2 - Digital trace data (1/2)

Diego Alburez-Gutierrez
MPIDR

European Doctoral School of Demography 2020-21

02 March 2021

MAX-PLANCK-INSTITUT
FÜR DEMOGRAFISCHE
FORSCHUNG

MAX PLANCK INSTITUTE
FOR DEMOGRAPHIC
RESEARCH

Agenda

1. Q&A
2. Introduction to digital trace and marketing data
3. **Break**
4. Example 1: Migration
5. Example 2: Internet users
6. Discussion

Q&A

- Questions about the final assignment
- Issues accessing the data
- Other?

Digital traces are incidental to our online presence

- Digital breadcrumbs are unavoidable
- Pre-GDPR, largely unchecked
- Marketing-led
- Not collected for social-scientific research

Some data sources (1)

- Marketing platforms
  - Facebook/Instagram/WhatsApp API
  - Linkedin API

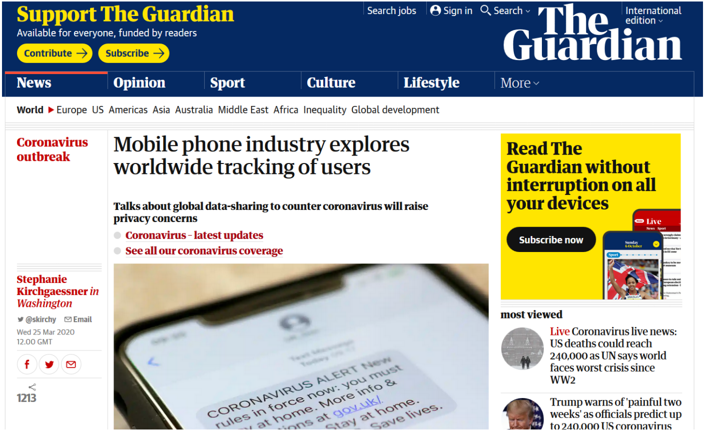
Some data sources (2)

- Online platforms and communication
  - Twitter (API)
  - Google Trends
  - Email, IP address, mobile phones

Some data sources (3)

- Internet of Things
  - Activity trackers and wearable medical devices
  - Wearable sensors

A contemporary issue



Support The Guardian

Available for everyone, funded by readers

Contribute → Subscribe →

Search jobs Sign in Search International edition

The Guardian

News Opinion Sport Culture Lifestyle More

World ▶ Europe US Americas Asia Australia Middle East Africa Inequality Global development

Coronavirus outbreak

Mobile phone industry explores worldwide tracking of users

Talks about global data-sharing to counter coronavirus will raise privacy concerns

- Coronavirus – latest updates
- See all our coronavirus coverage

Stephanie Kirchgaessner *in Washington*

@skirchy Email

Wed 25 Mar 2020 12.00 GMT

1213

Read The Guardian without interruption on all your devices

Subscribe now

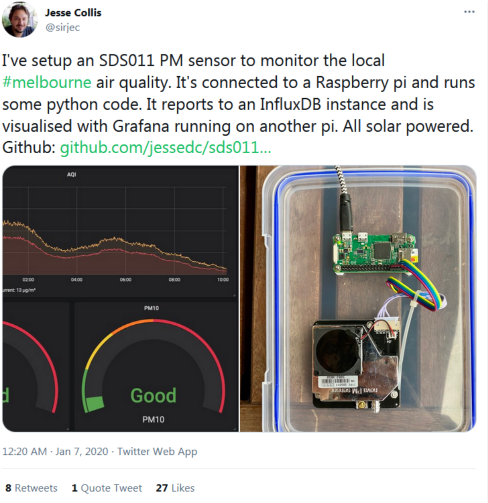
most viewed

Live Coronavirus live news: US deaths could reach 240,000 as UN says world faces worst crisis since WW2

Trump warns of 'painful two weeks' as officials predict up to 240,000 US coronavirus

https://www.theguardian.com/world/2020/mar/25/mobile-phone-industry-explores-worldwide-tracking-of-users-coronavirus

