Computational approaches to estimating key demographic quantities in data-sparse settings

Leverhulme Centre for Demographic Science

Casey F. Breen [1]

[1]University of Oxford | Nuffield College

December 14, 2023

Motivation

- Our ability to study demographic questions with important scientific and policy implications is often limited by **data availability**
- **Computational demography** offers techniques to address this data sparsity

Estimating death rates in complex humanitarian emergencies using the network method

Casey F. Breen [1]

[1]University of Oxford | Nuffield College

December 14, 2023

Joint with: Dennis Feehan (UC Berkeley), Saeed Rahman (Reach Initiative), Christina Kay (Reach Initiative), Joeri Smits (Reach Initiative), Steve Ahuka (University of Kinsasha)

Motivation for study

- Death rates are critical for assessing the severity of a crisis and effectively allocating resources
- How do you best estimate death rates in **humanitarian emergencies**?
- We can't use conventional methods such as retrospective household survey
  - Civil war, earthquake, etc.

Roadmap

1. High-level overview: **network method**
2. Case study in Democratic Republic of the Congo

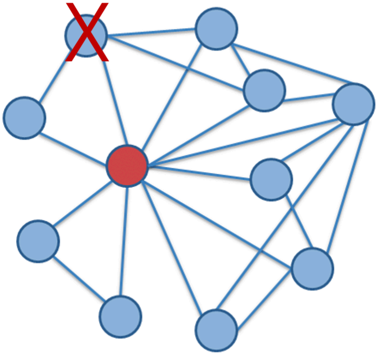
Overview of Network Method

$$M_\alpha = \frac{D_\alpha}{N_\alpha} \quad (1)$$

where

- M_α is the death rate (for subgroup α)
- D_α is the number of deaths
- N_α is the person days of exposure

Harness social networks



People can report valuable information about mortality among their social network

Network method

$$\widehat{M}_\alpha = \frac{D_\alpha}{N_\alpha} \quad (2)$$

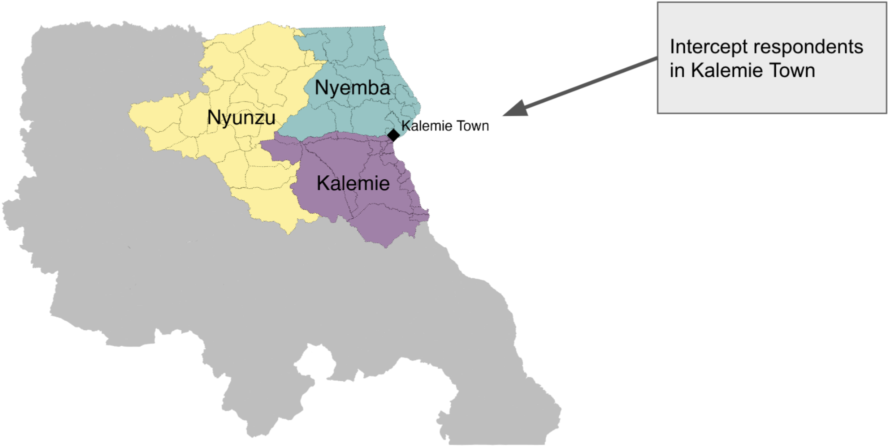
$$= \frac{\sum_{i \in s} y_{i,D_a}}{\sum_{i \in s} \widehat{d}_{F_\alpha,F}} \quad (3)$$

where

- $\sum_{i \in s} y_{i,D_a}$ is the total number of people in respondents' personal network who have died in time window
- $\sum_{i \in s} \widehat{d}_{F_a,F}$ is the total number of people in respondents' personal network

Case Study: Democratic Republic of the Congo

Tanganyika Province, Three Zone De Santes



Key design question – how do we pick a network tie?

- Too big a network: respondents can't report accurately (everyone you have ever talked to)
- Too small a network: we need a much bigger sample size (e.g, parents, siblings)

Formative research (qualitative)

- **Goal**:
  - What deaths can respondents report on accurately?
  - What reference data and reporting window should we chose?
-
- 8 focus groups + 20 individual interviews

Group	**Age**	**Gender**	**Location**
1	<45	Male	Urban
2	<45	Female	Urban
3	<45	Male	Rural
4	<45	Female	Rural
5	45+	Male	Urban
6	45+	Female	Urban
7	45+	Male	Rural
8	45+	Female	Rural

Best option — kin and neighbor networks

Module	Group	Notes
Neighbor	Respondent's Household	
Neighbor	1st Closest Neighbor Household	
Neighbor	2nd Closest Neighbor Household	
Neighbor	3rd Closest Neighbor Household	
Neighbor	4th Closest Neighbor Household	
Neighbor	5th Closest Neighbor Household	
Kin	Respondent's Grandchildren	
Kin	Respondent's Children	
Kin	Respondent's Siblings	
Kin	Respondent's Cousins	
Kin	Respondent's Aunts/Uncles	
Kin	Respondent's Parents	
Kin	Respondent's Grandparents	

Data Collection (Partnership with REACH Initiative)

- Interview respondents at major transit hubs in Kalemie Town
- 8 enumerators, 2 research managers
  - Daily check-ins
  - Weekly data review
- Continuous data collection for 4 months (March 1st 2023 - July 1st 2023)
- Quota sample by geographic region + gender (non-probability)

Data Collection

Interviews conducted at transit hubs such as markets, ports, taxi stations, health clinics, etc.

