Understanding Patterns of Human Mobility at Different Time Scales

Lee Fiorio, Emilio Zagheni, Guy Abel, Johnathan Hill,
Gabriel Pestre, Emmanuel Letouzé, Jixuan Cai

September 30, 2017

Abstract

Recent decades have seen an explosion in the quantity of behavioral data generated by human interaction with digital devices. A growing body of literature has focused on the value and potential pitfalls of leveraging these "digital trace data" to analyze social processes including human migration and mobility, but blind spots remain. One challenge, well known to migration scholars, is to standardize and compare different kinds of movements across different time and geographic scales. In this paper, we develop a methodology for parsing the population-level migration signal from individual-level point-in-time data using flexible time-scales. We propose a stochastic model for simulating patterns in digital trace data and test it against three datasets: geo-tagged Tweets and Gowalla check-ins in the U.S.; cell phone call detail records in Senegal. Similar patterns observed across all three empirical datasets demonstrate the utility of our approach for studying migration via digital trace data.

Introduction

The two main ways in which humans move around–mobility and migration–are typically analyzed as distinct processes, through different methods and literatures. This is in part due to differences in the types of data most relevant and available to study short-term mobility and relocation. In this paper we leverage geo-located data from two social media platforms, as well

as cell-phone data known as call detail records (CDRs) to better characterize and standardize, with a view to comparing and contrasting them, patterns of short-term internal mobility and long-term internal migration in the U.S. and Senegal.

We propose a stochastic model of mobility and relocation and evaluate its properties via microsimulation. In particular, we study how mobility and migration rates vary in response to changes in the definitions of *duration* (the span of time used to establish residency) and *interval* (the temporal distance between two points of reference used to evaluate relocations). Then we test our model on geo-located data from Twitter and Gowalla for the U.S., and CDRs for Senegal.

This paper builds on previous demographic work that relies on register data (e.g., [16]) as well as our previous work with Twitter data [7, 20]. In this article, we deepen the underlying theoretical model as well as expand the empirical basis of the analysis.

First, we offer some background on the approaches developed in demography and in social informatics to measure and model mobility and migration. Second, we describe our theoretical model and the outcomes of the microsimulation. Third, we present our data sets. Finally, we show the results of the empirical analysis as well as discuss them in the broader context of studying human movement.

Findings from this paper demonstrate our ability to isolate long-term migration trends from very large datasets of discrete, individual-level location information. What is presented here is not the application of old methodologies to newer, bigger data but an attempt to design new methodologies that take full advantage of the kinds of datasets made increasingly available to today's digital world. As such, the objective of this research is show that with a handful of simple concepts and standardization procedures, it is possible to study human mobility and migration together as a system of population movement.

Motivation

Despite its importance in human development, migration has received less attention than on the other two drivers of long term demographic change, fertility and mortality. Migration is difficult to study due to issues of data availability and reliability, compounded by divergences, and at times incon-

sistencies, in its very definition. Human populations have always been on the move and a significant body of research has sought to understand the drivers and effects of population movements at the individual, household and societal levels; but the diversity of ways in which people move, be it to commute to work every day or permanently relocate far away, has made this analysis extremely difficult.

The spread of telecommunication technologies in the last twenty years has resulted in a growing number of very large datasets of individual-level mobility information that come in the form of "digital traces" [9]. These new data are a boon to migration scholars. In their most essential form, these datasets are large but simple: hundreds of millions, if not billions of observations consisting of ⟨`individual id, time stamp, location`⟩. Digital trace data can originate from many different sources: call detail records, location information included in social media posts, meta-data associated with smart phone applications or log-ons to email, and so on. Though there is a growing body of literature on the use of these data in the social sciences [8, 11, 6], there is a general lack of standards or best practices for how they can be used to generate population-level estimates of migration [12, 14].

An obvious challenge is the fact that these data reflect all kinds of movements – commuting, vacations, travel for business, travel for study, visits to family – and not just information on definite relocations. To be precise, digital trace data are qualitatively and quantitatively different from administrative data like address registries, arguably the gold standard of migration data. With registry data, an assumption can be made that each observation is unequivocally a change in residence which, given a geography and time scale, could be classified as migration. With digital trace data, an extra methodological step must be taken to parse residency information from the roaring sea of individual-level mobility.

Coming up with a holistic approach to population movement is a daunting task, but we argue and attempt to show that it is nevertheless possible. Specifically, we argue and try to show how the totality of these large data – that is, the short-term mobility information and the long-term migration information taken together – can be leveraged to improve population movement estimation and analysis. We see four main arguments for bringing big data to bear on this critical issue.

1. *More timely estimates:* A clear advantage that big data have over traditional survey techniques is that they are, in theory, more readily

available. Though researchers may not always have access, these data are collected in real time as individuals place calls, post on social media or use applications on the web or on smart phones. Developing standards and efficient methodologies for parsing the relocation or migration signal from large individual-level time-and-place datasets would allow for more timely migration estimates. In comparison to surveys of migration which take a long time to design, implement, and analyze, estimates generated from big data could provide a near instantaneous first-look into migration as it happens.

2. *More flexible estimates and harmonization methods:* Given that different national governments have different legal definitions of migration [4][3] and given that surveys are limited in practice in the number of questionnaire items they can contain, digital trace data are advantageous in that they consist of a relatively stable population onto which different migration definitions can be applied. Information gathered from an analysis of big data could form the empirical basis for harmonization strategies between official statistics that use different duration definitions for migration.

3. *Better forecasting:* Though we distinguish between short-term mobility and long-term migration in practice, they both come from a singular data generating process: people moving around in time and space. In theory, every long-term relocation first appears in these data as a short-term move, and thus changes to migration estimates defined by longer term residences are likely first signaled by changes in migration estimates defined by shorter term residencies. Estimating baseline short-term rates and their variances could potentially allow for the identification of anomalous changes indicative of changes that will eventually manifest themselves in the long-term.

