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

Casey F. Breen[†‡§] Masoomali Fatehkia[¶] Jiani Yan[†§] Xinyi Zhao[§‖]

Douglas R. Leasure[‡§] Ingmar Weber[**] Ridhi Kashyap[†‡§]

Draft Version: December 8, 2024

Abstract

The digital revolution has brought many societal and economic benefits, yet digital access remains unequal, especially by gender in low- and middle-income countries (LMICs). While national-level estimates are increasingly available, reliable subnational estimates of digital gender inequalities are lacking, which are crucial for monitoring gaps and implementing targeted interventions within the global sustainable development goals that emphasize closing inequalities both between and within countries. We develop subnational estimates of internet and mobile adoption by gender for 2,158 regions across 118 LMICs, using machine-learning algorithms applied to Facebook user counts, geospatial data, development indicators, and population data, calibrated and assessed against household survey data from 31 LMICs. Our findings reveal substantial disparities both between and within countries: for women in the lowest human development countries, over 30% of inequality is within-country, and in more than 20 countries, gender gaps exceed 30 percentage points between regions. These disparities contribute to a global context where women are 21% less likely to use the internet and 17% less likely to own a mobile phone, corresponding to over 385 million fewer women owning a mobile phone and over 360 million fewer women using the internet, with significant implications for policy formulation and targeted infrastructure investments.

Keywords

Digital divide, sustainable development, machine learning, low- and middle-income countries

*Address correspondence: casey.breen@demography.ox.ac.uk and ridhi.kashyap@nuffield.ox.ac.uk. For helpful discussions and feedback, we thank Dennis Feehan, Michelle Niemann, participants of the PAA 2024 "Digital Demography" session, the EPC 2024 "Modelling Subnational and Spatial Variation" session, the Digital Gender Gaps workshop, the Johns Hopkins Population Health seminar series, and the Leverhulme Centre for Demographic Science seminar series. The authors acknowledge financial support from the Bill and Melinda Gates Foundation (INV-045370) and Leverhulme Trust (Grant RC-2018-003) for the Leverhulme Centre for Demographic Science.

[†]Department of Sociology, University of Oxford, Oxford OX1 1JD, United Kingdom

[‡]Nuffield College, University of Oxford, Oxford OX1 1NF, United Kingdom

[§]Leverhulme Centre for Demographic Science, University of Oxford, Oxford OX1 1JD, United Kingdom

[¶]Qatar Computing Research Institute, Hamad bin Khalifa University, Doha, Qatar

[‖]Department of Digital and Computational Demography, Max Planck Institute for Demographic Research, Rostock 18057, Mecklenburg-Vorpommern, Germany

[**]Saarland Informatics Campus, Saarland University, Saarbrucken 66123, Germany

1 Introduction

The digital revolution has yielded major societal and economic benefits in low- and middle-income country (LMIC) settings. Internet and mobile technologies are powerful mediums for boosting social connectivity (Masi et al., 2011; Findlay, 2003), promoting social learning, and providing access to new information channels (DiMaggio and Hargittai, 2001; Kashyap et al., 2023). Increasing digital adoption also has other "digital dividends," such as job creation (Hjort and Poulsen, 2019), better educational outcomes (Kho, Lakdawala and Nakasone, 2018), and improved economic growth (Czernich et al., 2011). Digital technologies also have the potential to empower women across many domains and reduce gender inequalities by providing access to information, networks, and vital services that lead to higher contraceptive uptake (Rotondi et al., 2020), increased labor market and economic opportunities (Dettling, 2017; Garcia et al., 2018; Nkoumou Ngoa and Song, 2021; Amber and Chichaibelu, 2023), and improved child and maternal health (Lagan, Sinclair and Kernohan, 2010; Lund et al., 2014; Rotondi et al., 2020; Bashingwa et al., 2021). The benefits of such digital technology are generally greatest in the most unequal, disadvantaged regions (Rotondi et al., 2020).

Yet the benefits of this digital revolution have accrued unevenly. Over 2.6 billion people have never accessed the internet, and the majority of the unconnected are women and girls (ITU, 2022). This *digital divide* by gender is an increasingly salient dimension of contemporary population inequality and is especially pronounced in LMICs. Reliable quantitative estimates of digital gender inequalities are key for tracking progress on and implementing targeted policies and interventions 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.

Data on digital adoption in LMIC settings are significantly lacking. The availability of national-level estimates of digital gender gaps has improved (Fatehkia, Kashyap and Weber, 2018; Kashyap et al., 2020; ITU, 2022), but to date, there are no subnational estimates of digital adoption by gender for the majority of LMICs in the world. Past estimates of digital

adoption have typically been based on probabilistic household surveys (Cohen and Adams, 2011; World Bank Group, 2016), which either lack gender disaggregation or are underpowered for subnational analyses. In other contexts, such as poverty (Blumenstock, Cadamuro and On, 2015; Pokhriyal and Jacques, 2017), wealth (Chi et al., 2022), and population mapping (Boo et al., 2022), "big data" derived from satellite, social media and mobile phone records have been used to overcome data gaps in survey-based approaches. The potential of non-traditional sources for mapping gender inequality indicators at subnational geographical resolution has yet to be explored.

Subnational estimates are critical because internet and mobile phone adoption may vary substantially within countries, just as key factors associated with digital adoption, such as economic and educational development, vary (Permanyer and Smits, 2020). These subnational inequalities in human development are often largest at low levels of human development and also substantial at middle levels. When taking subnational variation into account, inequalities in human development among LMICs are approximately double those when only accounting for national-level variation (Permanyer and Smits, 2020). Moreover, factors associated with digital adoption for women and men can differ, as economic development does not necessarily weaken gender inequalities and facilitate women's empowerment linearly (Jayachandran, 2015). As development programs are increasingly deployed through digital means, and are often targeted in local geographies, subnational estimates are critical to monitoring their progress and understanding how digital inequalities affect sustainable development.

2 Results

Here we introduce a new approach to estimating subnational digital adoption and gender gaps by applying machine learning algorithms to social media data, big geospatial data, and development indicators. We focus on internet adoption, defined as having used the internet in the past 12 months, and mobile phone ownership, defined as the proportion of the population who own a mobile phone. We train and assess the performance of these algorithms using "ground truth" data from subnationally-representative Demographic and

Health Surveys (DHS) from 500 regions across 31 LMICs. We use this approach to estimate digital adoption for the present day at the Administrative Level 1 (admin-1) for 2,158 regions in 118 LMICs where sufficient data are available to make estimates. We focus on LMICs as national-level adoption in these settings is low, digital gender gaps disfavoring women are large (Fatehkia, Kashyap and Weber, 2018; Kashyap et al., 2020), and gender-disaggregated data at finer geographical resolution on digital adoption are limited. Our results reveal large disparities in digital adoption between and within countries, with implications for real-time monitoring of global SDGs, policy formulation, and infrastructure investment.

Our general approach is illustrated in Figure 1D. We use survey-based indicators of internet usage and mobile phone ownership for men and women combined with a set of satellite-based geospatial data, development indicators, and social media user count data to train a machine learning model. The social media data we use are Facebook monthly active user counts by gender obtained from the public Facebook Marketing API. The availability of 31 recent DHS surveys (Figure 1A) provides us good coverage of ground truth data on internet and mobile adoption by gender to calibrate and assess our models. The near global coverage and more timely, higher-frequency availability of the geospatial and social media features provides a basis for extrapolation to locations without DHS data and produce estimates for the present ("nowcasting") . Using this approach, we expand our geographical coverage of digital adoption estimates to 2,158 regions in 118 LMICs across the globe for 2024. Figure 1B shows our estimates of internet adoption for women in 2024; Figure 1C is an enlargement of estimates for Nigeria.