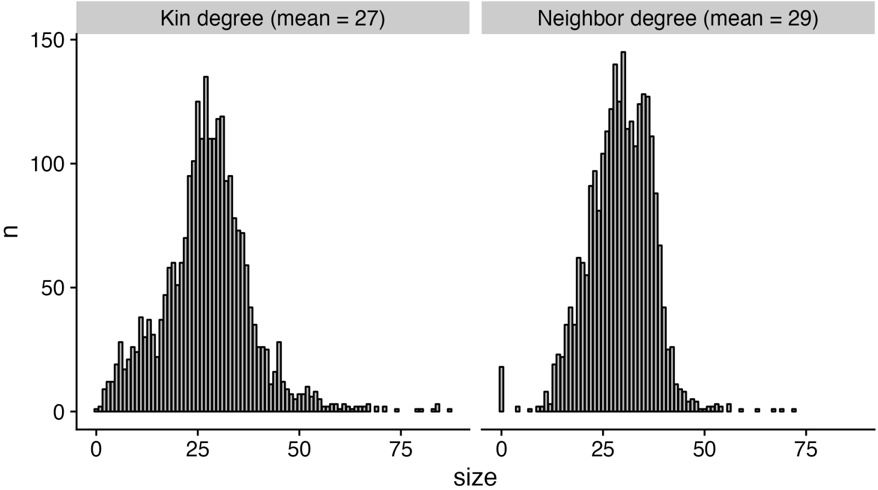
Informed consent on paper form, survey administered on smartphone

Interviews by month (N = 2,650)

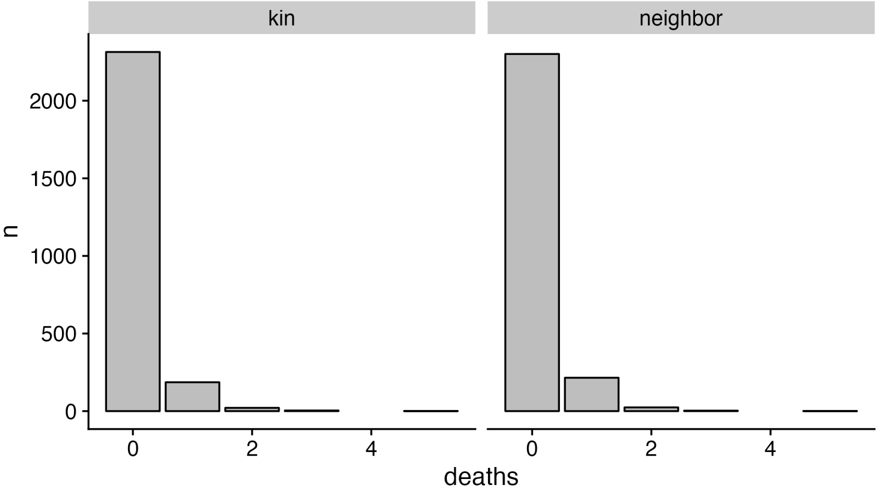
Month	Kalemie	Nyemba	Nyunzu
March	203	198	200
April	201	203	202
May	216	221	232
June	228	237	204

Note: All respondents report on deaths since January 1st, 2023

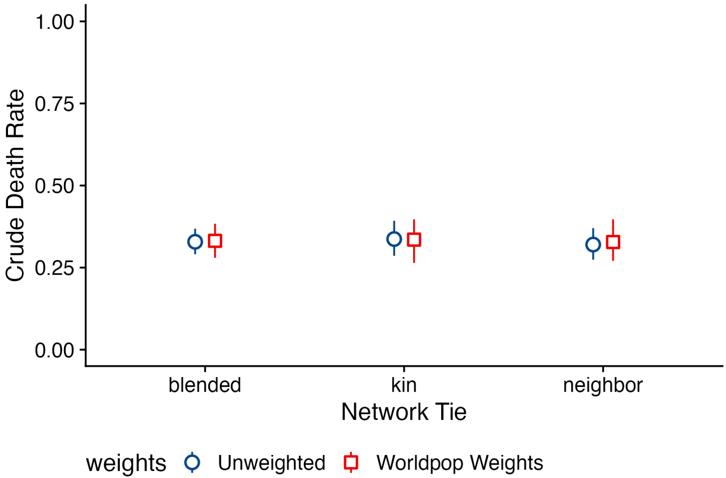
Network size – distribution



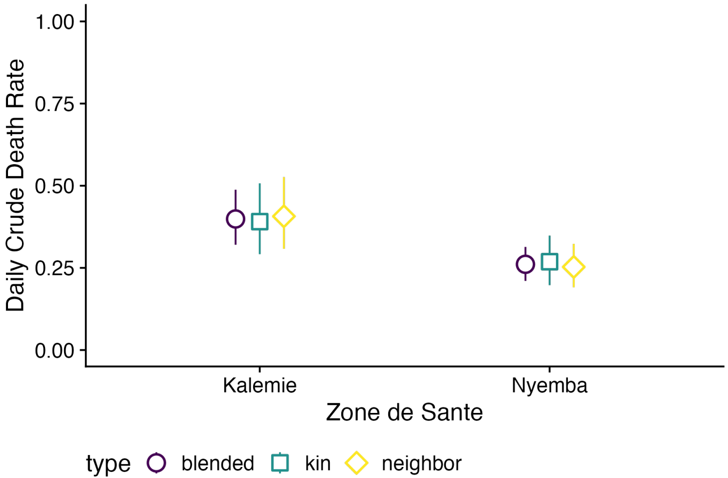
Deaths reported per interview – distribution