4. *Identifying generalizable patterns in human mobility:* In the sociological and geographic literature there is a theorized link between short-term mobility and long-term migration between particular places. The basic premise is that higher connectivity (e.g. flows of capital, flows of information) and greater commonality (e.g. shared language, history, religion) between two places are associated with higher rates of migration between those places[15]. So it follows that short-term mobility rates

(business travel, vacation travel, family travel) should be positively associated with long-term migration rates. With these large datasets, we can test this hypothesis and begin to explore the nuances of this relationship. For example, we can ask: are there bilateral flows for which short-term mobility is high but long-term migration is relatively low or vice versa? What would such a pattern say about the relationship between two places and how might it impact the measurement and forecasting of the corresponding flows?

In this paper we will focus on arguments (2) and (4). To do so, we propose a stochastic model and test its validity using three datasets: Twitter, Senegal (Orange) and Gowalla. Within each dataset, we calculate many different migration rates from roughly the same body of individuals, every rate with a slightly different temporal definition. This allows us to identify the relationships between different kinds of rates: e.g. one year vs. six month, illustrate the overall pattern and demonstrate the potential for producing standardized rates from digital trace data.

Background

Any study of migration must be explicit about its terminology. While we argue that there is a singular process – the movement of individuals through time and space – that underlies the phenomenon of human migration, we also acknowledge that this process manifests itself in a number of conceptually distinct kinds of data. It should be noted here that this paper focuses on the sensitivity of migration estimation methods to changes in time scale but not in geographic scale. In all discussions made here of movement, mobility, and migration, we hold geography fixed at a very coarse level: aggregates of states in the U.S. context and regions in the Senegalese context. This simplifies the data considerably. We are not interested in determining where a person is except at a very high level (e.g. are they calling from New England? or do their check-ins occur mostly in the western half of the Midwest? and so on.)

Move, Mover, and Transition Data

Setting aside at first the nuances of how 'move' is defined, migration data at their most basic take on one of two forms: 'move' data and 'mover' data [18].

In the former, the unit of observation is a movement: e.g. a count of total airline traffic over a given period. In the latter, the unit of observation is the individual: e.g. a count of individuals who flew at least once over a given period. To study migration probabilistically from a population perspective, mover data are preferred as they allow for the estimation of exposures to the risk of migration [10].

If possible, further refinements can be made to mover data to explicitly estimate the count of individuals who have relocated from a specific place to another over a given period. These are what are known as transition data [17], direction or timing, transition data estimate the number of people who were observed at a different location at the start and end of an interval. Transition data are the preferred data for the study of migration systems as they allow for the estimation of bilateral migration flows between geographic areas. These data are commonly collected in survey data with questionnaire items that ask about the respondent's location a year prior.

Migration and Time Scale

As is apparent from the above description of migration data, the interval over which migration is observed is of critical importance. The length and timing of this period with respect to the calendar year impact the corresponding number of moves, movers, and transitions. However, as was alluded to earlier, the issue of how to define a move is non-trivial and requires an additional but distinct conceptualization around the issue of time. Arguably the most crucial aspect of a migration definition, particularly in the eyes of government entities, is not the timing of move (e.g. whether a move occurred in one year or another) but the length or duration of the residency (e.g. whether residency in a place was maintained for a given span of time).

In registries, the conceptual distinction between these two issues of time still exists, but it can seem redundant. For example, if we know from an address registry that an individual has lived at a location for exactly one year then we also know they lived somewhere else one year ago. That being said, the reverse might not hold: if we know know that an individual lived at a different location one year ago, we do not necessarily know how long they have been at their current place – there could have been multiple moves or the move could have happened very recently. Thus, in order to study transitions using registry data, it is still necessary to be explicit about both the interval and residency criteria for defining a migrant.

This problem is amplified in the digital trace data discussed in this paper: time-and-location data observed at the individual level. Unlike registry data, these data do not directly contain information about residency. Each observation is simply a record of a person at a particular location at a particular time. If we compare two observations of the same individual and find the individual in two different locations, we do not know whether the movement is a vacation or evidence of a permanent relocation. Thus, we must start by inferring residency from the data, then compare inferred residencies at the start and end of an interval in order to estimate migration transitions. As such, we are extra explicit about how we conceptualize time scale in the estimation of migration. We separate 'duration,' the period used for estimating residency, from 'interval,' the period used for estimating whether residency has changed.

The Concept of Duration

On its own, a single data point from one of our datasets contains very little information about the residency of a given user. For this reason, we group together multiple data points by user based on time. More specifically, to infer the residency of a user, we take their modal location over a given time frame. In the case of a tie, we take the first observed location (e.g. if we see a user at three locations, A, B, C with in the following pattern ACABBC, we would assign them to A because that was the first location to hit the maximum observed of two). We call the time frame used to infer residency the *duration*. Figure 1 illustrates how duration is used to estimate migration from user time-lines.