2.1 Accuracy of predictions

To assess the predictive accuracy of our machine learning models, we compare our ground-truth data from DHS surveys to our model predictions for each admin-1 unit using leave-one-country-out cross-validation (LOCO-CV). In this cross-validation approach, we hold out data from one country at a time. We train our model on ground truth data from all countries not withheld, then make predictions for the hold-out country. This process is repeated for each country. The resulting estimates give insight into model performance in countries where we have no ground truth data. This is a contrasting and more stringent approach to standard

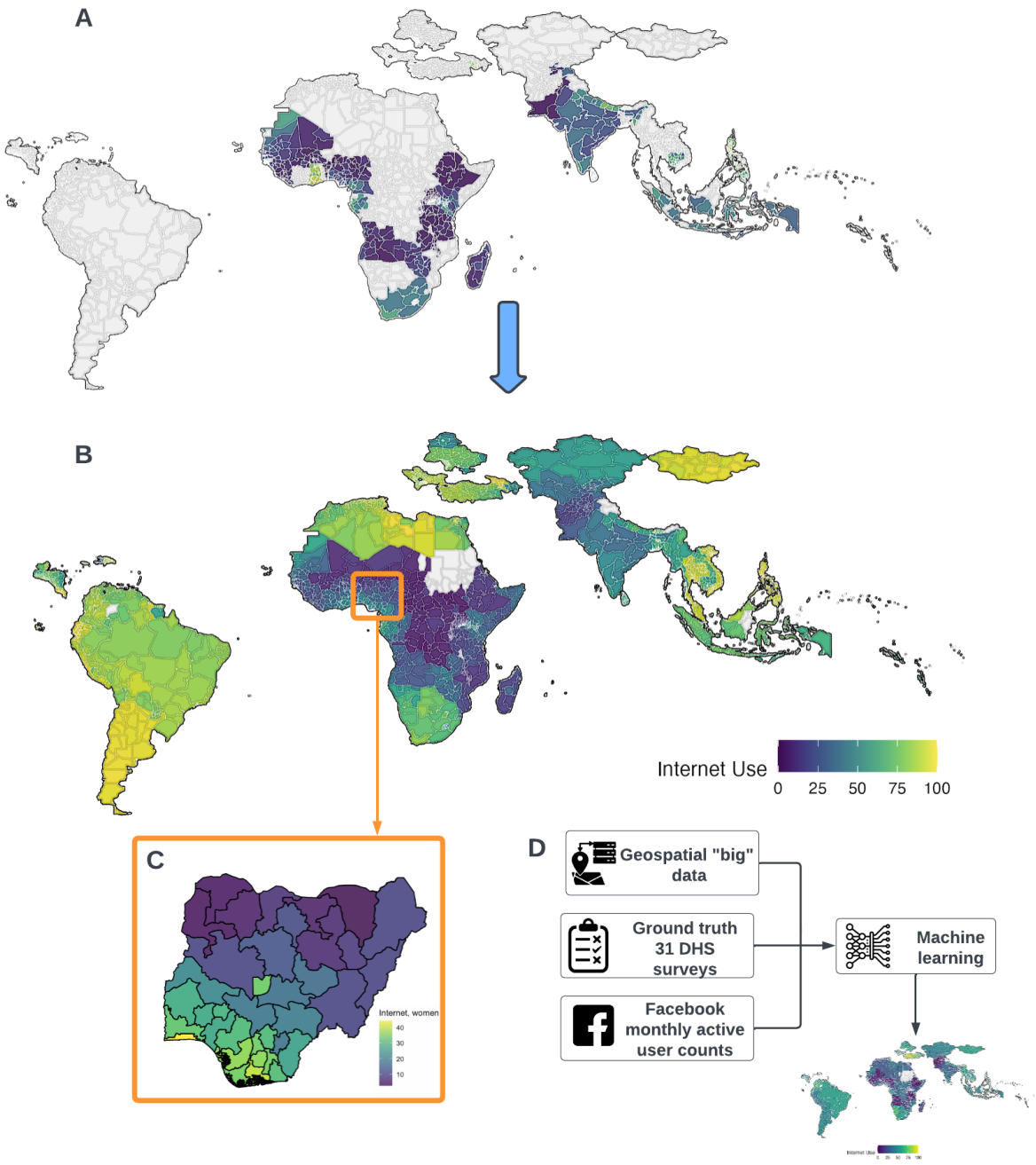


Figure 1: Description of general approach. *(A)* The 31 countries with available ground truth data. *(B)* Model-based estimates of contemporary internet adoption for women across 118 countries. *(C)* Enlargement of predictions for Nigeria. *(D)* Illustration of inputs to the machine learning models for predicting digital adoption indicators.

10-fold cross-validation where we leave out one fold (one-tenth) of the dataset at a time instead of entire countries. LOCO-CV is more stringent because leaving out entire countries

prevents any data from those countries from influencing the training process; even having some data from a country can give insights into levels of adoption within each country.

Figure 2 plots our model-based predicted values against our observed ground truth values across all subnational units for the different indicators of internet and mobile adoption by gender, and the gender gap of both. Our predictions closely align with the ground truth where these are available.[1] On average, our predictions are better for women's internet adoption than men's. Our predictive accuracy is lowest for the gender gaps, which reflects more uncertainty in our ground truth data (for more details, see Appendix C.3). For each estimate, we also calculate a corresponding estimate of uncertainty (Appendix C.2). Our predicted error is smallest in low-adoption settings, which indicates that our approach is able to identify areas that are lagging behind.

[1]To quantify the agreement between our predictions and ground truth, we report both the coefficient of determination (R^2) and the Pearson correlation coefficient (r). The coefficient of determination (R^2) represents the proportion of total variation in the dependent variable that is predictable from the independent variable.

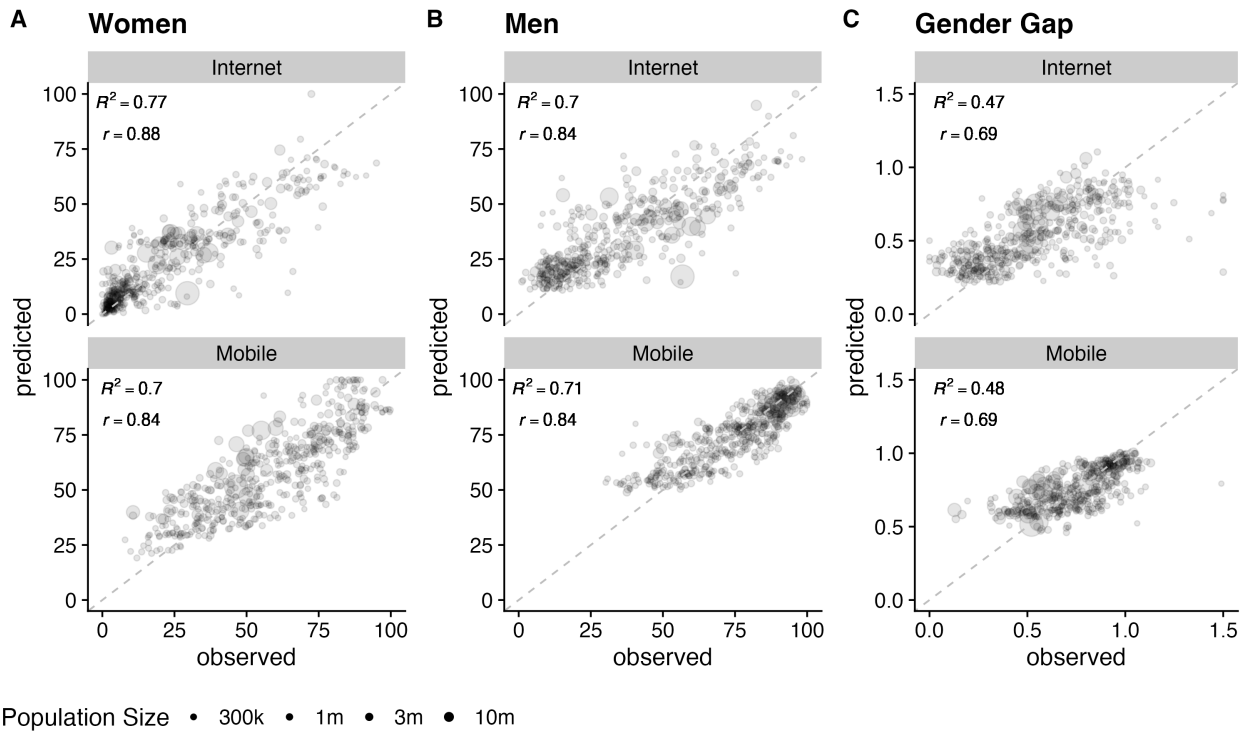


Figure 2: Summary of model performance using leave-one-country-out cross-validation (LOCO-CV). *(A)* Scatterplot of predicted and observed values for internet adoption and mobile phone ownership for women. *(B)* Scatterplot of predicted and observed values for mobile phone ownership and internet penetration for men. *(C)* Scatterplot of predicted vs. observed values of the gender gap for internet and mobile phone adoption. The R^2 values report the coefficient of determination and the r values report the Pearson Correlation Coefficient.

2.2 Within-country results

Figure 3 shows the within-country accuracy of our estimates of internet adoption for women. Within countries, there is subnational heterogeneity in ground-truth rates of internet penetration, as we illustrate using three examples from the African continent. In Nigeria, 55% of women in the relatively affluent and urban southwestern state of Lagos have accessed the internet in the past 12 months, while less than 1% of women have accessed the internet in the rural northern state of Kebbi (Figure 3A), highlighting the importance of considering the subnational context. Within countries, the model-based estimates align closely with the observed ground truth.

