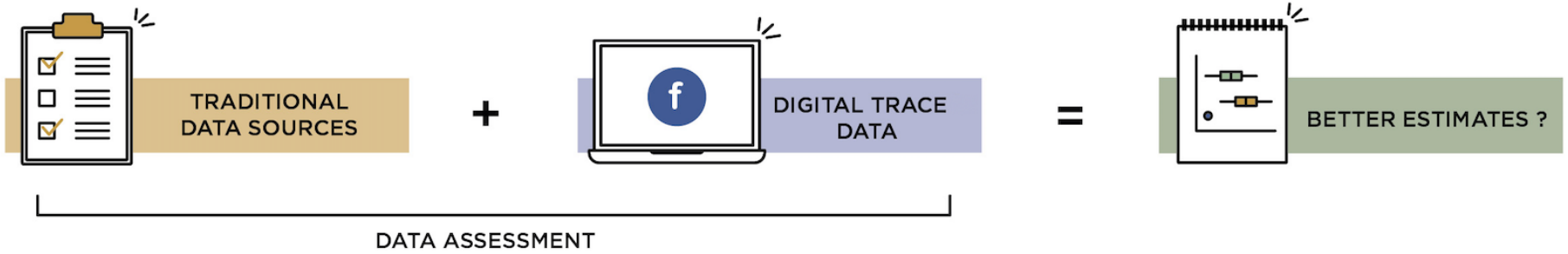
Online Surveys and Post-Stratification

Francesco Rampazzo
Career Development Fellow
in Marketing and Consumer Demography

Saïd Business School and
Leverhulme Centre for Demographic Science,
and Nuffield College
University of Oxford

SICSS-Oxford 2022

My research

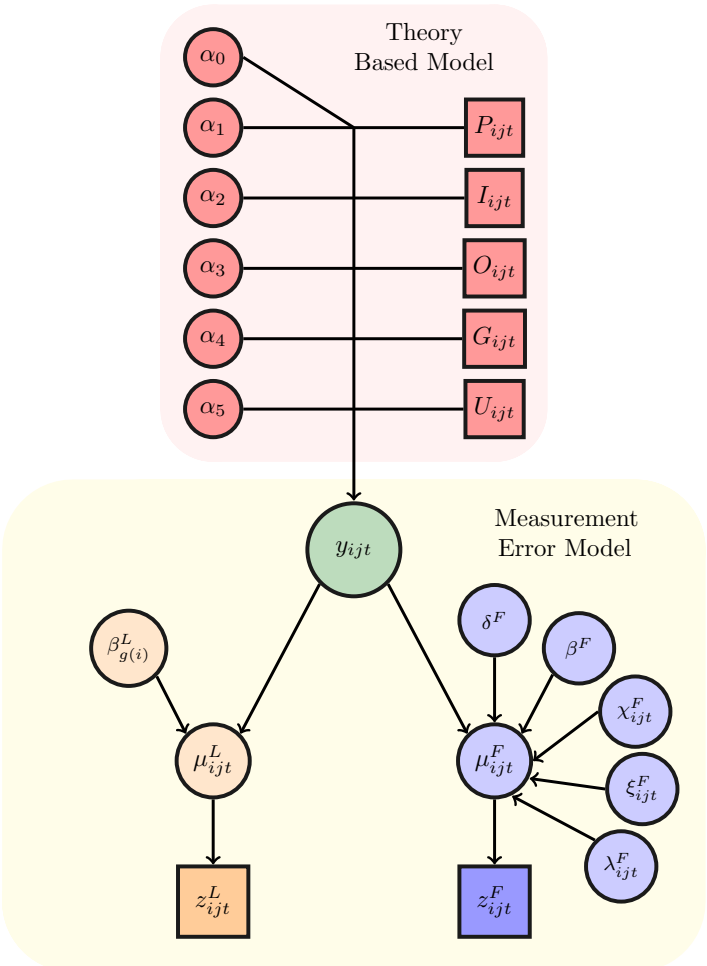


I like to **combine traditional and new sources of data** for topics related to **fertility, migration, and transition to adulthood**.

Example of my research

Bayesian Hierarchical model inspired by the Integrated European Migration Model (IMEM) (Raymer et al. 2013) repurposed to combine digital traces and traditional data from surveys.

The aim of the model is to obtain a *true* estimate of the number of migrants in a country.



Rampazzo, F., Bijak, J., Vitali, A. , Weber, I. , and Zagheni E. (Forthcoming). *A Framework for Estimating Migrant Stocks Using Digital Traces and Survey Data: an Application in the United Kingdom*. Demography.

`https://bit.ly/fb-migration-uk`

Collaborators and co-authors


Leah Rosenzweig


Katherine Hoffman Pham


Parrish Bergquist


Matto Mildenberg


André Grow


Daniela Perrotta


Emanuele Del Fava


Jorge Cimentada


Sofia Gil-Clavel


René Flores


Ilana Ventura


Ingmar Weber


Emilio Zagheni

What's a survey?

A **survey** is a **systematic method** for gathering information from a **sample** for then **extending the results** to a **larger population**.

Neil Malhotra
@namalhotra

Interesting story from the world of opt-in samples. Prolific went viral on Tik Tok. The sample became 80% female, and the average age is 21.

Traduci il Tweet

10:59 PM · 13 ago 2021 · Twitter Web App

36 Retweet **12** Tweet di citazione **200** Mi piace

Total Error with Big Data

Amaya A., Biemer P.P., and Kinyon D. (2020). **Total Error in a Big Data World: Adapting TSE Framework to Big Data**. Journal of Survey Statistics and Methodology 8, 89-119.

Error components	**Error causes**
Coverage error	Undercoverage, overcoverage, duplicates
Sampling error	Large, but non probabilistic samples
Specification error	Not clear definitions
Nonresponse/missing data error	Missing data counfounded with undercoverage
Measurement/content error	Selection in what we see
Processing error	Download and cleaning phases
Modeling/estimation error	Unknown errors
Analytic error	Interpretation

Advertisment!!!

Two steps:

- Quota Sampling
- Post-Stratification

Quota Sampling

Political Science Research and Methods (2020), **8**, 558– 564
doi:10.1017/psrm.2018.49

PS RM

RESEARCH NOTE

Quota sampling using Facebook advertisements

Baobao Zhang[1*], Matto Mildenberger[2], Peter D. Howe[3], Jennifer Marlon[4], Seth A. Rosenthal[4] and Anthony Leiserowitz[4]

[1]Department of Political Science, Yale University, New Haven, CT, USA, [2]Department of Political Science, University of California, Santa Barbara, Santa Barbara, CA, USA, [3]Department of Environment and Society, Utah State University, Logan, UT, USA and [4]School of Forestry and Environmental Studies, Yale University, New Haven, CT, USA
*Corresponding author. Email: baobao.zhang@yale.edu

Abstract
Researchers in different social science disciplines have successfully used Facebook to recruit subjects for their studies. However, such convenience samples are not generally representative of the population. We developed and validated a new quota sampling method to recruit respondents using Facebook advertisements. Additionally, we published an R package to semi-automate this quota sampling process using the Facebook Marketing API. To test the method, we used Facebook advertisements to quota sample 2432 US respondents for a survey on climate change public opinion. We conducted a contemporaneous nationally representative survey asking identical questions using a high-quality online survey panel whose respondents were recruited using probability sampling. Many results from the Facebook-sampled survey are similar to those from the online panel survey; furthermore, results from the Facebook-sampled survey approximate results from the American Community Survey (ACS) for a set of validation questions. These findings suggest that using Facebook to recruit respondents is a viable option for survey researchers wishing to approximate population-level public opinion.

Generally, it is **not** possible to **randomly sample** on social media platforms.

Quota Sampling a version of stratified random sampling, however it is **not** a **probabilistic sample**.

Post-Stratification

Models to **correct the estimates** through **weights** that re-balance the estimates towards the population under study.

International Journal of Forecasting 31 (2015) 980–991

Contents lists available at ScienceDirect

International Journal of Forecasting

journal homepage: www.elsevier.com/locate/ijforecast

Forecasting elections with non-representative polls

Wei Wang [a,*], David Rothschild [b], Sharad Goel [b], Andrew Gelman [a,c]

[a] Department of Statistics, Columbia University, New York, NY, USA
[b] Microsoft Research, New York, NY, USA
[c] Department of Political Science, Columbia University, New York, NY, USA

