Social Media Data and Demographic Research: Some Examples and Directions

Emilio Zagheni

ALAP and ABEP Conference
Foz do Iguacu, PR, Brazil, October 17, 2016

Organization: IUSSP Panel on Big Data and Population Processes, with support from PAA and Hewlett Foundation

Recap

We went over:

1. An introduction to Application Programming Interfaces (APIs)
2. Using Face++ to estimate demographic attributes based on profile pictures
3. Collecting Twitter data
4. Collecting data from Public Facebook Pages
5. A simple Sentiment Analysis

Where to now?

- Literature and data not always on the demographer's radar

Where to now?

- Literature and data not always on the demographer's radar
- Three possible directions for the future:
  1. New questions that could not be addressed because of lack of data
  2. Extract information from large, but biased, data using traditional social science tools
  3. Use Demographic methods to study populations of digital objects

Literature often off the demographer's radar

- Computer Science Conference proceedings, like the International Conference on Web and Social Media: `http://www.aaai.org/Library/ICWSM/icwsm-library.php`

Literature often off the demographer's radar

- Computer Science Conference proceedings, like the International Conference on Web and Social Media: `http://www.aaai.org/Library/ICWSM/icwsm-library.php`
- More via the ACM digital library: `http://dl.acm.org/`

Literature often off the demographer's radar

- Computer Science Conference proceedings, like the International Conference on Web and Social Media: `http://www.aaai.org/Library/ICWSM/icwsm-library.php`
- More via the ACM digital library: `http://dl.acm.org/`
- EPJ Data Science: `https://epjdatascience.springeropen.com/`

Special Collections

- Social Science Research Journal: Special issue on Big Data in the Social Sciences http://www.sciencedirect.com/science/journal/0049089X/59

Special Collections

- Social Science Research Journal: Special issue on Big Data in the Social Sciences `http://www.sciencedirect.com/science/journal/0049089X/59`
- The Annals of the American Academy of Political and Social Science. Special issue on Big Data in Digital Environments `http://ann.sagepub.com/content/659/1.toc`

Examples of data sources

- Archive of Twitter stream
  https://archive.org/details/twitterstream
- Public data from Enigma: http://www.enigma.io
- Stanford Large Network Dataset Collection:
  http://snap.stanford.edu/data/
- Yahoo! Labs datasets:
  http://webscope.sandbox.yahoo.com
- Yelp academic dataset:
  https://www.yelp.com/academic_dataset

1. New questions that could not be addressed with traditional data

"From Migration Corridors to Clusters",

by Messias, Benevenuto, Weber, and Zagheni 2016

Goals

- Use pseudo-migration histories for Google+ users to identify features of migration systems
- Key question: How are countries connected by people who have lived in multiple countries? (migration histories are typically not available in standard surveys)
  - Migration systems are typically identified by looking at changes over time in bilateral flows of migrants
  - "the trouble with this approach is that the system becomes little more than a summary of flows." - Bakewell (2013)

$\Rightarrow$ We consider a new dimension of migration systems: the frequency of people who have lived in 3 distinct countries

No obvious relationship between bilateral and 'trilateral' flows

		Countries Lived In				Bilateral Flows
		A	B	C	D	
Scenario 1	M1	x	x	x		(A,B), (A,C), (B,C)
	M2	x			x	(A,D)
	M3		x		x	(B,D)
	M4			x	x	(C,D)
Scenario 2	M1		x	x	x	(B,C), (B,D), (C,D)
	M2	x	x			(A,B)
	M3	x		x		(A,C)
	M4	x			x	(A,D)

Google+ Data Set

- Data originally collected by Gabriel Magno in 2012 to study gender differences in online social networks
- We considered the Google+ field ("Places where I lived") mapped to countries
- We used the subset of users who have lived in at least 2 countries ($n \approx 1.6$ million users). 270,000 users have lived in 3 countries.