The Concept of Interval

Once we have made estimates of each individual's location given a variety of reference points and duration definitions, we can then compare their location at two points in time to determine whether a transition has occurred. The *interval* is the time spanning the two reference points. An interval of a year, for example, means that we are interested in whether an individual's location at time t is different than their location at time t + one year. In order to avoid double counting observations, we constrain the duration so that it is always less than or equal to the interval. So in the case where we want to know the migration rate given an interval of one year and a duration of six

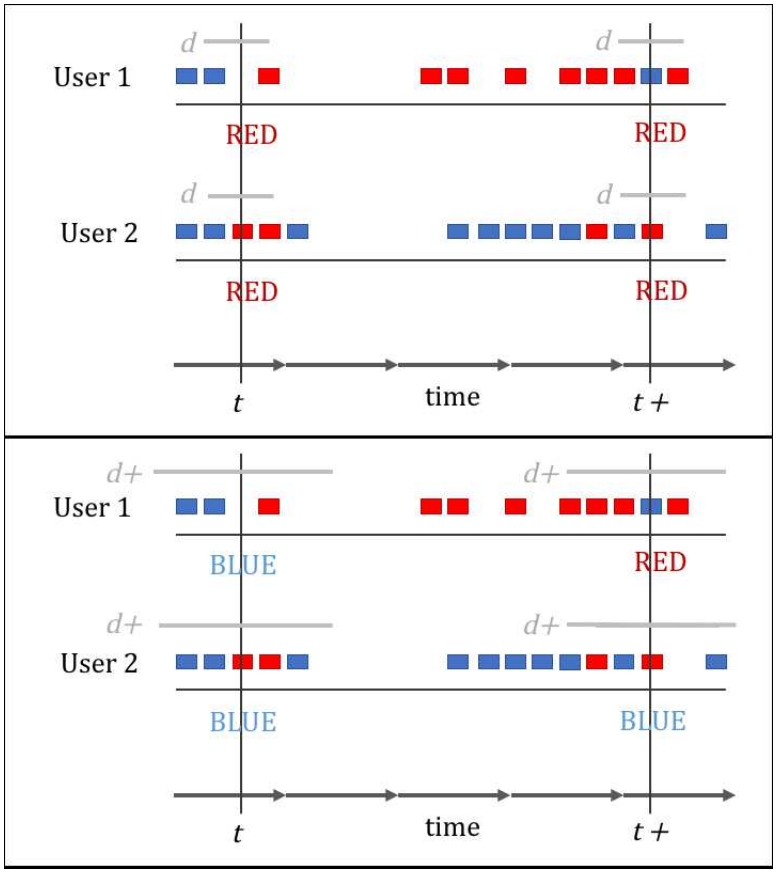

Figure 1: *To estimate whether a transition occurred for a given user, we must infer their residency at two points in time using a given duration. This figure illustrates how the inferred residency and thus the estimated transition might change as the duration increases. Each user posts either from the blue area or the red area. To determine whether the users have changed locations between reference points t and t+, we must infer their residency around these points with a duration. In the upper panel, we use buffer d; in the lower panel we use buffer d+.*

months, for example, we infer each individual location at time t using the time frame [(t – 3 months), (t + 3 months)] and at time t + one year using the time frame [(t + one year – 3 months), (t + one year + 3 months)] and then see if the two locations are the same. Figure 2 shows how interval is used to estimate migration from user time-lines.

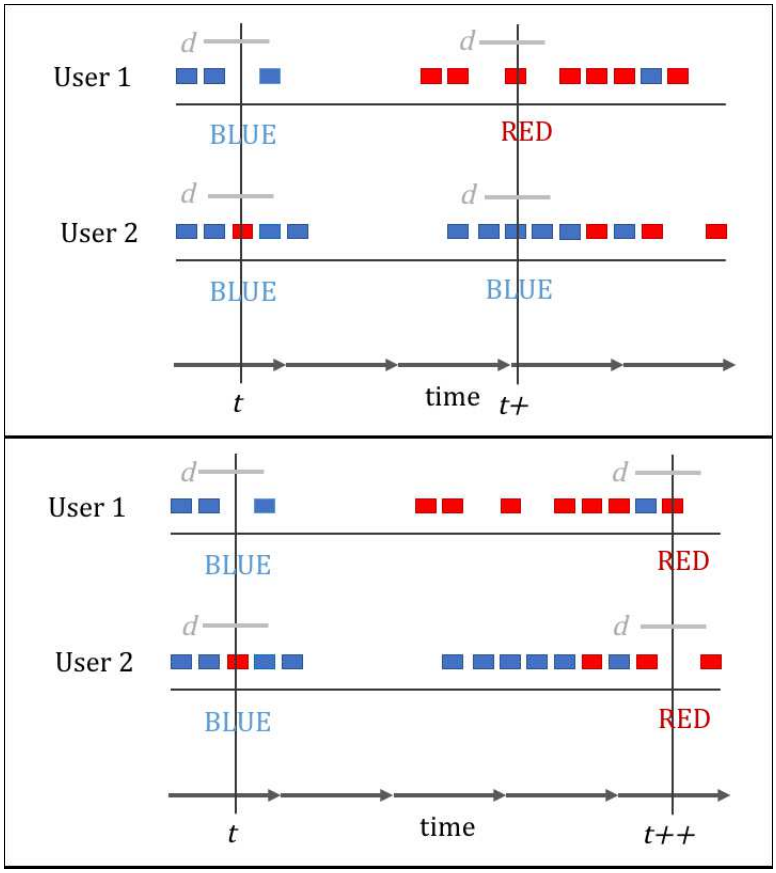

Figure 2: *To estimate whether a transition has taken place, we chose a period of interest. We can vary the length of this period, or interval, to achieve different kinds of rates. This figure illustrates how estimates of transition might change as the interval increases by showing for two users at different reference points determined by the length of the interval, t to t+ or t to t++.*

The Concept of Start

At a high level, there are similarities between the duration/interval approach for estimating migration and the age-period-cohort framework underlying the demographic life table. The most straightforward link is between interval and cohort. If we take a population of individuals and infer their respective residencies at a particular time t_0, we can treat them as a cohort in which having the same residency at the end of the interval is survivorship. As we

extend the interval forward in time (to t_1, t_2, ..., t_n) there is cumulative exposure to the risk of a change in residency. In this comparison, duration corresponds to age. Residential survivorship over a given interval can require a duration of an equivalent length to be maintained. To round out the analogy, we introduce the concept of 'start' as a rough approximation of period. When we fix a reference or 'start' point we will observe all kinds of movement around that point. For some 'start' points we might expect higher rates of movement, either in the short term (seasonal travel patterns) or in the long term (changing levels of permanent or semi-permanent migration). In the U.S. context, for example, a start around the Christmas and New Year holiday will capture a high amount of short term travel.