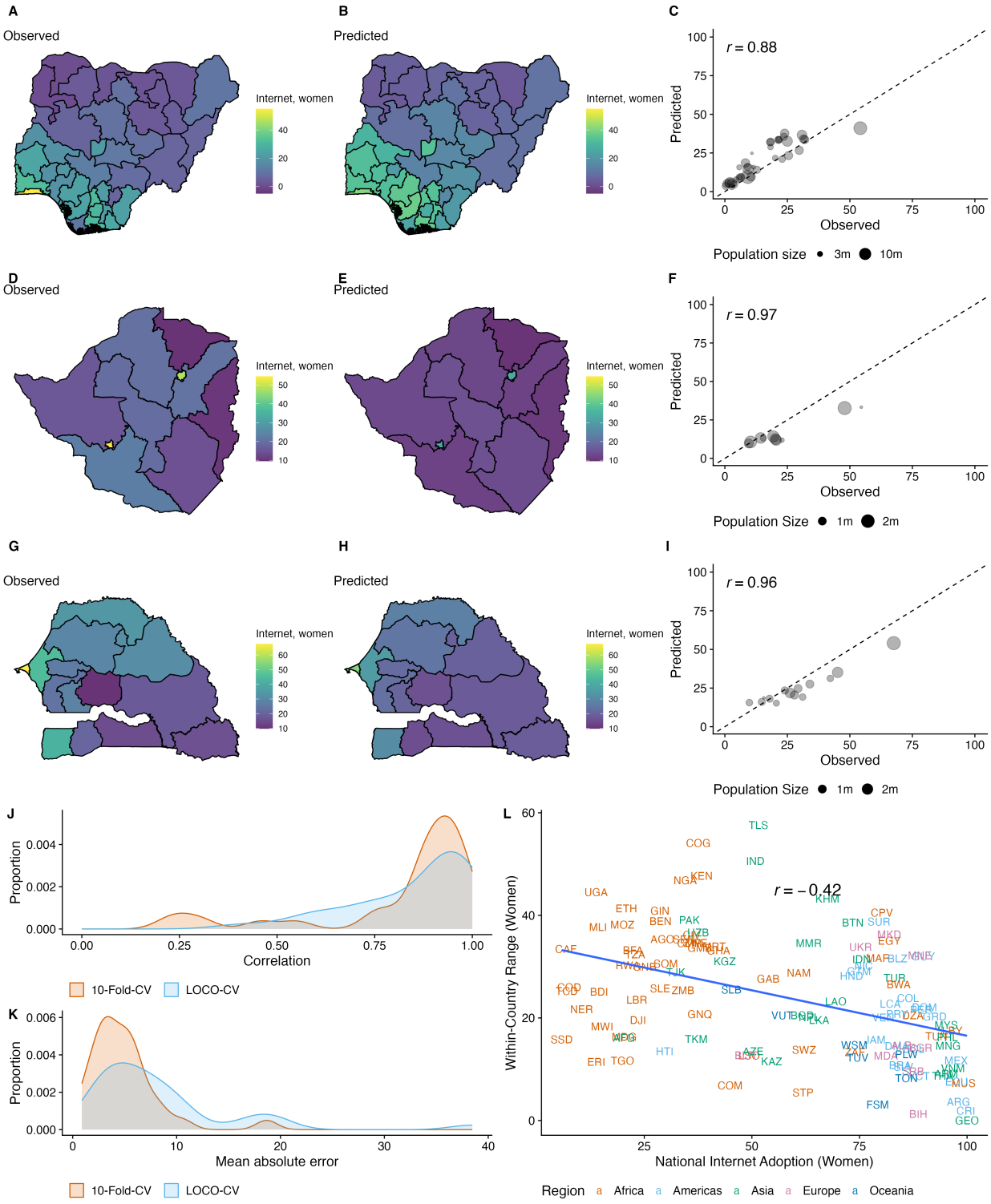


Figure 3: Performance of model predicting internet use among women, assessed using leave-one-country-out cross-validation (LOCO-CV). *(A–C)* Observed and predicted estimates of internet adoption for women for Nigeria. *(D–F)* Observed and predicted estimates of internet adoption for women for Zimbabwe. *(G–I)* Observed and predicted estimates of internet adoption for women for Senegal. *(J-K)* Distribution of within-country mean absolute error (MAE) and correlation between observed and predicted values for both standard 10-fold CV and LOCO-CV. *(L)* Relationship between national-level estimates and within-country range (difference between top and bottom subnational units) for female internet penetration.

For Zimbabwe (Figure 3D), our model is able to accurately estimate internet adoption even in the small geographic provinces of Harare and Bulawayo, which have nearly triple the rate of internet penetration of neighboring regions. In Senegal (Figure 3G), our estimates slightly under-predict overall internet penetration, but closely capture the overall pattern of adoption. Figure 3J shows the within-country distribution of the Pearsons Correlation Coefficient (r) and Figure 3K shows the within-country distribution of mean absolute error between our predictions and ground truth using both standard 10-fold cross-validation and LOCO-CV. For a comparison of our model-based estimates and ground truth for each country, see Figure A13. Finally, Figure 3L shows the within-country range in internet adoption among women at different levels of national adoption for our 2024 estimates, indicating that within-country variation is generally larger at lower levels of adoption. This pattern is consistent across other adoption indicators (Figure A16).

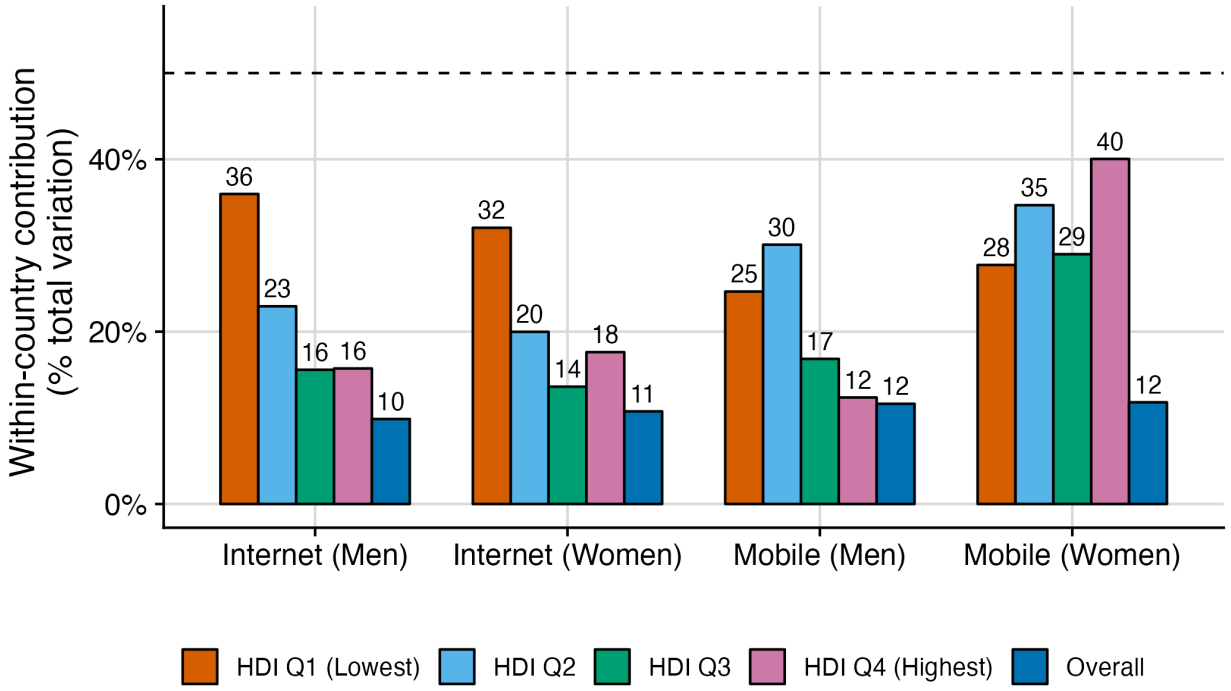


Figure 4: The relative contribution of within-country variation in internet and mobile adoption to total variation by human development index (HDI) quantile. Dashed line represents equal contribution of within-country and between-country variation to overall inequality.