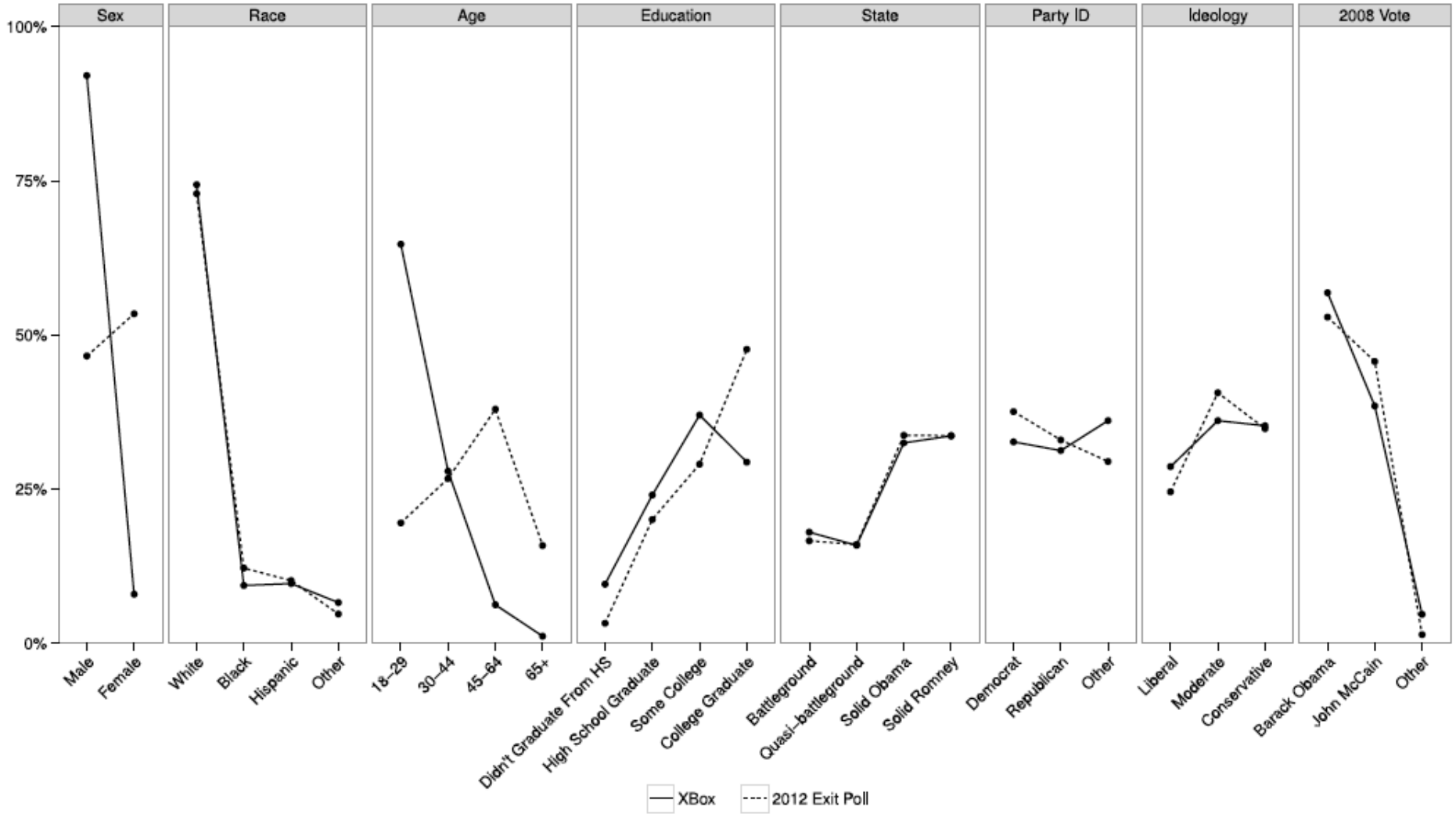
ARTICLE INFO

Keywords:
Non-representative polling
Multilevel regression and poststratification
Election forecasting

ABSTRACT

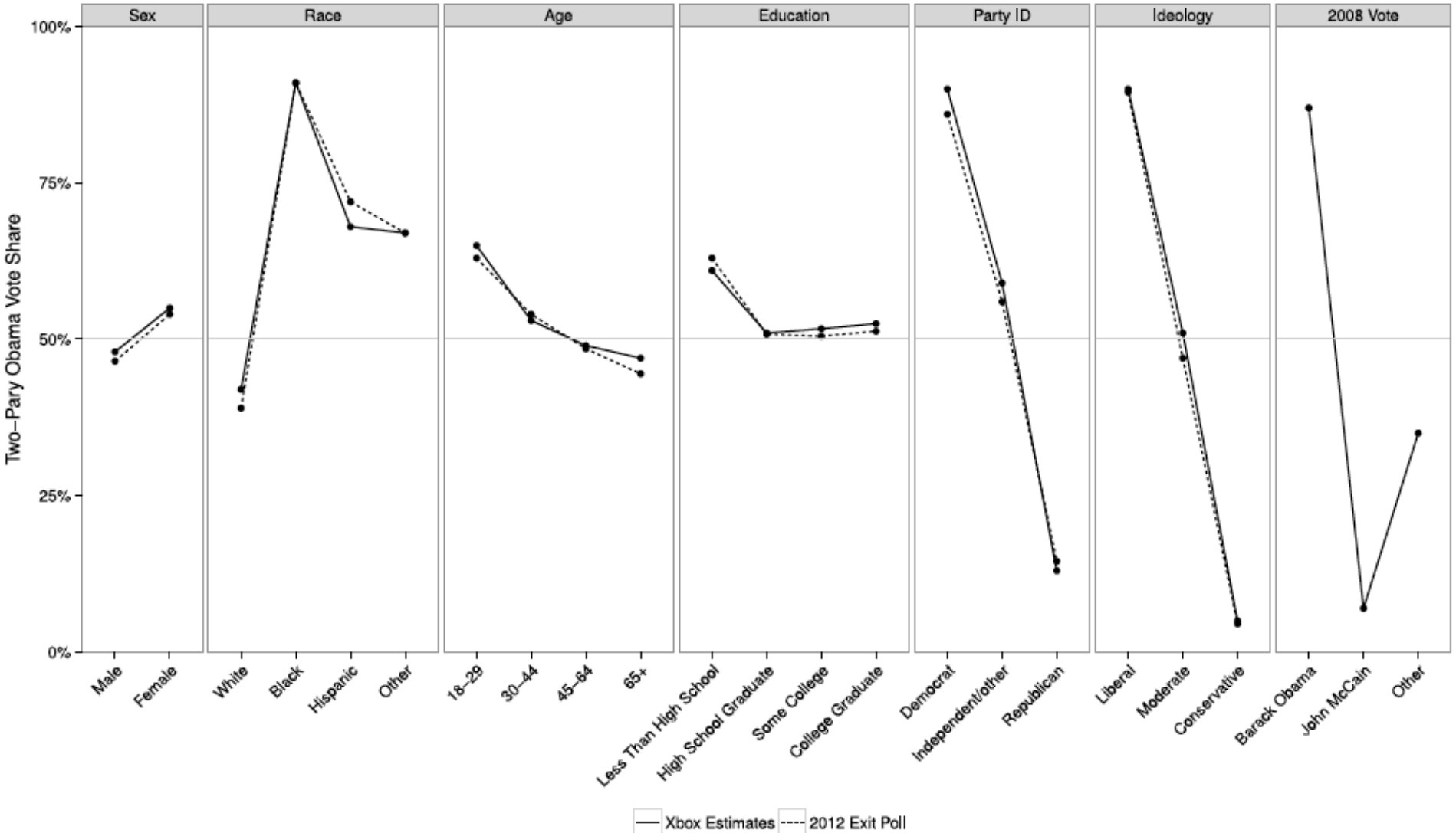
Election forecasts have traditionally been based on representative polls, in which randomly sampled individuals are asked who they intend to vote for. While representative polling has historically proven to be quite effective, it comes at considerable costs of time and money. Moreover, as response rates have declined over the past several decades, the statistical benefits of representative sampling have diminished. In this paper, we show that, with proper statistical adjustment, non-representative polls can be used to generate accurate election forecasts, and that this can often be achieved faster and at a lesser expense than traditional survey methods. We demonstrate this approach by creating forecasts from a novel and highly non-representative survey dataset: a series of daily voter intention polls for the 2012 presidential election conducted on the Xbox gaming platform. After adjusting the Xbox responses via multilevel regression and poststratification, we obtain estimates which are in line with the forecasts from leading poll analysts, which were based on aggregating hundreds of traditional polls conducted during the election cycle. We conclude by arguing that non-representative polling shows promise not only for election forecasting, but also for measuring public opinion on a broad range of social, economic and cultural issues.

Xbox



A comparison of the demographic, partisan, and 2008 vote distributions in the Xbox dataset and the 2012 electorate (as measured by adjusted exit polls). As one might expect, the sex and age distributions exhibit considerable differences.

Xbox – Post-Stratified



Comparison of the two-party Obama vote share for various demographic subgroups, as estimated from the 2012 national exit poll and from the Xbox data on the day before the election.

Through the **right considerations** **non-probabilistic samples** can be used to **estimate quantities generalisable** to the **entire population**.

Questions?

Can we use

Facebook

to conduct surveys?

Facebook Advertising Platform

https://www.facebook.com/adsmanager

Add locations in bulk

Age

18 - 65+

Gender

All genders

Detailed targeting

Include people who match

Behaviours > Ex-pats

Lived in Poland (formerly Ex-pats – Poland)

Audience definition

Your audience selection is fairly broad.

Potential reach: 380,000 people

Estimated daily results

Reach

3.9K-11K

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

Were these estimates helpful?

