Age-disaggregated subnational patterns of internet and mobile phone adoption

Michael Y.C. Chong[1,2,3] and Ridhi Kashyap[1,2,3]

[1]Department of Sociology, University of Oxford, United Kingdom
[2]Nuffield College, University of Oxford, United Kingdom
[3]Leverhulme Centre for Demographic Science, University of Oxford, United Kingdom

May 2026

Abstract

Digital technologies are transforming population and development processes, yet access remains uneven across geography, gender, and age – particularly in low- and middle-income countries (LMICs). While younger individuals are typically more connected, most existing evidence on these age patterns comes from high-income countries, leaving uncertain whether similar dynamics hold in LMICs. Whether digital gender gaps persist among younger cohorts in LMICs is also unclear. Reliable subnational, age-disaggregated data on who is connected remain scarce, limiting understanding of demographic digital divides and progress toward the Sustainable Development Goals. This paper introduces a two-stage approach to estimate subnational age- and gender-specific patterns of internet and mobile adoption across LMICs. We first smooth sparse survey data from the Demographic and Health Surveys (DHS) and Multiple Indicators Cluster Surveys (MICS) across ages to generate age curves of digital adoption for subnational areas. We then apply a machine learning model integrating social media, geospatial, and development covariates to predict adoption where survey data are unavailable. The results will provide new global, age-structured subnational estimates, advancing digital demography by revealing demographic dimensions of digital inequality.

1 Introduction

Digital technologies are transforming how individuals access information, access goods and services and connect with each other, with significant implications for population and development outcomes [18, 9, 16]. Despite their spread, access to these technologies continues to be shaped by geography, gender and age. Low- and middle-income countries (LMICs) have a larger share of their populations unconnected to the internet compared with high-income countries [10, 23], and women in LMICs in particular are less likely to use the internet and personally own mobile phones [8, 12, 3]. Age differences in internet use generally point to younger populations being more connected than older populations, with age gaps shrinking over time. Yet, much of our current understanding of age patterns of digital connectivity draws on the experience of high-income countries [4, 21].

Despite growing policy attention to universal digital access, reliable demographically disaggregated data on technology adoption remain limited, especially at subnational levels in LMICs [3, 19]. Understanding who is connected, and how patterns of digital adoption vary within and between countries, is essential for monitoring progress toward the Sustainable Development Goals (SDGs) – particularly those related to education, gender equality, decent work, and reducing inequalities (Goals 4, 5, 8, and 10) [20].

Recent work has sought to address these data gaps by using a machine learning approach that integrates population, social media and geospatial covariates to generate harmonised indicators of subnational internet and mobile adoption by gender for LMICs [3]. This work highlights that within-country inequalities in internet and mobile adoption are often as large as between country inequalities, and especially large at lower levels of human development. Yet, whether large subnational disparities prevail even among younger cohorts experiencing more rapid digital diffusion in LMICs remains poorly understood. Similarly, while [3] identify persistent gender gaps in internet and mobile adoption within countries, to what extent gender gaps are smaller among younger cohorts is unclear.

This paper proposes a new framework for estimating subnational age- and gender-specific patterns of internet and mobile adoption for LMICs. We develop a two-stage approach, which first smooths sparse survey data across ages and then applies an ML model to predict observed digital adoption patterns from surveys from a combination of social media, geospatial and development covariates. This approach allows us to estimate smoothed age-specific internet and mobile adoption curves, and expand coverage of digital adoption indicators beyond the 33 countries for which survey data are available to 117 LMICs. By combining the reliability of demographic surveys with the coverage of digital and geospatial data, this study contributes to emerging work in digital demography that leverages new data sources to understand population dynamics and the demographic dimensions of digital in-/exclusion [11].

2 Data

We combine ground-truth data from established household surveys with geographically-aligned development indicators and social media trace data to perform our estimation.

