Using Twitter Data to Estimate the Relationships between Short-term and Long-term Migration:

Initial forays into a general model

Lee Fiorio - Geography
University of Washington
fiorio@uw.edu

CENTER FOR STUDIES IN DEMOGRAPHY & ECOLOGY

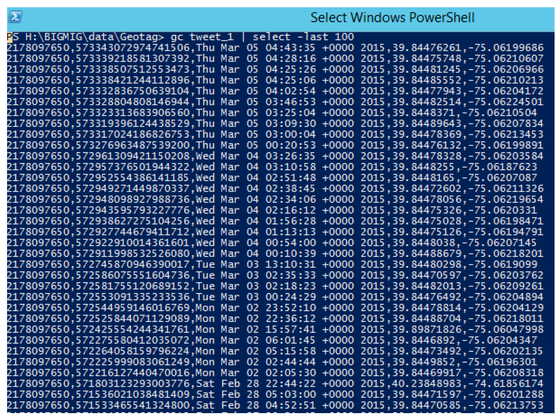
ABSTRACT Migration estimates are sensitive to definitions of time interval and duration. For example, when does a tourist become a migrant? As a result, harmonizing across different kinds of estimates or data sources can be difficult. Moreover in countries like the United States, that do not have a national registry system, estimates of internal migration typically rely on survey data that can require over a year from data collection to publication. In addition, each survey can ask only a limited set questions about migration (e.g., where did you live a year ago? where did you live five years ago?). We leverage a sample of geo-referenced Twitter tweets for about 62,000 users, spanning the period between 2010 and 2016, to estimate a series of US internal migration flows under varying time intervals and durations. Our findings, expressed in terms of `migration curves', document, for the first time, the relationships between short-term mobility and long-term migration. The results open new avenues for demographic research. More specifically, future directions include the use of migration curves to produce probabilistic estimates of long-term migration from short-term (and vice versa) and to nowcast mobility rates at different levels of spatial and temporal granularity using a combination of previously published American Community Survey data and up-to-date data from a panel of Twitter users.

MOTIVATION (AND SOME CAVEATS) Geo-referenced social media data offer a unique opportunity to study migration due to their sheer size. Thousands of geo-observations clustered within thousands of users allow for experimentation on the sensitivity of migration estimates to key concepts in the measurement of migration: duration (the length of stay) and interval (the reference period). The purpose of this project is to explore whether these sensitivities themselves can tell us anything interesting about the overlying migration patterns observed. We are not at the point in this research to claim that migration estimates generated in this analysis can be linked to migration estimates observed among the US population at large. Obviously there are biases in our estimates. Instead, we are attempting to develop a method for describing and quantifying the internal consistency of migration estimates within a population.

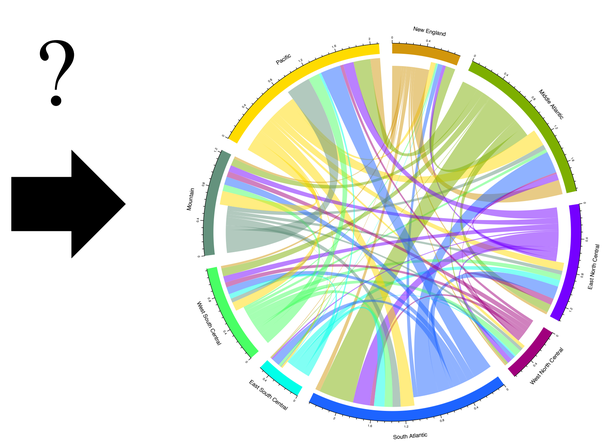
RESEARCH QUESTIONS

- What are the ways to estimate flows between regions using individual level spatial-temporal data (i.e. observation = <person, time, place>?
- What are the concepts from migration theory that must be invoked? What are the limits to applying these concepts to these data?
- What can the internal logic of a given method tell us about migration behavior? Can this logic be used towards a general model with applications in projection and/or data harmonization?