N.B. Download Facebook's data

pySocialWatcher
Matheus Araujo
https://github.com/maraujo/pySocialWatcher

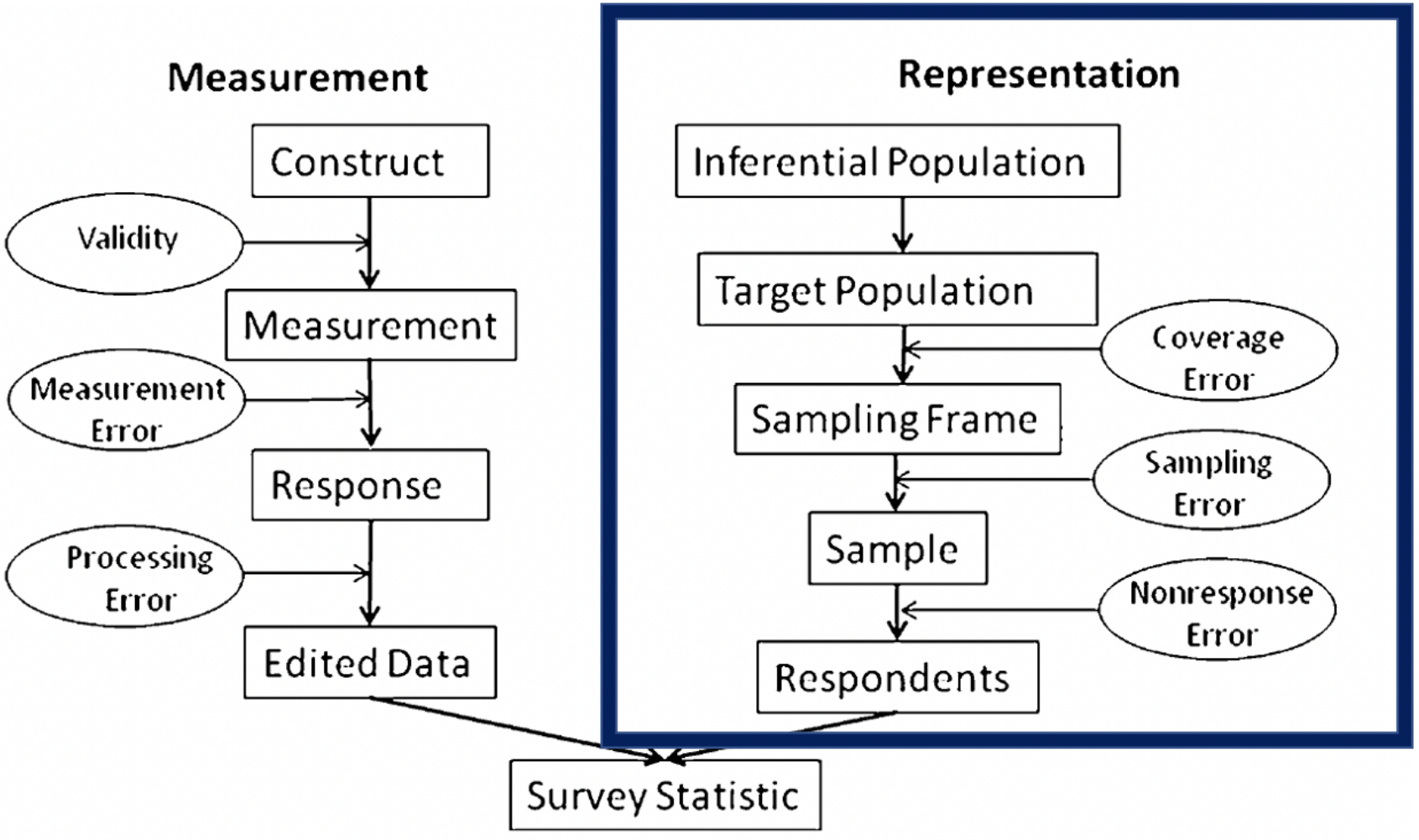
R Estimates FB Ads
Connor Gilroy
https://github.com/ccgilroy/r-estimates-fb-ads

Using FB API
Sofia Gil
https://github.com/SofiaG11/Using_Facebook_API

Example of Ads (1)

Example of Ads (2)