Our ground-truth data come from 33 Demographic and Health Surveys (DHS) [2] conducted between 2017 and 2024 and 24 Multiple Indicator Cluster Surveys (MICS) [1] conducted between 2017 and 2023. In the present paper, we consider the data from the DHS only, and MICS data will be incorporated at a later stage. Both sets of household surveys are widely used to understand social, economic, and health disparities due to their harmonized survey questions and sampling design. Our sample includes the surveys which measure individual-level internet use and mobile phone adoption, and have geographic information allowing estimation at subnational admin-1 levels. The individual-level data also include age and gender, which enable us to produce gender- and age-disaggregated estimates.

From the DHS microdata, we use individual-level survey responses of men and women aged 15-49 to the question of having used the internet within the past 12 months to measure internet adoption, and responses to the question of owning a mobile phone to measure mobile phone adoption.

We use a number of development indicators as covariates to provide additional information when survey data is noisy or unavailable. Development and gender-equality indicators come from the Global Data Lab [17], and include measures of human development, gender

development, income, and education and national and subnational levels. We also use a map of nighttime light map data from the Earth Observation group [7, 6] as a proxy for local development, which measures the radiance of lights at night using satellite imagery. We summarize nighttime light data at the subnational level by taking the mean radiance within each subnational unit.

Finally, we include social media trace data as covariates in our estimation. The Facebook Marketing API provides advertisers estimates of their potential ad reach through queries of the number of target users with specific attributes. This tool has been used by researchers in several contexts to study population movement [13, 15], social dynamics [5], and indeed the adoption of digital technologies [3, 12]. Using the Facebook Marketing API, we collected monthly active user (MAU) counts disaggregated by region, gender, age, and mode of access. Counts of users who access Facebook through a Wi-Fi network are used as a covariate in the estimation of internet adoption, while counts of users who access Facebook through a cellular network are used as a covariate for mobile phone ownership.

3 Methods

Estimating subnational age-specific rates in this context is challenging for two reasons. First, in some regions, population counts are small and lead to noisy data, particularly when disaggregated into age groups. Second, previous work by Breen et al. [3] has suggested the presence of non-linear relationships between digital adoption and development indicators.

To address the first challenge, a common strategy in small area and global health estimation is to employ hierarchical models and other smoothing structures to share information between populations and age groups, but the ability to incorporate non-linear and higher-order relationships is limited in a traditional Bayesian modelling framework. On the other hand, the naïve application of machine learning methods that may be more adept at capturing complex relationships may struggle with properly accounting for uncertainty arising from small counts and producing estimates that have a sensible age structure.

We therefore propose a two-stage procedure to estimate digital adoption rates that are smoothed over age. In short, the broad idea is to first perform a dimension reduction to the set of age-specific rates before applying a machine learning model to relate the reduced dataset to the covariates.

Step 1: Dimension reduction to spline coefficients

In the first (dimension reduction) step, we fit a generalized additive model with hierarchically-modelled trends over age. For each admin-1 region, the relationship of the internet or mobile adoption over age is modelled as the sum of a global trend, a country-level trend, a country-gender specific trend, and a subnational region-gender specific trend. Each of these are parameterized using a spline basis, for instance, using cubic regression splines [22]. We are specifically interested in the coefficients of the splines as a way to summarize the curve over age.

Explicitly, let y indicate individual responses (e.g. own a mobile phone or not) and be indexed by i, and let $c[i]$, $r[i]$, $g[i]$, and $x[i]$ index the country, subnational region, gender, and age of the individual respectively. Then we model

$$\begin{aligned} y_i &\sim \text{Bernoulli}(p_{c[i],r[i],g[i],x[i]}) \\ \text{logit}(p_{c,r,g,x}) &= \mu + s(x) + t_c(x) + u_{c,g}(x) + v_{c,r,g}(x), \end{aligned} \tag{1}$$

where s, t, u, and v are smooth functions represented using a spline basis. For example the subnational region-gender-specific trend v can be written