The analogy between duration/interval and the age-period-cohort framework has its limits. First, unlike true survivorship, it is possible for a migration transition rate to decrease as the interval increases if individuals return to their original location at the end of an interval after a period of being away. Second, taking the analogy literally grows the risk of overlooking the actual age and cohort effects of migration: migration is highest for young adult ages and certain generations might exhibit higher levels of migration than others. But there are also strengths to thinking about our problem in this way. Just like with the age-period-cohort framework, it is highly difficult to isolate the effect of duration, interval or the start point. These concepts must therefore be jointly considered.

The Modeling Approach

In this paper, we are interested in investigating the relationship between duration and interval, setting aside the seasonal or more long term migration trends that would be indicated by the start point. What follows is a discussion of a model that we argue replicates the basic logic of migration as it relates to duration and interval. The approach taken here is to simulate data that exhibits the kinds of patterns we expect to see in the empirical data. The analysis is preliminary but we feel is has great potential.

Theoretical Considerations

The relationship between a six month migration rate, a one year migration rate, and a five year migration rate has been demonstrated to be non-linear.

Generally speaking, we expect interval and migration rate to be positively associated. As the interval grows so does the exposure to the risk of migrating and the migration rate. However, due to return migration, the rate of increase should slow as the interval increases past a certain point. On the other hand, at very short intervals we expect randomness and perhaps very high rates of movement to be observed. Put together, we expect a U-shaped curve as the interval and duration increase from small (i.e. a week) to large (i.e. a year).

To help illustrate these ideas, consider the following scenario. For a particular population, the risk of changing location permanently, r, is distributed evenly across a year such that the number of people who experience a permanent move from one week to the next is roughly the same for any pair of consecutive weeks. Now imagine that for this same population there is some constant but independent probability, p, of traveling on any given week such that traveling on n consecutive weeks is equivalent to p^n. Formally the math here gets messy rather quickly – does someone experiencing a move also experience the risk of travel? – but the idea is that risk of traveling might be high for any given week and mask the on-going, but more subtle long-term migration trend. That being said, the probability of an individual spending the majority of their time traveling falls rapidly with increased duration while the risk of relocating accumulates throughout the year. As such, if we use a small interval and duration to estimate migration, then we will capture some long term movement but a good deal of travel. If we use a large interval and duration to estimate migration we will capture a good deal of long term movement but a small amount of travel. The low point on the U-shaped curve of migration rate with respect interval/duration will be where the effect of travel subsides and the effect of long-term mobility picks up.

The Simulation Model

The goal of our simulation model is to evaluate the macro-demographic consequences of simple behavioral rules. More specifically, we would like to test simple rules in the way people move across space and determine their residence are consistent with patterns of migration rates that we expect to observe in empirical data across a number of settings.

We simulate data that take the form of tuples: ⟨`individual id, time stamp, location`⟩. The structure of each tuple is simple and is intended to mimic the type of meta data that we obtain from geo-located social media

data or call detail records. A single tuple does not provide much information for the study of migration and mobility. However, a series of these tuples over time, for the same individual, offers insights into mobility patterns and relocations. The underlying assumption of our model is that each individual has a latent characteristic: a home location that conditions their mobility behavior. Individuals will be observed most often in their home location; however, if they are observed away from home for a long enough time, then it may be assumed that their home location has changed.

In our approach we simulate time-lines for a population of n individuals. Each individual has known location, l, at each unit of time 1, 2, ..., t such that an individual, i, can be represented by a vector:

$$\{l_{i,1}, l_{i,2}, ..., l_{i,t}\}$$

where $l_{i,t}$ is the location of individual i at time t.

In order to focus on insight over complexity, we build a model in which there are only two possible locations, 1 or 0. The probability that an individual, i, is observed at either 1 or 0 at time t is represented by a simple Bernoulli random variable conditional on the individual's 'home' attribute, which can also only be 1 or 0. This gives us two conditions:

$$P(l_{i,t}|home = 1) = \begin{cases} p, & \text{for } l_{i,t} = 1 \\ 1 - p, & \text{for } l_{i,t} = 0 \end{cases}$$

and

$$P(l_{i,t}|home = 0) = \begin{cases} p, & \text{for } l_{i,t} = 0 \\ 1 - p, & \text{for } l_{i,t} = 1 \end{cases}$$

To model long-term relocation, we add an additional feature. If an individual is observed 'away' from 'home' for k consecutive observations, then the probabilities associated with being observed in the location designated as 'away' become those previously associated with being in the location once designated as 'home':

$$\begin{cases} \text{if } l_{i,t+1} = ... = l_{i,t+k} = 0 | home = 1, & \text{then } 0 \rightarrow home \\ \text{if } l_{i,t+1} = ... = l_{i,t+k} = 1 | home = 0, & \text{then } 1 \rightarrow home \end{cases}$$

As an illustrative example, if the probability of being observed in the 'home' location, p is equal to 0.9 and we use a threshold $k = 5$ to establish

a relocation, it means that the probability of observing a relocation over the course of a period of 104 consecutive time intervals is approximately 0.001. This is because the probability of observing 5 consecutive locations away from home is small: it is 0.1^5. However the individual is 'at risk' of relocating 100 times over a period of 104 consecutive time intervals.

An Example Simulation

Figure 3 shows the results from the model for 500 simulated timelines, with the probability of being observed at home, p, equal to 0.9 and the long-term move threshold, k, equal to 5.