Anyone can collect data these days



Jesse Collis
@sirjec

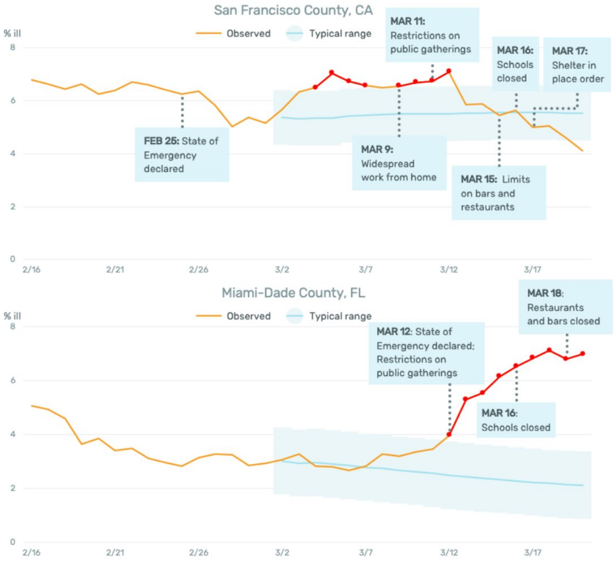
I've setup an SDS011 PM sensor to monitor the local #melbourne air quality. It's connected to a Raspberry pi and runs some python code. It reports to an InfluxDB instance and is visualised with Grafana running on another pi. All solar powered. Github: github.com/jessedc/sds011...

12:20 AM · Jan 7, 2020 · Twitter Web App

8 Retweets **1** Quote Tweet **27** Likes

Source: https://twitter.com/sirjec/status/1214325789707005953?s=20

Smart thermometer to track body temperature


San Francisco County, CA

FEB 25: State of Emergency declared

MAR 9: Widespread work from home

MAR 11: Restrictions on public gatherings

MAR 15: Limits on bars and restaurants

MAR 16: Schools closed

MAR 17: Shelter in place order

Miami-Dade County, FL

MAR 12: State of Emergency declared; Restrictions on public gatherings

MAR 16: Schools closed

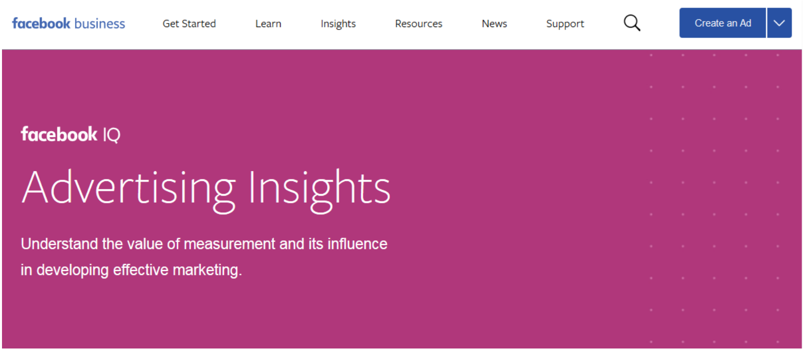
MAR 18: Restaurants and bars closed

https://twitter.com/ASlavitt/status/1241569790151593986

Some data sources

1. Marketing platforms
   - Facebook/Instagram/WhatsApp API
   - Linkedin API
2. Online platforms and communication
   - Twitter (API)
   - Google Trends
   - Email, IP address, mobile phones
3. Internet of Things
   - Activity trackers and wearable medical devices
   - Wearable sensors

