Mapping subnational gender gaps in internet and mobile adoption using social media data *†

Casey F. Breen‡ Masoomali Fatehkia§ Jiani Yan‡

Douglas R. Leasure‡ Ingmar Weber¶ Ridhi Kashyap‡

Draft Version: October 25, 2023

Abstract

The digital revolution has ushered in tremendous societal and economic benefits. Yet access to digital technologies such as mobile phones and internet remains highly unequal, especially by gender in the context of low- and middle-income countries. Reliable, quantitative estimates of digital gender inequalities are essential for monitoring gaps and implementing targeted interventions within the global sustainable development goals. While national-level estimates are available for many countries, subnational estimates are critical since internet and mobile phone adoption vary substantially by geography. Here we develop estimates of internet and mobile adoption by gender and digital gender gaps at the subnational level for 874 regions in 55 countries across the African continent, a context where digital penetration is low and national-level gender gaps disfavouring women are large. We construct these estimates by applying machine-learning algorithms to Facebook audience counts derived from the platform's marketing application programming interface (API), geospatial and population data. We train and assess the performance of these algorithms using "ground truth" data from nationally-representative household survey data from 19 countries in Africa. Our results reveal striking disparities in access to mobile and internet technologies between and within countries, with implications for policy formulation and infrastructure investment.

---

*Preliminary. Please do not cite or redistribute. Address correspondence: casey.breen@sociology.ox.ac.uk and ridhi.kashyap@sociology.ox.ac.uk.

†This work received funding from the Bill and Melinda Gates Foundation (INV-045370) and Leverhulme Trust (Grant RC-2018-003) for the Leverhulme Centre for Demographic Science.

‡University of Oxford

§Qatar Computing Research Institute

¶Saarland University

1 Introduction

The digital revolution has yielded major societal and economic benefits. Internet and mobile technologies enhance information access (DiMaggio and Hargittai, 2001; Kashyap et al., 2023), bolster social connectivity (Masi et al., 2011; Findlay, 2003), increase economic prosperity (Aker and Mbiti, 2010; Hjort and Poulsen, 2019), and expand access to key services like mobile banking (Suri and Jack, 2016). Yet the benefits of this digital revolution have accrued unevenly. An estimated 2.7 billion people have never accessed the internet (Union, 2022), and of these the majority are women and girls. In terms of mobile access, over 130 million more men than women own mobile phones (GSMA, 2023). This *digital divide* by gender is an increasingly salient dimension of population inequality in the modern world.

The gender digital divide is especially pronounced in low- and middle-income countries (LMICs). Reliable quantitative estimates of digital gender inequalities are key for tracking progress on and implementing targeted policies and intervention in the context of the global sustainable development goals (SDGs). Reducing inequalities in access to digital technologies by gender is a target within SDG 5 on gender equality, while digital literacy is a core part of SDG 4 on the right to education. While the availability of national-level estimates of digital gender gaps has improved (Fatehkia, Kashyap and Weber, 2018; Kashyap et al., 2020), subnational estimates remain sparse. Subnational estimates however are critical since internet and mobile phone adoption vary within countries, and geographically granular estimates are relevant for monitoring progress and developing targeted interventions. As development programmes increasingly become digital (e.g. mHealth), understanding which social groups and regions stand to benefit from them is central to promoting sustainable development. Past subnational estimates of digital adoption are typically based on probabilistic household surveys or censuses (Cohen and Adams, 2011; World Bank Group, 2016), but often lack gender disaggregation. Moreover, as subnational estimation requires larger sample sizes, these conventional methods are often slow and expensive to implement (Rojas, 2015). To date, there are no subnational estimates of digital gender gaps in the majority of LMICs in the world.

To help address this challenge, we construct estimates of digital adoption by gender and digital gender gaps in Africa by applying machine-learning algorithms to social media data together with population and development indicators. The social media data that we use are gender-disaggregated, subnational Facebook audience counts derived from the Facebook marketing API. We train and assess the performance of these algorithms using "ground truth" data from nationally-representative Demographic and Health Surveys (DHS) from 19 countries in Africa. Our analyses focuses on Africa as this is the context where national-level digital gender gaps disfavouring women are large (Fatehkia, Kashyap and Weber, 2018; Kashyap et al., 2020), and subnational data on digital inequalities by gender across the whole continent are limited. The availability of recent DHS data across the continent provides us good coverage of ground truth data to train and test our models to assess the validity of our approach, and expand geographical coverage of subnational digital gender gaps to 55 countries and four territories across the African continent. Our results reveal striking geographical disparities in access to internet technology between and within countries, with implications for policy formulation and infrastructure investment.

2 Background

2.1 Benefits of digital technology

Digital technologies affect health and overall well-being through many channels (Hjort and Poulsen, 2019; Suri and Jack, 2016; World Bank Group, 2016; Kashyap et al., 2023). The internet and mobile phones are powerful mediums for boosting social connectivity, social learning, and access to economic services such as mobile banking (Unwin, 2009; DiMaggio and Hargittai, 2001; Suri and Jack, 2016). Increasing internet adoption also has other "digital dividends"— it creates new jobs (Hjort and Poulsen, 2019), improves educational outcomes (Kho, Lakdawala and Nakasone, 2018), increases social capital (Kharisma, 2022), and impacts demographic processes such as fertility (Billari, Giuntella and Stella, 2019) and migration (Pesando et al., 2021). Digital technologies also have the potential to empower women (Dettling, 2017; Lund et al., 2014; Lagan, Sinclair and Kernohan, 2010; Rotondi et al., 2020). Mobile phone usage is associated with lower gender inequality, higher contraceptive uptake, and lower child and maternal mortality (Rotondi et al., 2020). Notably, these benefits are often greatest in the most unequal, disadvantaged areas.

2.2 Gender-based digital disparities

Large inequality persists in access to and usage of digital technologies. Factors like education, age, class, and race, as well as their intersections, play a significant role in determining who gets access to these technologies and how they use them (Muschert, 2013). Although the accessibility gap has declined or disappeared in most high-income countries, gaps persist in the majority of low- and middle-income countries (Kashyap, 2021).

This digital inequality is highly gendered. More than 250 million more men than women have accessed the internet (Union, 2017), and 130 million more men than women own mobile phones (GSMA, 2023). These digital gender gaps reflect broader structural inequality in institutional sectors such as the education system and labor markets (Hilbert, 2011; Robinson et al., 2015). In addition to institutional sexism, culture is also key in determining women's access to digital technologies. In many strongly patriarchal countries, access to such technologies is mediated by men who often limit women's access (Abu-Shanab and Al-Jamal, 2015).

2.3 Big data innovations for development indicators

The data ecosystem for measuring population and development indicators has increasingly expanded with the growing use of digital technologies across the world, which have generated new streams of digital trace and geospatial data (Kashyap, 2021). Researchers have taken advantage of this new data ecosystem in different ways to measure population and development processes, such as to predict wealth for microregions from mobile metadata (Blumenstock, Cadamuro and On, 2015; Chi et al., 2022), assess air quality after wildfires using sattelite imagery (Burke et al., 2023), and predict well-being from tweets (Resce and Maynard, 2018). Despite weaknesses of these new data resources, such as issues of bias and non-representativeness, and lack of transparency about the algorithms that often generate them (Lazer et al., 2014), their high-frequency and real-time characteristics, as well as often better geographical resolution, makes them a promising data source to predict the present ("nowcasting") (Salganik, 2018).

Facebook's advertisement audience size estimates — freely available through Facebook's marketing application interface (API) — provide researchers with counts of Facebook users by geographic area and sociodemographic characteristics, such as gender and age. Researchers have used these audience count data to study migration (Zagheni, Weber and Gummadi, 2017; Rampazzo et al., 2021), population displacement (Leasure et al., 2023), wealth inequalities (Fatehkia et al., 2020), population health (Araujo et al., 2017), and most relevantly, gender inequality in access to the internet and mobile phones at the country-level (Kashyap et al., 2020; Fatehkia, Kashyap and Weber, 2018). These Facebook audience count data can serve as a type of "digital census" of the platform allowing researchers to look both at overall counts of users and differential rates of use across sociodemographic groups.