Figure 4 shows the proportion contribution of within-country inequality in internet and mobile adoption to total inequality within different human development index (HDI) quan-

tiles for contemporary estimates based on most recent features for 118 LMICs (further details described in Appendix C.4). Subnational inequality accounts for a large proportion of the overall variation in adoption, especially at lower levels of human development. For internet adoption in countries with the lowest levels of human development, the within-country inequality accounts for 32% (women) and 36% (men) of the overall inequality.

2.3 Feature sets

Next, we compare the performance of models trained on different feature sets to give insight into the most important features for our models. We test three sets of features: features constructed from Facebook monthly active user counts ("Facebook features"), features derived from geospatial, satellite, and population data ("offline features"), and our full set of features. Figure 4B shows the R^2 value (coefficient of determination) based on models using each different set of features and LOCO-CV.

For mobile phone outcomes, the models trained using only Facebook features generally had the worst performance, whereas for internet outcomes, the offline models had the worst performance. Our model trained with the full set of Facebook and offline features generally had the best overall performance. These results suggest that combining offline features with social media data generally produces the most accurate estimates of digital adoption.

Figure 4A shows the most important features in our model predicting internet adoption for women; Figure A19 shows the most important features for all indicators.[2] Proxies of overall economic development and human development at the subnational level, together with national-level Facebook penetration features, emerge as the most important features. Other important features include the subnational education index and subnational Facebook penetration features. These are consistent with prior research modelling digital gender gaps at the national level (Kashyap et al., 2020; Fatehkia, Kashyap and Weber, 2018). While income indices are the most important predictor for overall levels of adoption, educational, human development, and gender development indices are more important for predicting

gender gaps.

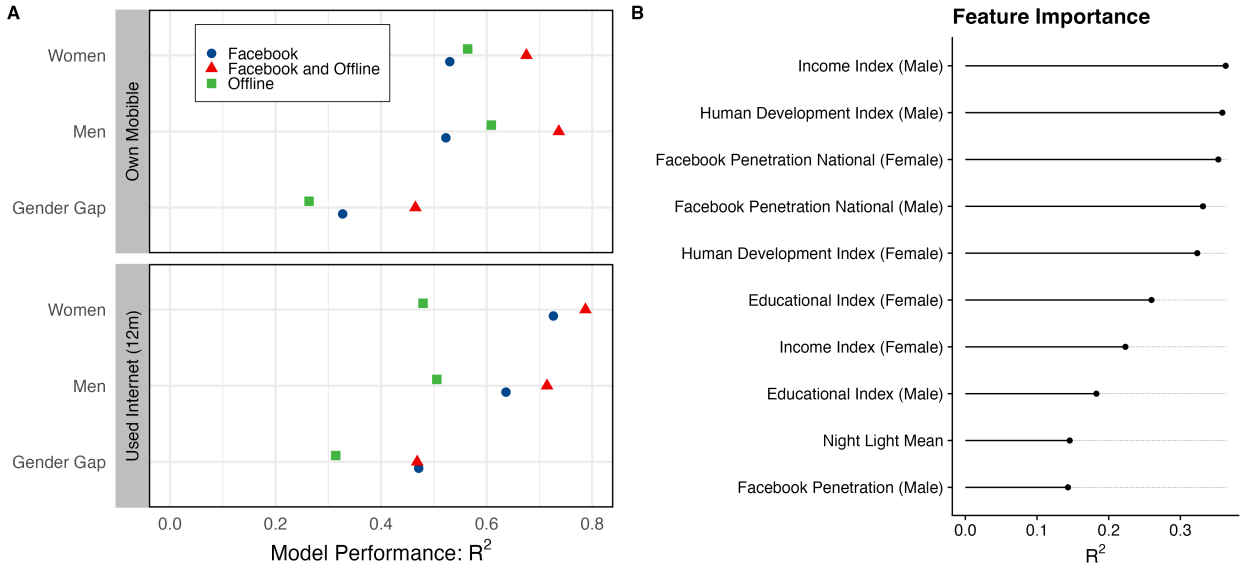


Figure 5: Model performance and feature importance. *(A)* The R^2 values from leave-one-country-out cross-validation (LOCO-CV) using Facebook predictors, offline predictors, and both Facebook and offline predictors. *(B)* The 10 top features with strongest predictive power for internet penetration for women, as measured by R^2 for a univariate regression. All features are at a subnational resolution unless explicitly noted.

3 Discussion

Although access to digital technology is an increasingly important dimension of population inequality, and acknowledged as a key indicator within the global sustainable development goals (SDGs),[3] consistent and reliable subnational estimates of internet or mobile adoption are lacking, particularly in LMICs. Here, we demonstrate a new approach for estimating subnational levels of internet adoption and mobile phone ownership by gender by applying machine learning algorithms to Facebook user counts, geospatial data, development indicators, and population composition data. Our approach enables us to significantly expand subnational estimates from 500 regions in 31 LMICs, for which DHS ground truth are available, to 2,158 regions in 118 LMICs. These estimates for all regions to-

[2]Given the number of highly correlated predictors (e.g., Facebook internet penetration for men and Facebook internet penetration for women), we calculate feature importance as the R^2 for a univariate regression, assessed using using LOCO-CV.

gether are provided freely alongside this study, and will be publicly available and updated on www.digitalgendergaps.org.

Our results highlight the importance of focusing on the subnational context. In nearly 40 countries, the gap between the highest and lowest subnational units with respect to female internet adoption exceeds 30 percentage points, with an average gap of 24 percentage points. This subnational variation is especially pronounced in countries with very low levels of human development, where within-country inequality can account for over 30% of total inequality. This underscores the importance of subnational resolution for a more complete understanding of inequality in digital adoption, especially in countries with the lowest overall levels of human development. Our estimates provide a valuable lens for researchers and policy makers through which to assess areas that stand to benefit from the increasing rollout of digital programs and services within global development policies, as well as those at risk of being left behind.

There are several promising avenues for further research to address some of our study's limitations. First, as shown in Figure A13, there is some heterogeneity across countries in the overall accuracy of our predictions. Future work could consider settings where predictions do better or worse and diagnose factors that explain these residuals. For instance, even though DHS surveys use consistent questions and implementation, there may be country-level differences in respondents' ability to understand and accurately answer questions about internet usage. Second, although we have temporally aligned our predictors with our ground truth as closely as possible, this alignment is not always perfect. Third, while Facebook is currently the world's largest social media platform, and is especially dominant in many LMICs, its popularity may decline in the future. This approach could be expanded to include social media data from additional platforms, though the ability to do so is limited by public access to platform user count data. Finally, if countries with DHS surveys systematically differ from those without DHS surveys in ways related to digital adoption, our LOCO-CV performance metrics might misrepresent our model's accuracy. External comparisons with a

[3]The three most relevant sustainable development goals that our work feeds into are SGD 4 on education, SDG 5 on gender equality and women empowerment, and SDG 17 on revitalizing global partnership for sustainable development. For a more complete explanation of each SDG and the most relevant indicator, see Appendix B.

high-quality, subnationally representative and gender-disaggregated census or survey would be a valuable validation of our estimates, but these data for the time being are significantly lacking. Our estimates, when aggregated to the national-level, show strong and consistent correlations with independent estimates from the Multiple Indicator Cluster Surveys (MICS) (see Figure A7). We call upon more data collection efforts within global surveys to monitor digital inequalities at subnational levels.

Our approach addresses the call to usher in a "data revolution for sustainable development" by integrating social media, geospatial, and population data to monitor progress on the SDGs with better geographical and temporal resolution (IEAG, 2014). We contribute to a growing body of work that emphasizes how machine learning approaches together with new sources of data can provide large-scale, cost-effective, and complementary measurement approaches to survey and field-based approaches for SDG monitoring (Fatehkia, Kashyap and Weber, 2018; Boo et al., 2022) and poverty and economic inequality mapping (Pokhriyal and Jacques, 2017; Chi et al., 2022; Mirza et al., 2021; Dong, Ratti and Zheng, 2019). We show how these innovations can be applied to map gender-disaggregated indicators at finer geographical resolution, which is crucial for reducing inequalities both within- and between-countries. While ground truth data are essential for training our models, the incorporation of higher-frequency social media and geospatial data enables us to provide a contemporary estimate of digital gender gaps. Moving forwards, our continuous FB collections and regular updates will allow us track progress on digital gender inequalities across all LMICs with a monthly resolution ("nowcasting"). Alongside the geographical breadth, this subnational pipeline can also be usefully applied to assess the impacts of policies and interventions in specific cases and contexts.

4 Materials and Methods

For this study, we use three main sources of data: ground truth from Demographic and Health Surveys (DHS), offline features, and Facebook features. Our Facebook features are constructed from the Facebook monthly active user counts obtained from the Facebook Marketing API. Our offline features are constructed from population density data, satel-

lite imagery, and subnational 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 DHS surveys in 31 LMICs.