$$v_{c,r,g}(x) = \sum_{j=1}^{J_v} b_j(x)\beta^{(v)}_{c,r,g,j} \tag{2}$$

where $\{b_j\}_{j=1}^{J_v}$ are the J_v basis functions and each corresponding $\beta^{(v)}_{c,r,g,j}$ is a coefficient to be estimated. Each subnational region-gender-curve (net of the national trends) can then be summarized from 35 single-year age groups to only J_v parameters. In our preliminary testing (see Section 4), we find that $J_v = 4$ is enough to capture a satisfactory amount of variation in our data.

We note that the model is expressed with this hierarchical structure to the age curves with independent control over each level of the hierarchy. At this step of the estimation, the hierarchical structure is useful from the modelling perspective in order to share information between small populations. However, in practice we may choose identical spline bases so that the coefficients can be collapsed and that each subnational region-gender group is described by a single vector of spline coefficients.

Step 2: Machine learning on covariates

Following the reduction of the data to spline coefficients, we use a machine learning approach to learn the relationship of the covariates to the spline coefficients. In recent related work by Breen et al. [3] an ensemble machine learning ("Superlearner" [14]) is used, whereby a number of candidate algorithms are trained, assessed, and weighted relative to performance in order to produce a final prediction.

In this work, we plan to use a similar approach, with the important distinction that the estimand for each subnational region – the vector of spline coefficients $[\beta^{(v)}_{c,r,g,j}]_{j=1}^{J^v}$ – is multidimensional, which rules out certain candidate methods typically included in the ensemble learner. We will therefore adapt the approach with candidate methods capable of learning multivariable outcomes.

After a machine learning model has been trained, predictions for subnational regions without data can be produced, and finally age-curves of digital adoption can be constructed from the predicted spline coefficients.

Validation and Interpretation

An important feature of this proposed method is the ability to tune the dimension reduction process, which introduces a tradeoff between the two steps. By increasing the number of basis splines, we may fit the ground-truth data more closely, at the cost of increasing the dimension of the machine learning problem. Validation of the estimates will therefore be carried out for both stages.

At the first step of estimation, we are primarily concerned with determining a minimum number of basis splines to satisfactorily describe age patterns in the DHS data. Testing a range of values, we will assess the variation unexplained by the resulting age curves in order to quantify the tradeoff.

At the second step of estimation, we are interested in the machine learning model's prediction accuracy. Here we will perform cross-validation exercises to assess whether the

machine learning model is capable of recovering spline coefficient values that are close to the model fitted values.

Combining information from these two exercises, we will be able to estimate the total amount of error as the sum of the representation and observation error from the first step and prediction error from the second step.

4 Preliminary results

4.1 Age patterns of digital indicators

When disaggregated into granular age groups, subnational mobile phone ownership rates and internet use rates can be unstable over age due to small counts. Figures **??** and **??** shows rates calculated from DHS data for subnational regions in three countries. In each case, the curve representing the fit of the spline model is shown.

We note some common age-patterns in the digital adoption rates. First, rates of mobile phone ownership are typically low in adolescence and increase until approximately age 25. Rates plateau across mid- and older-adult ages, or show a modest decrease in older ages. Meanwhile, the proportion of individuals who report internet use is low among the youngest age groups, increasing with age before peaking at approximately age 20-30, then typically declining among older ages.