While the above-mentioned research has highlighted the value of data from the Facebook marketing API for monitoring national-level digital gender inequality, there are currently no estimates of digital gender gaps at the subnational level. Whether methods using the Facebook marketing API developed for the national-level can be extended for generating subnational estimates for this indicator, but also potentially also for other population and development indicators, remains unexplored. Subnational estimates are crucial for several reasons. First, there is often large amounts of geographic hetereogeneity: countries may exhibit significant regional disparities in infrastructure, education, overall development, as well as social norms (Michalopoulos and Papaioannou, 2014), which in turn can create large variation in digital adoption by gender. This variation is obscured in a national-level estimate. Second, for effective targeted policy, infrastructure enhancement, and intervention strategies, it is essential to identify subnational areas with low digital connectivity rates, and if these rates vary differentially by gender.

3 Data

For this study, we employ three sources of data. For our predictive models, we use both "online" and "offline" features. Our "online features" are variables generated from data on

Facebook Monthly Active Users (MAUs) (e.g., fraction of male users over age 13, fraction of female users over age 13) from the marketing API. Our "offline" features are a set of variables on population density and indices on human development, education, and income. To train and calibrate our models, we use ground-truth data on internet use and mobile phone ownership from 19 Demographic and Health Surveys in Africa.

3.1 Ground truth data on internet and mobile access

Our ground-truth data comes from 19 Demographic and Health Survey (DHS) conducted between 2015–2019, i.e. from phase seven onward in the DHS programme when the digital measures were first included in the DHS. The DHS surveys are representative at the first administrative subnational level and collect individual-level data about both internet usage and mobile phone ownership for both men and women. We combine these DHS estimates with population estimates from WorldPop (WorldPop, 2023) to obtain estimates of the percent of men and women aged 15–49 who (1) own a mobile phone; (2) have accessed the internet in the past 12 months; (3) who have ever accessed the internet. We also calculate the gender gap, defined as:

$$\text{Gender Gap} = \frac{\mathrm{I}_f/\mathrm{I}_m}{\mathrm{Pop}_f/\mathrm{Pop}_m} \tag{1}$$

where for a specific indicator I (e.g., mobile phone ownership or internet use in the past 12 months), I_f is the number of female users aged 15–49, I_m is the number of male users aged 15–49, Pop_f is the total population of women aged 15–49, and Pop_m is the total male population aged 15–49.

3.2 Facebook Audience Counts

To obtain counts of Facebook monthly active users, we query the Facebook Marketing API. The Facebook Marketing API provides estimates of the number of daily or monthly active users disaggregated by characteristics such as gender, age, and access device type in a given geographic boundary (e.g., country or state). We used an adapted version of the pysocialwatcher package (Araujo et al., 2017) to collect information on digital connectivity at the GADM-1 level.[1] GADM1 regions largely correspond to the first administrative subnational region of a country. We define all online features as gender-specific fractions, or as gender gaps (female-to-male ratios) (see Table 1). For example, the 'All Devices Gender Gap' variable refers to the female-to-male ratio of Facebook users in a given GADM-1 unit across all devices. The 13+ FB penetration variable corresponds to the proportion of female Facebook users relative to the female population in the same GADM-1 unit.

Variable Name	Type	Source	Country (N)	Subnational (N)
Perc Ever Used Internet 15-49 FM Ratio	Offline	DHS	19	309
Perc Ever Used Internet 15-49 Men	Offline	DHS	19	309
Perc Ever Used Internet 15-49 Wom	Offline	DHS	20	319
Perc Owns Mobile Phone 15-49 FM Ratio	Offline	DHS	19	309
Perc Owns Mobile Phone 15-49 Men	Offline	DHS	19	309
Perc Owns Mobile Phone 15-49 Wom	Offline	DHS	20	319
Perc Used Internet Past Year 15-49 FM Ratio	Offline	DHS	19	308
Perc Used Internet Past Year 15-49 Men	Offline	DHS	19	309
Perc Used Internet Past Year 15-49 Wom	Offline	DHS	20	319
All Devices Age 13+ GG	Online	FB marketing API	57	813
FB Penetration 13+ Female	Online	FB marketing API	57	844
FB Penetration 13+ Male	Online	FB marketing API	57	844
iOS 13+ Female Fraction	Online	FB marketing API	57	781
iOS 13+ Male Fraction	Online	FB marketing API	57	813
WiFi Age 13+ Female Fraction	Online	FB marketing API	57	781
WiFi Age 13+ Male Fraction	Online	FB marketing API	57	813
X4G Network Age 13+ Female Fraction	Online	FB marketing API	57	781
X4G Network Age 13+ Male Fraction	Online	FB marketing API	57	813
FB Rural WiFi Mean (Pop Weighted)	Offline	FB marketing API	50	764
Educational Index Females	Offline	Subnational Dev. Database	50	782
Educational Index Males	Offline	Subnational Dev. Database	50	782
Income Index Females	Offline	Subnational Dev. Database	50	782
Income Index Males	Offline	Subnational Dev. Database	50	782
Subnational GDI	Offline	Subnational Dev. Database	50	782
Subnational HDI Females	Offline	Subnational Dev. Database	50	782
Subnational HDI Males	Offline	Subnational Dev. Database	50	782
WPop 2020 Age 15-49 Female Fraction	Offline	WorldPop	58	869
WPop 2020 Age 15-49 Male Fraction	Offline	WorldPop	58	869
WPop 2020 Pop Density	Offline	WorldPop	59	874
Nightlights DHS Year Mean Pop Weighted	Offline	Earth Observation Group	58	869

Table 1: List of features used in the analysis with their predictor type.

[1]GADM, the Database of Global Administrative Areas, is a publicly-available, high-resolution database of country administrative areas. When boundaries are available in the FB marketing API that match the GADM-1 boundaries, we use use the default FB boundaries. In situations where we do not use any boundaries available in Facebook that match the GADM-1 boundaries, we instead create custom polygons to match the GADM-1 boundaries. We collected estimates on gender, age, device type, and other indicators.

4 Methods

We model three different outcomes (mobile phone ownership, used internet in the past 12 months, and used internet ever), three different indicators (percent of men, percent of women, and the Female-Male gender gap), and three different types of predictive models (online predictors, offline predictors, and online and offline predictors). In total, we fit 27 separate models.

4.1 Machine learning approach

We use a machine learning approach for prediction. We predict each of these separate indicators using a combination of online and offline features. Flexible machine learning algorithms are appealing in this setting because of their ability to detect interactions, model higher order effects, and better handle multiple, highly-correlated predictors (Rose, 2013; Puterman et al., 2020). Machine learning approaches have been applied for similar predictions setting, such for small-area estimation of wealth (Blumenstock, Cadamuro and On, 2015; Chi et al., 2022).

For most prediction tasks, it is impossible to know a priori which algorithm will have the best performance. To overcome this, we use Superlearning—also known as weighted ensembling or stacking—a method for combining many machine learning algorithms into a single algorithm (Van der Laan, Polley and Hubbard, 2007). The motivation behind Superlearning is that a weighted combination of different algorithms may outperform any single algorithm by smoothing out limitations of any specific algorithm. The Superlearner algorithm selects the best weighted combination of algorithms using a k-fold cross-validation procedure to minimize cross-validated risk (Van der Laan, Polley and Hubbard, 2007). For our Superlearner, we use a range of popular machine learning algorithms: random forests, generalized linear regression, gradient boosting machines, lasso regression, elastic net regression, polynomial splines regression, ridge regression, and extreme gradient boosting machines.