Ground truth data on internet and mobile access Our ground-truth data comes from 31 DHS surveys conducted between 2015–2022 covering 500 subnational units. We include surveys from DHS Phase 7 onwards, when questions on digital adoption were first added to the DHS questionnaire. There are several reasons DHS surveys provide the best source of ground truth data for measuring digital adoption. First, DHS surveys are subnationally representative at the admin-1 level and have a well-established and vetted survey design and probability sample. Second, unlike many censuses which ask about digital adoption at the household level, DHS surveys collect information at the individual level about digital adoption for both men and women. The individual-level data allow us to calculate gender-specific rates of digital adoption for internet and mobile technologies. Finally, across countries, the DHS program uses harmonized survey questions and sampling design, allowing for cross-national comparisons.

We use DHS microdata to obtain estimates of the percent of men and women aged 15–49 who own a mobile phone and have accessed the internet in the past 12 months.[4] We also calculate the gender gap, defined as:

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

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.

Facebook monthly active users To obtain counts of Facebook monthly active users (MAU), we query the public Facebook Marketing API. The Facebook Marketing API pro-

[4]We focus on internet use in the past 12 months, rather than having ever used the internet, because we are interested in measuring how many people have reliable access to the internet.

vides estimates of the number of daily or monthly active users disaggregated by characteristics such as gender, age, and device type (e.g., Android, IOS) within a given geographic boundary. To query the Facebook Marketing API, we use the pysocialwatcher package (Araujo et al., 2017), and collect counts of monthly active users by gender and device type at the admin-1 level.[5]

We use these MAU counts to construct several different Facebook features. Our primary features are Facebook penetration by gender, defined as the proportion of women (or men) aged 13+ who used Facebook in the past month within a given admin-1 unit. Additionally, we create features corresponding to both the gender-specific gaps (female-to-male ratios) and the fraction of users who accessed Facebook through different access devices (e.g., iOS device). Finally, we include three national-level Facebook features on adoption by gender and gender gaps. For the full set of Facebook features used in our models, see Table A1.

Temporal alignment of features with ground truth Our DHS surveys spanned the years 2015-2022, a period under which digital adoption increased. To the extent possible, we temporally aligned our features with the year of our ground truth observations as closely as possible. For instance, to align our features temporally with ground truth from the 2018 Nigeria DHS, we calculate the population-weighted nightlights feature for Nigeria using nightlights data from 2018 and population data from 2018. When perfect temporal agreement between feature and observed ground truth is not feasible, we used the closest available year to the DHS survey year. For our Facebook features, we constructed national-level features from our regular data collections spanning 2019–2024. Since the Facebook Marketing API cannot retrieve historical MAU counts, we linearly imputed national MAU counts back to 2015 to align temporally with the year of our ground truth DHS surveys. Our subnational Facebook features are based on an April 2024 collection. To achieve temporal alignment, we rescaled the 2024 subnational MAU counts using an adjustment factor, calculated as the ratio of the national MAU counts for the year of interest to the national MAU counts in 2024 (see Appendix A for details). To further capture changes in the underlying ground truth

[5]GADM, the Database of Global Administrative Areas, is a publicly-available database of country administrative areas. When boundaries are available in the FB marketing API that match the GADM-1 boundaries, we use the default FB boundaries. When no boundary is available in Facebook that match the GADM-1 boundaries (30%), we create custom shapefiles to match the GADM-1 boundaries.

over time, we include a feature corresponding to the first year a DHS survey was in the field. For prediction of contemporary (2024) levels of adoption and gender gaps, we used the most current available version of all features.

Machine learning approach We use a machine learning approach for prediction. We predict each of these separate indicators using a combination of Facebook 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 (Lundberg, Brand and Jeon, 2022). Machine learning approaches have been applied for similar prediction settings for LMICs, such as for small-area estimation of wealth and poverty (Blumenstock, Cadamuro and On, 2015; Chi et al., 2022).

We use ensemble Superlearning—also known as weighted ensembling or stacking—a method for combining multiple machine learning algorithms into a single algorithm (Van der Laan, Polley and Hubbard, 2007). The motivation behind ensemble Superlearning is that a weighted combination of different algorithms may outperform any single algorithm by smoothing out limitations of any specific algorithm. The ensemble Superlearner algorithm selects the best weighted combination of algorithms using a cross-validation procedure to minimize overfitting risk (Phillips et al., 2023). For our ensemble Superlearner, we use a library of widely-used 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 (see Table A4 for a description of each algorithm and its weight towards the final ensemble Superlearner algorithm).

In total, we model six different outcomes: adoption levels by gender and gender gaps for both internet penetration and mobile phone ownership. We also generate an accompanying estimate of uncertainty for each prediction (Appendix C.2). For more detailed technical information on our machine learning approach, see our completed REFORMS checklist (Kapoor et al., 2023), a resource for promoting transparency and reproducibility in machine learning science.

Cross-validation To evaluate the performance of our model, we use both standard 10-fold cross-validation (10-fold CV) and leave-one-country-out cross-validation (LOCO-CV). For 10-fold CV, we randomly split our sample into ten separate folds. We trained our models on 9 folds and made predictions on a 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 31 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. LOCO-CV addresses concerns of geographical independence, providing a more stringent assessment of the model's geographical robustness.

We use these two separate cross-validation designs as they provide different perspectives. The LOCO-CV imitates a setting where we have no ground truth data for an entire country, while standard cross-validation is helpful for approximating how our model would perform in settings where ground truth data is available for most, but not all, subnational regions in a country. In comparison to 10-fold CV, LOCO-CV predictions show more conservative estimates of model performance (Figure A8).

4.1 Performance Metrics

We use several different model performance metrics to evaluate model performance. 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 is defined 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, $\bar{y}$ is the mean of the observed values, and n is the total number of observations. The R^2 value, or coefficient of determination, quantifies the proportion of variance in the dependent variable explained by the model. An R^2 value of 0 means the predictions are no better than using the mean of the outcome, whereas a value of 1 signifies perfect predictions.

As an alternative metric for assessing model fit, we use mean absolute error (MAE):

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

The 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 R^2 and MAE 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.

References

Amber, Hina and Bezawit Beyene Chichaibelu. 2023. "Narrowing the Gender Digital Divide in Pakistan: Mobile Phone Ownership and Female Labor Force Participation." *Review of Development Economics* 27(3):1354–1382.

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.

Bashingwa, Jean Juste Harrisson, Diwakar Mohan, Sara Chamberlain, Salil Arora, Jai Mendiratta, Sai Rahul, Vinod Chauhan, Kerry Scott, Neha Shah, Osama Ummer, Rajani Ved, Nicola Mulder, Amnesty Elizabeth LeFevre and Kilkari Impact Evaluation Team. 2021. "Assessing Exposure to Kilkari: A Big Data Analysis of a Large Maternal Mobile Messaging Service across 13 States in India." *BMJ global health* 6(Suppl 5):e005213.

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

Boo, Gianluca, Edith Darin, Douglas R. Leasure, Claire A. Dooley, Heather R. Chamberlain, Attila N. Lázár, Kevin Tschirhart, Cyrus Sinai, Nicole A. Hoff, Trevon Fuller, Kamy Musene, Arly Batumbo, Anne W. Rimoin and Andrew J. Tatem. 2022. "High-Resolution Population Estimation Using Household Survey Data and Building Footprints." *Nature Communications* 13(1):1330.

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.

Coyle, Jeremy, Nima Hejazi, Ivana Malenica, Oleg Sofrygin and Rachael Phillips. 2021. "Sl3: Modern Super Learning with Pipelines." Zenodo.

Czernich, Nina, Oliver Falck, Tobias Kretschmer and Ludger Woessmann. 2011. "Broadband Infrastructure and Economic Growth." *The Economic Journal* 121(552):505–532.

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.

Dong, Lei, Carlo Ratti and Siqi Zheng. 2019. "Predicting Neighborhoods' Socioeconomic Attributes Using Restaurant Data." *Proceedings of the National Academy of Sciences* 116(31):15447–15452.

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.

Garcia, David, Yonas Mitike Kassa, Angel Cuevas, Manuel Cebrian, Esteban Moro, Iyad Rahwan and Ruben Cuevas. 2018. "Analyzing gender inequality through large-scale Facebook advertising data." *Proceedings of the National Academy of Sciences* 115(27):6958–6963.

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

IEAG. 2014. "A World That Counts: Mobilising the Data Revolution for Sustainable Development." *Report prepared for the UN Secretary General by the Independent Expert Advisory Group on a Data Revolution for Sustainable Development* .