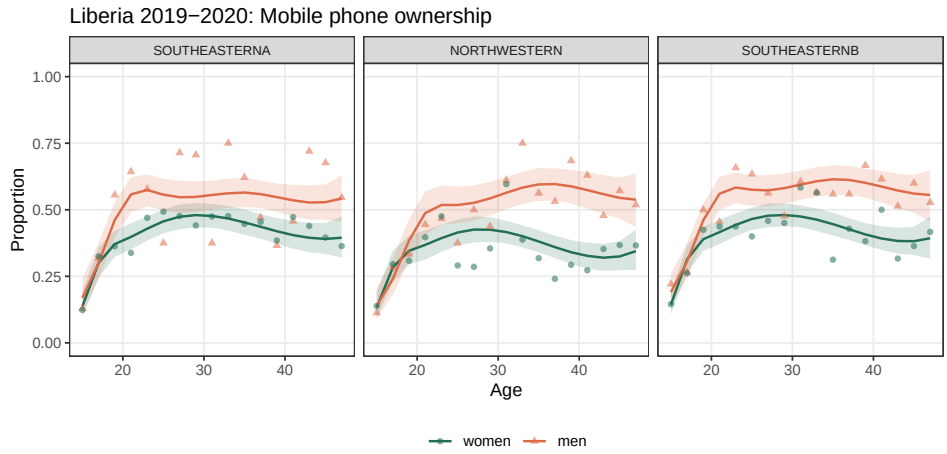
Beyond these broad patterns, there is considerable heterogeneity across geography. Between subnational regions within a country, rates for both mobile phone ownership and internet use vary widely. For example, in Ethiopia, nearly all adults in capital region of Addis Ababa own a mobile phone, compared to fewer than half in some rural regions.

Recent available data has allowed for temporal comparisons between rounds of the DHS. Analyzing patterns across age provides insight into adoption patterns and the diffusion of digital technologies across the population. Figure 3 shows estimates of internet use in subnational regions of Nigeria in 2018 and 2024. In the north-east and north-west regions where rates were the lowest among women, progress was fastest among young adults, but slow in the 15 year-old age group. Meanwhile, in regions that began with higher rates in 2018, progress is more even throughout the age curve, although increases are still the smallest at the 15 year-old age curve.

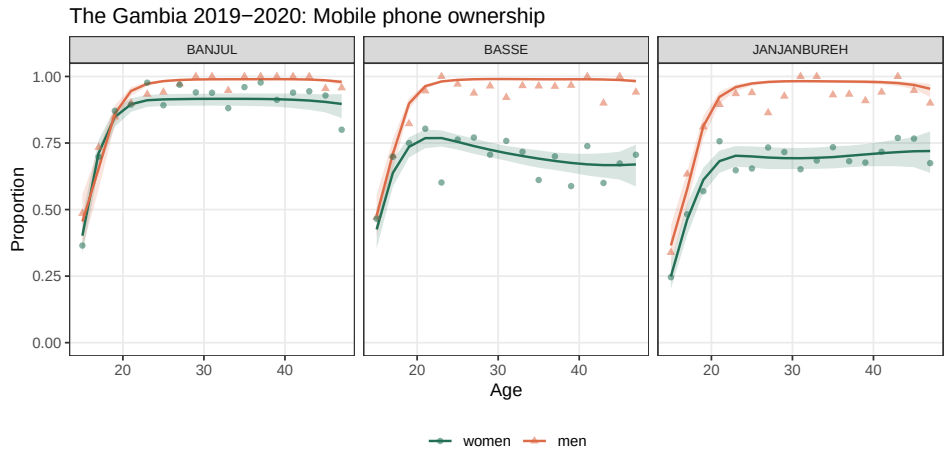
4.2 Modelling and prediction

Figure 4 compares the prediction performance of machine learning methods on a leave-country-out cross-validation task. LASSO regression outperforms more complex methods for this task, but prediction performance varies across age. The 17-20 age groups are the most difficult to predict. The highest mean average error in left-out prediction performance is about 21 percentage points in the 18-19 age groups, and about 15 percentage points in the 47+ age group.

Among the available predictor variables, education index and Facebook marketing API audience metrics are the most consistently selected predictors by the LASSO algorithm. Figure **??** shows empirical patterns between the fitted spline coefficients and these predictor variables.