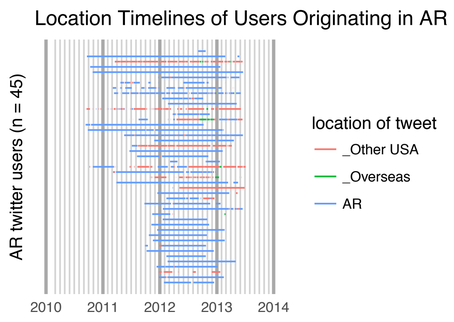
1,000,000s of Geo-referenced Tweets



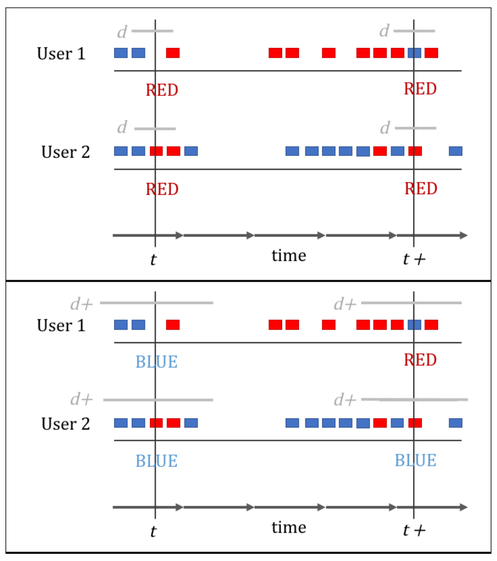
Mobility Flows Between Regions



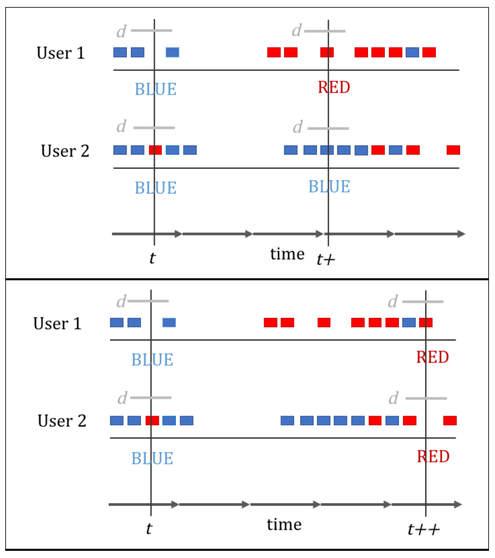
User Timelines



The Concept of Duration



The Concept of Interval



APPROACH Each tweet is geo-coded into one of nine U.S. Census Divisions and time-coded into a week bin. Tweets are grouped together within user and sorted by time (far left).

Next, each user is assigned a location at a particular time by calculating the modal location of their tweets within a buffer or *duration* around the selected time (middle).

