarreal A, Frias-Martinez V (2019) Characterization of internal migrant behavior in the immediate post-migration period using cell phone traces. In: Proceedings of the tenth international conference on information and communication technologies and development. ICTD'19. ACM, New York, pp 4–1412
16. Messias J, Benevenuto F, Weber I, Zagheni E (2016) From migration corridors to clusters: the value of google+ data for migration studies. 1607.00421
17. Rodriguez M, Helbing D, Zagheni E et al (2014) Migration of professionals to the us. In: International conference on social informatics. Springer, Berlin, pp 531–543
18. Dredze M, Paul MJ, Bergsma S, Tran H (2013) Carmen: a Twitter geolocation system with applications to public health. In: AAAI workshop on expanding the boundaries of health informatics using AI (HIAI), vol 23, p 45
19. Malik MM, Lamba H, Nakos C, Pfeffer J (2015) Population bias in geotagged tweets. In: Standards and practices in large-scale social media research: papers from the 2015 ICWSM workshop. AAAI Press, Menlo Park, pp 3–759
20. Jurgens D, Finethy T, McCorriston J, Xu YT, Ruths D (2015) Geolocation prediction in Twitter using social networks: a critical analysis and review of current practice. In: Ninth international AAAI conference on web and social media
21. Graham M, Hale SA, Gaffney D (2014) Where in the world are you? Geolocation and language identification in Twitter. Prof Geogr 66(4):568–578. https://doi.org/10.1080/00330124.2014.907699
22. Bakerman J, Pazdernik K, Wilson A, Fairchild G, Bahran R (2018) Twitter geolocation: a hybrid approach. ACM Trans Knowl Discov Data 12(3):1–17
23. International Organization for Migration (2019) Who is a migrant. Available at https://www.iom.int/who-is-a-migrant
24. Torpey JC (2018) The invention of the passport: surveillance, citizenship and the state. Cambridge University Press, Cambridge
25. Abel GJ, Sander N (2014) Quantifying global international migration flows. Science 343(6178):1520–1522. https://doi.org/10.1126/science.1248676
26. Poorthuis A, Zook M (2017) Making big data small: strategies to expand urban and geographical research using social media. J Urban Technol 24(4):115–135
27. Ruths D, Armstrong C (2020) The residence history inference problem. 2003.04155
28. Liao Y, Yeh S, Jeuken GS (2019) From individual to collective behaviours: exploring population heterogeneity of human mobility based on social media data. EPJ Data Sci 8(1):34
29. Favell A, Feldblum M, Smith MP (2007) The human face of global mobility: a research agenda. Society 44(2):15–25
30. Mau S (2010) Social transnationalism: lifeworlds beyond the nation-state. Routledge, Abingdon
31. Massey DS, Arango J, Hugo G, Kouaouci A, Pellegrino A (1999) Worlds in motion: understanding international migration at the end of the millennium: understanding international migration at the end of the millennium. Clarendon Press, Oxford