Using online marketing tools for demographic research

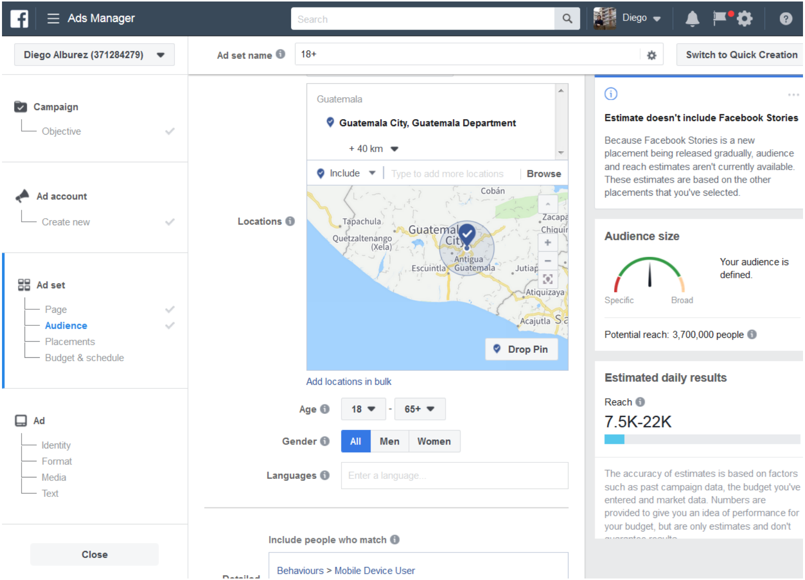


facebook IQ

Advertising Insights

Understand the value of measurement and its influence in developing effective marketing.

‘Audience estimates’: FB users in Guatemala



Diego Alburez (371284279)

Ad set name: 18+

Campaign
- Objective ✓

Ad account
- Create new ✓

Ad set
- Page ✓
- Audience ✓
- Placements
- Budget & schedule

Ad
- Identity
- Format
- Media
- Text

Locations: Guatemala
- Guatemala City, Guatemala Department
- + 40 km

Include | Type to add more locations | Browse

Add locations in bulk

Age: 18 - 65+

Gender: [✓ All] Men Women

Languages: Enter a language...

Include people who match

Detailed: Behaviours > Mobile Device User

Estimate doesn't include Facebook Stories

Because Facebook Stories is a new placement being released gradually, audience and reach estimates aren't currently available. These estimates are based on the other placements that you've selected.

Audience size

Specific — Broad

Your audience is defined.

Potential reach: 3,700,000 people

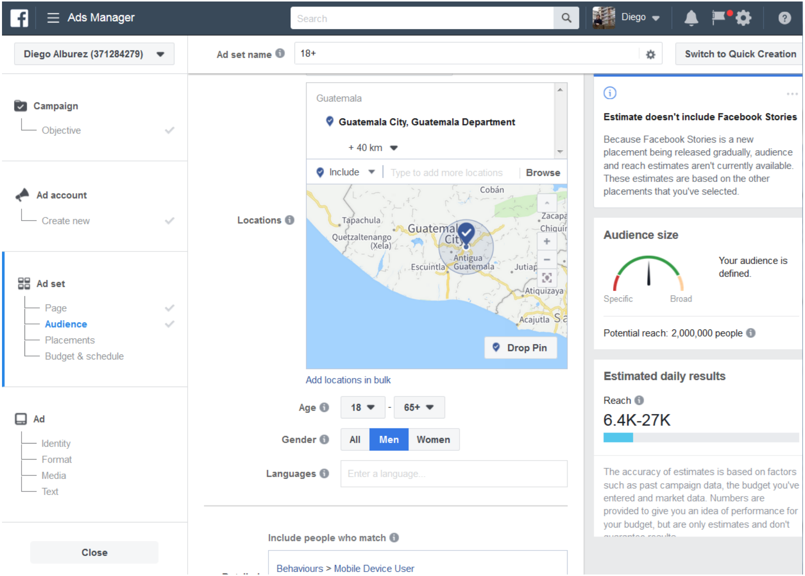
Estimated daily results

Reach

7.5K-22K

The accuracy of estimates is based on factors such as past campaign data, the budget you've entered and market data. Numbers are provided to give you an idea of performance for your budget, but are only estimates and don't guarantee results.

Male FB users, aged 18+ in Guatemala City



Diego Alburez (371284279)

Ad set name: 18+

Locations: Guatemala — Guatemala City, Guatemala Department + 40 km

Age: 18 - 65+

Gender: All **Men** Women

Languages: Enter a language...