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

Jayachandran, Seema. 2015. "The Roots of Gender Inequality in Developing Countries." *Annual Review of Economics* 7(Volume 7, 2015):63–88.

Kapoor, Sayash, Emily Cantrell, Kenny Peng, Thanh Hien Pham, Christopher A. Bail, Odd Erik Gundersen, Jake M. Hofman, Jessica Hullman, Michael A. Lones, Momin M. Malik, Priyanka Nanayakkara, Russell A. Poldrack, Inioluwa Deborah Raji, Michael Roberts, Matthew J. Salganik, Marta Serra-Garcia, Brandon M. Stewart, Gilles Vandewiele and Arvind Narayanan. 2023. "REFORMS: Reporting Standards for Machine Learning Based Science.".

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.

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.

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.

Lundberg, Ian, Jennie E. Brand and Nanum Jeon. 2022. "Researcher Reasoning Meets Computational Capacity: Machine Learning for Social Science." *Social Science Research* 108:102807.

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.

Mirza, M. Usman, Chi Xu, Bas Van Bavel, Egbert H. Van Nes and Marten Scheffer. 2021. "Global Inequality Remotely Sensed." *Proceedings of the National Academy of Sciences* 118(18):e1919913118.

Nkoumou Ngoa, Gaston Brice and Jacques Simon Song. 2021. "Female Participation in African Labor Markets: The Role of Information and Communication Technologies." *Telecommunications Policy* 45(9):102174.

Permanyer, Iñaki and Jeroen Smits. 2020. "Inequality in Human Development across the Globe." *Population and Development Review* 46(3):583–601.

Phillips, Rachael V, Mark J van der Laan, Hana Lee and Susan Gruber. 2023. "Practical Considerations for Specifying a Super Learner." *International Journal of Epidemiology* 52(4):1276–1285.

Pokhriyal, Neeti and Damien Christophe Jacques. 2017. "Combining Disparate Data Sources for Improved Poverty Prediction and Mapping." *Proceedings of the National Academy of Sciences* 114(46).

Program, DHS. 2023. "Guide to DHS Statistics - DHS-8.".

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.

UN. 2016. Global Sustainable Development Report. Technical report Technical report, Division of Economic and Social Affairs.: .

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.

Acknowledgements

For helpful discussions and feedback, we thank Dennis Feehan, participants of the PAA 2024 "Digital Demography" session, the EPC 2024 "Modelling Subnational and Spatial Variation" session, the Digital Gender Gaps workshop, the Johns Hopkins Population Health seminar series, and the Leverhulme Centre for Demographic Science seminar series.

Funding

The authors acknowledge financial support from the Bill and Melinda Gates Foundation (INV-045370) and Leverhulme Trust (Grant RC-2018-003) for the Leverhulme Centre for Demographic Science.

Author contributions

- Conceptualization: IW, RK
- Methodology: CB, MF, IW, RK
- Data Preparation: CB, JY, MF, XZ, DL
- Analysis: CB, MF, RK
- Supervision: CB, RK
- Writing – original draft: CB, RK
- Writing – review and editing: CB, DL, IW, RK
- Funding: RK

Competing interests

The authors declare they have no competing interests.

Data and Materials Availability

All data and code needed to evaluate the conclusions in the paper, and our estimates of digital adoption, are available on a dedicated replication repository: https://doi.org/10.17605/OSF.IO/5E8WF.

Supplementary Information

Table of Contents

A Feature and Ground Truth Construction **24**

A.1 Constructing Facebook Features 24

A.2 DHS Processing 25

A.3 Handling Missing Values 26

B Sustainable development goals **30**

C Modeling **30**

C.1 Superlearner Weights 30

C.2 Quantifying Uncertainty 31

C.3 Bootstrap Analysis of Uncertainty in Ground Truth 32

C.4 Decomposing Within and Between Country Variance 33

D Additional Results **35**

E REFORMS Checklist **48**

E.1 Study Goals 48

E.2 Computational Reproducibility 48

E.3 Data Quality 49

E.4 Data Pre-processing 50

E.5 Modeling 51

E.6 Data Leakage 53

E.7 Metrics and Uncertainty 53

E.8 Generalizability and Limitations 54

A Feature and Ground Truth Construction

To construct our features and ground truth measures of digital adoption, we used several different data sources. A full list of features and ground truth outcomes is shown in Table A1. An overview of the processing steps taken to create different features is shown in Table A2.

A.1 Constructing Facebook Features

We collected Facebook monthly active user (MAU) counts using the Facebook (Meta) Marketing API. This API, which is publicly accessible, is designed to provide advertisers with tools to help them advertise across Meta platforms. We used the "Ad Account Delivery Estimate" endpoint, which allows users to query MAU counts in a given location with specific attributes (e.g., gender, age, access device type, etc.). Importantly, the Marketing API only provides current MAU counts and cannot be used to query MAU counts from the past.

To systematically query the Marketing API, we used an adapted version of the publicly available pysocialwatcher package (Araujo et al., 2017). This software is a wrapper that automates repeated queries to the Marketing API based on a user-provided set of locations and attributes. To construct the Facebook features for our models, we collected MAU counts for each admin-1 region by gender (female and male) and access type (iOS, Wi-Fi, and 4G+ mobile network). We collected these data with a daily temporal resolution beginning in February 2024. The subnational features used in this paper are based on averaged MAU counts (obtained daily) from April 25th, 2024 to May 14th, 2024.

We used these subnational MAU counts to construct several different features. First, we calculated overall subnational Facebook penetration features by gender by combining averaged MAU counts with current population estimates from Worldpop. Second, we used these penetration features to calculate the female to male ratio in Facebook penetration. Finally, we calculated the subnational gender-specific fraction of overall users who access Facebook through different means (e.g., fraction of women who accessed Facebook through

Wi-Fi).

In addition to subnational Facebook features, we constructed national-level Facebook penetration metrics using MAU counts from 2019 to 2024. Given the stable, positive trend in national-level MAU counts in most countries, we fit a country-level generalized linear model (GLM) with a log link and a linear trend to our data from 2019 to 2024. We used these models to impute back to temporarily align our national-level Facebook features with earlier DHS surveys, when necessary. To achieve temporal alignment for subnational features, we rescaled the 2024 subnational MAU counts using an adjustment factor, calculated as the ratio of the national MAU count for the year ground truth was observed to the national MAU counts in 2024. We treat Facebook access method (e.g., WiFi, 4g+, etc.) features as time invariant. We used both the national-level and subnational-level features to provide a more comprehensive, accurate, and validated model that captures both national and regional heterogeneity in adoption.

The full set of Facebook features used in our model is shown in Table A1. We currently have ongoing subnational- and national-level collections. We plan to update the models monthly to incorporate the most current Facebook features to predict contemporary ("now-cast") levels and gaps in digital adoption.

A.2 DHS Processing

We calculated all ground truth estimates of internet and mobile adoption using DHS surveys. We restricted to DHS surveys between 2015 and 2022 that have digital adoption information available for both men and women. For most DHS surveys, the universe of eligible respondents is men and women aged 15–49. However, the 2018 Indonesia DHS survey only included married men aged 25–54. We exclude this survey. In total, we used 31 DHS surveys corresponding to 1,523,349 interviews (Table A3).

In most DHS surveys, the first administrative boundaries generally align with GADM1 boundaries. In cases of disagreement, we used location of GPS clusters to map individuals onto their corresponding GADM1 administrative boundaries.

A.3 Handling Missing Values

We selected a parsimonious set of features to ensure broad coverage across LMICs. For the 500 admin-1 units used to train our machine learning model, there were only 8 missing values, all for the gender development index in Guinea. We imputed these missing values using median feature value. In the full set of 2,158 subnational units, there were no missing values for our key Facebook features. For units with missing education or income indices (approximately 9% of countries), we imputed missing values for a feature using the median overall value.

We do not make predictions for countries with no available Facebook MAU counts are available: American Samoa, China, Cuba, Fiji, French Southern Territories, Kosovo, Marshall Islands, Mayotte, North Korea, Papua New Guinea, Russia, Réunion, Saint Helena, Ascension, and Tristan da Cunha, Seychelles, Sudan, Western Sahara.