The first panel plots migration rates against interval for all possible durations and starts. For example, the rates estimated for interval == 12 correspond to all 12 week intervals in the data regardless of start date and duration. Note, duration cannot be larger than the interval because then the two periods used to estimate residency at either side of the interval would overlap. So for the box plot associated with interval == 12, duration must be less than or equal to 12. The second panel shows migration rate and duration. Again, all intervals and starts are included. So if duration is two, for example, this means that the period used to estimate residency at either end of the interval is two weeks, regardless of the length of the interval or the start. Similarly, the third panel shows migration rate and start. All intervals and durations are included.

The simulated model does what we intended it to do. In particular, the plot for interval shows rates starting high for very small intervals, then decreases before finally beginning to rise and level off. There is more variation with the duration plot, but the relationship between duration and migration rate appears to be negative as hypothesized–the larger the window for inferring residency, the less travel-related noise and the lower the rate. The start appears to also have a negative relationship with migration. This is an unintended consequence of how the model was set up and will be explored in future iterations of the model.

Data

The empirical portion of the analysis was conducted on three datasets: two sets of social media data in the U.S., from Twitter and Gowalla, and call

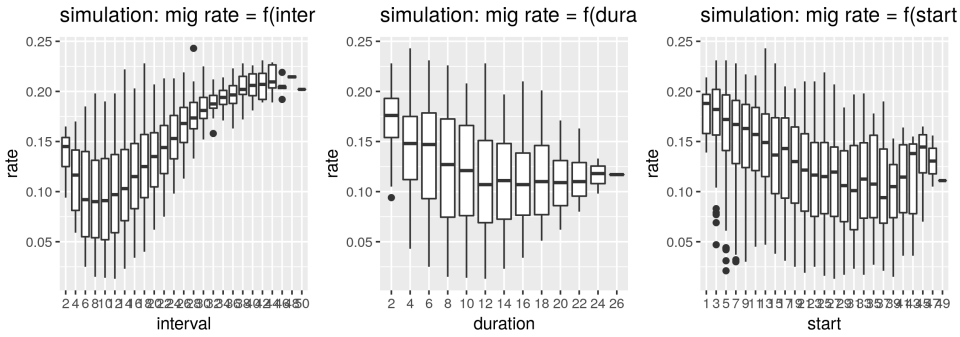

Figure 3: *Interval, Duration and Start for simulated data*

detail records from Senegal. In this section we describe the three datasets. In the following section we repeat the interval-duration-start analysis on these datasets and compare the patterns observed with those from the model.

Geo-located Twitter data for the U.S.

The Twitter data from this analysis consist of roughly 570 million geo-located tweets from 2.9 million Twitter users. We extracted our collection of Twitter users from a long-term archive of the 1% Twitter stream sample[1]. Only user IDs with at least one geo-tagged tweet from within the U.S. during 2011 to 2016 were included. Then, to expand the dataset, we used the Twitter API to crawl the timeline of every qualified user in our collection for their recent geo-tagged tweet history. The analysis of Twitter data was conducted at the U.S. Census Division level of which there are nine for the whole country[1].

Call Detail Records for Senegal

Our analysis makes use of anonymized Call Detail Records (CDR) produced for Orange's 2014 Data for Development (D4D) Challenge, which consisted of phone calls and SMS exchanges between more than 9 million Orange customers in Senegal between 1 January 2013 to 31 December 2013. The CDRs were divided into the three following datasets:

- Dataset 1: One year of site-to-site traffic for 1666 sites on an hourly basis;

[1]https://www.census.gov/geo/img/webatlas/Division.png

- Dataset 2: Fine-grained mobility data (site level) on a rolling 2-week basis with bandicoot behavioral indicators at individual level for about 300,000 randomly sampled users;

- Dataset 3: One year of coarse-grained (3rd administrative level) mobility data with bandicoot behavioral indicators at individual level for about 150,000 randomly sampled users.[5]

In order to prevent re-identification of customers, the datasets are coarsened in at least one dimension (either spatially or temporally) or cover a limited time period. Additionally, only users meeting the following criteria were included in the datasets:

Only users meeting the following criteria were included in the dataset:

1. Users having interactions on more than 75% of days in the given period.

2. Users having had an average of fewer than 1000 interactions per week (since users with more than 1000 interactions per week were presumed to be machines or shared phones).[5]

This paper makes use of Dataset 3, which is made up of 561,025,219 records produced by the calls of 146,352 randomly selected Orange subscribers in the 2013 calendar year. Each record provides a numerical pseudonym representing the user who placed or received the call, the timestamp of the call, and the *arrondissement* in which the user was located at the time of the call.

The D4D dataset divides Senegal's territory into 123 *arrondissements*, which can be grouped into 45 *départements* or 14 *régions*. The latter, the 14 regions of Senegal, are the geographic divisions used in this paper.

Gowalla's check-ins

Finally, we repeated our analysis on geo-located data captured through the *check-in* feature of mobile geo-social network, Gowalla, which allowed users to share their location on various other mobile applications and websites. Per check-in, Gowalla stored user identification information, a timestamp, and latitude/longitude coordinate data. This information is sufficient for observing location changes of individuals over time. Therefore, Cho et al. [2] were able to develop a model of human mobility using check-in data they collected

between February 2009 and October 2010. This dataset included 6,442,890 check-ins generated from 107,092 unique users. Through the Stanford Large Network Dataset Collection [13], we were able to assess the same dataset for our own analysis. As with the U.S. Twitter data, we use U.S. Census Divisions as the geographic unit of analysis, and estimate internal migration that occurred between divisions.

Preliminary Results from Empirical Data

We perform the same analysis that we conducted on our simulation model on the three datasets discussed above: Senegal call detail records, U.S. geotagged Twitter posts, and U.S. Gowalla check-ins.