Include people who match

Behaviours > Mobile Device User

Estimate doesn't include Facebook Stories

Because Facebook Stories is a new placement being released gradually, audience and reach estimates aren't currently available. These estimates are based on the other placements that you've selected.

Audience size

Your audience is defined.

Potential reach: 2,000,000 people

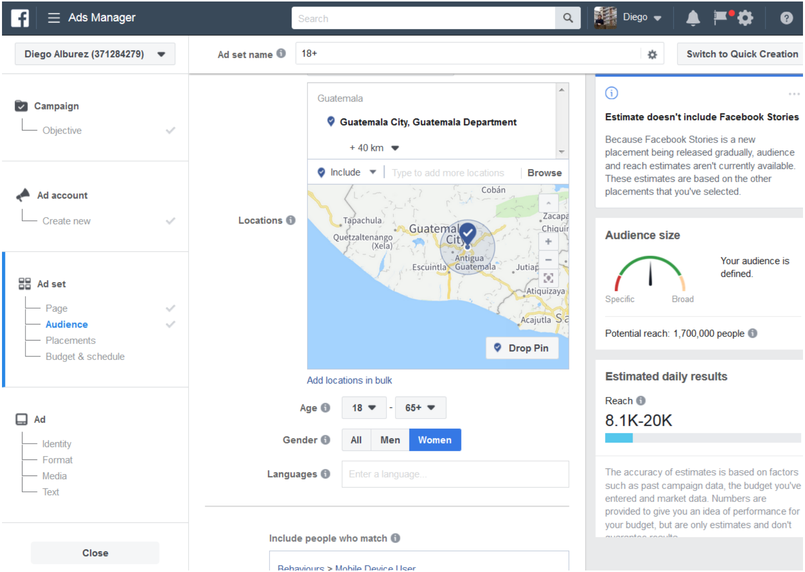
Estimated daily results

Reach

6.4K-27K

The accuracy of estimates is based on factors such as past campaign data, the budget you've entered and market data. Numbers are provided to give you an idea of performance for your budget, but are only estimates and don't guarantee results.

Female FB users, aged 18+ in Guatemala City



Diego Alburez (371284279)

Ad set name: 18+

Locations: Guatemala — Guatemala City, Guatemala Department + 40 km

Include | Type to add more locations

Age: 18 - 65+

Gender: All | Men | **Women**

Languages: Enter a language...

Include people who match: Behaviours > Mobile Device User

Estimate doesn't include Facebook Stories

Because Facebook Stories is a new placement being released gradually, audience and reach estimates aren't currently available. These estimates are based on the other placements that you've selected.

Audience size

Your audience is defined.

Potential reach: 1,700,000 people

Estimated daily results

Reach

8.1K-20K

The accuracy of estimates is based on factors such as past campaign data, the budget you've entered and market data. Numbers are provided to give you an idea of performance for your budget, but are only estimates and don't guarantee results.

Facebook marketing platforms and APIs

- GUI vs API (by hand or programmatically)
- Sofia Gil's tutorial: https://github.com/SofiaG1l/Using_Facebook_API
- For python users, Carol Coimbra's: https://github.com/carolcoimbra/facebook-ads

Group discussion

FB audience estimates are used for **micro-targeted advertisment**.

- *A marketing strategy that uses digital trace to segment audiences into small groups for content targeting.*

💡

1. How can it be used for demographic research?
2. What are the pros and cons of using it?

Good practices for digital demography

1. Acknowledge non-representativeness
2. Use IRL data to compare and completment
3. Account for drifting and algorithmic confounding (observing a casino?)
4. Think of ethics, be transparent and upfront

Break

Example 1: Migration

Group discussion

💡

We'll review two studies. Identify the

1. **strengths**
2. **weaknesses**

of their reliance on digital trace data.