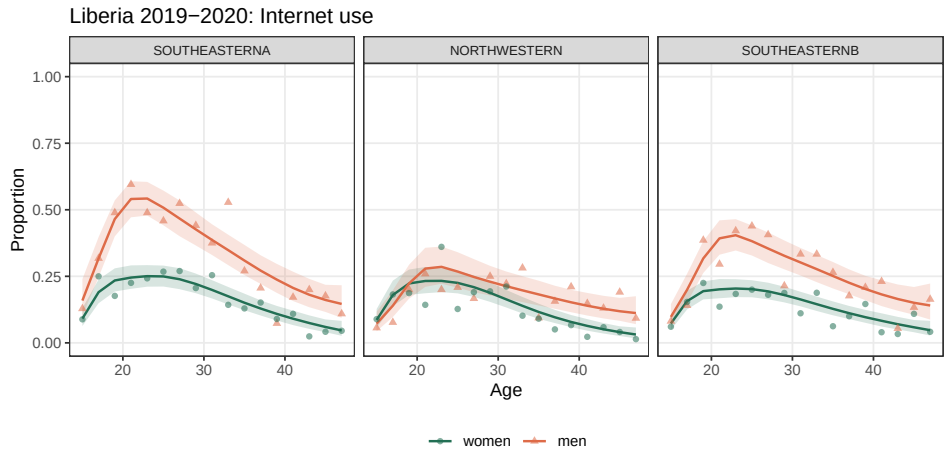


(a) Estimates for Liberia.

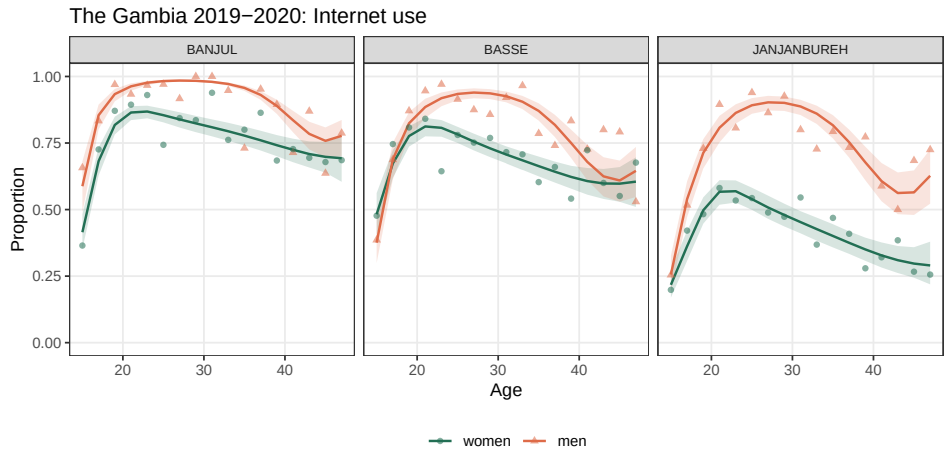


(b) Estimates for the Gambia.

Figure 1: Mobile phone ownership estimates for select subnational regions in Liberia and the Gambia. Points represent observed survey proportions, lines represent posterior medians and shaded regions represent 90% credible intervals.



(a) Estimates for Liberia.



(b) Estimates for the Gambia.

Figure 2: Internet use estimates for select subnational regions in Liberia and the Gambia. Points represent observed survey proportions, lines represent posterior medians and shaded regions represent 90% credible intervals.