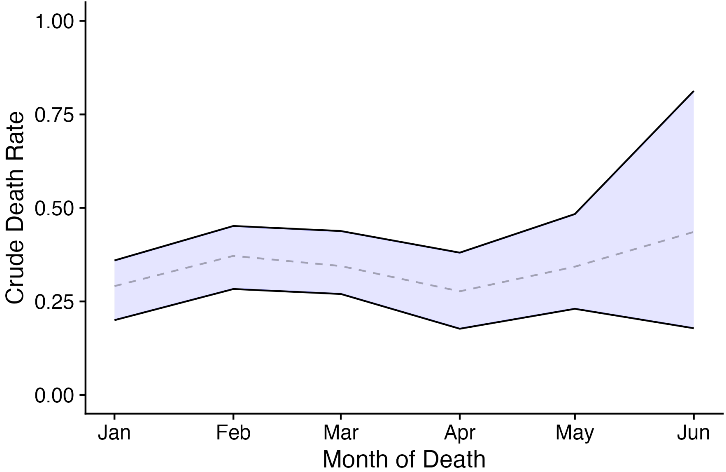
CDR estimates from network method



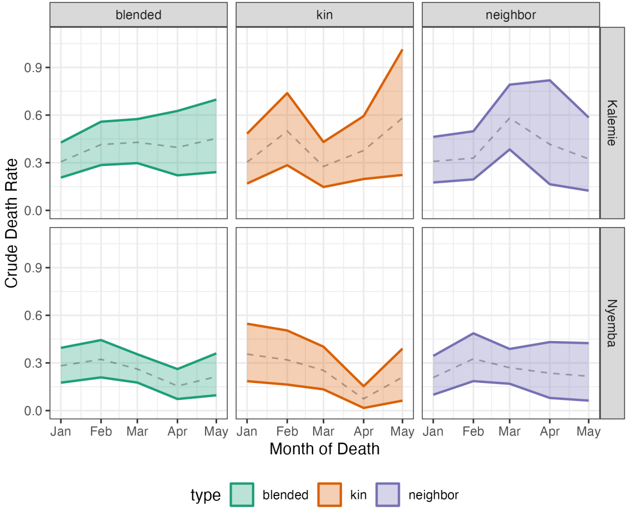
CDR estimates from network method



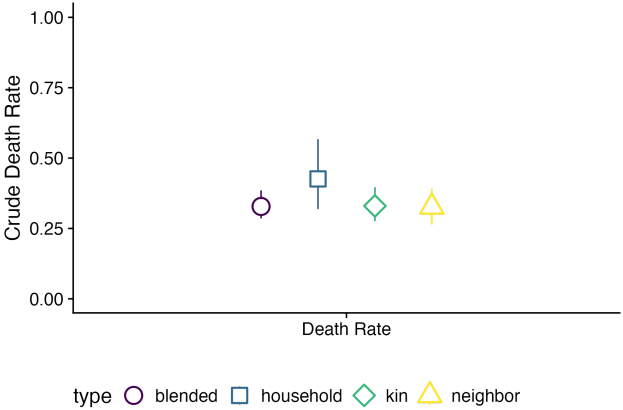
Change over time - blended estimate



Change over time



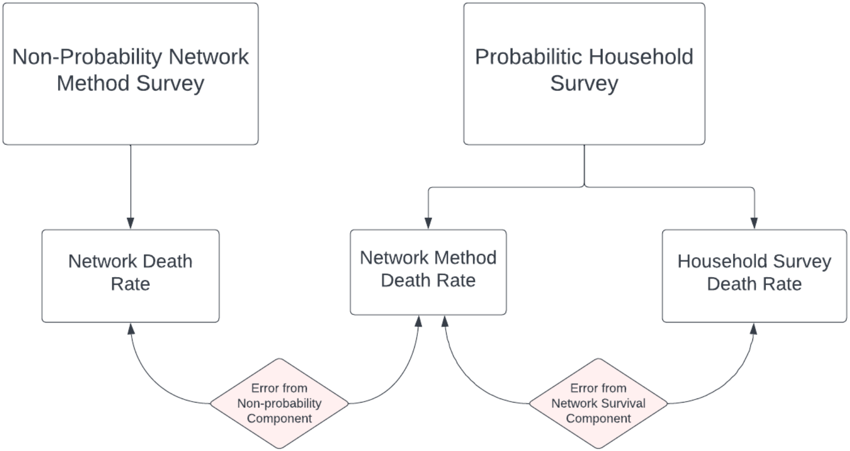
Internal Consistency Check - Household Reports



Next Steps

- Compare estimates to **probability-based**, retrospective household survey (N = 2,970)
- Decompose potential sources of **error**

Study design

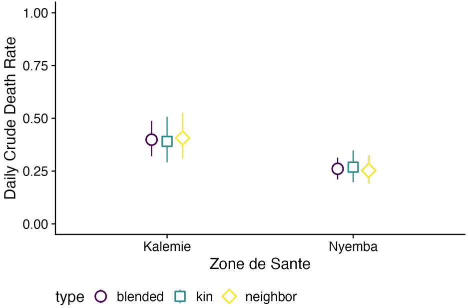


Conclusions

- Network method can **detect large changes** in humanitarian settings, helpful for monitoring mortality
- Highly contextual – **requires** localized knowledge of social networks, diffusion of info about deaths, etc.

Thank You

- Questions?



caseyfbreen

casey.breen@demography.ox.ac.uk

Roadmap

1. Overview of **digital gender gaps** project
2. Our approach to using **social media data** to predict subnational digital gender gaps
3. Overview of subnational estimates

Benefits of digital revolution

- The **digital revolution** has ushered in tremendous societal and economic benefits
  - Lower gender inequality, lower maternal/child mortality, higher contraception (Rotondi et al., 2020)
  - Boost social connectivity, social learning, access to vital services (Unwin, 2009; DiMaggio and Hargittai, 2001; Suri and Jack, 2016)
  - Increases levels of education, economic benefits (Hjort and Poulsen, 2019; Kho, Lakdawala and Nakasone, 2018; Kharisma, 2022)
- Benefits are often greatest in the most unequal, disadvantaged areas