Focus on Representation



Inferential Population

=

Global Population

Aim is to use Facebook to infer the Global Population

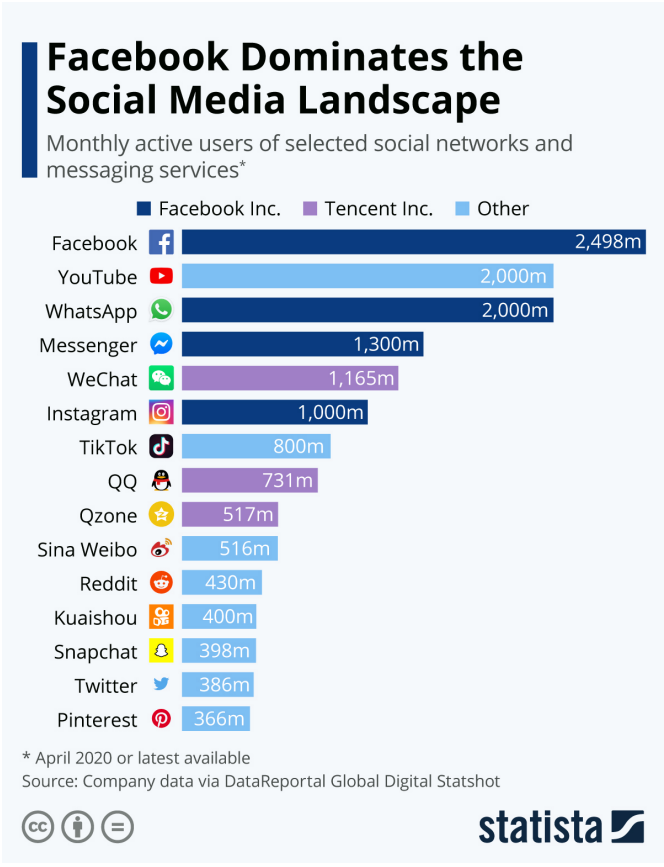
Target Population

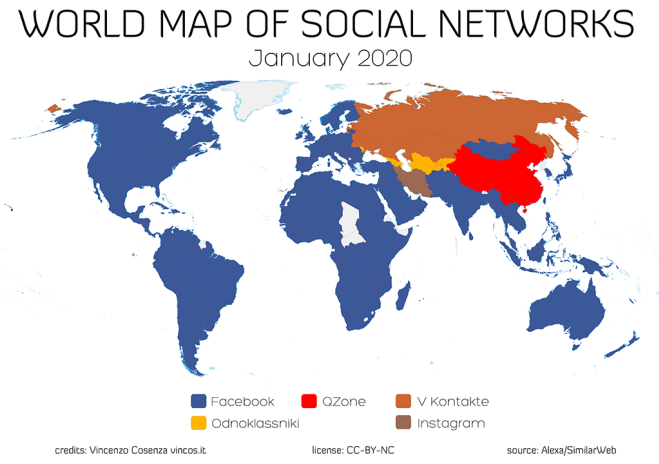
Analysis of the distribution

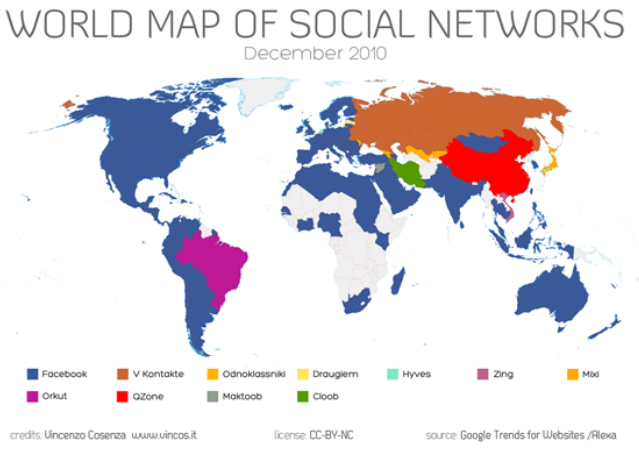
of the Facebook users

in the key variables

Use of Social Media

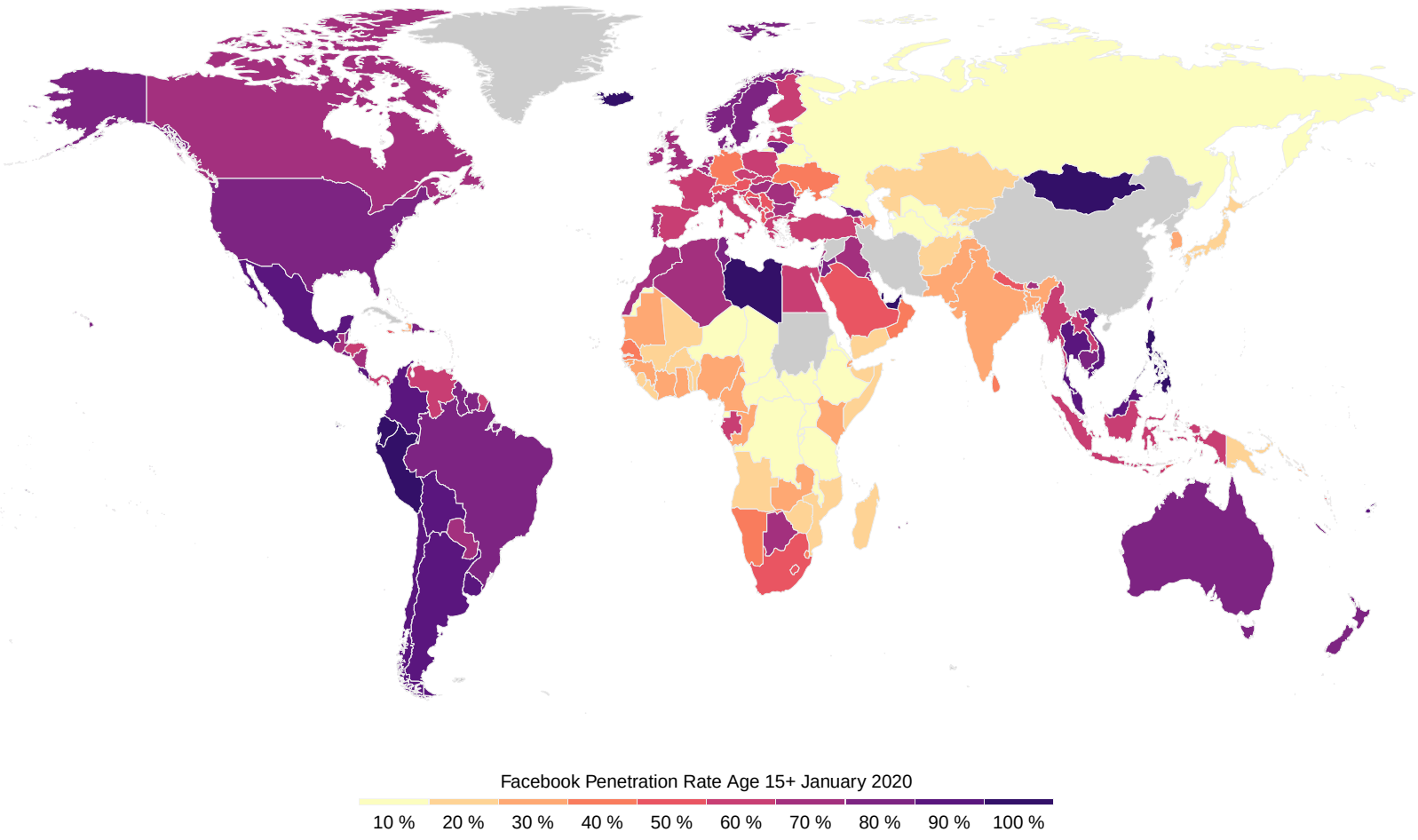






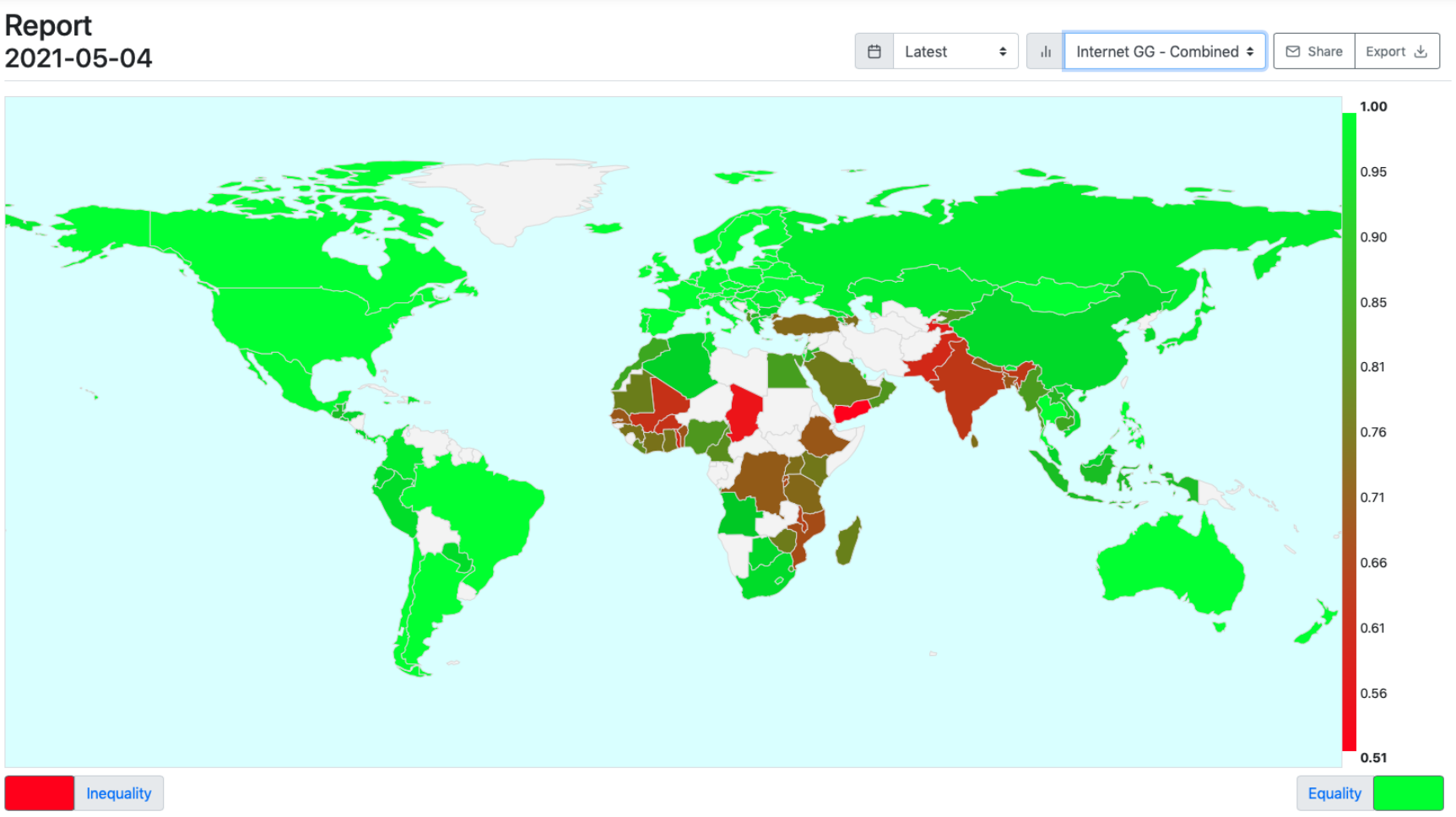
https://www.statista.com/chart/5194/active-users-of-social-networks-and-messaging-services/
https://vincos.it/world-map-of-social-networks/

Use of Facebook in the World



Facebook penetration rates among users age 15 years and older around the world. The rate is computed as of January 2020 with Facebook and United Nations data (UN 2019, Ševčiková 2020). The Facebook data is courtesy of Palotti et al. 2020.

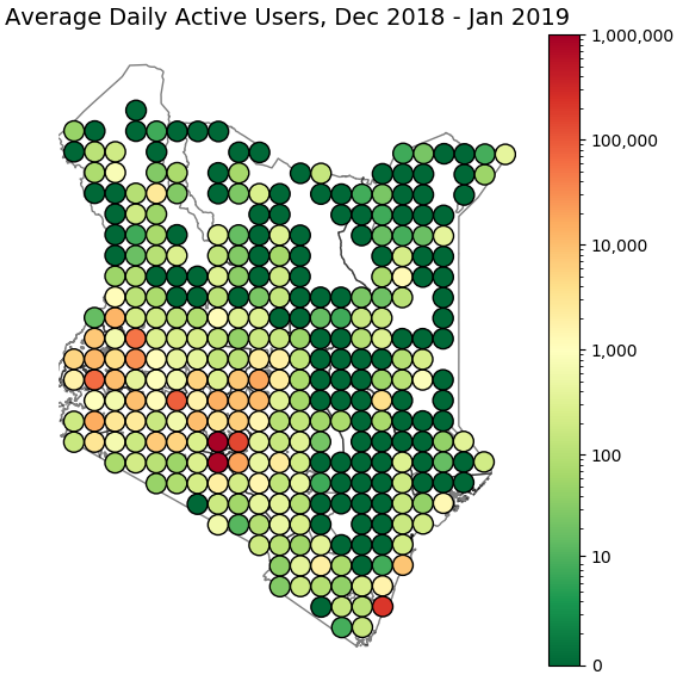
Gender Index

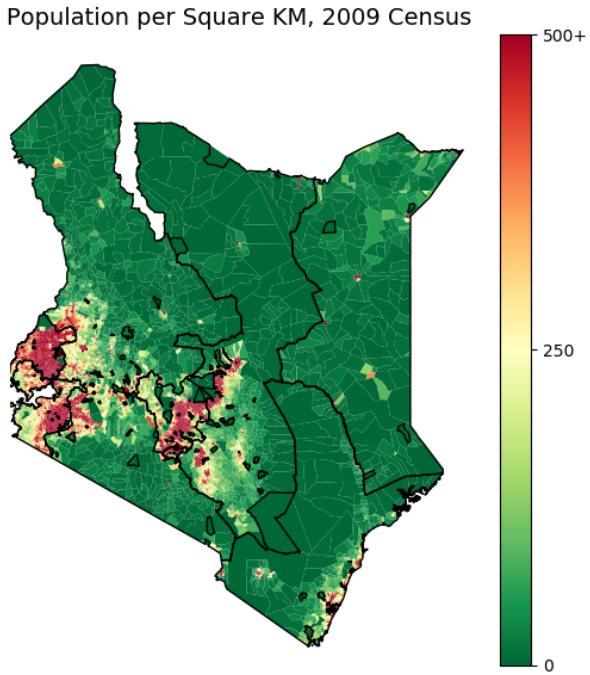


https://www.digitalgendergaps.org/

Also in the South of the World

Examples of Facebook data in Kenya and comparison with Census data from 2009.