Algorithm	Description
glm	Generalized Linear Model
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
xgboost	Extreme Gradient Boosting
SuperLearner	Ensemble method combining multiple learning algorithms

Table 2: Machine learning algorithms

4.2 Cross-validation

To evaluate the performance of our model, we use 10-fold cross-validation and leave-one-country-out cross-validation (LOCO-CV). For conventional 10-fold cross-validation, we randomly split our sample into ten separate folds. We trained our models on 9 folds and made predictions on single hold-out fold; we repeated this process for each fold. We use the predictions on all held-out folds to estimate several model performance metrics.

For LOCO-CV, we split the sample into 19 separate folds defined by country. Holding out all subnational units in a given country ("hold-out partition"), we fit our models on the rest of our dataset ("training partition"). We then use our models to predict on the held-out subnational units of that country. This process is iterated for each country in the dataset, ensuring that every country's subnational units serve as a hold-out set. We use the predictions on all held-out units to estimate model performance metrics. By holding out data from a single country during training, LOCO-CV tests the model's capability to handle inter-country variability and minimizes overfitting risks specific to individual countries. Contrary to standard 10-fold cross-validation, LOCO-CV addresses concerns of geographical independence, providing a more stringent assessment of the model's geographical robustness. In comparison to 10-fold cross-validation, LOCO-CV predictions show more conservative estimates of predictive fit (see Figure A6).

4.3 Performance Metrics

We use several different to assess model performance metrics. First, we use R^2, the coefficient of determination. Given a set of observed values $\{y_1, y_2, \ldots, y_n\}$ and a set of predicted values $\{\hat{y}_1, \hat{y}_2, \ldots, \hat{y}_n\}$, the R^2 value can be computed as:

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2} \tag{2}$$

Where y_i is the observed value for the i-th observation; $\hat{y}_i$ is the predicted value for the i^{th} observation; and $\bar{y}$ is the mean of the observed values. As an alternative metric for assessing model fit, we use mean average error (MAE):

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i| \tag{3}$$

The R^2 value, or coefficient of determination, quantifies the proportion of variance in the dependent variable explained by the model, ranging between 0 and 1; a higher value suggests a better fit. The Mean Absolute Error (MAE) provides an absolute measure of the average prediction error in the dependent variable's units, with a lower MAE indicating better model accuracy. Using both metrics is advantageous: while R^2 offers a relative measure of fit, MAE yields a direct interpretation of prediction error magnitude, and is more robust to outliers. Together, they offer a more comprehensive assessment of model performance than either metric alone.

5 Results

Figure 1 illustrates our main result: our model-based approach for estimating subnational gender gaps greatly expands our geographic coverage of digital gender gaps. Panels (A), (C), and (E) show our ground-truth indicators of mobile phone ownership from the DHS surveys. Our ground truth data cover approximately one-third of countries in the African continent. In Panels (B), (D), and (F), we present our model-based indicators of mobile phone ownership from our superlearner online-offline model, capturing almost all countries in Africa, and strong predictive performance (see Table A3 for comparison across different algorithms). Qualitatively, our model-based predictions broadly track our observed ground truth. In short, our model-based approach allows for a three-fold increase in geographic coverage and approximates our observed rates of mobile phone ownership reasonably well. Similar patterns also apply to the internet use outcomes (see Figure A7), for which we also obtain similar expansion of geographical coverage for the indicator. Notably, overall levels of internet usage are on average lower than mobile phone ownership.

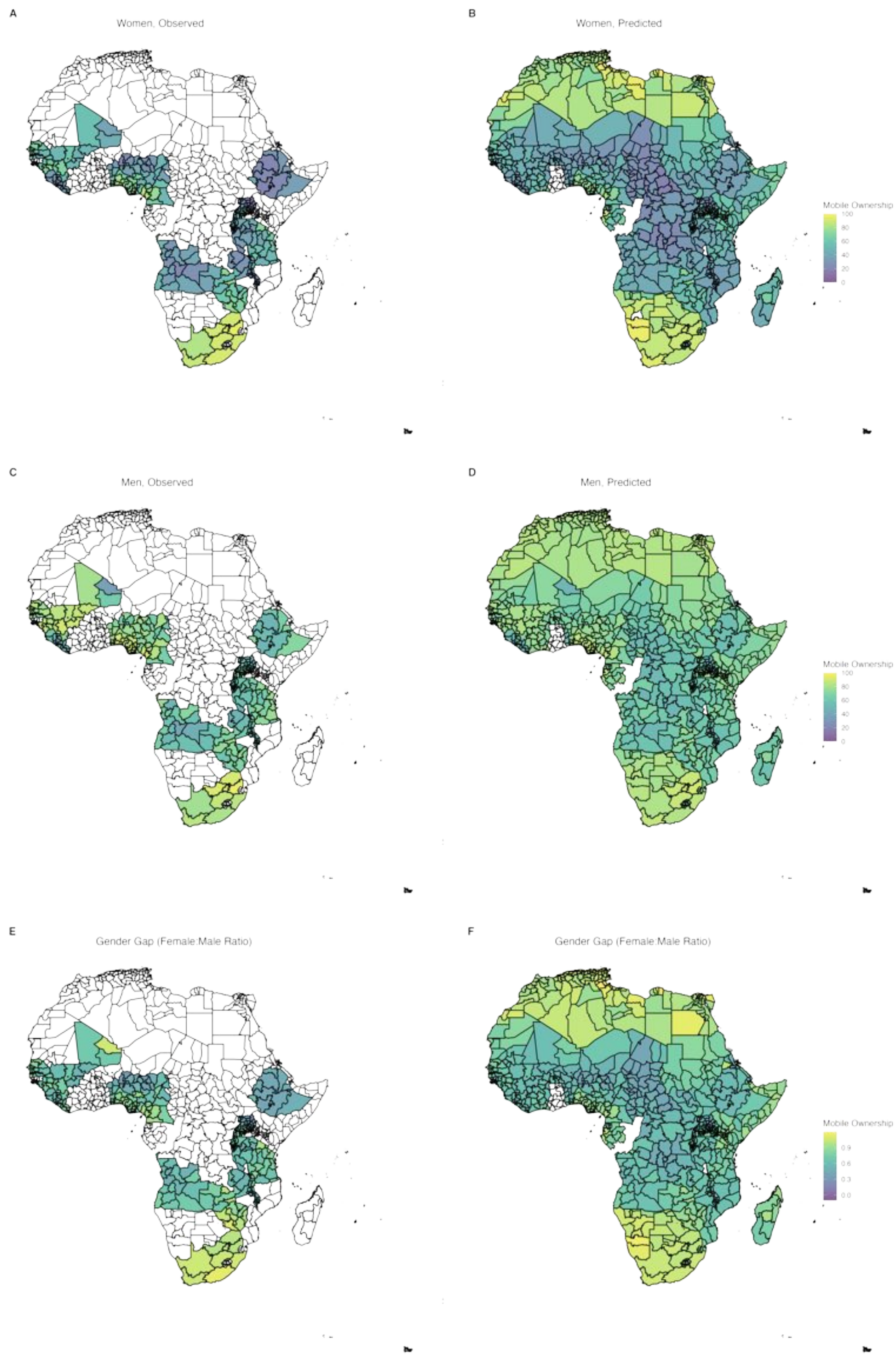

Figure 1: **Panel (A)**, **Panel (C)**, **Panel (E)** show survey-based 'ground truth' estimates of mobile phone ownership indicators for 19 countries. **Panel (B)**, **Panel (D)**, **Panel (F)** show model-based estimates of the mobile phone ownership digital gender gaps for 55 countries and 4 territories.