Variable Name	Type	Source	Country (N)	Subnational (N)
Perc Used Internet Age 15-49 Wom	Ground truth	DHS	34	603
Perc Owns Mobile Age 15-49 Wom	Ground truth	DHS	36	622
Perc Used Internet Age 15-49 Men	Ground truth	DHS	31	501
Perc Owns Mobile Age 15-49 Men	Ground truth	DHS	32	513
Perc Used Internet Age 15-49 FM Ratio	Ground truth	DHS	31	500
Perc Owns Mobile Age 15-49 FM Ratio	Ground truth	DHS	32	513
NL Mean	Offline	NASA Earth Observations	118	2158
Pop Density	Offline	Worldpop	118	2158
DHS Year	Offline	Worldpop	118	2158
Pop All Age Female	Offline	Worldpop	118	2158
Pop All Age Male	Offline	Worldpop	118	2158
Subnational GDI	Offline	Subnational Dev. Database	103	1998
Subnational HDI Males	Offline	Subnational Dev. Database	103	1998
Subnational HDI Females	Offline	Subnational Dev. Database	103	1998
Educational Index Females	Offline	Subnational Dev. Database	104	2004
Educational Index Males	Offline	Subnational Dev. Database	104	2004
Income Index Females	Offline	Subnational Dev. Database	108	2039
Income Index Males	Offline	Subnational Dev. Database	108	2039
Internet Speed	Offline	Speedtest Global Index	105	2040
Continent	Offline	Constructed	118	2158
HDI National	Offline	Subnational Dev. Database	118	2158
GDI National	Offline	Subnational Dev. Database	110	2095
FB Penetration 13+ Male 2024	Facebook	FB Marketing API	118	2158
FB Penetration 13+ Female 2024	Facebook	FB Marketing API	118	2158
FB Age 13+ Gender Gap 2024	Facebook	FB Marketing API	118	2158
iOS Age 13+ Female Fraction 2024	Facebook	FB Marketing API	118	2158
iOS Age 13+ Male Fraction 2024	Facebook	FB Marketing API	118	2158
WiFi Age 13+ Female Fraction 2024	Facebook	FB Marketing API	118	2158
WiFi Age 13+ Male Fraction 2024	Facebook	FB Marketing API	118	2158
4G+ Age 13+ Female Fraction 2024	Facebook	FB Marketing API	118	2158
4G+ Age 13+ Male Fraction 2024	Facebook	FB Marketing API	118	2158
FB Penetration 13+ Male 2024 (National)	Facebook	FB Marketing API	117	2145
FB Penetration 13+ Female 2024 (National)	Facebook	FB Marketing API	117	2145
FB Age 13+ Gender Gap (National)	Facebook	FB Marketing API	117	2145

Table A1: List of feature and ground truth measures used in the analysis and their source.

Data	**Data source**	**Processing Steps**
Ground truth estimates of digital adoption	DHS Program	1. Match the cluster of DHS survey units and GADM1 units. 2. Calculate population-weighted measures of the mobile phone and internet features at the GADM1-level.
Population estimates by sex and age	WorldPop	1. Estimate population for each sex and age group in DHS year. 2. Aggregate the estimated population for the target age group (15-49). 3. Calculate the population density.
Development indicators	Global Data Lab	1. Match the GDL units and GADM1 (Exact name matching; Fuzzy name matching; Geo-matching). 2. Weight GDL variables in each GADM1 based on population.
Nightlight data	Earth Observation Group	1. Resample the population raster data. 2. Weight the nightlight in each GADM1 using the estimated population.
Facebook monthly active user counts	Facebook marketing API	1. Resample the population raster data. 2. Weight the nightlight in each GADM1 using the estimated population.

Table A2: Overview of data processing steps.

Country	Country Code	Start Year	End Year	# Women	# Men
Angola	AO	2015	2016	14379	5684
Armenia	AM	2015	2016	6116	2755
Benin	BJ	2017	2018	15928	7595
Burkina Faso	BF	2021	2021	17659	7720
Burundi	BU	2016	2017	17269	7552
Cambodia	KH	2021	2022	19496	8825
Cameroon	CM	2018	2019	14677	6978
Ethiopia	ET	2018	2018	15683	12688
Gabon	GA	2019	2021	11043	6894
Gambia	GM	2019	2020	11865	4636
Guinea	GN	2018	2018	10874	4117
Haiti	HT	2016	2017	15513	9795
India	IA	2019	2021	724115	101839
Kenya	KE	2022	2022	32156	14453
Liberia	LB	2019	2020	8065	4249
Madagascar	MD	2021	2021	18869	9037
Malawi	MW	2015	2016	24562	7478
Mali	ML	2018	2018	10519	4618
Mauritania	MR	2019	2021	15714	5673
Nepal	NP	2078	2079	14845	4913
Nigeria	NG	2018	2018	41821	13311
Pakistan	PK	2017	2018	15068	3691
Rwanda	RW	2019	2020	14634	6513
Senegal	SN	2019	2019	8649	3365
Sierra Leone	SL	2019	2019	15574	7197
South Africa	ZA	2016	2016	8514	3618
Tanzania	TZ	2015	2016	13266	3514
Timor-Leste	TL	2016	2016	12607	4622
Uganda	UG	2016	2016	18506	5336
Zambia	ZM	2018	2019	13683	12132
Zimbabwe	ZW	2015	2015	9955	8396
Total				1,206,822	316,527

Table A3: The 31 DHS surveys used to construct ground truth estimates of digital adoption.

B Sustainable development goals

The United Nations sustainable development goals (SDG) are a set of 17 goals and targets to guide international development policy. Originally established in 2015, the SDGs were established to reduce poverty, hunger, AIDS, and discrimination against women and girls (UN, 2016). Each sustainable development goal has a set of accompanying targets and indicators for tracking development. The estimates produced in this study most directly contribute to monitoring and tracking the following indicators:

- Goal 5: Achieve gender equality and empower all women and girls
  - 5.b.1 Proportion of individuals who own a mobile telephone, by sex
- Goal 4: Ensure inclusive and equitable quality education and promote lifelong learning opportunities for all
  - 4.4.1 Proportion of youth and adults with information and communications technology (ICT) skills, by type of skill
- Goal 17: Strengthen the means of implementation and revitalize the Global Partnership for Sustainable Development
  - 17.8.1 Proportion of individuals using the Internet

A full list of sustainable development goals are available on the UN Website.

C Modeling

C.1 Superlearner Weights

The set of machine learning algorithms included in our ensemble Superlearner models and their respective weights are presented in Table A4. Each algorithm's weight represents its contribution to the overall ensemble Superlearner predictions. These weights are estimated using non-negative least squares regression; for more details on practical considerations of implementing a Superlearner, see Phillips et al. (2023).

The algorithms that are most heavily weighted in our final ensemble models are random forest and gradient boosting machines (GBM). This suggests that tree-based models, which are well-suited for handling non-linear relationships, interactions, and heterogeneity in the

data, are the best performing class of models for these prediction tasks.

		Internet			Mobile		
Algorithm	Description	Women	Men	Ratio	Women	Men	Ratio
glm	Generalized Linear Model	0.00	0.06	0.17	0.06	0.31	0.19
lasso	Lasso Regression	0.23	0.27	0.00	0.00	0.00	0.00
ridge	Ridge Regression	0.19	0.00	0.11	0.21	0.00	0.00
elastic_new	Elastic Net with 50% L1 Ratio	0.00	0.00	0.00	0.00	0.06	0.00
poly_spline	Polynomial Spline	0.10	0.09	0.10	0.00	0.22	0.11
random forest	Random Forest with 100 Trees	0.54	0.22	0.23	0.52	0.05	0.26
gbm	Gradient Boosted Machine	0.00	0.04	0.38	0.26	0.37	0.00
xgb	Extreme Gradient Boosting	0.00	0.33	0.00	0.00	0.00	0.44
SuperLearner	Ensemble Model	–	–	–	–	–	–

Table A4: Full set of machine learning algorithms used in the Superlearner and their relative contribution to the final Superlearner model.

C.2 Quantifying Uncertainty

To estimate uncertainty for each subnational unit, we regressed the absolute residual against a set of all observable variables for each subnational unit with ground truth. We used a non-negative least squares regression to ensure the resulting estimates are non-negative. This model was used to predict the absolute residual size for all subnational units. Qualitatively, our predicted error is largest in high-adoption settings and smallest in low-adoption settings (Figure A6).