Tracking the digital divide

- Access to digital technologies such as mobile phones and internet remains **highly unequal**
  - Especially in low- and middle-income countries
  - Especially among women
- **UN Sustainable Development Goals (SDGs)**: Reducing inequalities in access to digital technologies by gender (SDG5) and reducing digital literacy gaps (SDG4)

Digital gender gaps project overview

1. **Data infrastructure**: Map and understand gender gaps in digital connectivity and social media use
   - **Today - subnational estimates**
2. **Impacts research**: impacts of digital information and capabilities on women's economic and social empowerment outcomes
   - Cross-national, comparative perspective (low- and middle- income countries)

Original “impacts” research

Using Facebook ad data to track the global digital gender gap

Masoomali Fatehkia [a], Ridhi Kashyap [b], Ingmar Weber [c]

Show more

Add to Mendeley Share Cite

Regular article | Open access | Published: 29 July 2021

Analysing global professional gender gaps using LinkedIn advertising data

Ridhi Kashyap & Florianne C. J. Verkroost

EPJ Data Science **10**, Article number: 39 (2021) | Cite this article

RESEARCH ARTICLE | SOCIAL SCIENCES |

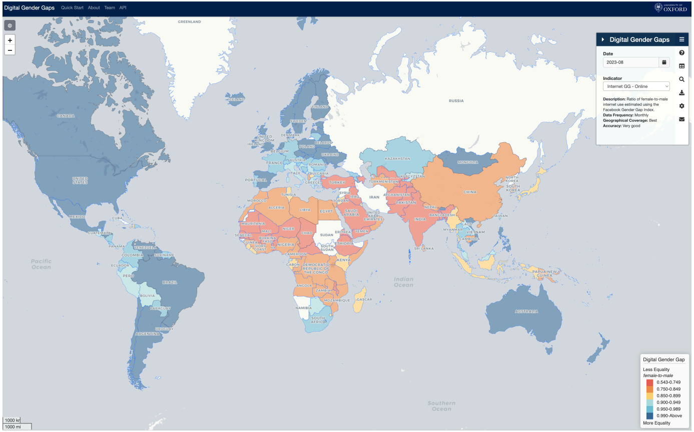
Leveraging mobile phones to attain sustainable development

Valentina Rotondi, Ridhi Kashyap, Luca Maria Pesando, +1, and Francesco C. Billari Authors Info & Affiliations

Edited by Barbara Entwisle, University of North Carolina at Chapel Hill, Chapel Hill, NC, and accepted by Editorial Board Member Mary C. Waters April 6, 2020 (received for review May 30, 2019)

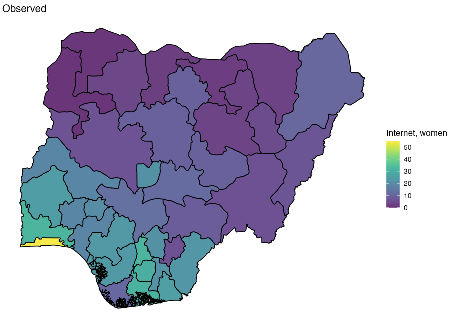
June 1, 2020 | 117 (24) 13413-13420 | https://doi.org/10.1073/pnas.1909326117

Data infrastructure – digitalgendergaps.org



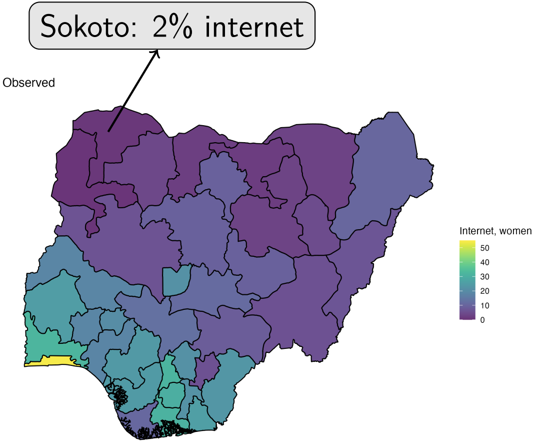
(Kashyap et al., 2020)