Interval, Duration, and Start

Figure 4 shows the preliminary results from our analysis. A general pattern appears to hold across the three datasets. The level of migration rate observed increases with interval, though in none of the three datasets does this relationship appear to be linear. Unlike in our simulation model in which the rates estimated at very small intervals are high, in the Senegal CDRs, U.S. Twitter and U.S. Gowalla data, the rates observed at very small intervals tend to be the lowest of those observed.

Duration is the length of time used to impute residency on either side of the interval. The level of migration rate observed appears to be slightly negatively associated with duration; however, the much more prominent feature appears to be the negative relationship between the *variance* of migration rate observed and the duration. This makes sense. Rates calculating using small durations pick up both permanent migrations and short term moves which are much more variable and spread unevenly over the calendar year.

The plots of migration rate and start date show a fairly muted relationship. There appears to be considerable seasonal variability, but because the rates for a given start shown in this plot are for all combinations of interval and duration, there is a good deal of variation associated with any given start date.

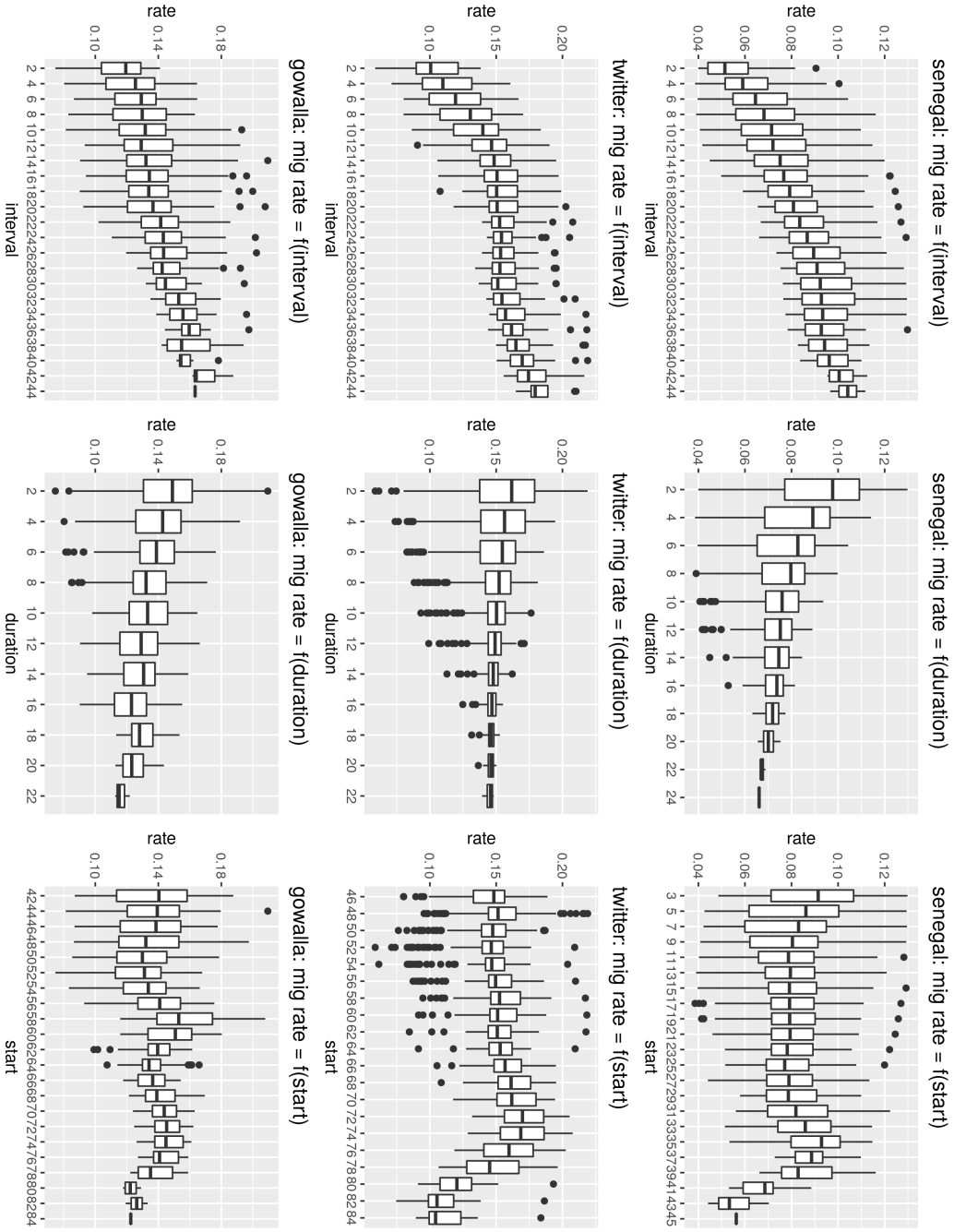

Figure 4: *Interval, Duration and Start for the three datasets*

Smoothing Out Seasonality

One of the largest sources of discrepancy between the patterns observed in the simulated data and the patterns observed in the three empirical datasets is the level of noise associated with seasonal travel. Our simulation model did not attempt to reproduce seasonal variations in small to medium term migration rates, instead proposing a constant likelihood of short-term travel over time. Yet, it appears from all three empirical datasets – Orange, Twitter and Gowalla – that the seasonal signal is capable of large variations in migration estimates. In some circumstances, especially when the residency criteria (duration) is small, this seasonal noise appears large enough as to mask the influences associated cumulative risk of long-term moves. This finding makes a good deal of sense, especially in the case of geo-tagged Twitter data given that recent work has demonstrated that users are more likely to include a location in their tweets when they are traveling [19].