Next, we compare the performance of models trained on on different features sets (e.g., online features, offline features, online and offline features). Figure 2 shows the R^2 value for our superlearner algorithm using each different set of features measured with leave-one-country-out cross-validation (LOCO-CV). The modeled trained using only "online" predictors from Facebook (blue points) generally had the best performance. Models trained only with the offline features (green points) had the worst overall performance, and models trained using online and offline features (red points) generally had slightly lower performance than models trained exclusively with the online features. Across all models, adding in the online features led to a substantial increase in the predictive accuracy of the model. When examining model performance across LOCO-CV and 10-fold CV, we generally find higher R-squared values with 10-fold CV, as shown in Figure A6. With 10-fold CV, we also find that the online-offline feature set performs the best more consistently than is the case with LOCO-CV. This suggests that LOCO-CV may minimize potential overfitting that a larger feature set offers.

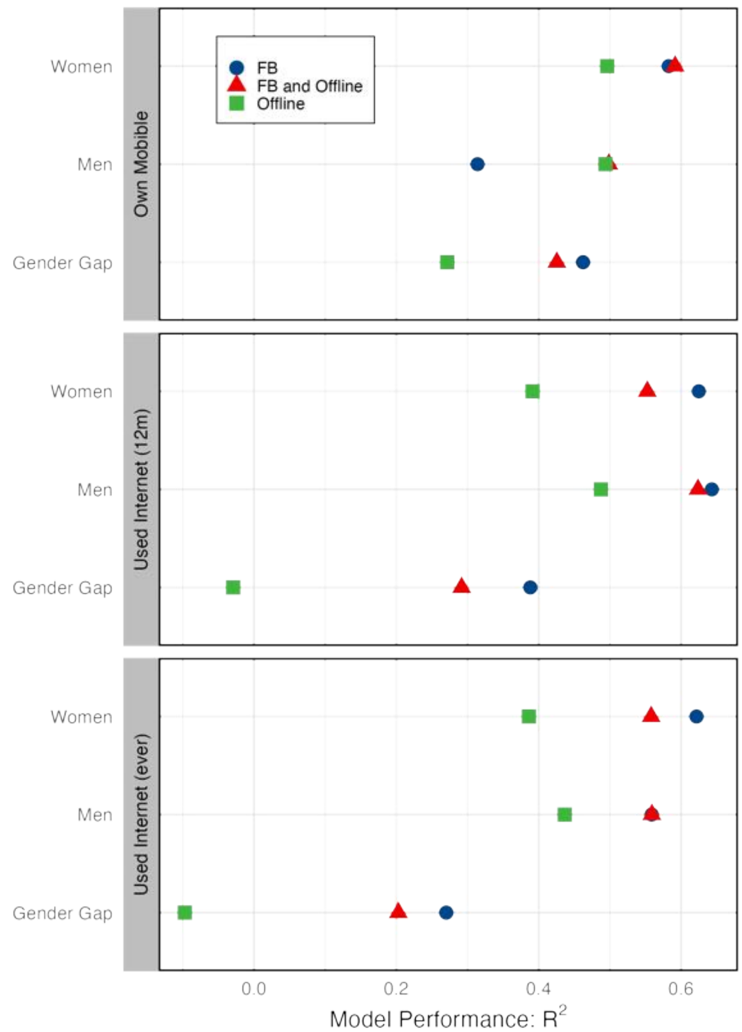

Figure 2: For each indicator, the R^2 from leave-one-country-out cross-validation using online predictors, offline predictors, and online and offline predictors.

To further assess the predictive accuracy of our machine learning models, we compared our 'ground-truth' data from the DHS surveys to our model predictions for each GADM-1 subnational unit from leave-one-country-out cross-validation (LOCO-CV). Figure 3 shows the observed vs. predicted values of the mobile phone ownership indicators for each GADM-1 subnational unit. The correlation between the predicted and observed value is highest for women ($R = 0.74$) and lowest for the gender gap ($R = 0.62$). The gender gap is intuitively a noisier metric to predict, as the underlying "ground truth" data is likely to have more uncertainty, as it is the ratio of two separate estimates, both with sampling uncertainty. We would therefore not expect a perfect correlation between our observed and modeled estimates. In addition, we note that while this plot shows the average correlation pooled across all countries, there is substantial country-level heterogeneity in the accuracy of our predictions (Figure A9), a point we intend to explore in more depth as we extend this work.

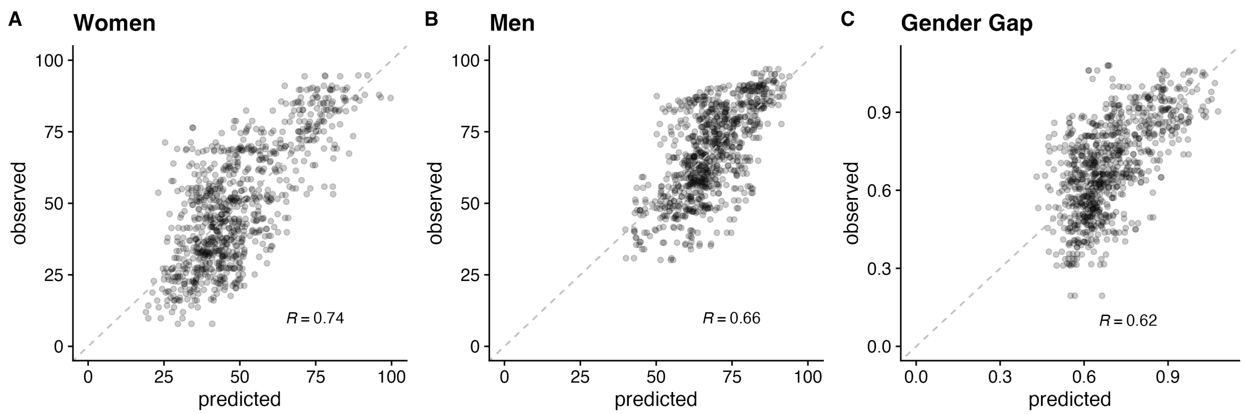

Figure 3: **Panel (A)** shows the predicted vs. observed model mobile phone ownership for women. **Panel (B)** shows mobile phone ownership for men. **Panel (C)** shows the mobile mobile phone ownership gap, defined as the ratio of female mobile phone users to male mobile phone users

Figure 4 shows the performance of our approach for estimating internet use (past 12 months) in Nigeria. Several insights emerge from this figure. First, there is large subnational heterogeneity in the underlying ground-truth data. Nearly 55% of women in the relatively affluent and urban state of Lagos have accessed the internet in the past 12 months, while less than 1% of women have accessed internet in the rural state of Kebbi. This highlights the importance of considering the subnational context. Second, the model-based estimates align closely with the observed predictions; the correlation between the model-based estimates and the observed ground-truth is $R = 0.88$. Finally, the error in the predictions (Panel C) displays some geographic clustering. These same patterns are observable in our predictions of female mobile phone ownership in Nigeria (see Figure A8).