Adoption of digital technology varies geographically

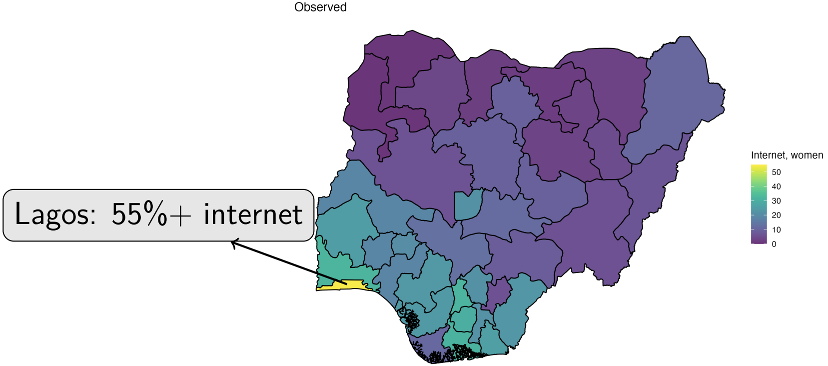


Source: Nigeria, Demographic and Health Survey

Women using internet, past 12 months

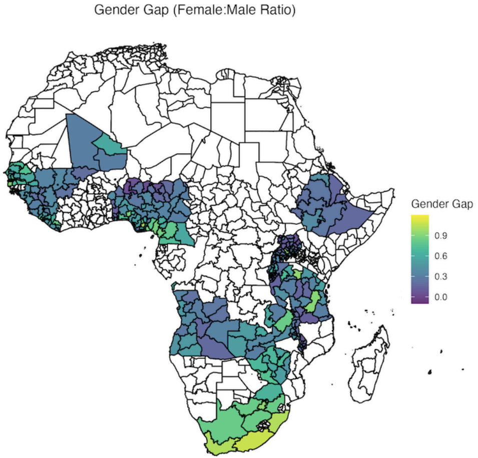


Women using internet, past 12 months



Develop subnational estimates of adoption

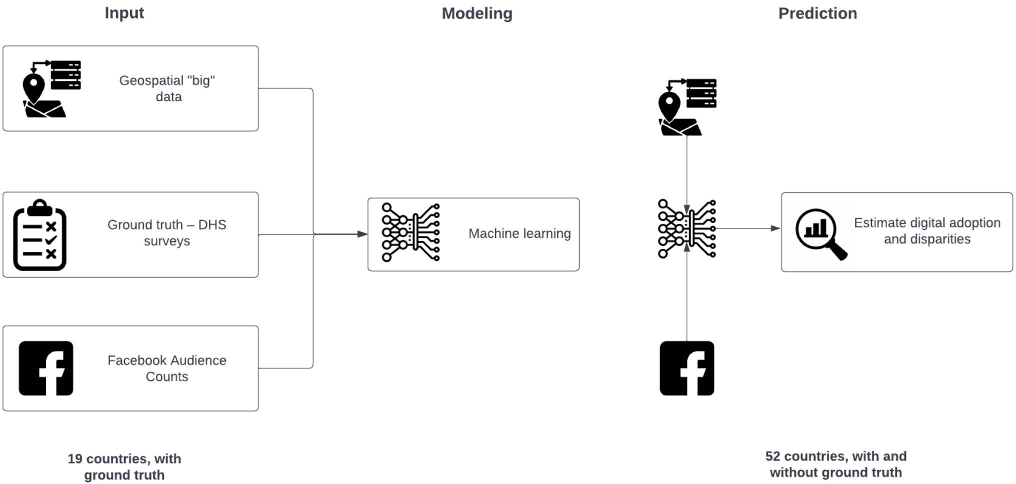
- **Goal**: Develop estimates of internet and mobile adoption by gender and digital gender gaps
- First GADM1 subnational level
  - N = 874



Prediction framework - theoretical background

- Digital gender gaps will be shaped by overall levels of economic development and digital infrastructure
- **Patriarchal** norms and beliefs will moderate this relationship

Overview of approach

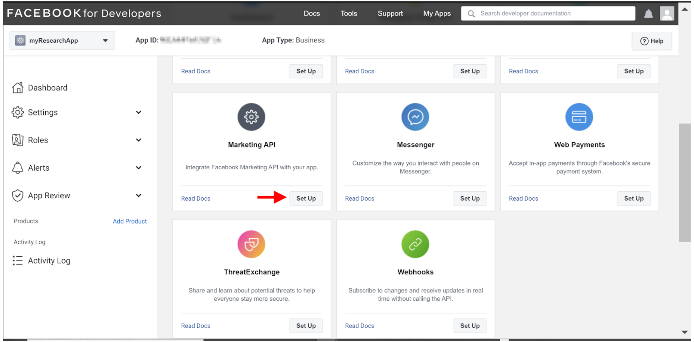


Ground truth – Demographic and Health Surveys (DHS)

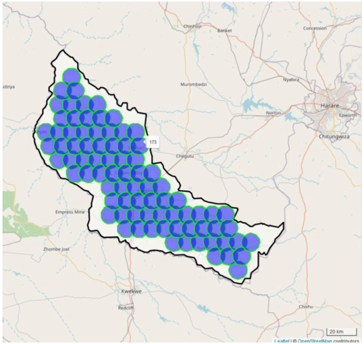
- Household surveys representative at the first subnational level
  - Standardized sample design, questionnaire, implementation, etc.
  - Questions on individual-level internet use and mobile phone use (wave 7 onwards)
- Focus on 19 different DHS surveys, 2016-2020

Facebook audience counts

- Collected through public marketing API
- Specify geographic region (FB template or custom region)
- Disaggregated counts by gender, age, device type, etc.



Facebook audience counts 'online predictors'



- Collected in 2021:
  1. Facebook penetration 13+ female
  2. Facebook penetration 13+ male
  3. Facebook audience 13+ gender gap
  4. iOS 13+ female fraction
  5. iOS 13+ male fraction
  6. WiFi age 13+ female fraction
  7. WiFi age 13+ male fraction
  8. 4G network age 13+ female fraction
  9. 4G network age 13+ male fraction
  10. FB rural WiFi mean (pop weighted)

Geospatial and population data

- Include 'offline' predictors that are uniformly available and consistent across subnational units
  - Satellite-derived nighlights data
  - Population density
  - Subnational education index, income index, human development index (HDI), gender development index (GDI)

Full set of offline predictors

Variable Name	Source	Country (N)
Educational Index Females	Subnational Dev. Database	50
Educational Index Males	Subnational Dev. Database	50
Income Index Females	Subnational Dev. Database	50
Income Index Males	Subnational Dev. Database	50
Subnational GDI	Subnational Dev. Database	50
Subnational HDI Females	Subnational Dev. Database	50
Subnational HDI Males	Subnational Dev. Database	50
WPop 2020 Age 15-49 Female Frac	WorldPop	58
WPop 2020 Age 15-49 Male Frac	WorldPop	58
WPop 2020 Pop Density	WorldPop	59
Nightlights Mean Pop Weighted	Earth Observation Group	58

Outcomes of interest (from DHS)

Indicators	Women	Men	Gender Gap
Mobile Phone Ownership	✓	✓	✓
Internet Use, Past 12 Mo	✓	✓	✓

Defining a Digital Gender Gap

$$\text{Gender Gap} = \frac{\text{Indicator}_f/\text{Indicator}_m}{\text{Pop}_f/\text{Pop}_m} \tag{4}$$

where

- Indicator_f is the number of female (male) users aged 15–49 (e.g., internet, past 12 months)
- Pop_f is the total population of women (men) aged 15–49