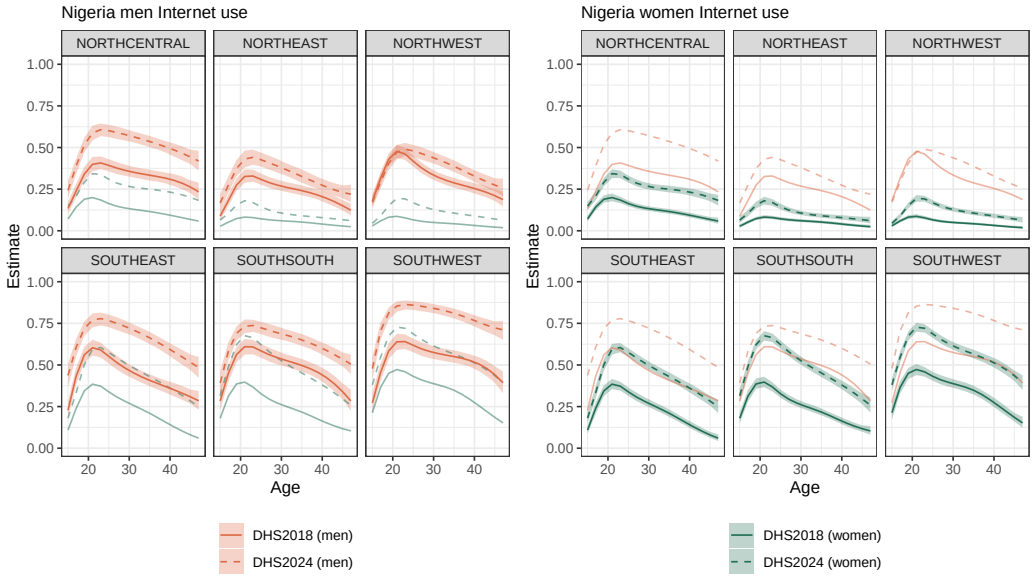


Figure 3: Internet use estimates for Nigeria from the 2018 and 2024 DHS. Lines represent posterior medians and shaded regions represent 90% credible intervals. Median curves representing the womens' estimates are shown as faded lines for reference in the mens' panels and vice versa.

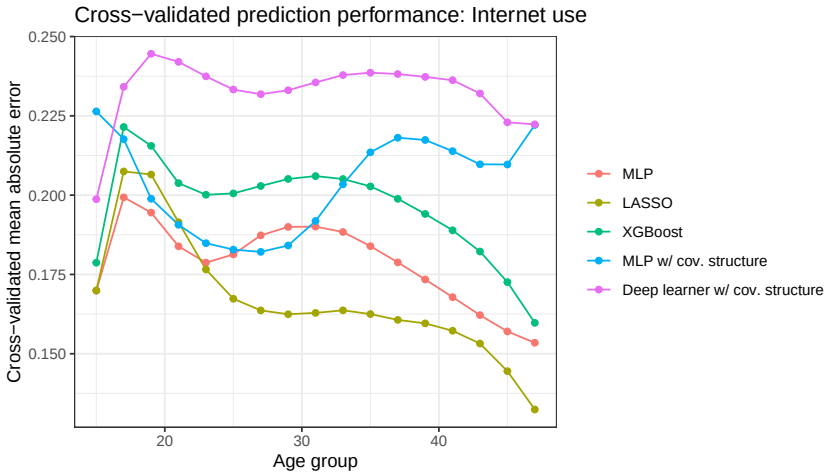


Figure 4: Comparison of cross-validation performance of machine learning models. Points represent the mean absolute error from the observed survey proportion in subnational regions of left-out countries.