Rosenzweig, R., Bergquist, P., Hoffmann Pham, K., Rampazzo, F., and Mildenberger, M. (2020). *Survey sampling in the Global South using Facebook advertisements*. SocArXiv.

Coverage Error

Coverage of Facebook affected by:

- Age
- Gender
- Education
- Country

Sampling Frame

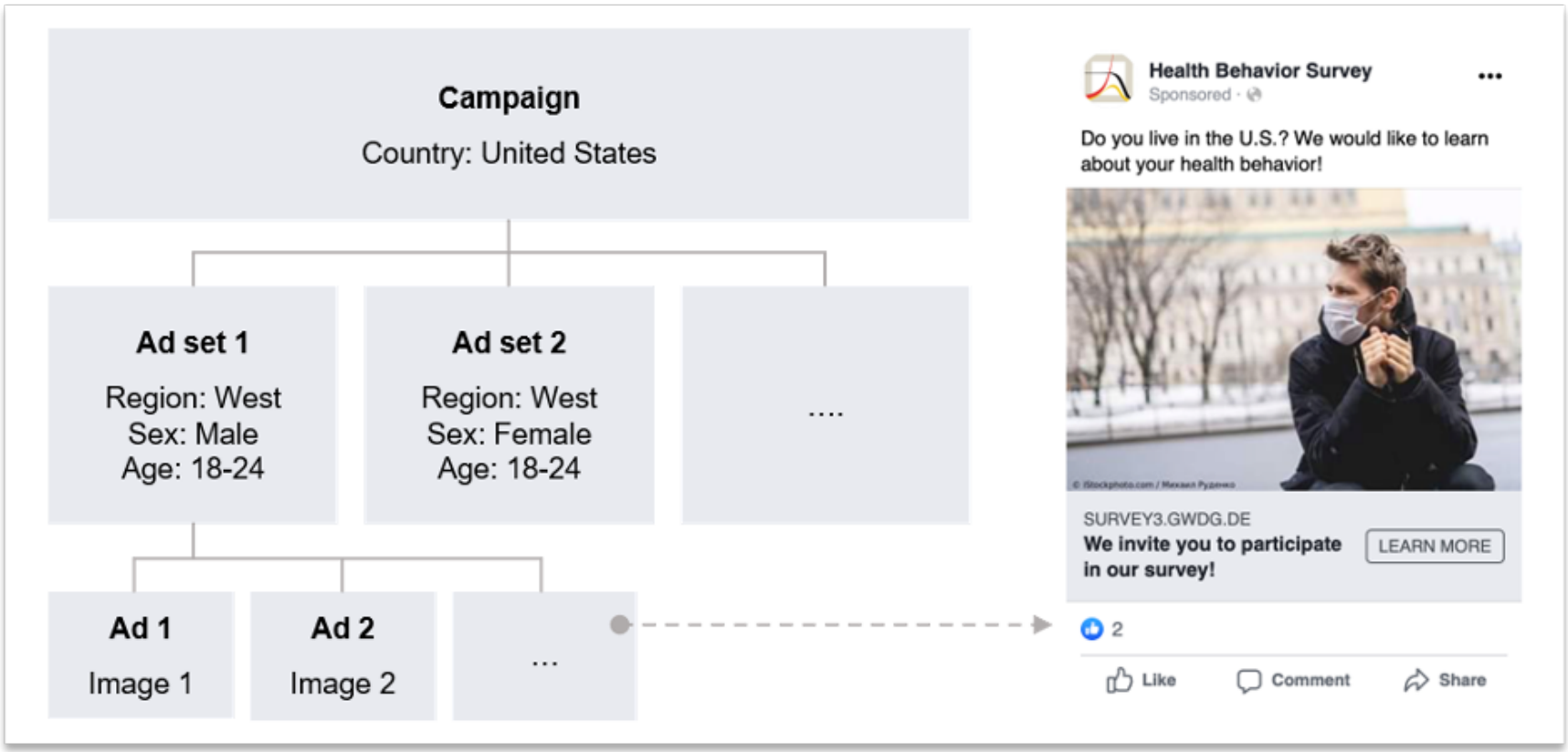
Quotas of Facebook users

distributed as the Inferential Population.

Structure of the Ads Campaigns

Stratum:

- **Sex** (Male and Female),
- **Age** (18-24, 25-44, 44-64, 65+),
- **Regions of residence** (NUTS1/US Census regions).



Grow, A., Perrotta, D., Del Fava, E., Cimentada, J., Rampazzo, F., Gil, S. and Zagheni, E. (2020). *Addressing Public Health Emergencies via Facebook Surveys: Advantages, Challenges, and Practical Considerations*. Journal of Medical Internet Research, 22(12):e20653.

Ads of the Surveys

Pictures

Images Used in the FB ads campaing.

1 – Male athlete
©Adobe Stock/grki

2 – Group of athletes
©Adobe Stock/nd3000

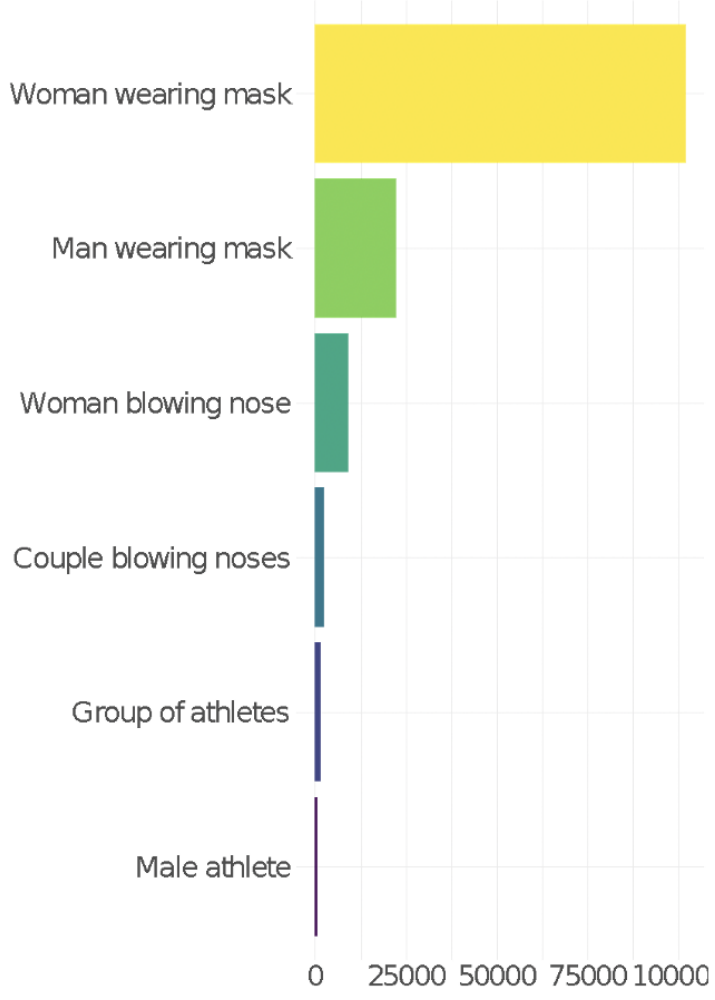
3 – Woman blowing nose
©iStockphoto/Goodboy Picture Company

4 – Couple blowing noses
©iStockphoto/Goodboy Picture Company

5 – Woman wearing mask
©Adobe Stock/shintartanya

6 – Man wearing mask
©iStockphoto/Михаил Руденко

Number of respondents in each image.