Illustrative example: Fewer people have lived in all these three countries than expected from bilateral flows and a baseline model

$$\overbrace{\underbrace{\text{Brazil - USA;}}_{46{,}784}\underbrace{\text{Mexico - USA;}}_{67{,}065}\underbrace{\text{Brazil - Mexico;}}_{14{,}593}}^{1{,}386}$$

- Expected ranking for people who have lived in the 3 countries based on bilateral flows of Google+ users = # 12
- Actual ranking in Google+ data set = # 80

Illustrative example: Fewer people have lived in all these three countries than expected from bilateral flows and a baseline model

$$\overbrace{\underbrace{\text{Brazil - USA;}}_{46{,}784}\underbrace{\text{Mexico - USA;}}_{67{,}065}\underbrace{\text{Brazil - Mexico;}}_{14{,}593}}^{1{,}386}$$

- Expected ranking for people who have lived in the 3 countries based on bilateral flows of Google+ users = # 12
- Actual ranking in Google+ data set = # 80

⇒ Question: What makes this group of countries different or special?

- For more information, see the paper: https://arxiv.org/pdf/1607.00421.pdf

- For more information, see the paper: `https://arxiv.org/pdf/1607.00421.pdf`
- The data set is freely available: `www.dcc.ufmg.br/~fabricio/migration-dataset/`

2. Extract information from biased data

Inferring migration/mobility patterns from Twitter Data,

Zagheni, Garimella, Weber and State 2014

Geo-located Twitter data

Geo-located Twitter data

- We collected tweets for $\approx$ 500,000 Twitter users with at least one geolocated tweet between May 2011 and April 2013 in OECD countries

Geo-located Twitter data

- We collected tweets for $\approx$ 500,000 Twitter users with at least one geolocated tweet between May 2011 and April 2013 in OECD countries
- We split the time period in intervals of 4 months each.

Geo-located Twitter data

- We collected tweets for $\approx$ 500,000 Twitter users with at least one geolocated tweet between May 2011 and April 2013 in OECD countries
- We split the time period in intervals of 4 months each.
- We considered only the subsample of users who posted more than 3 geolocated tweets for each of the periods of 4 months. $\Rightarrow$ the sample size reduces to $\approx$ 15,000

- We had a sample of geo-located Twitter tweets (geographic coordinates)

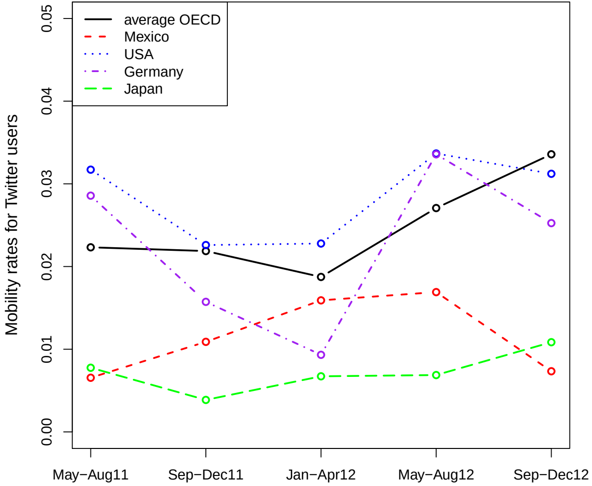
- We had a sample of geo-located Twitter tweets (geographic coordinates)
- No demographic information

- We had a sample of geo-located Twitter tweets (geographic coordinates)
- No demographic information
- No official statistics to calibrate the model

- We had a sample of geo-located Twitter tweets (geographic coordinates)
- No demographic information
- No official statistics to calibrate the model

$\Rightarrow$ We proposed a difference-in-differences approach to estimate trends

Geographic mobility from geolocated Twitter data



Source: Zagheni, Garimella, Weber and State, WWW'14

Consider the following model for how the data are generated

Consider the following model for how the data are generated

$$\underbrace{y_i^t}_{\substack{\text{Observation from}\\ \text{social media}\\ \text{for location i}}} = \underbrace{n}_{\substack{\text{bias for}\\ \text{location i}}} + \underbrace{x_i^t}_{\substack{\text{“true” rate}\\ \text{for location i}}}$$

and

$$\underbrace{y_z^t}_{\substack{\text{Observation from}\\ \text{social media}\\ \text{for location z}}} = \underbrace{m}_{\substack{\text{bias for}\\ \text{location z}}} + \underbrace{x_z^t}_{\substack{\text{“true” rate}\\ \text{for location z}}}$$