That being said, one of the primary purposes of this research is to examine the sensitivity of migration estimates to issues related to time scale and to demonstrate the potential for studying population movement: human mobility and migration together. As we know from the duration plots from Figure 4, the larger the duration, the less variation in the rate. So we conduct an additional analysis that investigates our ability to smooth out seasonal variation with increased duration. We do this with a series of plots made by subsetting the estimates by duration, grouping together sets of estimates with the same start date and plotting the rate by the date associated with the end of the interval (start date plus interval length) to illustrate true period effects. Results from this analysis can be see in Figure 5. For example, the plot in the upper left hand corner is for all rates calculated from the Orange-Senegal call data using a duration (residency criteria) of four weeks. The upper most line in this plot shows all rates calculated with duration == 4 with the start of the interval at week 7. As you follow this line left to right, the changes in the height of the line are changes that can be associated with the increased interval. These plots illustrate the seasonal variations there are associated with the period specific to end of the interval. The line just below that shows the same but is of rates calculated with the start of the interval at week 9, and so on. The three columns of this figure are estimates calculated with a duration of four, twelve and twenty weeks respectively. From the common shapes produced by these plots when aligned by end point or period, it is clear that is possible to identify period-specific trends using rates

estimated with different intervals.

For the Senegal-Orange and Gowalla, this analysis is limited by the shortened time scale of the data, about a year in both cases. As such, there are a limited number of rates that can be calculated using a duration of 20 weeks (remember: we use 20 weeks to impute residency at both the start and end of the interval). But certainly moving from a four week duration to a twelve week duration in Senegal-Orange and Gowalla results in much smoother rates. The Twitter data are more interesting because they covers a larger amount of time, 2011 to 2016. Notably, with a duration of four weeks, the seasonal variation in rates estimated from Twitter are much more variable than those estimated from the Senegal-Orange data. One obvious point to make here is that cell-phone data, being more temporally fine grained and perhaps less biased towards travel than Twitter data, leads to smoother estimates. But even so, we are able to mostly smooth out the seasonal noise in the Twitter data by increasing the duration to 20 weeks.

Discussion and Next Steps

The results presented here are preliminary, but they already demonstrate useful and seemingly common patterns across three large digital trace datasets of three different kinds (a social media micro blog platform, cell-phone CDRs, a social media check-in site) in two internal migration contexts (Senegal and the United States). There appears to be a consistent positive association between interval and the migration rate. Unlike what we hypothesized and simulated with our model, rates calculated using very small intervals are not always elevated. Instead, small interval/small duration rates tend to be the most variable. There may be a slight negative association between duration and migration rate, but there is a much stronger negative association between duration and the variance of the migration rate. One source of this error is the strong seasonal signal apparent in all three datasets. The specific start/end of an interval matters a good deal to the level of migration estimated, especially when the duration is small.

As next steps, we plan to refine the modeling approach so that it retains its conceptual simplicity but more closely replicates the patterns observed in the empirical data. We also plan on linking U.S. internal migration rates from Twitter and Gowalla together and, if possible, to American Community Survey estimates of internal migration between divisions.