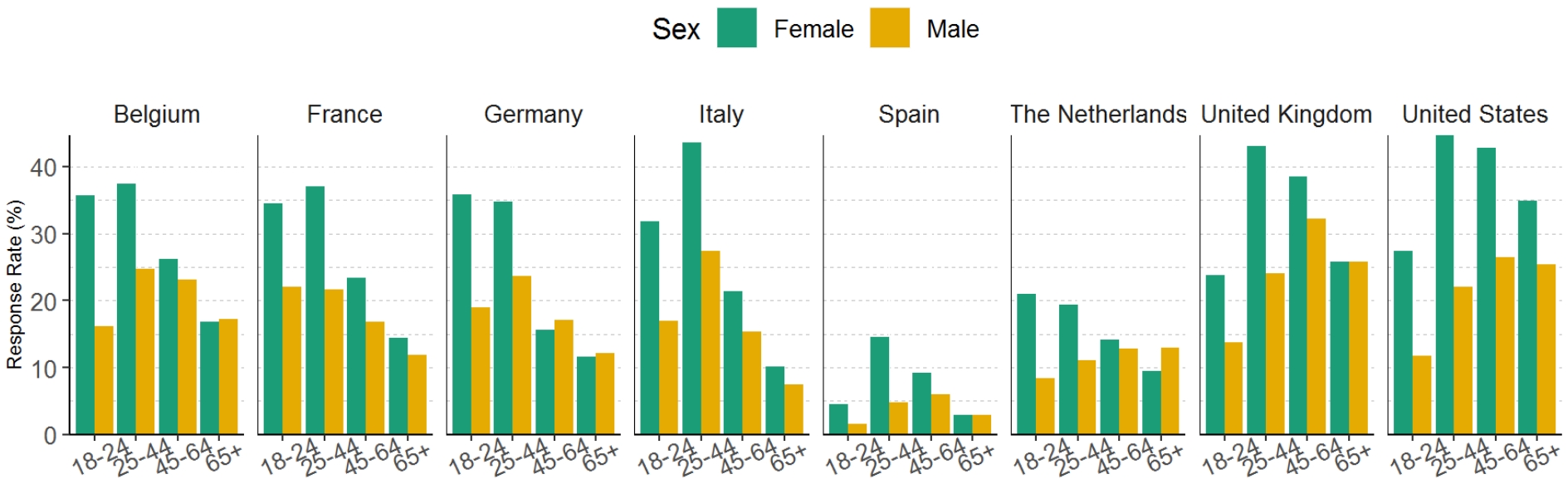
Sampling Error affected by:

- Self-selection of the users
- Facebook's Algorithm
- Trust
- Incentive

Sample

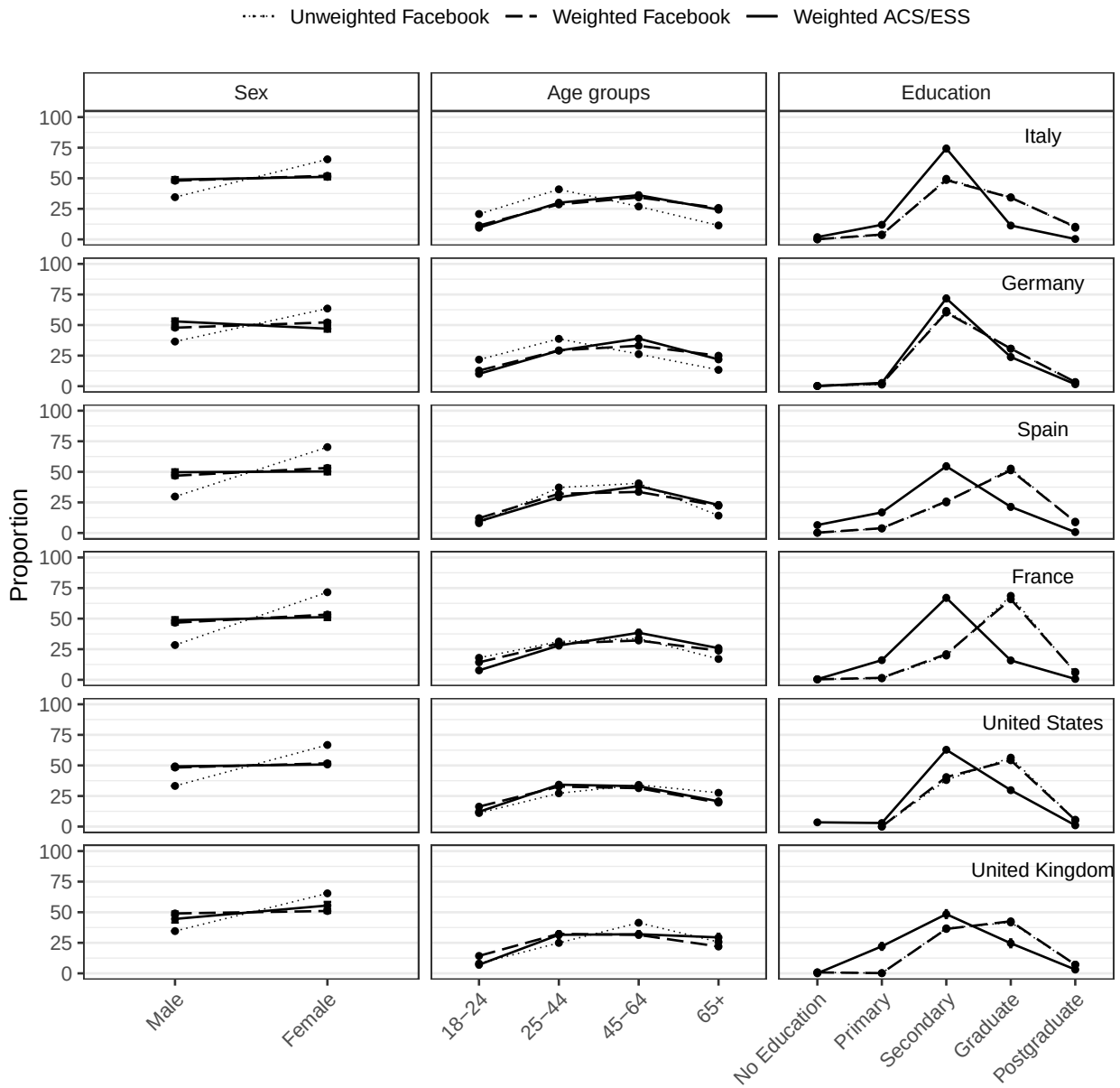
Response Rate by Sex

- **Response Rate** (% Facebook users who completed the questionnaire after clicking on the ad) generally reasonable with some variability
- **Observation 1:** Spain
- **Observation 2:** Female $>$ Male

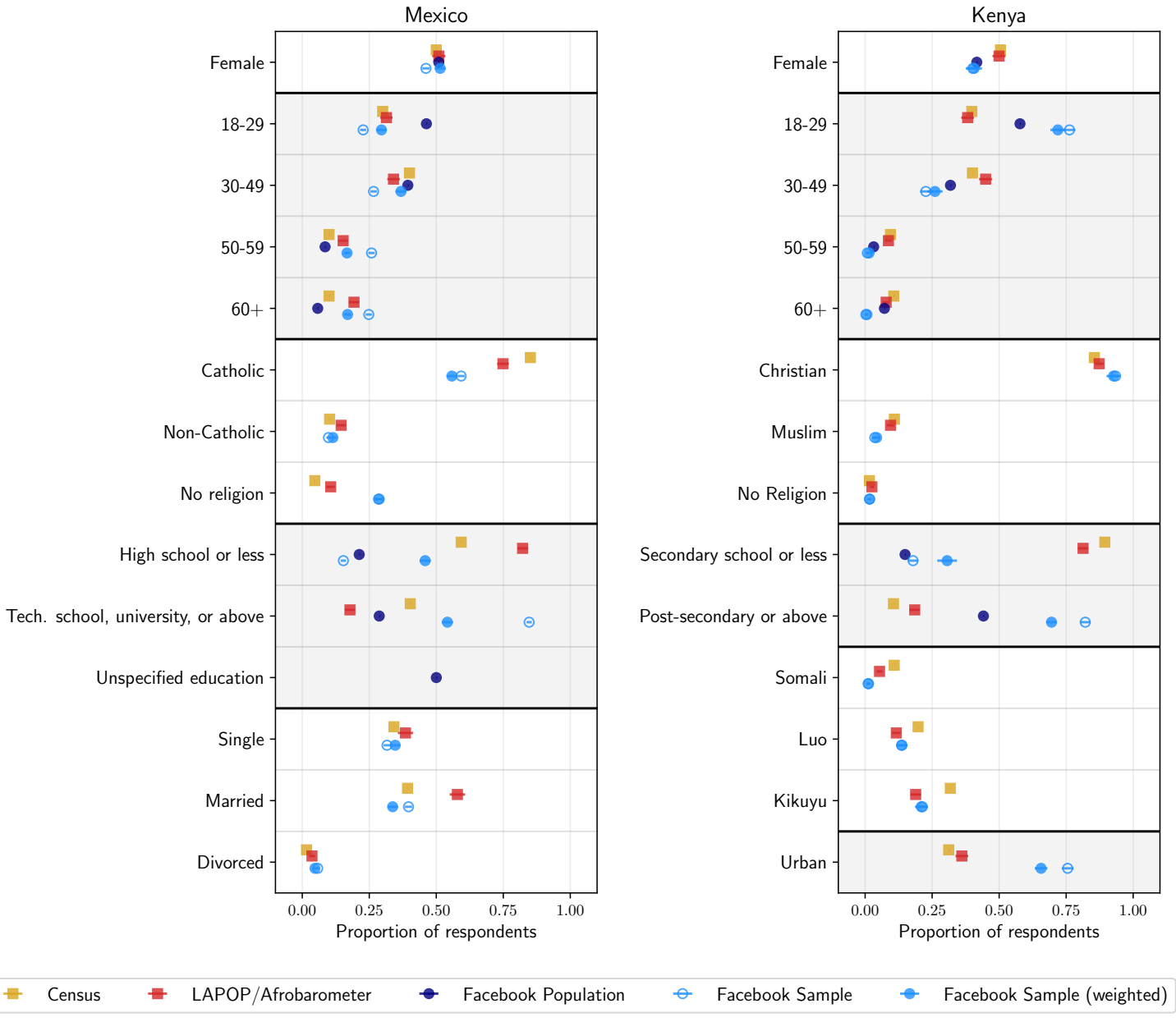


Perrotta, D., Grow, A., Rampazzo, F. et al. *Behaviours and attitudes in response to the COVID-19 pandemic: insights from a cross-national Facebook survey*. EPJ Data Sci. 10, 17 (2021). https://doi.org/10.1140/epjds/s13688-021-00270-1