Consider the following model for how the data are generated

$$\underbrace{y_i^t}_{\substack{\text{Observation from} \\ \text{social media} \\ \text{for location i}}} = \underbrace{n}_{\substack{\text{bias for} \\ \text{location i}}} + \underbrace{x_i^t}_{\substack{\text{“true” rate} \\ \text{for location i}}}$$

and

$$\underbrace{y_z^t}_{\substack{\text{Observation from} \\ \text{social media} \\ \text{for location z}}} = \underbrace{m}_{\substack{\text{bias for} \\ \text{location z}}} + \underbrace{x_z^t}_{\substack{\text{“true” rate} \\ \text{for location z}}}$$

Additive bias different across regions, but constant (or changes by the same amount across regions) over short periods of time

Assume that we knew the ‘true’ rates (x) for France and Spain

$x_{FR}^{t+1} = 0.7$	$x_{SP}^{t+1} = 0.5$
$x_{FR}^{t} = 0.5$	$x_{SP}^{t} = 0.4$

Let’s define δ^{t+1} as the differential in the variation of these quantities of interest between time t and $(t+1)$

$$\delta^{t+1} = \underbrace{(x_{FR}^{t+1} - x_{FR}^{t}) - (x_{SP}^{t+1} - x_{SP}^{t})}_{\text{difference in the increments}} = ?$$

Assume that we knew the 'true' rates (x) for France and Spain

$x_{FR}^{t+1} = 0.7$	$x_{SP}^{t+1} = 0.5$
$x_{FR}^{t} = 0.5$	$x_{SP}^{t} = 0.4$

Let's define δ^{t+1} as the differential in the variation of these quantities of interest between time t and $(t+1)$

$$\delta^{t+1} = \underbrace{(x_{FR}^{t+1} - x_{FR}^{t}) - (x_{SP}^{t+1} - x_{SP}^{t})}_{\text{difference in the increments}} =?$$

$$\delta^{t+1} = (0.7 - 0.5) - (0.5 - 0.4) =$$

Assume that we knew the ‘true’ rates (x) for France and Spain

$x_{FR}^{t+1} = 0.7$	$x_{SP}^{t+1} = 0.5$
$x_{FR}^{t} = 0.5$	$x_{SP}^{t} = 0.4$

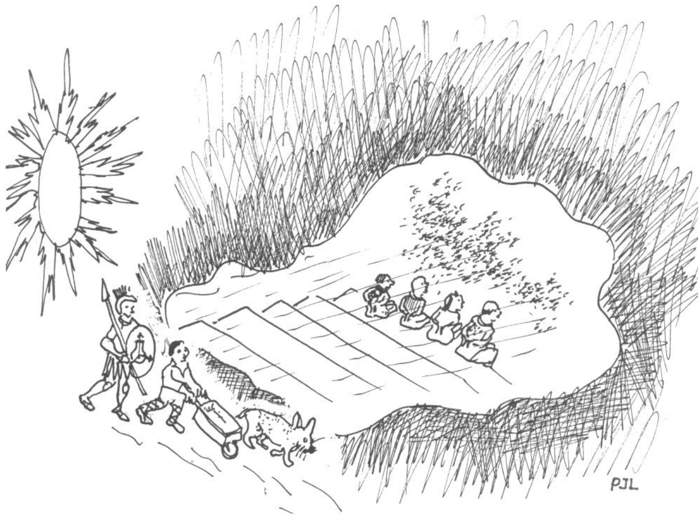
Let’s define δ^{t+1} as the differential in the variation of these quantities of interest between time t and $(t+1)$

$$\delta^{t+1} = \underbrace{(x_{FR}^{t+1} - x_{FR}^{t}) - (x_{SP}^{t+1} - x_{SP}^{t})}_{\text{difference in the increments}} =?$$

$$\delta^{t+1} = (0.7 - 0.5) - (0.5 - 0.4) =$$

$$= 0.2 - 0.1 = 0.1$$

Plato's allegory of the Cave



Plato's Allegory of the Cave

All we see is a distorted image (y) of the 'true' rates (x)