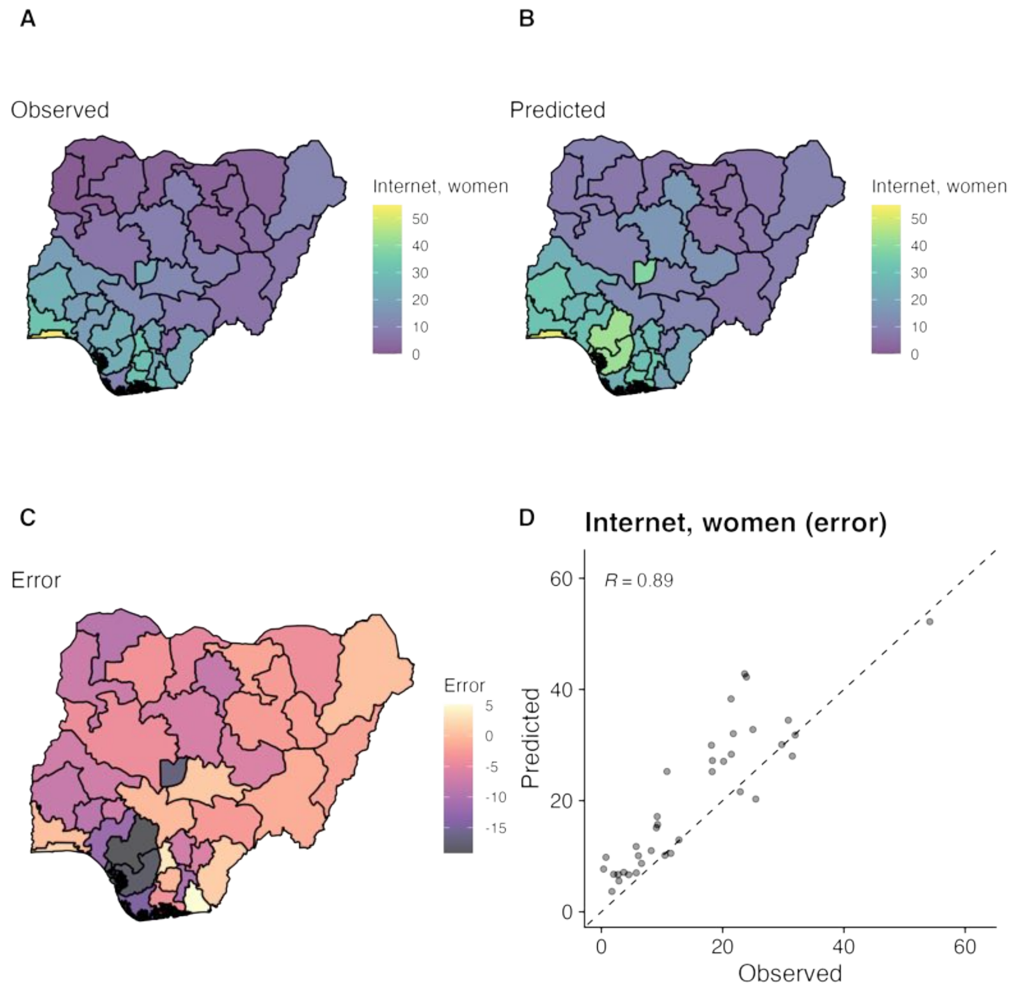

Figure 4: For women in Nigeria, the observed rate of internet use **(Panel A)**, model-based predictions of rate of internet use **(Panel B)**, and the error between our observed and predicted values **(Panel C, Panel D).**

We investigate the relationship between overall levels of mobile phone ownership and the mobile phone gender gap by comparing rates of male mobile phone ownership to mobile gender gaps at the GADM-1 level. Figure 5 shows there is a clear linear relationship between rates of male mobile phone ownership and the mobile phone gender gap: as rates of mobile phone ownership increases for men, the mobile gender gap declines. Yet there is also substantial variation in this broad trend, suggesting that institutional and cultural factors likely mediate the relationship between overall rates of mobile phone ownership and gender gaps.

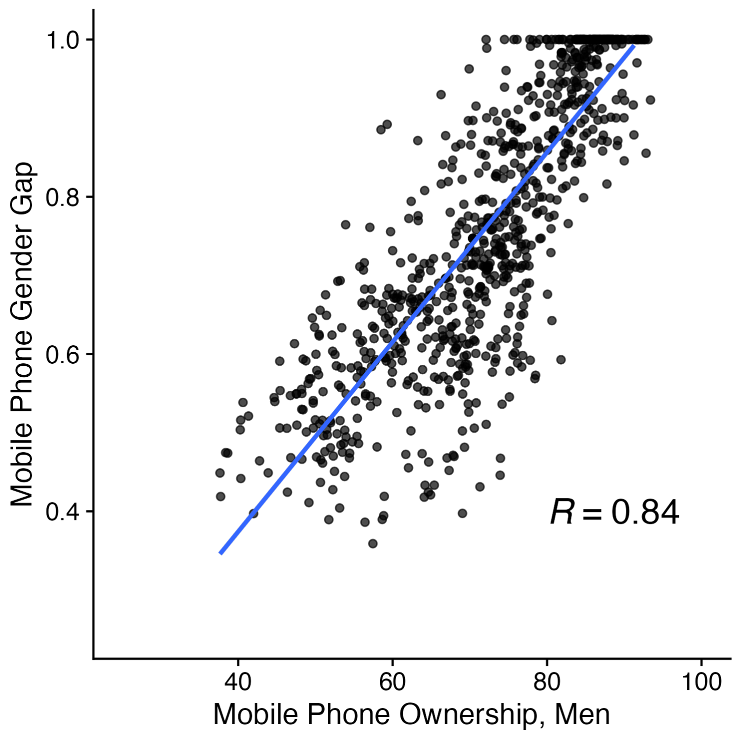

Figure 5: Scatterplot of the level of male mobile phone ownership vs. mobile phone gender gap. The mobile phone gender gap is capped at 1.

6 Discussion

Gender-based disparity in access to digital technology is an increasingly important dimension of population inequality. Yet tracking and measuring this important indicator is often challenging due to data limitations. Here, we demonstrate a new approach to estimating subnational indicators of digital gender gaps using Demographic and Health Surveys paired with aggregate Facebook audience count data derived from the platform's marketing API.

Together, our results demonstrate the promise of using Facebook audience count data combined with population and development indicators for making subnational predictions on digital adoption by gender for the continent of Africa. Our results suggest that there is substantial variation in access to internet and mobile access across the African continent. The more affluent Northern and Southern Africa have much higher rates of internet and mobile penetration, with overall levels of both being higher for men than women. The middle of Africa, and especially Sub-Saharan Africa have the lowest internet penetration and also the largest gender gaps. This broad pattern is also reflected in the mobile gender gap. Especially in Southern Africa, there is close to parity between ownership of mobile phone. At the subnational level, there is much geographic heterogeneity. This is apparent in both the ground-truth and the modeled estimates.

There are several promising avenues for further research that we will expand on. First, as shown in Figure A9, we are better at predicting the ground truth in some countries and settings than others. In our next steps, we intend to examine where our predictions do better or worse and diagnose factors that explain these residuals. Second, the models presented here do not explicitly account for the hierarchical structure of the data; in next steps we will explore the value of explicitly modeling the hierarchical structure of these data (e.g., subnational units nested within countries). Third, our ground truth training data is from the Demographic and Health Surveys, which were collected between 2016 and 2019, while our estimates of Facebook Audience size were collected in September 2021. This continuity between these timescales could be modeled or otherwise adjusted for. Finally, our leave-one-country-out cross-validation strategy, while more conservative than traditional k-fold cross validation, may not perfectly capture how our model would perform on other countries we have no DHS data for. For instance, if countries that had a DHS survey varied systematically from countries that do not in a way that influenced the predictiveness of our models, our LOCO-CV metric might overstate our model's performance.

References

Abu-Shanab, Emad and Nebal Al-Jamal. 2015. "Exploring the Gender Digital Divide in Jordan." *Gender, Technology and Development* 19(1):91–113.

Aker, Jenny C. and Isaac M. Mbiti. 2010. "Mobile Phones and Economic Development in Africa." *Journal of Economic Perspectives* 24(3):207–232.

Araujo, Matheus, Yelena Mejova, Ingmar Weber and Fabricio Benevenuto. 2017. Using Facebook Ads Audiences for Global Lifestyle Disease Surveillance: Promises and Limitations. In *Proceedings of the 2017 ACM on Web Science Conference*. WebSci '17 New York, NY, USA: ACM pp. 253–257.

Billari, Francesco C., Osea Giuntella and Luca Stella. 2019. "Does Broadband Internet Affect Fertility?" *Population Studies* 73(3):297–316.