Comparing Facebook and European Social Survey (ESS)



Mexico and Kenya

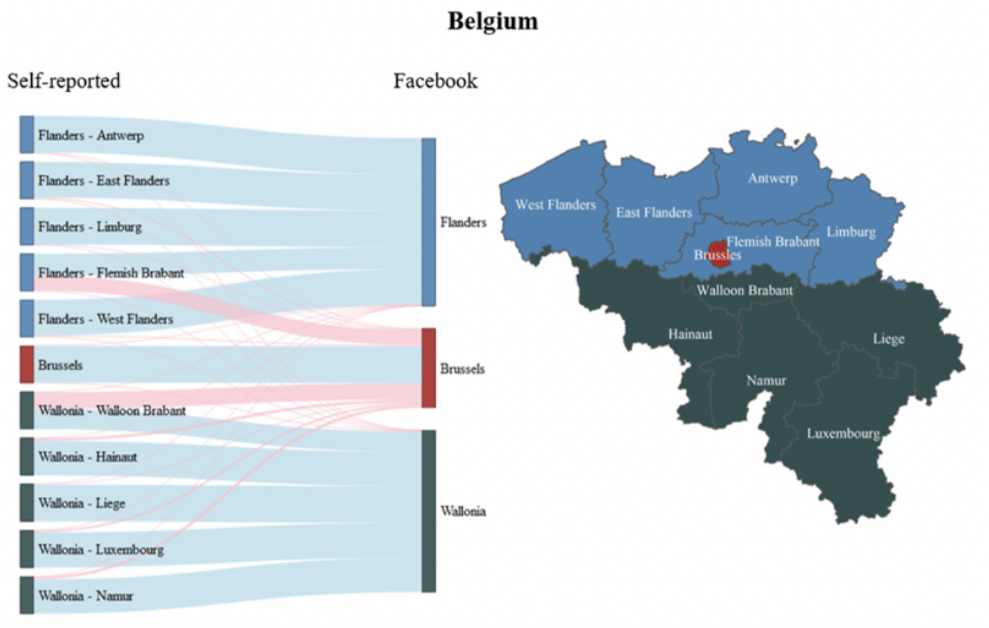


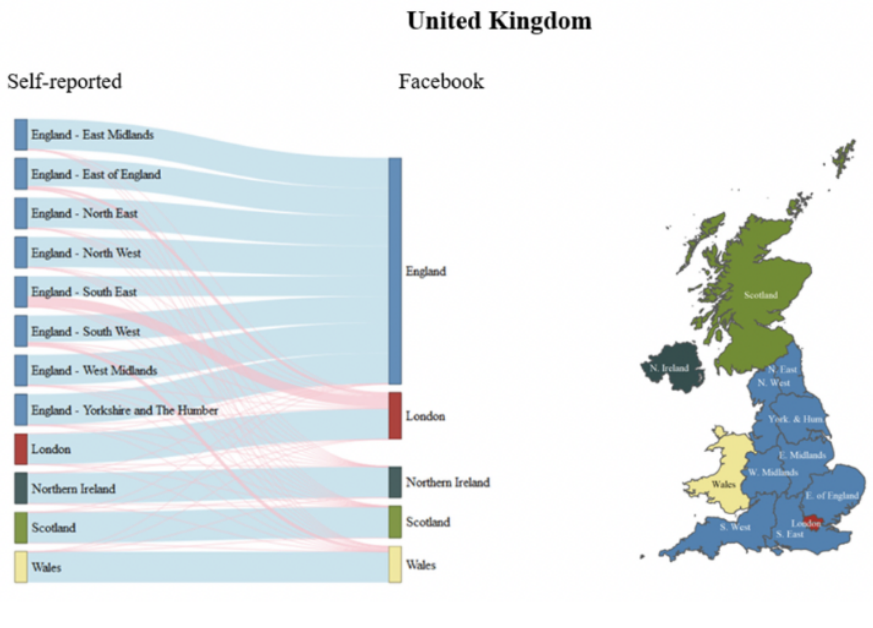
Do we trust Facebook's Targeting?

On **137,224** questionnaires about **90%** was **correctly** identified for **age, sex, and place of residence** (comparison between ads targeting and questionnaires data).

Place of Residence

Specification error in places of residence in the **United Kingdom** and **Belgium**.





To sum up...

	Errors	
Coverage	**Sampling**	**Specification**
Age	Self-selection	Non-response per strata
Sex	Facebook's Algorithm	Area of Residence
Education	Trust	
Location	Incentive	

Questions?

Advantages

- Global Sample
- Targeting for demographic characteristics
- Fast and not too expensive
- Not only Facebook, AmazonTurk or Panels (e.g. Prolific, Qualtrics...)

Limitation

Transparency

Definition **Algorithm** **Ethics**

Limitation

Transparency

Definition **Algorithm** **Ethics**

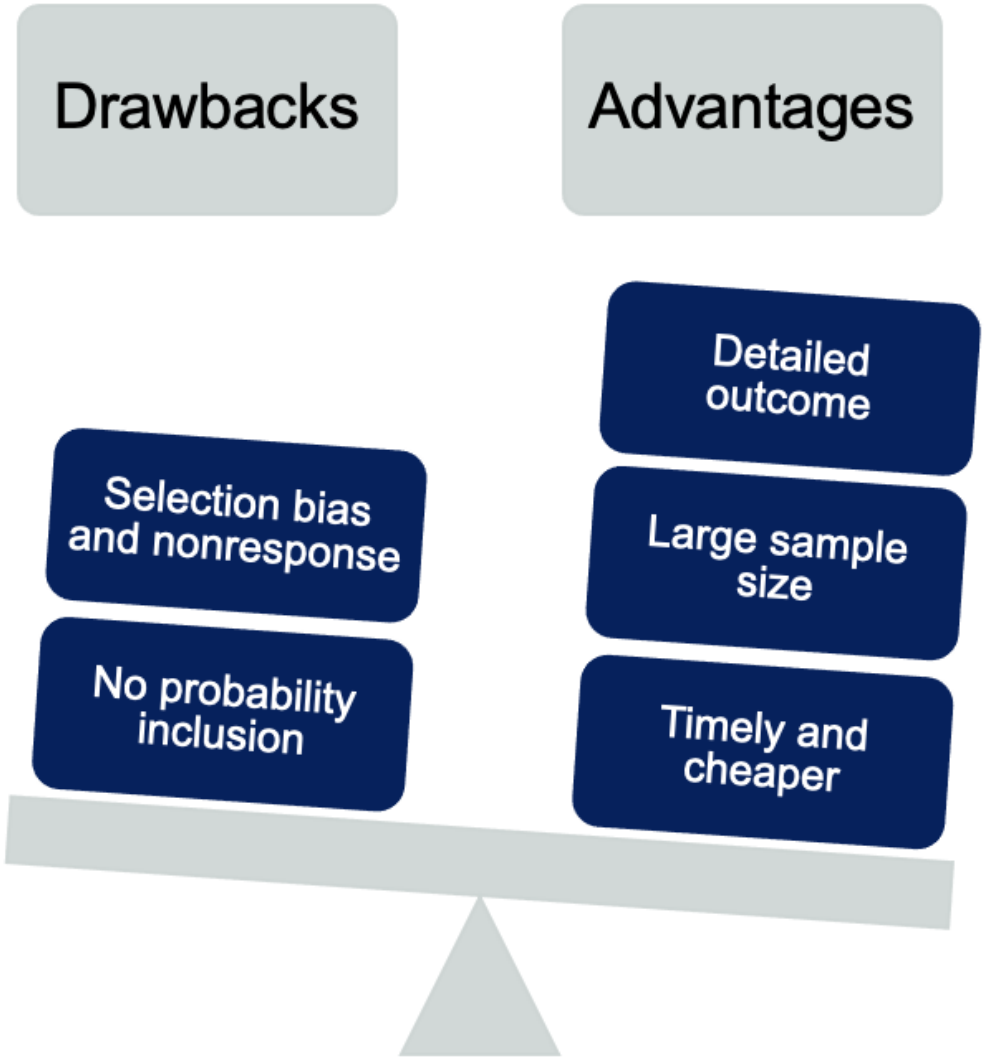
We cannot control the sample.

Questions?

Examples of Non-Probability Samples

- **Convenience Sampling**
  - mall intercept samples
  - volunteer samples
- **Sample matching**
  - quota sampling
- **Network sampling**
  - snowball
  - respondent-driven sampling (RDS, Heckathorn 1997)
  - network sampling with memory (NSM, Mouw and Verdery 2012)

The Two Sides of the Coin



How to Make Inference from NPS?

- Building **pseudo-weights** (Elliot 2009, Elliot and Valliant, 2017)
  - Estimate probability of inclusion relying on a **true probability sample** or **census** with common variables that explain selection mechanism.
- Using **post-stratification** (Little 1993)
  - Adjust for selection bias and correct for imbalances (e.g., by nonresponse) in the sample composition, relying on information from a **true probability sample** or **census**

Assumption behind post-stratification

1. We can describe the population in terms of **K categorical variables** – each variable having J_k categories –, so that the population can be represented by $j = \prod_{k=1}^{K} j_k$ cells.