Research at a glance

- RQ: Estimate out-migration from Puerto Rico in the months after 2017 Hurricane Maria
- Data: FB advertising platform and American Community Survey (ACS)
- Findings:
  - Oct 2017 to Jan 2018: 17.0% increase in Puerto Rican migrants (185K people)
  - Jan to March 2018: 1.8% decrease (return migration)
  - Flows by age, sex, and US State

Alexander, M., Polimis, K. and Zagheni, E. (2019), The Impact of Hurricane Maria on Out-migration from Puerto Rico: Evidence from Facebook Data. Population and Development Review, 45: 617-630.

Sanity checks

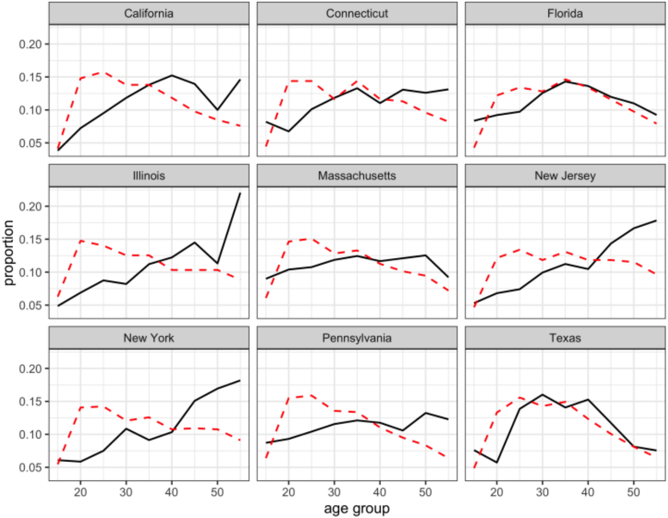


Figure 1: Age distribution of Puerto Rican migrants in FB data (red dashed line) and American Community Survey (black solid line).

Population increase

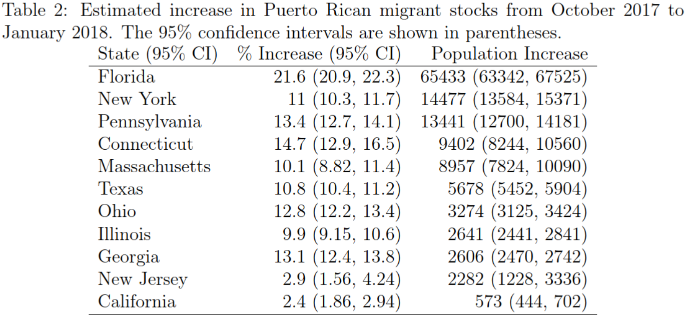


Table 2: Estimated increase in Puerto Rican migrant stocks from October 2017 to January 2018. The 95% confidence intervals are shown in parentheses.

State (95% CI)	% Increase (95% CI)	Population Increase
Florida	21.6 (20.9, 22.3)	65433 (63342, 67525)
New York	11 (10.3, 11.7)	14477 (13584, 15371)
Pennsylvania	13.4 (12.7, 14.1)	13441 (12700, 14181)
Connecticut	14.7 (12.9, 16.5)	9402 (8244, 10560)
Massachusetts	10.1 (8.82, 11.4)	8957 (7824, 10090)
Texas	10.8 (10.4, 11.2)	5678 (5452, 5904)
Ohio	12.8 (12.2, 13.4)	3274 (3125, 3424)
Illinois	9.9 (9.15, 10.6)	2641 (2441, 2841)
Georgia	13.1 (12.4, 13.8)	2606 (2470, 2742)
New Jersey	2.9 (1.56, 4.24)	2282 (1228, 3336)
California	2.4 (1.86, 2.94)	573 (444, 702)

Alexander, M., Polimis, K. and Zagheni, E. (2019), The Impact of Hurricane Maria on Out-migration from Puerto Rico: Evidence from Facebook Data. Population and Development Review, 45: 617-630.