Blumenstock, Joshua, Gabriel Cadamuro and Robert On. 2015. "Predicting Poverty and Wealth from Mobile Phone Metadata." *Science* 350(6264):1073–1076.

Burke, Marshall, Marissa L. Childs, Brandon de la Cuesta, Minghao Qiu, Jessica Li, Carlos F. Gould, Sam Heft-Neal and Michael Wara. 2023. "The Contribution of Wildfire to PM2.5 Trends in the USA." *Nature* pp. 1–6.

Chi, Guanghua, Han Fang, Sourav Chatterjee and Joshua E. Blumenstock. 2022. "Microestimates of Wealth for All Low- and Middle-Income Countries." *Proceedings of the National Academy of Sciences* 119(3):e2113658119.

Cohen, Robin A. and Patricia F. Adams. 2011. "Use of the Internet for Health Information: United States, 2009." *NCHS data brief* (66):1–8.

Dettling, Lisa J. 2017. "Broadband in the Labor Market: The Impact of Residential High-Speed Internet on Married Women's Labor Force Participation." *ILR Review* 70(2):451–482.

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

Fatehkia, Masoomali, Isabelle Tingzon, Ardie Orden, Stephanie Sy, Vedran Sekara, Manuel Garcia-Herranz and Ingmar Weber. 2020. "Mapping Socioeconomic Indicators Using Social Media Advertising Data." *EPJ Data Science* 9(1):22.

Fatehkia, Masoomali, Ridhi Kashyap and Ingmar Weber. 2018. "Using Facebook Ad Data to Track the Global Digital Gender Gap." *World Development* 107:189–209.

Findlay, Robyn A. 2003. "Interventions to Reduce Social Isolation amongst Older People: Where Is the Evidence?" *Ageing and Society* 23(5):647–658.

GSMA. 2023. The Mobile Gender Gap Report. Technical report.

Hilbert, Martin. 2011. "Digital Gender Divide or Technologically Empowered Women in Developing Countries? A Typical Case of Lies, Damned Lies, and Statistics." *Women's Studies International Forum* 34(6):479–489.

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

Kashyap, Ridhi. 2021. "Has Demography Witnessed a Data Revolution? Promises and Pitfalls of a Changing Data Ecosystem." *Population Studies* 75(sup1):47–75.

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.

Kashyap, Ridhi, R Gordon Rinderknecht, Aliakbar Akbaritabar, Diego Alburez-Gutierrez, Sofia Gil-Clavel, André Grow, Jisu Kim, Douglas R Leasure, Sophie Lohmann, Daniela Veronica Negraia et al. 2023. Digital and Computational Demography. In *Research Handbook on Digital Sociology.* Edward Elgar Publishing pp. 47–85.

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.".

Lagan, Briege M., Marlene Sinclair and W. George Kernohan. 2010. "Internet Use in Pregnancy Informs Women's Decision Making: A Web-Based Survey." *Birth (Berkeley, Calif.)* 37(2):106–115.

Lazer, David, Ryan Kennedy, Gary King and Alessandro Vespignani. 2014. "The Parable of Google Flu: Traps in Big Data Analysis." *Science* 343(6176):1203–1205.

Leasure, Douglas R, Ridhi Kashyap, Francesco Rampazzo, Claire A Dooley, Benjamin Elbers, Maksym Bondarenko, Mark Verhagen, Arun Frey, Jiani Yan, Evelina T Akimova et al. 2023. "Nowcasting Daily Population Displacement in Ukraine through Social Media Advertising Data." *Population and Development Review* .

Lund, Stine, Birgitte B. Nielsen, Maryam Hemed, Ida M. Boas, Azzah Said, Khadija Said, Mkoko H. Makungu and Vibeke Rasch. 2014. "Mobile Phones Improve Antenatal Care Attendance in Zanzibar: A Cluster Randomized Controlled Trial." *BMC Pregnancy and Childbirth* 14(1):29.

Masi, Christopher M., Hsi-Yuan Chen, Louise C. Hawkley and John T. Cacioppo. 2011. "A Meta-Analysis of Interventions to Reduce Loneliness." *Personality and social psychology review : an official journal of the Society for Personality and Social Psychology, Inc* 15(3):10.1177/1088868310377394.

Michalopoulos, Stelios and Elias Papaioannou. 2014. "National Institutions and Subnational Development in Africa." *The Quarterly Journal of Economics* 129(1):151–214.

Muschert, Glenn W., Massimo Ragnedda, ed. 2013. *The Digital Divide: The Internet and Social Inequality in International Perspective.* London: Routledge.

Pesando, Luca Maria, Valentina Rotondi, Manuela Stranges, Ridhi Kashyap and Francesco C Billari. 2021. "The Internetization of International Migration." *Population and Development Review* 47(1):79–111.

Puterman, Eli, Jordan Weiss, Benjamin A. Hives, Alison Gemmill, Deborah Karasek, Wendy Berry Mendes and David H. Rehkopf. 2020. "Predicting Mortality from 57 Economic, Behavioral, Social, and Psychological Factors." *Proceedings of the National Academy of Sciences* 117(28):16273–16282.

Rampazzo, Francesco, Jakub Bijak, Agnese Vitali, Ingmar Weber and Emilio Zagheni. 2021. "A Framework for Estimating Migrant Stocks Using Digital Traces and Survey Data: An Application in the United Kingdom." *Demography* 58(6):2193–2218.

Resce, Giuliano and Diana Maynard. 2018. "What Matters Most to People around the World? Retrieving Better Life Index Priorities on Twitter." *Technological Forecasting and Social Change* 137:61–75.

Robinson, Laura, Shelia R. Cotten, Hiroshi Ono, Anabel Quan-Haase, Gustavo Mesch, Wenhong Chen, Jeremy Schulz, Timothy M. Hale and Michael J. Stern. 2015. "Digital Inequalities and Why They Matter." *Information, Communication & Society* 18(5):569–582.

Rojas, Guillermo. 2015. "Harnessing Technology to Streamline Data Collection.".

Rose, Sherri. 2013. "Mortality Risk Score Prediction in an Elderly Population Using Machine Learning." *American Journal of Epidemiology* 177(5):443–452.

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.

Salganik, Matthew J. 2018. *Bit by Bit: Social Research in the Digital Age.* Princeton University Press.

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

Union, International Telecommunication. 2017. Fast-Forward Progress Leveraging Tech to Achieve the Global Goals. Technical report.

Union, International Telecommunication. 2022. Bridging the Gender Divide. Technical report.

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).

World Bank Group. 2016. *World Development Report 2016: Digital Dividends*. Washington, DC: World Bank.

WorldPop. 2023. "Open Spatial Demographic Data and Research." https://www.worldpop.org/.

Zagheni, Emilio, Ingmar Weber and Krishna Gummadi. 2017. "Leveraging Facebook's Advertising Platform to Monitor Stocks of Migrants." *Population and Development Review* 43(4):721–734.