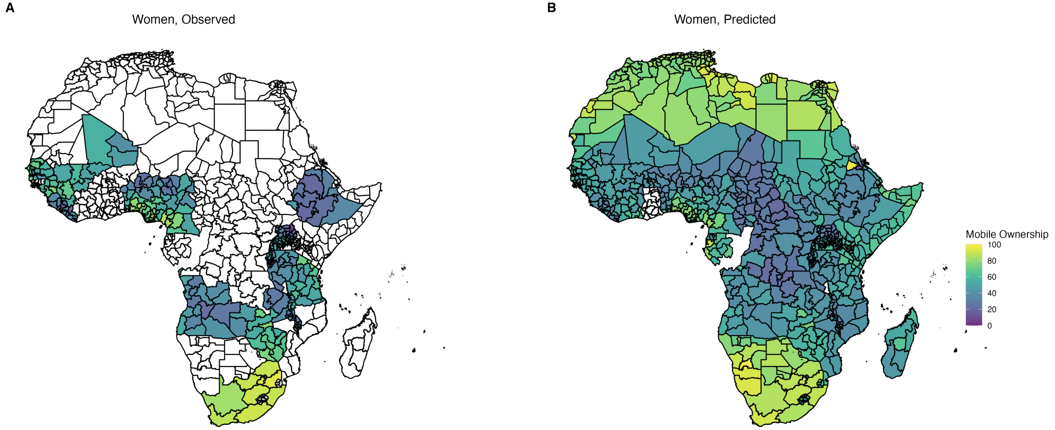
Machine Learning Strategy

- How do you pick the **best** machine learning algorithm?
- Fit lots of algorithms, see which have the best performance
- Ensemble learning to combine algorithms and tests performance using cross-validation to estimate mean squared error for each algorithm (Van der Laan, Polley and Hubbard, 2007)

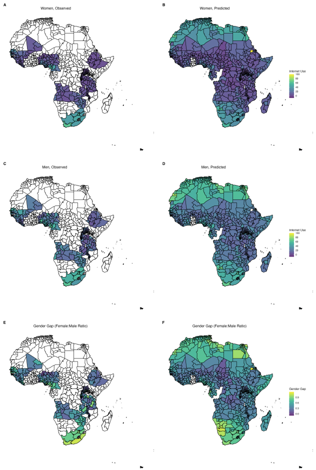
Machine Learning Algorithms Considered

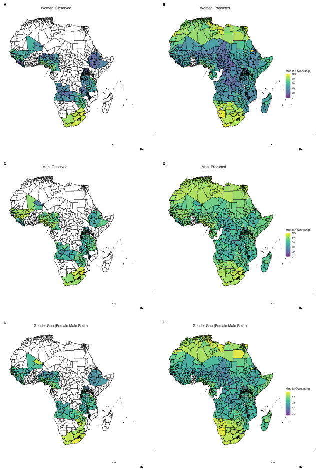
Algorithm	Description
glmnet (Lasso)	Lasso Regression
glmnet (Ridge)	Ridge Regression
glmnet (Elastic Net)	Elastic Net with 50% L1 Ratio
polspline	Polynomial Spline
ranger	Random Forest with 100 Trees
gbm	Gradient Boosted Machine
glm	Generalized Linear Model
xgboost	Extreme Gradient Boosting
SuperLearner	Ensemble method combining multiple learning algorithms

Greatly expanded coverage of digital technology adoption



Similar overall patterns for internet and mobile

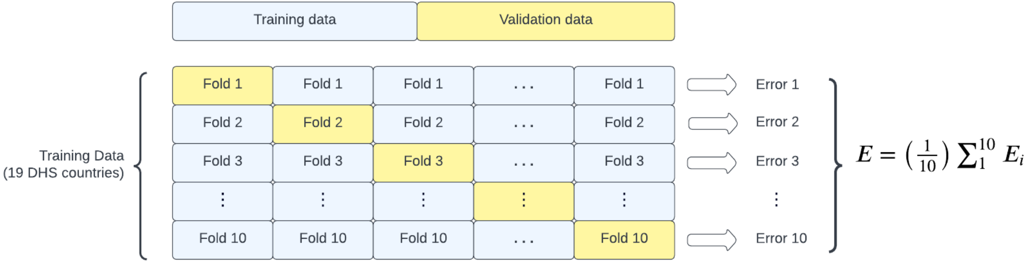




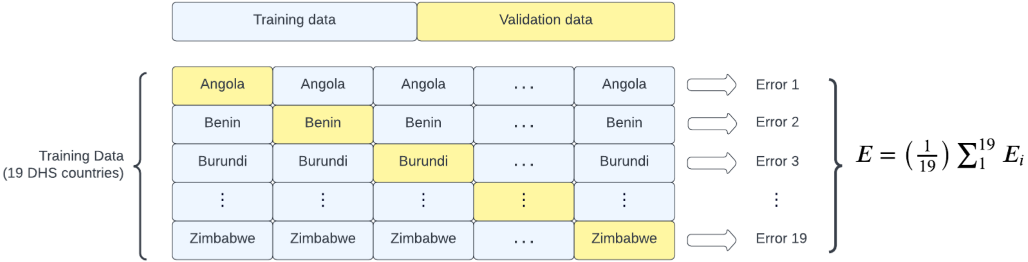
Testing model performance

- How do we assess model performance?
- **Cross-validation** using 19 countries with ground truth data