Percent change by age groups

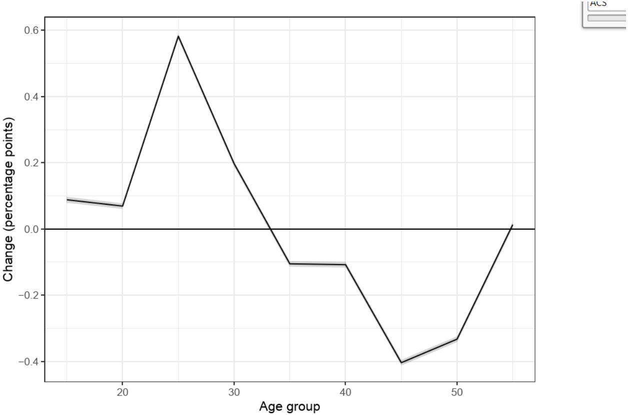


Figure 3: Estimated change in Puerto Rican migrant age distribution from October 2017 to January 2018.

Alexander, M., Polimis, K. and Zagheni, E. (2019), The Impact of Hurricane Maria on Out-migration from Puerto Rico: Evidence from Facebook Data. Population and Development Review, 45: 617-630.

Example 2: Digital use

Summary

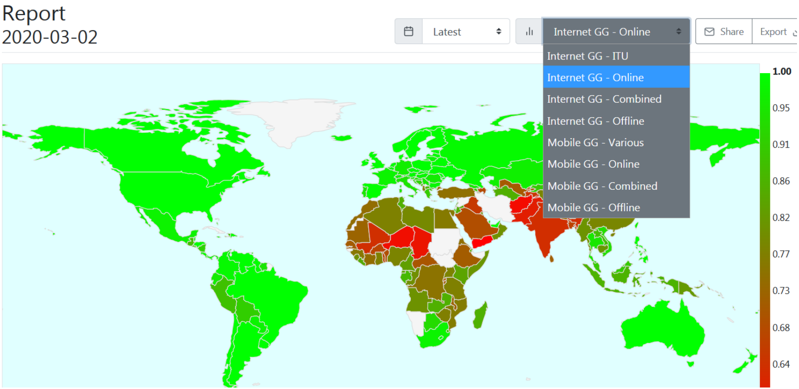
- RQ: Predict internet and mobile phone use gender gaps
- Data: FB advertising platform and indicators from offline sources
- Estimating rates: Facebook Gender Gap Index:

$$\frac{\text{Female to male gender ratio of people with characteristic}}{\text{Female to Male gender ratio of the population}}$$

- Findings:
  - FB measure explained 69% of ground-truth variance
  - Online+offline measure: best estimates

Fatehkia, M., Kashyap, R., and Weber, I. (2018). Using Facebook ad data to track the global digital gender gap. World Development 107:189–209.

Measuring the gender gap in real-time



Report
2020-03-02

Internet GG - Online

https://www.digitalgendergaps.org/data/?report=2020-03-02

Discussion

Group discussion

💡

We'll review two studies. Identify the

1. **strengths**
2. **weaknesses**

of their reliance on digital trace data.

Strengths and weaknesses: Puerto Rico migration

- Con: No ‘ground-truth’ data (?)
- Con: Non-representative sample
- Con: Algorithmic drifting
- Pro: Real-time data (no delay as in official data)
- Adjust for bias: Difference-in-difference to

Strengths and weaknesses: Digital gender gap

- Pro: Nowcasting at sub-national level
- Con: Non-representative
- Pro: ‘Ground-truth’ data: Internet Gender Gap Index
- Con: No data for China (FB penetratio: 0.2%)
- Adjust for bias: correction factor (internet penetration)

Challenges going ahead

Whoever you are... I've always depended on the kindness of strangers.

— Blanche DuBois, A Streetcar Named Desire

1. Ensuring sustainable data access
2. Addressing systematic bias
3. No information information about algorithms that companies use internally (eg. rounding errors)
4. Privacy and ethical digital research

Zuboff, S. (2015). Big other: Surveillance capitalism and the prospects of an information civilization. Journal of Information Technology 30(1):75–89.

Make yourself heard!

📢

1. What are the main ethical concerns when using digital trace data?
2. Do all/any apply to digital demographers?
3. How can we minimise risk for users?