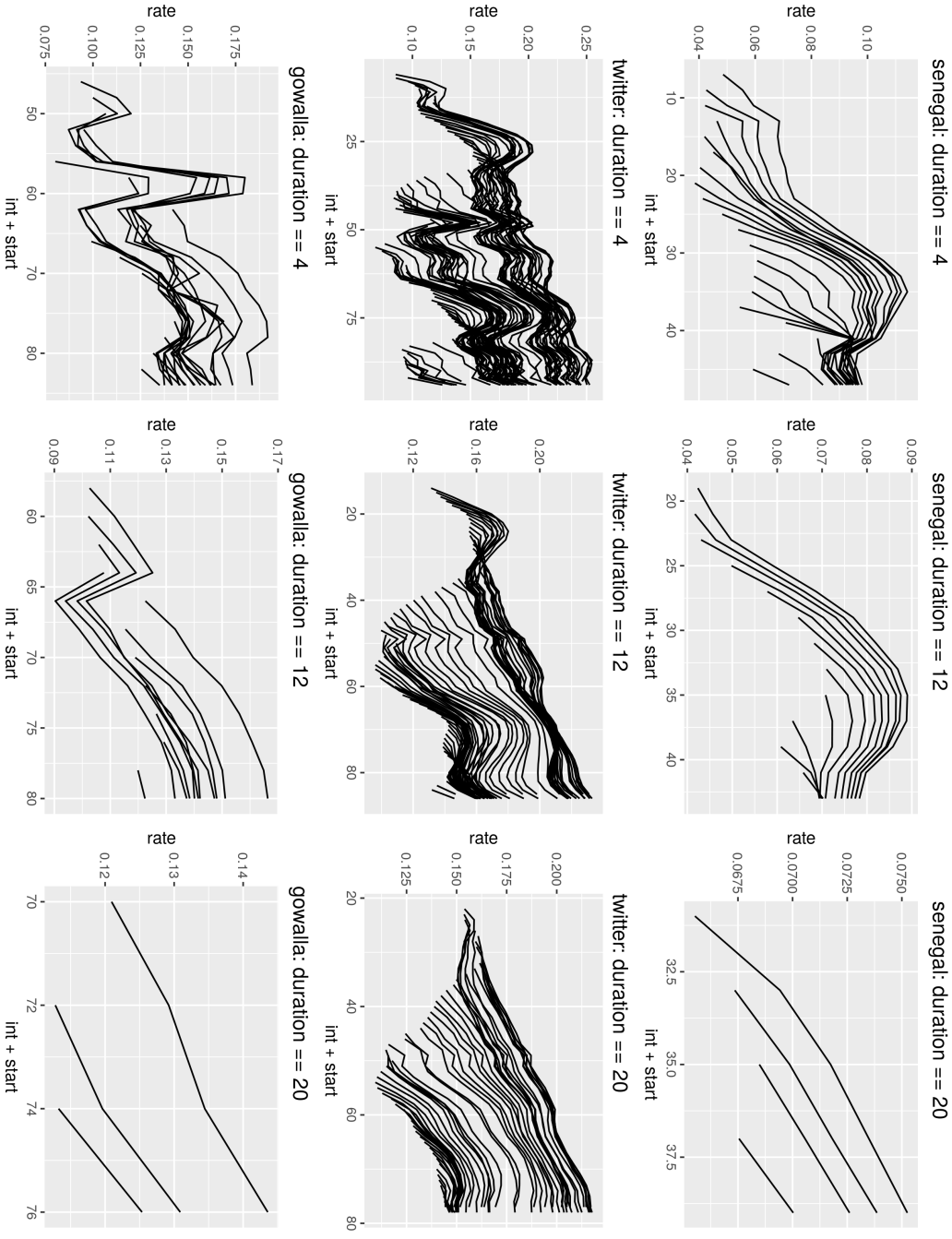

Figure 5: *Plotting rates so that the end of the intervals line up*

There are obviously representative biases in the three datasets discussed in this paper, particularly in the Twitter and Gowalla data. That being said, we hope to have demonstrated here that the unique features of digital trace data are enough to justify investigating them on their own merits. Studying population movement as a whole–that is, mobility and migration together–with these data can be achieved with a standardization procedure based on simple concepts: interval, duration, and start. Results presented from this paper suggest with further experimentation this approach can be used to refine harmonization techniques between official statistics and to develop a theory of population movement.

References

[1] Twitter Archivearchive.org of twitter 1% sample. `https://archive.org/details/twitterstream`. Accessed: 2016-09-30.

[2] Eunjoon Cho, Seth A Myers, and Jure Leskovec. Friendship and mobility: user movement in location-based social networks. In *Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 1082–1090. ACM, 2011.

[3] European Commission. Regulation (ec) no 862/2007 of the european parliament and of the council of 11 july 2007 on community statistics on migration and international protection, 2007. `http://eur-lex.europa.eu/LexUriServ/LexUriServ.do?uri=OJ:L:2007:199:0023:0029:EN:PDF`.

[4] Joop De Beer, James Raymer, Rob Van der Erf, and Leo Van Wissen. Overcoming the problems of inconsistent international migration data: A new method applied to flows in europe. *European Journal of Population/Revue européenne de Démographie*, 26(4):459–481, 2010.

[5] Yves-Alexandre de Montjoye, Zbigniew Smoreda, Romain Trinquart, Cezary Ziemlicki, and Vincent D. Blondel. D4d-senegal: The second mobile phone data for development challenge. *CoRR*, abs/1407.4885, 2014.

[6] Pierre Deville, Catherine Linard, Samuel Martin, Marius Gilbert, Forrest R Stevens, Andrea E Gaughan, Vincent D Blondel, and Andrew J Tatem. Dynamic population mapping using mobile phone data. *Proceedings of the National Academy of Sciences*, 111(45):15888–15893, 2014.

[7] Lee Fiorio, Guy Abel, Jixuan Cai, Emilio Zagheni, Ingmar Weber, and Guillermo Vinué. 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, 2017.

[8] Vanessa Frias-Martinez, Victor Soto, Heath Hohwald, and Enrique Frias-Martinez. Characterizing urban landscapes using geolocated tweets. In *Privacy, Security, Risk and Trust (PASSAT), 2012 International Conference on and 2012 International Confernece on Social Computing (SocialCom)*, pages 239–248. IEEE, 2012.

[9] Fabien Girardin, Francesco Calabrese, Filippo Dal Fiore, Carlo Ratti, and Josep Blat. Digital footprinting: Uncovering tourists with user-generated content. *IEEE Pervasive computing*, 7(4), 2008.

[10] William Haenszel. Concept, measurement, and data in migration analysis. *Demography*, 4(1):253–261, 1967.

[11] Bartosz Hawelka, Izabela Sitko, Euro Beinat, Stanislav Sobolevsky, Pavlos Kazakopoulos, and Carlo Ratti. Geo-located twitter as proxy for global mobility patterns. *Cartography and Geographic Information Science*, 41(3):260–271, 2014.

[12] Christina Hughes, Emilio Zagheni, Guy J Abel, Alessandro Sorichetta, Arkadius Wi'sniowski, Ingmar Weber, and Andrew J Tatem. 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. 2016.

[13] Jure Leskovec and Andrej Krevl. SNAP Datasets: Stanford large network dataset collection. `http://snap.stanford.edu/data`, June 2014.

[14] Jianzheng Liu, Jie Li, Weifeng Li, and Jiansheng Wu. Rethinking big data: A review on the data quality and usage issues. *ISPRS Journal of Photogrammetry and Remote Sensing*, 115:134–142, 2016.

[15] Douglas S Massey, Joaquin Arango, Graeme Hugo, Ali Kouaouci, Adela Pellegrino, and J Edward Taylor. Theories of international migration: A review and appraisal. *Population and development review*, pages 431–466, 1993.

[16] Beata Nowok and Frans Willekens. A probabilistic framework for harmonisation of migration statistics. *Population, Space and Place*, 17(5):521–533, 2011.

[17] Philip H Rees. The measurement of migration, from census data and other sources. *Environment and Planning A*, 9(3):247–272, 1977.

[18] Andrei Rogers. The formal demography of migration and redistribution: measurement and dynamics. 1978.

[19] Dan Tasse, Zichen Liu, Alex Sciuto, and Jason I Hong. State of the geotags: Motivations and recent changes. In *ICWSM*, pages 250–259, 2017.

[20] Emilio Zagheni, Venkata Rama Kiran Garimella, Ingmar Weber, et al. Inferring international and internal migration patterns from twitter data. In *Proceedings of the 23rd International Conference on World Wide Web*, pages 439–444. ACM, 2014.