2. We know the composition of this population, i.e., how many people are in each population cell, based on a **true probability sample** or a **census**.

3. The people who didn't answer the survey are like those who did answer, i.e.,
   1. in cell j, people who were asked to participate are a random sample from cell,
   2. in cell j, people who answered the survey are a random sample of those who were asked.

Assumptions behind post-stratification

- The more cells I have, the more likely to draw a random sample from each cell. ✓
- The more cells I have, the more likely they become sparse. ✗
- What if assumption 3 fails?
  - I can use **multilevel regression** to predict the estimate of the outcome of interest in each cell, regardless of how little information the cell contains.

How to Make Inference from NPS?

- Building **pseudo-weights** (Elliot 2009, Elliot and Valliant, 2017)
  - Estimate probability of inclusion relying on a **true probability sample** or **census** with common variables that explain selection mechanism.
- Using **post-stratification** (Little 1993)
  - Adjust for selection bias and correct for imbalances (e.g., by nonresponse) in the sample composition, relying on information from a **true probability sample** or **census**
- Using **multilevel regression (MR) and poststratification (P)** (Gelman and Little 1997)
  - MR-step for small area/cells estimation by using (i) a **predictive model** with large number of covariates and (ii) **borrowing information** to deal with sparse cells.
  - P-step to adjust for selection bias and correct for imbalances (e.g., by nonresponse) in the sample composition, relying on information from a **true probability sample** or **census**.

Questions?

Multilevel regression and poststratification (MRP)

Social Sciences

- **Election forecasting**:
  - Wang et al. 2015 (polls on Xbox in the US)
  - Lauderdale et al. 2020 (YouGov polls in the UK and US)
  - Zahorski 2020 (polls on Viber and on the street in Belarus)
- **Public opinion**:
  - Lax and Phillips 2009 (on gay rights)
  - Pacheco and Maltby 2019 (on the Obamacare (ACA))

Epidemiology

- **Health quantities**:
  - Downes et al. 2018 (based on health survey)
  - Downes and Carlin 2020
- **COVID-19**:
  - Gelman and Carpenter 2020 (on estimating disease prevalence)
  - Barnwal et al. 2021 (on excess mortality in Bangladesh)
  - Breen et al. 2021 (on social contacts in the US)

Multilevel regression (MR)

We define a hierarchical regression model for the outcome of interest in cell j, which includes varying intercepts a_j^k, for $k = 1, \ldots, K$, and $j = 1, \ldots, J$:

$$g(\theta_j) = g(E[Y_{j[i]}]) = \beta_0 + X_j^T \beta + \sum_{k=1}^{K} \alpha_{j[i]}^k$$

The varying intercepts (or random effects) a_j^k are given independent normal prior distributions:

$$a_j^k \sim N(0, \sigma^2)$$

Multilevel regression (MR)

The varying intercepts $\alpha^k_{j[i]}$ have the effect of partially pooling each θ_j towards the estimate obtained by the model fitted to the whole data ($\bar{y} = \beta_0 + X_j^T \beta$), with sparse cells benefitting the most from this regularization.

$$\tilde{\theta}_j \approx \frac{n_j \sigma_\theta^2 \bar{y}_j + \sigma_j^2 \bar{y}_r}{n_j \sigma_\theta^2 + \sigma_j^2}$$

- n_j: sample size of cell j
- σ_j^2 : within-cell variance of cell j
- $sigma_\theta^2$: between-cell variance

- When $n_j \to 0$, $\tilde{\theta}_j \approx \bar{y}_r$, i.e., with sparse cells, the population estimate is pulled towards the overall uncorrected expected value, $\bar{y}_r$ (complete pool)
- When $n_j \to \infty$, $\tilde{\theta}_j \approx \bar{j}_r$, i.e., with large cells, the population estimate is pulled towards the cell-wise direct estimate, $\bar{y}_j$ (no pool)

Poststratification (P)

Using the known size of each cell j in the population, N_j, we can obtain the estimate of the outcome of interest in the population (or in any subpopulation s):

- The final MRP estimate is a weighted average of cell-wise estimates $\tilde{\theta}_j$, where smaller cells get downweighted and larger cells upweighted:

$$\tilde{\theta}^{MRP} = \frac{\sum_{j=1}^{J} N_j \tilde{\theta}_j}{\sum_{j=1}^{J} N}$$

$$\tilde{\theta}^{MRP} = \frac{\sum_{j=1}^{J} N_j \tilde{\theta}_j}{\sum_{j=1}^{J} N}$$

Post-stratification (P)

How do we choose the variables to construct the post-stratification cells?

- Post-stratification based on cells that are homogeneous w.r.t. outcome of interest Y reduces both variance and bias (Holt and Smith 1979).
- Post-stratification based on cells that are homogeneous w.r.t. the participation in the survey reduces the bias, but not necessarily the variance (Little 1986).

$\longrightarrow$ Choose variables that are predictive of either the survey outcome (primarily) or the response propensities (secondly) (Little and Vartivarian 2005).

Some References about Facebook and Surveys

- Zhang, B., Mildenberger, M., Howe, P. D., Marlon, J., Rosenthal, S. A., Leiserowitz, A. (2020). Quota sampling using Facebook advertisements. Political Science Research and Methods, 8(3), 558-564.
- Perrotta, D., Grow, A., Rampazzo, F. et al. *Behaviours and attitudes in response to the COVID-19 pandemic: insights from a cross-national Facebook survey*. EPJ Data Sci. 10, 17 (2021). https://doi.org/10.1140/epjds/s13688-021-00270-1
- Grow, A., Perrotta, D., Del Fava, E., Cimentada, J., Rampazzo, F., Gil, S. and Zagheni, E. (2020). *Addressing Public Health Emergencies via Facebook Surveys: Advantages, Challenges, and Practical Considerations*. Journal of Medical Internet Research, 22(12):e20653.
- Rosenzweig, R., Bergquist, P., Hoffmann Pham, K., Rampazzo, F., and Mildenberger, M. (2020). *Survey sampling in the Global South using Facebook advertisements*. SocArXiv.
- Grow, A., Perrotta, D., Del Fava, E., Cimentada, J., Rampazzo, F., Gil-Clavel, S., ... Weber, I. (2021). *How reliable is Facebook's advertising data for use in social science research? Insights from a cross-national online survey (No. WP-2021-006)*. Max Planck Institute for Demographic Research, Rostock, Germany.

Some References on Post-stratification

- Heckathorn, D. D. (1997). Respondent-driven sampling: A new approach to the study of hidden populations. Social Problems 44(2), 174–199.
- Mouw, T. and A. M. Verdery (2012). Network sampling with memory: A proposal for more efficient sampling from social networks. Sociological Methodology. 42(1), 206-256.
- Elliot, M. R. (2009). Combining data from probability and non-probability samples using pseudo-weights. Survey Practice 2(6).
- Elliott, M. R. and R. Valliant (2017). Inference for nonprobability samples. Statistical Science 32(2), 249-264.
- Little, R. (1983). Comment on "An evaluation of model-dependent and probability-sampling inferences in sample surveys", by M. H. Hansen, W. G. Madow and B. J. Tepping. Journal of the American Statistical Association 78, 797-799.
- Gelman, A. and T. C. Little (1997). Poststratification into many categories using hierarchical logistic regression. Survey Methodology 23, 127-135.
- Wang, W., D. Rothschild, S. Goel, and A. Gelman (2015). Forecasting elections with nonrepresentative polls. International Journal of Forecasting 31(3), 980-991.
- Lauderdale, B. E., D. Bailey, J. Blumenau, and D. Rivers (2020). Model-based pre-election polling for national and sub-national outcomes in the US and UK. International Journal of Forecasting 36, 399-413.
- Zahorski, A. (2020). Multilevel regression with poststratification for the national level viber/street poll on the 2020 presidential election in Belarus. https://arxiv.org/abs/2009.06615.
- Lax, J. and J. Phillips (2009a). Gay rights in the states: Public opinion and policy responsiveness. American Political Science Review 103, 367-386.
- Pacheco, J. and E. Maltby (2019). Trends in State-Level Opinions toward the Affordable Care Act. Journal of Health Politics, Policy and Law 44(5), 737-764.
- Downes, M., L. C. Gurrin, D. R. English, J. Pirkis, D. Currier, M. J. Spittal, and J. B. Carlin (2018). Multilevel regression and poststratification: A modeling approach to estimating population quantities from highly selected survey samples. American Journal of Epidemiology 187(8), 1780-1790.
- Downes, M. and J. B. Carlin (2020). Multilevel regression and poststratification versus survey sample weighting for estimating population quantities in large population health studies. American Journal of Epidemiology 189(7), 717-725.
- Gelman, A. and B. Carpenter (2020). Bayesian analysis of tests with unknown specificity and sensitivity. Journal of the Royal Statistical Society Series C (Applied Statistics) 69(5), 1269-1283.

Exercise

Example: web survey on COVID-19

`https://github.com/chiccorampazzo/SICSS-Oxford-2022`

Thanks!

Francesco Rampazzo
@chiccorampazzo
francesco.rampazzo@sbs.ox.ac.uk

Funded by the

LEVERHULME TRUST