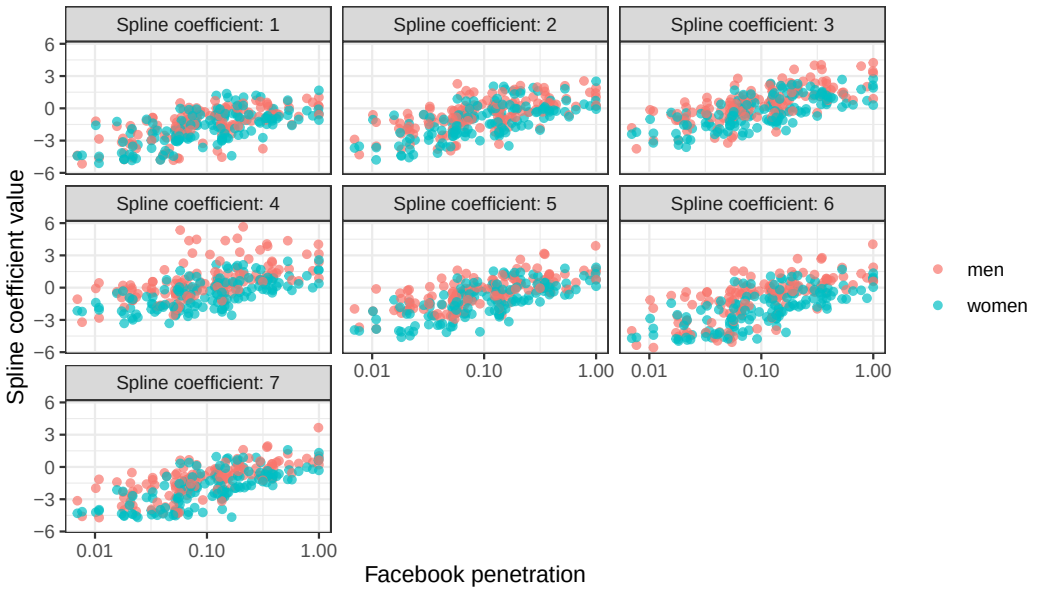


Figure 5: Empirical trends between population-standardized Facebook audience metrics and estimated spline coefficients. Spline coefficients are ordered in order of increasing age group such that the first spline coefficient primarily controls the youngest age groups.

5 Discussion and Future Directions

In this paper we explore subnational age patterns in internet use and mobile phone ownership in low- and middle-income countries. We employ a two-stage approach to estimating age-specific rates, in which we first smooth the age-specific proportions observed in survey data, then model the underlying parameters that govern the smooth curves. Our results provide new evidence on the patterns of early adoption of digital technologies, highlighting gaps in digital access and connectivity across age, gender, and geography.

Young adult populations aged 20-30 are usually the first adopters of digital technologies, and exhibit the highest rates of internet use. Digital connectivity is lower among older adults, and contrary to evidence from high-income countries, connectivity among younger groups in low- and middle-income countries can also be low. Temporal comparisons suggest that further adoption of digital technologies happens evenly across age, suggesting broad expansion beyond what would be expected solely by cohort ageing. However, it is concerning that gender gaps in both internet use and mobile phone ownership are often already visible at young ages. These estimates highlight important disparities in access to digital technologies, and with high granularity identify which groups are included in and excluded from the expansion of digital programs and services.

Accurate estimation of age-specific rates for out-of-sample countries remains challenging, but the relative strength of digital trace data as a predictor of population-level digital access suggests possible avenues for future modelling. Work on collecting more detailed signals is ongoing. While the model presently only uses a single gender-specific audience count for each subnational region, work is ongoing to collect and incorporate more detailed signals, including age-specific Facebook user counts as well as Instagram user counts, which may

help to account for varying popularity of platforms across age and geography.

If complex machine learning classifiers are unable to benefit from most of the predictor variables, a simpler modelling approach may be preferred. A hierarchical Bayesian approach for example, would allow for the two-stage estimation to occur simultaneously, yielding a more principled treatment of uncertainty. Early experimentation with a simple Bayesian model specification with only the non-age-disaggregated Facebook audience counts shows comparable performance to other methods.

References

[1] Anna Bolgrien et al. *IPUMS MICS Data Harmonization Code. Version 1.2 [Stata syntax]*. IPUMS: Minneapolis, MN. 2024. URL: https://doi.org/10.18128/D082.V1.2.

[2] Elizabeth Heger Boyle, Miriam King, and Matthew Sobek. *IPUMS-Demographic and Health Surveys: Version 11 [dataset]*. IPUMS and ICF. 2024. URL: https://doi.org/10.18128/D080.V11.