$y_{FR}^{t+1} = 0.2 + 0.7$	$y_{SP}^{t+1} = 0.1 + 0.5$
$y_{FR}^{t} = 0.2 + 0.5$	$y_{SP}^{t} = 0.1 + 0.4$

What is δ^{t+1}?

$$\delta^{t+1} = \underbrace{(y_{FR}^{t+1} - y_{FR}^{t}) - (y_{SP}^{t+1} - y_{SP}^{t})}_{\text{difference in the increments}} = ?$$

All we see is a distorted image (y) of the 'true' rates (x)

$y_{FR}^{t+1} = 0.2 + 0.7$	$y_{SP}^{t+1} = 0.1 + 0.5$
$y_{FR}^{t} = 0.2 + 0.5$	$y_{SP}^{t} = 0.1 + 0.4$

What is δ^{t+1}?

$$\delta^{t+1} = \underbrace{(y_{FR}^{t+1} - y_{FR}^{t}) - (y_{SP}^{t+1} - y_{SP}^{t})}_{\text{difference in the increments}} = ?$$

$$\delta^{t+1} = (0.9 - 0.7) - (0.6 - 0.5) =$$

All we see is a distorted image (y) of the 'true' rates (x)

$y_{FR}^{t+1} = 0.2 + 0.7$	$y_{SP}^{t+1} = 0.1 + 0.5$
$y_{FR}^{t} = 0.2 + 0.5$	$y_{SP}^{t} = 0.1 + 0.4$

What is δ^{t+1}?

$$\delta^{t+1} = \underbrace{(y_{FR}^{t+1} - y_{FR}^{t}) - (y_{SP}^{t+1} - y_{SP}^{t})}_{\text{difference in the increments}} = ?$$

$$\delta^{t+1} = (0.9 - 0.7) - (0.6 - 0.5) =$$

$$= 0.2 - 0.1 = 0.1$$

Same as before...

Difference in differences estimator

- To the extent that the bias is additive and, within each country, is constant over short periods of time, DiD estimates from social media data:

$$\delta^{t+1} = (y_i^{t+1} - y_z^{t+1}) - (y_i^t - y_z^t)$$

Difference in differences estimator

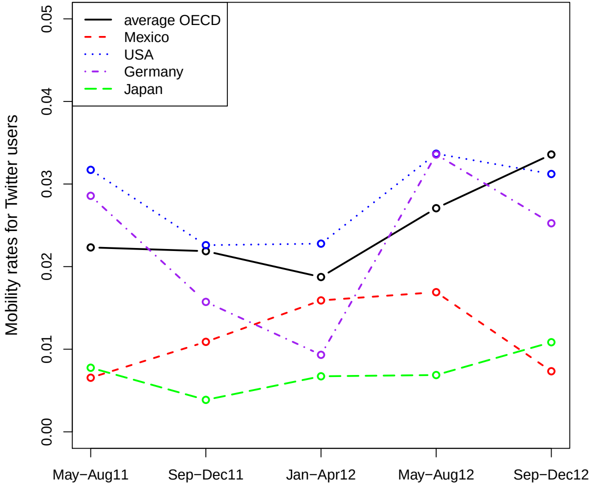
- To the extent that the bias is additive and, within each country, is constant over short periods of time, DiD estimates from social media data:

$$\delta^{t+1} = (y_i^{t+1} - y_z^{t+1}) - (y_i^t - y_z^t)$$

are good estimates of the underlying differential:

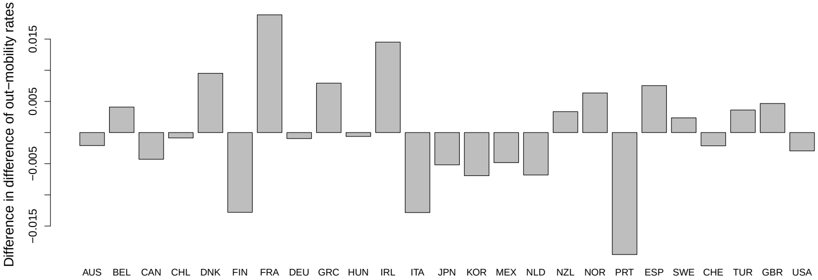
$$\delta^{t+1} = \underbrace{(x_i^{t+1} - x_i^t) - (x_z^{t+1} - x_z^t)}_{\text{difference in the increments}}$$