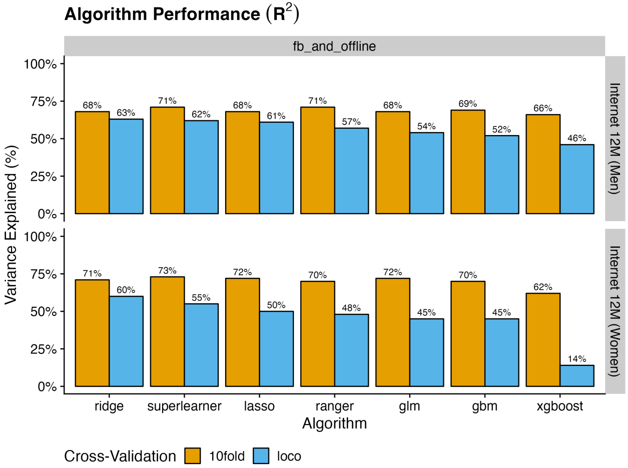
10-fold cross validation



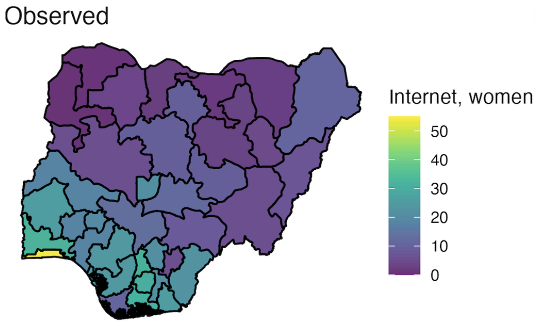
Leave-one-country-out cross validation

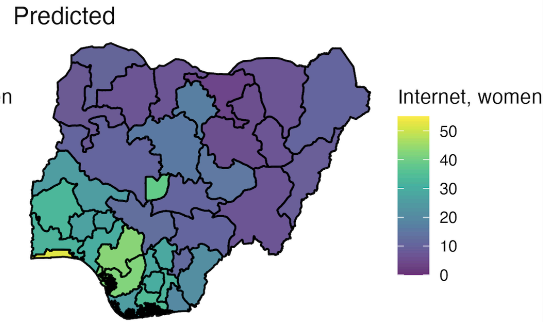


Model performance

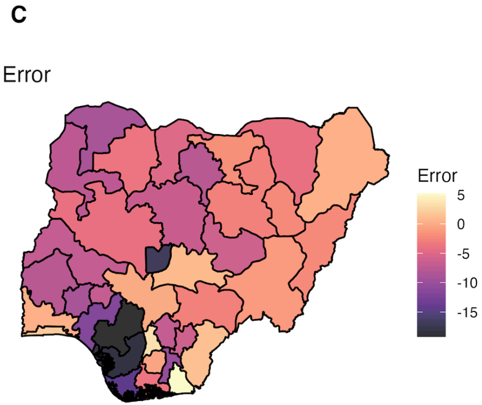


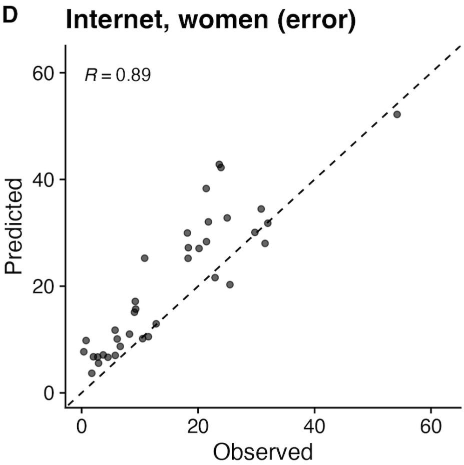
Results for Nigeria (Leave-one-country-out)