Supplemental Information

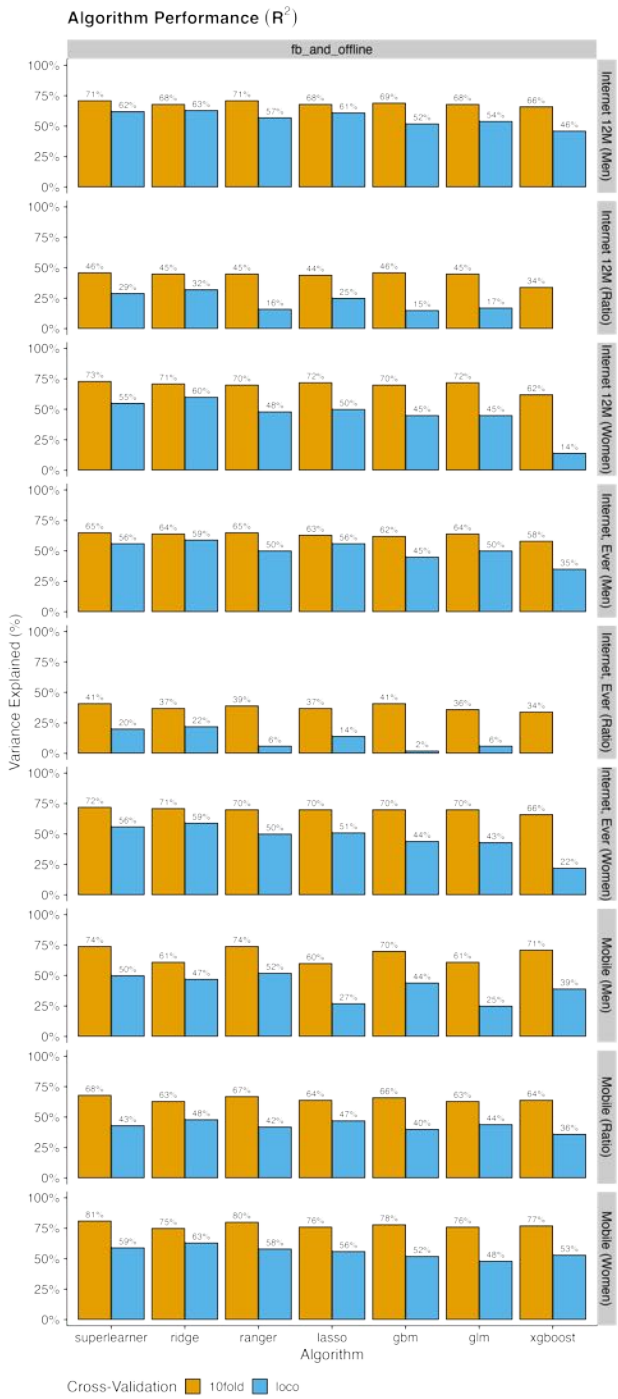

Figure A6: The R^2 from leave-one-country-out cross-validation and 10-fold cross-validation

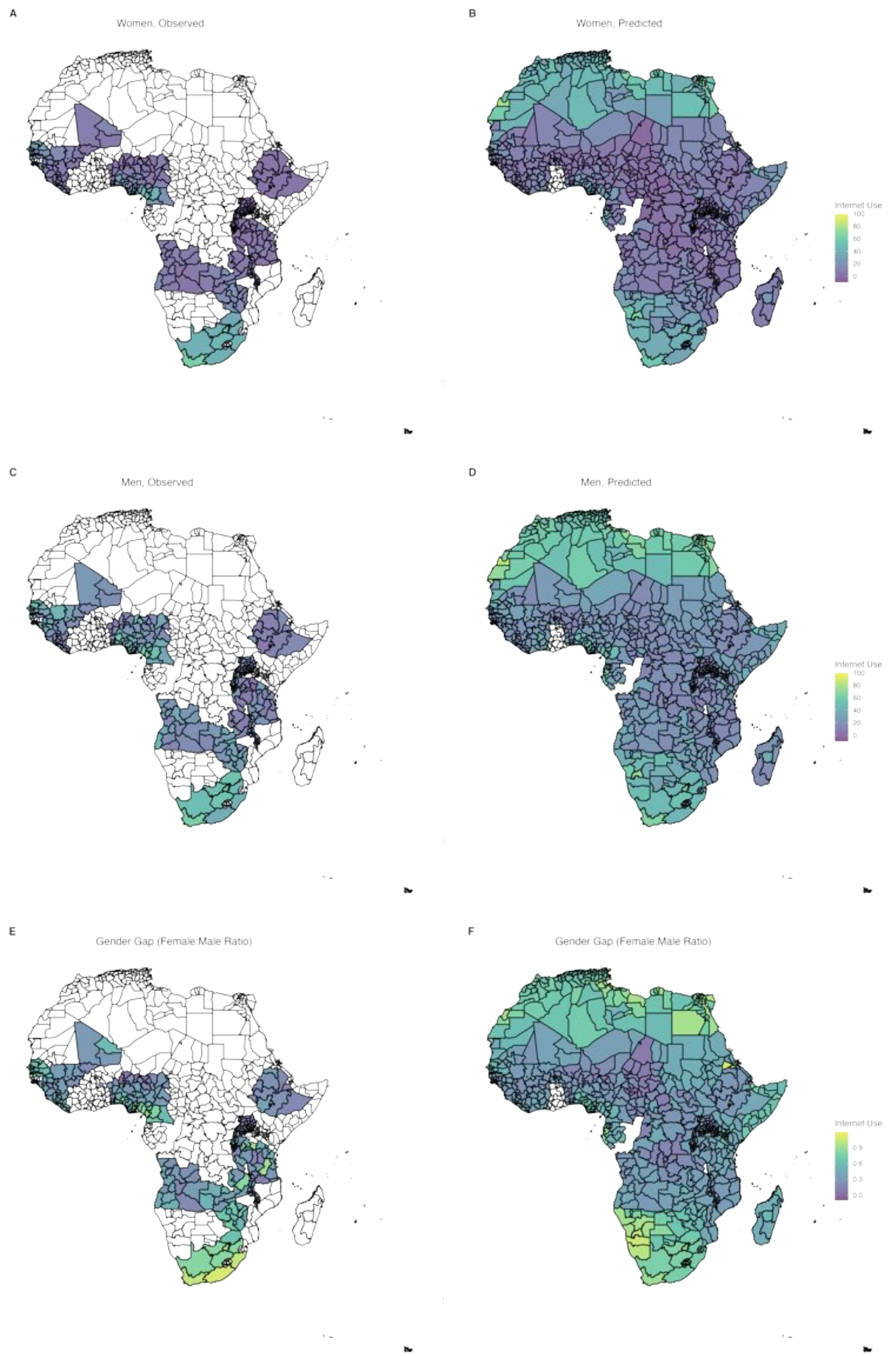

Figure A7: **Panel (A)**, **Panel (C)**, **Panel (E)** show survey-based 'ground truth' estimates of internet penetration (past 12 months) indicators for 19 countries. **Panel (B)**, **Panel (D)**, **Panel (F)** show model-based estimates of the internet use digital gender gaps for 55 countries and 4 territories.