- Additive values of the bias (m and n) cancel out



Source: Zagheni, Garimella, Weber and State, WWW'14

Twitter example



Source: Zagheni, Garimella, Weber and State, WWW'14

Remarks

If the bias is expected to be multiplicative:

$$\underbrace{y_i^t}_{\substack{\text{Observation from} \\ \text{social media} \\ \text{for location i}}} = \underbrace{n}_{\substack{\text{bias for} \\ \text{location i}}} \times \underbrace{x_i^t}_{\substack{\text{“true” rate} \\ \text{for location i}}}$$

Remarks

If the bias is expected to be multiplicative:

$$\underbrace{y_i^t}_{\substack{\text{Observation from}\\ \text{social media}\\ \text{for location i}}} = \underbrace{n}_{\substack{\text{bias for}\\ \text{location i}}} \times \underbrace{x_i^t}_{\substack{\text{“true” rate}\\ \text{for location i}}}$$

Use a logarithmic transformation

$$\log(y_i^t) = \log(n) + \log(x_i^t)$$

Remarks

If the bias is expected to be multiplicative:

$$\underbrace{y_i^t}_{\substack{\text{Observation from}\\ \text{social media}\\ \text{for location i}}} = \underbrace{n}_{\substack{\text{bias for}\\ \text{location i}}} \times \underbrace{x_i^t}_{\substack{\text{“true” rate}\\ \text{for location i}}}$$

Use a logarithmic transformation

$$\log(y_i^t) = \log(n) + \log(x_i^t)$$

Then use the difference-in-differences estimator on the logs:

$$\delta^{t+1} = [\log(y_i^{t+1}) - \log(y_z^{t+1})] - [log(y_i^t) - log(y_z^t)]$$

Addressing the issue of selection bias in Social Media data is an important area where demographers can contribute.

Addressing the issue of selection bias in Social Media data is an important area where demographers can contribute.

E.g., see this literature review ⇒ Zagheni and Weber (2015) Demographic Research with non-Representative Internet Data

3. Use demographic methods to study populations of digital objects

- We saw how Tweets are stored and how we can access them via the API

- We saw how Tweets are stored and how we can access them via the API
- These data can be useful to understand population processes

- We saw how Tweets are stored and how we can access them via the API
- These data can be useful to understand population processes
- Also, demographic methods can help us understand these data

- We saw how Tweets are stored and how we can access them via the API
- These data can be useful to understand population processes
- Also, demographic methods can help us understand these data
- An example: estimating Twitter growth rate from a cross section of Tweets

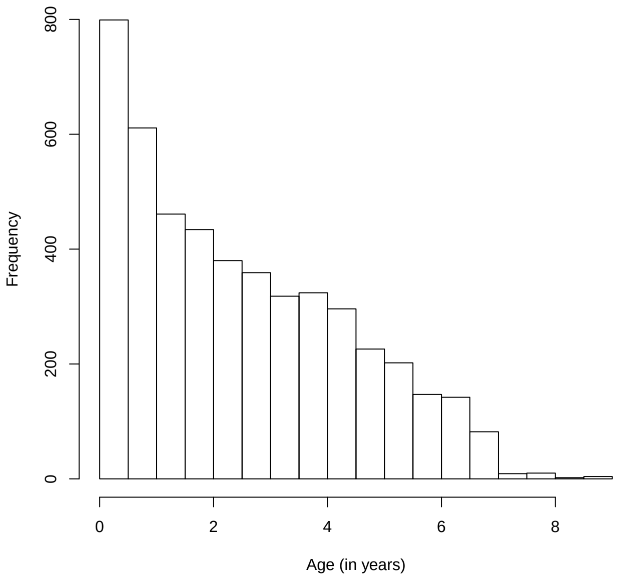
The U2 band has ‘lived” in Twitter more than 7 years