[3] Casey F Breen et al. "Mapping subnational gender gaps in internet and mobile adoption using social media data". In: *Proceedings of the National Academy of Sciences* 122.42 (2025), e2416624122.

[4] Paul DiMaggio and Eszter Hargittai. *From the 'Digital Divide' to 'Digital Inequality': Studying Internet Use as Penetration Increases*. en. 2001. DOI: 10.31235/osf.io/rhqmu. URL: https://osf.io/rhqmu (visited on 06/13/2025).

[5] Antoine Dubois et al. "Studying migrant assimilation through Facebook interests". In: *International conference on social informatics*. Springer. 2018, pp. 51–60.

[6] Christopher D Elvidge et al. "Annual time series of global VIIRS nighttime lights derived from monthly averages: 2012 to 2019". In: *Remote Sensing* 13.5 (2021), p. 922.

[7] Christopher D Elvidge et al. "Why VIIRS data are superior to DMSP for mapping nighttime lights". In: *Proceedings of the Asia-Pacific Advanced Network* 35.0 (2013), p. 62.

[8] GSMA. *The Mobile Gender Gap Report 2024*. Tech. rep. GSMA, 2024.

[9] Jonas Hjort and Lin Tian. "The economic impact of internet connectivity in developing countries". In: *Annual Review of Economics* 17 (2021).

[10] International Telecommunication Union. *Measuring Digital Development: Facts and Figures 2024*. Tech. rep. International Telecommunication Union, 2024.

[11] Ridhi Kashyap et al. "Digital and computational demography". In: *Research handbook on digital sociology*. Edward Elgar Publishing, 2023, pp. 47–85.

[12] Ridhi Kashyap et al. "Monitoring global digital gender inequality using the online populations of Facebook and Google". In: *Demographic Research* 43 (2020), pp. 779–816.

[13] Lindsay Katz, Michael Chong, and Monica Alexander. "Measuring short-term mobility patterns in North America using Facebook advertising data, with an application to adjusting COVID-19 mortality rates". In: *Demographic Research* 50 (2024), pp. 291–324.

[14] Mark J van der Laan, Eric C Polley, and Alan E Hubbard. "Super Learner". In: *Statistical Applications in Genetics and Molecular Biology* 6.1 (2007), pp. 1–23.

[15] Douglas R Leasure et al. "Nowcasting daily population displacement in Ukraine through social media advertising data". In: *Population and Development Review* 49.2 (2023), pp. 231–254.

[16] Valentina Rotondi et al. "Leveraging mobile phones to attain sustainable development". In: *Proceedings of the National Academy of Sciences* 117.24 (June 2020). Publisher: Proceedings of the National Academy of Sciences, pp. 13413–13420. DOI: 10.1073/pnas.1909326117. URL: https://www.pnas.org/doi/abs/10.1073/pnas.1909326117 (visited on 07/18/2023).

[17] Jeroen Smits and Iñaki Permanyer. "The Subnational Human Development Database". In: *Scientific Data* 6.1 (2019), pp. 1–15.

[18] Tavneet Suri and William Jack. "The long-run poverty and gender impacts of mobile money". In: *Science* 354.6317 (2016), pp. 1288–1292.

[19] UNICEF. *Bridging the Gender Digital Divide*. Tech. rep. UNICEF, 2023.

[20] United Nations. *Transforming our world: the 2030 Agenda for Sustainable Development*. https://sdgs.un.org/2030agenda. United Nations General Assembly, A/RES/70/1. New York, 2015.

[21] Jan van Dijk. *The Digital Divide*. Newark: Polity Press, 2020. ISBN: 978-1-5095-3446-3.

[22] Simon N Wood. *Generalized additive models: an introduction with R*. Chapman and Hall/CRC, 2017.

[23] World Bank. *Digital Progress and Trends Report 2023*. Tech. rep. World Bank, 2023.