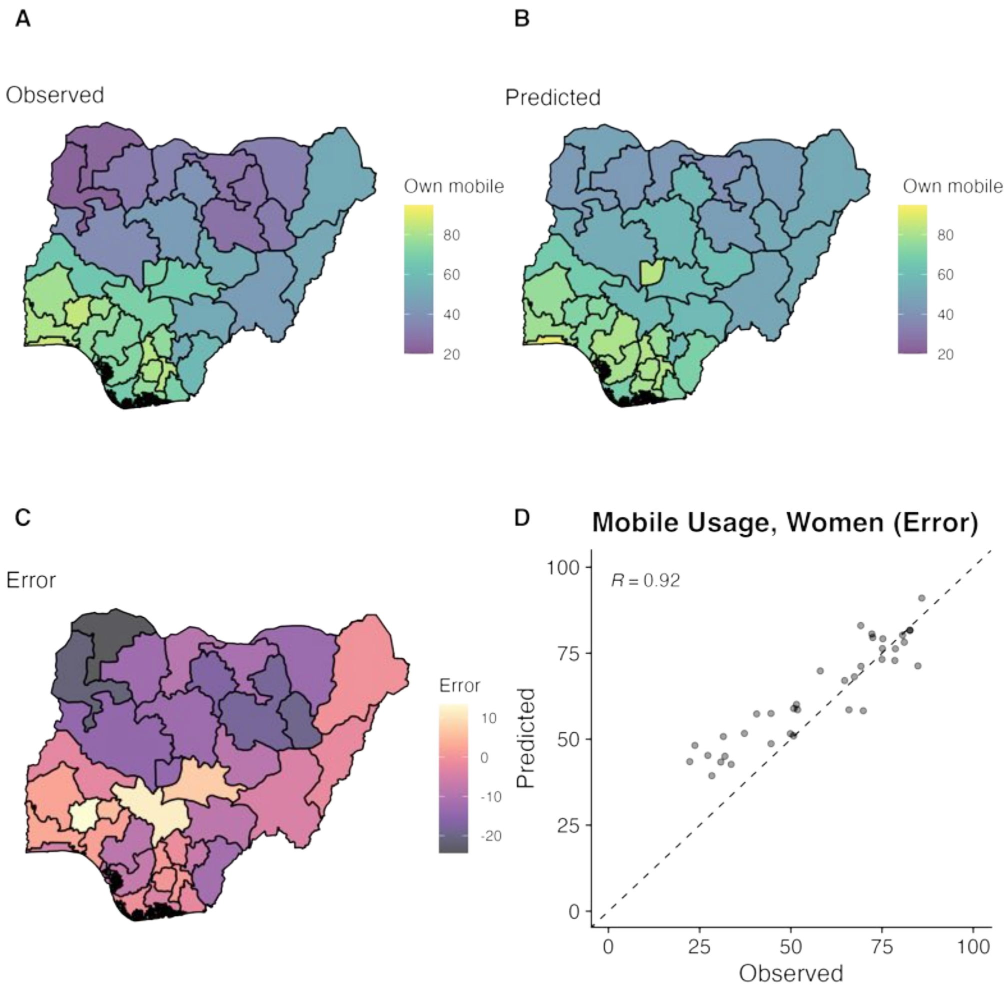

Figure A8: For women in Nigeria, the observed rate of mobile phone ownership **(Panel A)**, model-based predictions of rate of internet use **(Panel B)**, and the error between our observed and predicted values **(Panel C, Panel D).**

Indicator	**Detail**	**SuperLearner**			**Random Forest**			**Lasso**		
		R^2	**RMSE**	**MAE**	R^2	**RMSE**	**MAE**	R^2	**RMSE**	**MAE**
Owns Mobile Phone	Women	$0.61^\dagger$	$12.74^\dagger$	$10.21^\dagger$	0.58	13.16	10.95	0.51	14.25	11.61
	Men	0.51	10.79	$8.23^\dagger$	$0.52^\dagger$	$10.70^\dagger$	8.28	0.26	13.27	10.50
	Gender Ratio	$0.42^\dagger$	$0.14^\dagger$	$0.11^\dagger$	0.44	0.14	0.11	0.47	0.13	0.11
Accessed Internet (12 Months)	Women	$0.56^\dagger$	$9.49^\dagger$	$6.37^\dagger$	0.52	9.90	6.73	0.52	9.92	7.22
	Men	$0.63^\dagger$	$10.44^\dagger$	$7.59^\dagger$	0.59	10.89	7.95	0.59	10.96	8.21
	Gender Ratio	$0.29^\dagger$	$0.20^\dagger$	$0.15^\dagger$	0.18	0.22	0.17	0.26	0.20	0.16
Accessed Internet (Ever)	Women	$0.58^\dagger$	$9.79^\dagger$	$6.47^\dagger$	0.52	10.44	7.30	0.50	10.69	7.78
	Men	$0.58^\dagger$	$11.60^\dagger$	$8.53^\dagger$	0.53	12.28	9.20	0.56	11.90	8.99
	Gender Ratio	$0.22^\dagger$	$0.23^\dagger$	$0.16^\dagger$	0.07	0.25	0.18	0.14	0.24	0.17

Table A3: Model Performance by Outcome and Metric for countries with available ground-truth data. Dagger denotes the top-performing model my metric (highest R^2, lowest RMSE and MAE). Model performance was assessed with leave-one-country-out cross-validation.

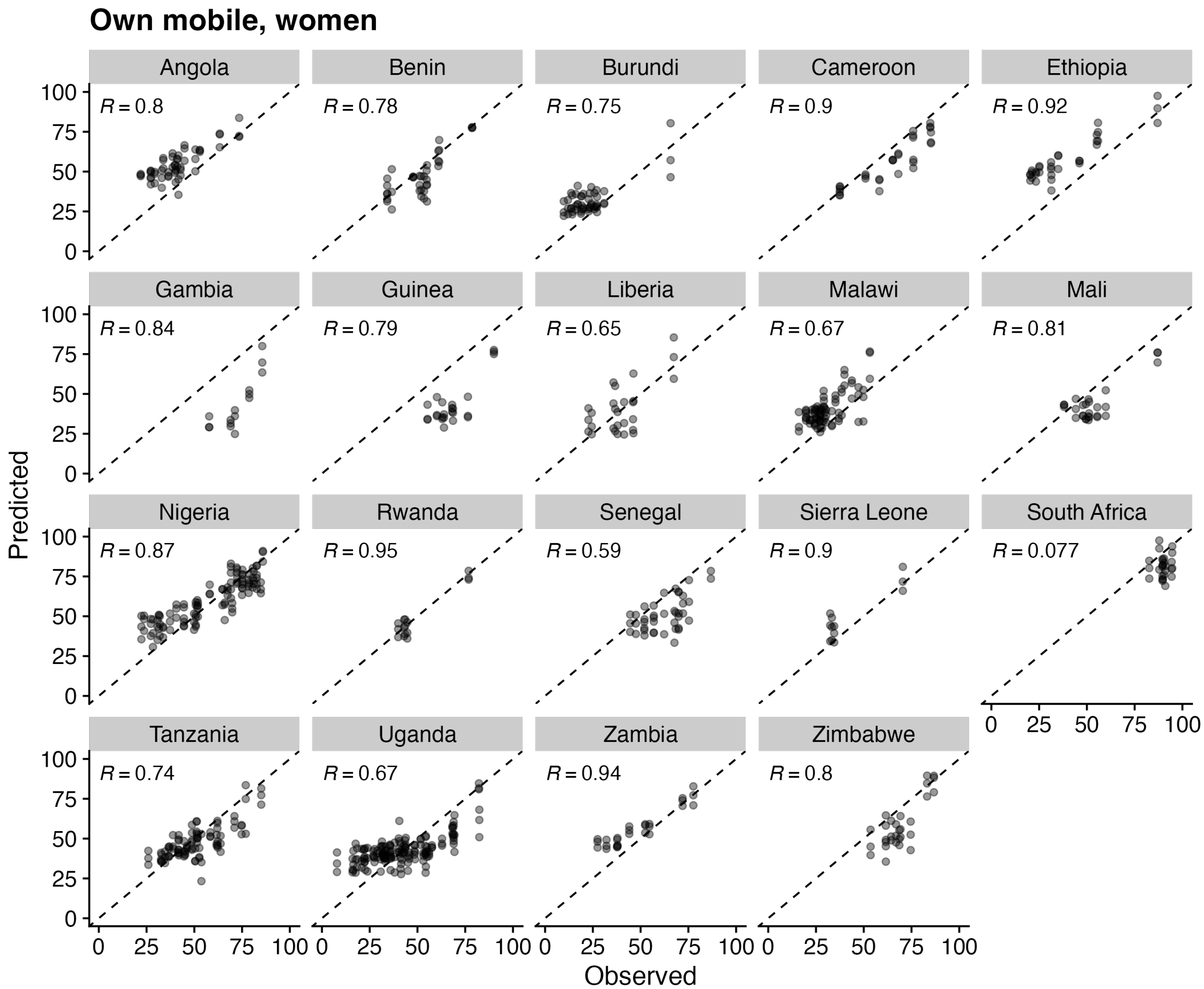

Figure A9: Comparison of observed ground-truth and predictions for percent of women who own mobile phones.