Robert Moffitt was “born” in Twitter a little over 2 years ago

Barack Obama has been on Twitter for more than 9 years

```
{ "created_at": "Wed Nov 07 04:16:18 +0000 2012",
  "id": 266031293945503744,
  "text": "Four more years. http://t.co/bAJE6Vom",
  "source": "web",
  "user": {
    "id": 813286,
    "name": "Barack Obama",
    "screen_name": "BarackObama",
    "location": "Washington, DC",
    "description": "This account is run by Organizing for Action staff.
        Tweets from the President are signed -bo.",
    "url": "http://t.co/8aJ56Jcemr",
    "protected": false,
    "followers_count": 40873124,
    "friends_count": 654580,
    "listed_count": 202495,
    "created_at": "Mon Mar 05 22:08:25 +0000 2007",
    "time_zone": "Eastern Time (US & Canada)",
    "statuses_count": 10687,
    "lang": "en" },
  "coordinates": null,
  "retweet_count": 783488,
  "favorite_count": 295026,
  "lang": "en"
}
```

Age distribution of a sample of active Twitter users (birth=signing up)



Estimating population growth rate from one census

- Problem: Given the number of individuals P_x at age x and P_y at age y, at time t, find the rate at which the births were increasing between years $t - x$ and $t - y$;
- Consider the situation where y is greater than x.

Estimating population growth rate from one census

- Problem: Given the number of individuals P_x at age x and P_y at age y, at time t, find the rate at which the births were increasing between years $t-x$ and $t-y$;
- Consider the situation where y is greater than x.
- We have

$$\underbrace{B(t-x)}_{\substack{\text{births at}\\\text{time t-x}}} \quad \underbrace{L_x}_{\substack{\text{fraction surviving}\\\text{x years to time t}}} \quad = \quad \underbrace{P_x}_{\substack{\text{Population size}\\\text{of age x at time t}}}$$

$$B(t-y)L_y = P_y$$

We have

$$B(t-x)L_x = P_x$$

$$B(t-y)L_y = P_y$$

We have

$$B(t-x)L_x = P_x$$

$$B(t-y)L_y = P_y$$

which can be expressed as:

$$\frac{B(t-x)}{B(t-y)}\frac{L_x}{L_y} = \frac{P_x}{P_y}$$

We have

$$B(t-x)L_x = P_x$$

$$B(t-y)L_y = P_y$$

which can be expressed as:

$$\frac{B(t-x)}{B(t-y)}\frac{L_x}{L_y} = \frac{P_x}{P_y}$$

or

$$\frac{B(t-x)}{B(t-y)} = \frac{P_x}{P_y}\frac{L_y}{L_x}$$

Let's then assume that the birth function is exponential:

$$B(t) = B(0)e^{rt}$$

Let's then assume that the birth function is exponential:

$$B(t) = B(0)e^{rt}$$

Then

$$\underbrace{\frac{B(t-x)}{B(t-y)}}_{e^{(y-x)r}} = \frac{P_x}{P_y}\frac{L_y}{L_x}$$

Let's then assume that the birth function is exponential:

$$B(t) = B(0)e^{rt}$$

Then

$$\underbrace{\frac{B(t-x)}{B(t-y)}}_{e^{(y-x)r}} = \frac{P_x}{P_y}\frac{L_y}{L_x}$$

Thus

$$r = \frac{1}{y-x} log(\frac{P_x}{P_y}\frac{L_y}{L_x})$$

Let's then assume that the birth function is exponential:

$$B(t) = B(0)e^{rt}$$

Then

$$\underbrace{\frac{B(t-x)}{B(t-y)}}_{e^{(y-x)r}} = \frac{P_x}{P_y}\frac{L_y}{L_x}$$

Thus

$$r = \frac{1}{y-x} log(\frac{P_x}{P_y}\frac{L_y}{L_x})$$

For the specific small Twitter sample described above we get $r \approx 0.3$

Toy example, but the message is that the demographer's toolbox can be relevant outside of standard applications