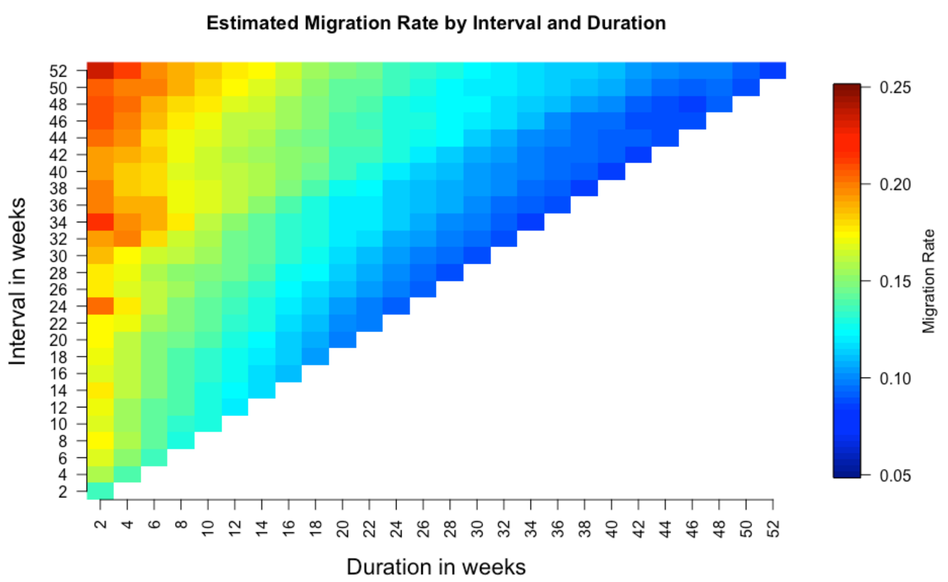
To estimate migration, a second point is selected further along in time at a specified *interval* (far right), and the location of the user at this time is estimated using the same duration. A ”migrant” is thus a user for which the modal location differs between the two times at either side of the interval.

With this approach comes two hypotheses. First, migration estimates will fall with increased duration, holding interval constant. This is because with increased duration, we expect fewer random, short moves to be mistaken for ”migration”. Second, migration estimates will rise with increased interval, holding duration constant. This is because increased interval means increased exposure to the risk of migration. A migration rate estimated using a one year interval should be larger than a migration rate estimated using a six month interval.

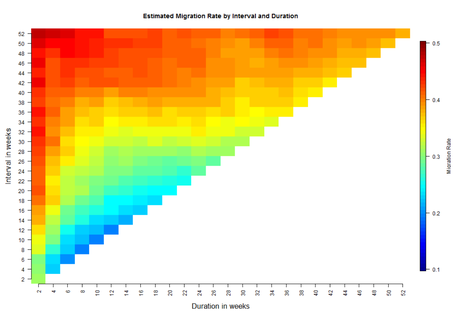
RESULTS Below is a contour plot generated from our sample of Twitter data. At a series of intervals and durations, we estimated whether each user in our data had moved. We then summed all movers and divided by the number of total users to get a “migration rate” for each interval-duration pair.

The figure below provides initial confirmation for our hypotheses. At nearly all intervals, increased duration results in declining migration rate (i.e. a change from warm colors to cool colors with movement along the x-axis). And at nearly all durations, increased interval results in increasing migration rate (i.e. a change from cool colors to warm colors with movement along the y-axis).

More importantly, there appears to be a more subtle pattern within the contour plot. This is what we want to explore further.



MODELING The next step in our analysis is develop a simple model that can reproduce the kinds of patterns observed in the contour plot generated from the Twitter data. This work is *very* preliminary. Using only three parameters in a simple loop in R, we can simulate data that comes close to reproducing the pattern of migration estimates from our results. There are obviously some issues here, namely, the intensity of the rates, but it seems as though it may be possible to reduce individual level spatial-temporal data into a set of parameters that summarizes the mobility and migration patterns captured therein.




```
make.city <- function(home, away, permanent.stay){
  city <- NULL
  for (j in 1:1000){
    location <- c(1,0) #two location options for each draw at time t
    away.ct <- 0 #count of the number of times a person is away
    person <- sample(location, 1, prob = c(home, 1-home)) #initial location
    for (i in 1:500){
      if (tail(person,1) == location[1]) { #if person was 'home' at t-1
        person <- c(person, sample(location, 1, prob = c(home, 1-home)))
        away.ct <- 0}
      else {
      away.ct <- away.ct + 1 #away count goes up
      person <- c(person, sample(location, 1, prob = c(away, 1-away)))}
      if (away.ct == permanent.stay){ #permanent stay condition triggered
        ab <- rev(location) #'away' becomes 'home'
        away.ct <- 0 #count goes to zero
      }
    }
    city <- rbind(city, person, deparse.level = 0)
  }
  return(city)
}
```