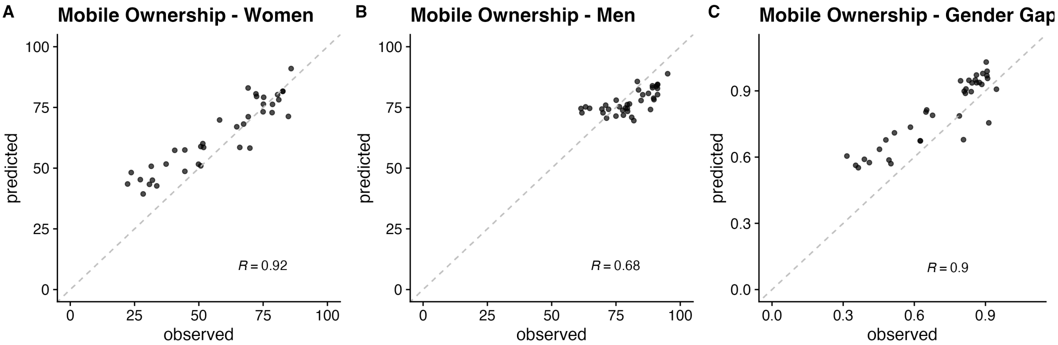


Assessing predictive accuracy

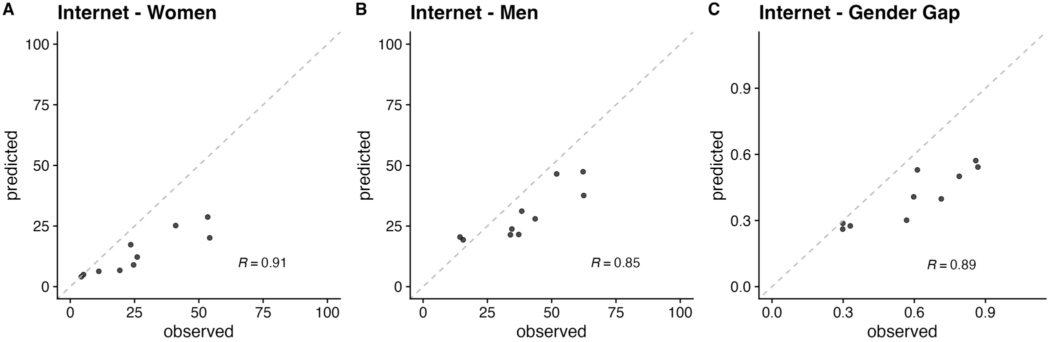




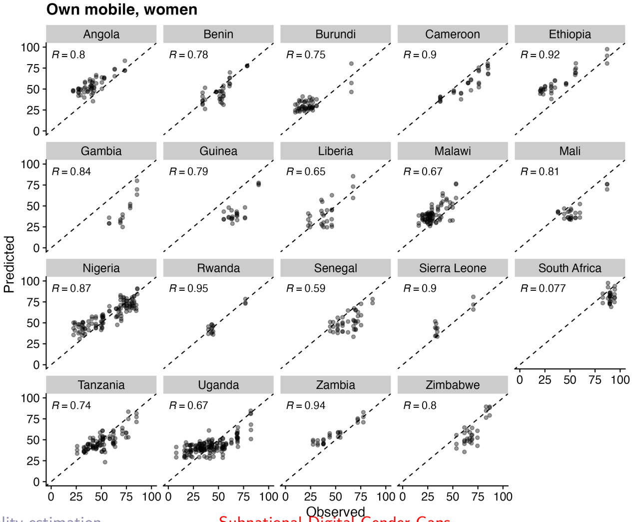
Overall predictiveness – mobile



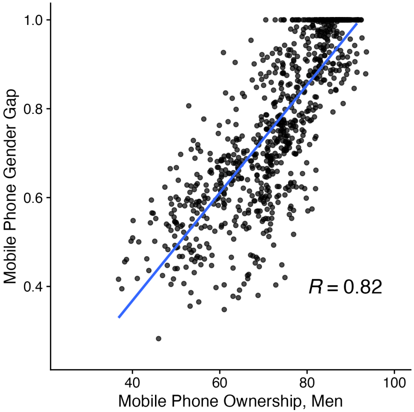
Overall predictiveness – internet



Large variation in predictive accuracy across countries



Relationship: levels of mobile phone penetration and gender gaps



Next steps and future opportunities

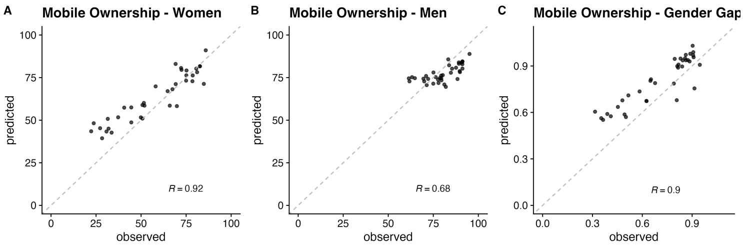
- Regular Facebook collections and pipeline to monitor trends over time
- Residual analysis + quantifying uncertainty: what factors explain where model does worse?

Summary

- Using Facebook audience counts **greatly expands** our ability to accurately predict digital gender gaps in countries with no ground truth
- Huge **disparities** in access to mobile and internet technologies between and within countries
- New opportunities to study **population-level impacts** of digital technology using these subnational estimates

Thank You

- Questions?



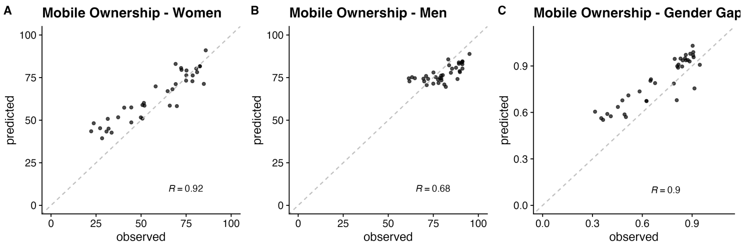
caseyfbreen
casey.breen@demography.ox.ac.uk

Conclusions

- Many long-standing **data sparsity** issues, previously didn't have techniques to overcome
- Computational methods + digital data source open up new **research possibilities**
  - More to come...

Thank You

- Questions?



caseyfbreen
casey.breen@demography.ox.ac.uk

References

DiMaggio, Paul and Eszter Hargittai. 2001. "From the 'Digital Divide' to 'Digital Inequality': Studying Internet Use as Penetration Increases." p. 25.

Hjort, Jonas and Jonas Poulsen. 2019. "The Arrival of Fast Internet and Employment in Africa." *American Economic Review* 109(3):1032–1079.

Kashyap, Ridhi, Masoomali Fatehkia, Reham Al Tamime and Ingmar Weber. 2020. "Monitoring Global Digital Gender Inequality Using the Online Populations of Facebook and Google." *Demographic Research* 43:779–816.

Kharisma, Bayu. 2022. "Surfing Alone? The Internet and Social Capital: Evidence from Indonesia." *Journal of Economic Structures* 11(1):8.

Kho, Kevin, Leah K Lakdawala and Eduardo Nakasone. 2018. "Impact of Internet Access on Student Learning in Peruvian Schools.".

Rotondi, Valentina, Ridhi Kashyap, Luca Maria Pesando, Simone Spinelli and Francesco C. Billari. 2020. "Leveraging Mobile Phones to Attain Sustainable Development." *Proceedings of the National Academy of Sciences* 117(24):13413–13420.

Suri, Tavneet and William Jack. 2016. "The Long-Run Poverty and Gender Impacts of Mobile Money." *Science* 354(6317):1288–1292.

Unwin, P. T. H. 2009. *ICT4D: Information and Communication Technology for Development*. Cambridge University Press.

Van der Laan, Mark J., Eric C. Polley and Alan E. Hubbard. 2007. "Super Learner." *Statistical Applications in Genetics and Molecular Biology* 6(1).