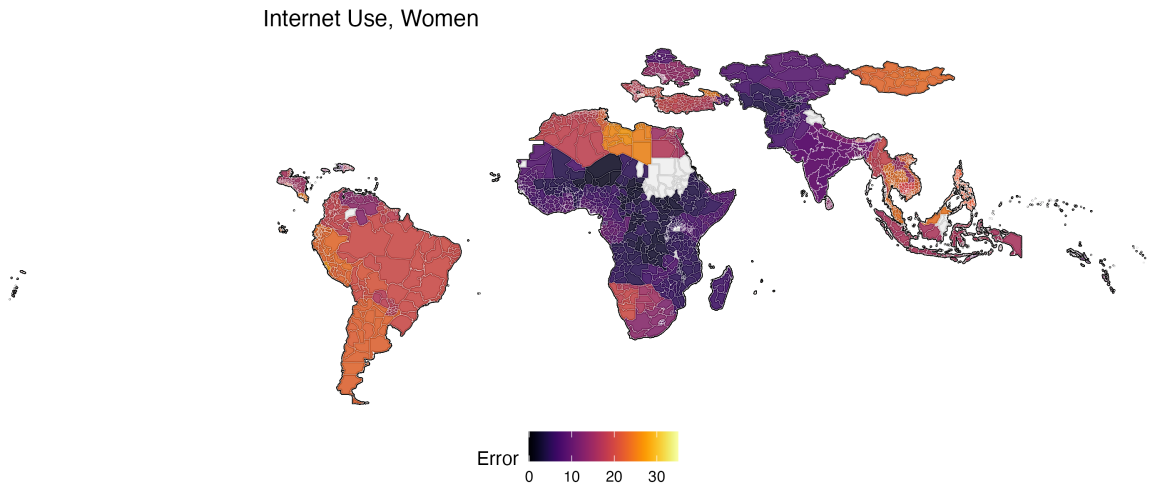


Figure A6: The predicted residual error in each subnational region

C.3 Bootstrap Analysis of Uncertainty in Ground Truth

Our underlying ground truth data, which come from DHS surveys, are subject to sampling error, especially at the admin-1 level. To quantify the impact of this sampling uncertainty on our analysis, we conduct a bootstrap analysis. Specifically, we repeatedly resample individual survey respondents with replacement within subnational units to create 10,000 synthetic bootstrap samples. We then compare our original estimates of digital adoption to the estimates from our 10,000 bootstrapped samples.

This process gives insight into how much uncertainty there is in our underlying estimates of digital adoption. The results shown in in Table A5 give insight into the upper bound of our estimated R^2 values. Notably, the R^2 value is lower for estimating gender gaps than overall levels of adoption.

Category	Q10	Q50	Q90	Mean
mobile (women)	0.929	0.981	0.995	0.966
mobile (men)	0.785	0.925	0.980	0.884
mobile (ratio)	0.683	0.903	0.981	0.843
internet (women)	0.958	0.987	0.996	0.981
internet (men)	0.914	0.974	0.992	0.962
internet (ratio)	0.392	0.837	0.958	0.712

Table A5: Percent of total variation explained (R^2) from 10,000 bootstrap resamples

C.4 Decomposing Within and Between Country Variance

To assess the relative contribution of within-country and between country variation to overall inequality, we used mean log deviation. Mean log deviation has appealing decomposability properties, and has been used elsewhere to quantify the additional inequality revealed at the subnational level (Permanyer and Smits, 2020). The mean log deviation (MLD) for all subnational units is defined as:

$$\text{MLD} = \frac{1}{N} \sum_{i=1}^{N} \log \left(\frac{\mu}{r_i} \right), \tag{4}$$

where N is the number of subnational units, μ is the overall mean rate (e.g., mean internet penetration rate), and r_i is the rate for a given subnational unit i. The MLD can be decomposed into within-country and between-country components:

$$\text{MLD} = \text{MLD}_W + \text{MLD}_B. \tag{5}$$

The overall mean rate μ can be expressed as:

$$\mu = \sum_{k=1}^{K} \frac{N_k}{N} \mu_k, \tag{6}$$

where K is the number of countries, N_k is the number of subnational units in country k, and μ_k is the mean rate for country k. The within-country MLD for each country k is:

$$\text{MLD}_k = \frac{1}{N_k} \sum_{i \in k} \log \left(\frac{\mu_k}{r_i} \right). \tag{7}$$

The overall within-country MLD is a weighted sum of the within-country MLDs:

$$\text{MLD}_W = \sum_{k=1}^{K} \frac{N_k}{N} \text{MLD}_k. \tag{8}$$

The between-country MLD is calculated based on the mean rates of the countries:

$$\text{MLD}_B = \sum_{k=1}^{K} \frac{N_k}{N} \log \left(\frac{\mu}{\mu_k} \right). \tag{9}$$

The resulting estimates of MLD_B and MLD_k indicate how much of the total inequality in the rate (e.g., internet penetration rate) is due to within-country disparities versus between-country disparities.

D Additional Results

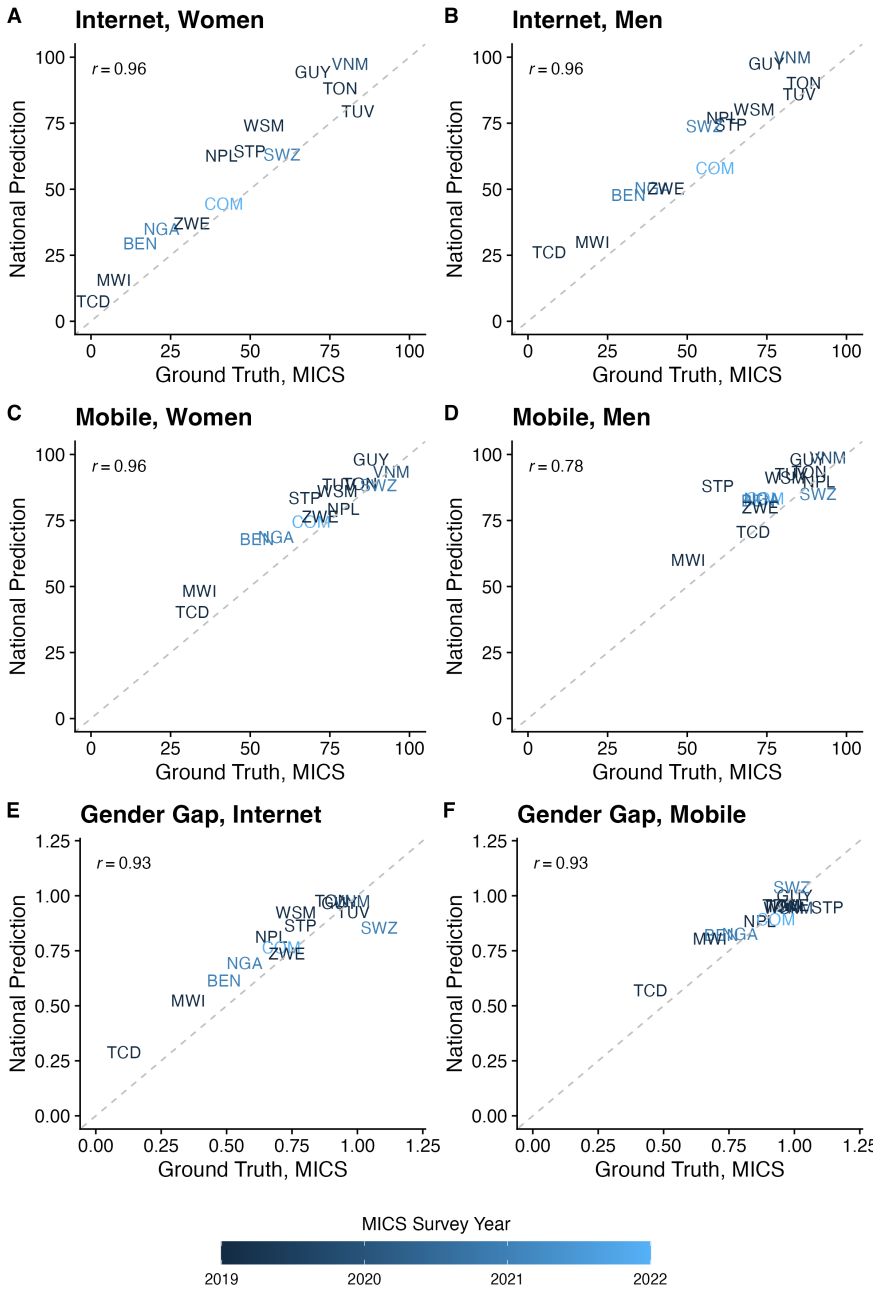


Figure A7: Comparison of national-level estimates with independent ground truth data from Multiple Indicator Cluster Surveys (MICS) provides a robust validation framework. Originally developed by UNICEF, MICS surveys capture key indicators related to development, policy evaluation, and the monitoring of Sustainable Development Goals (SDGs). These surveys are nationally representative, offering valuable independent estimates for comparison. National-level predictions are population-weighted averages of subnational predictions.

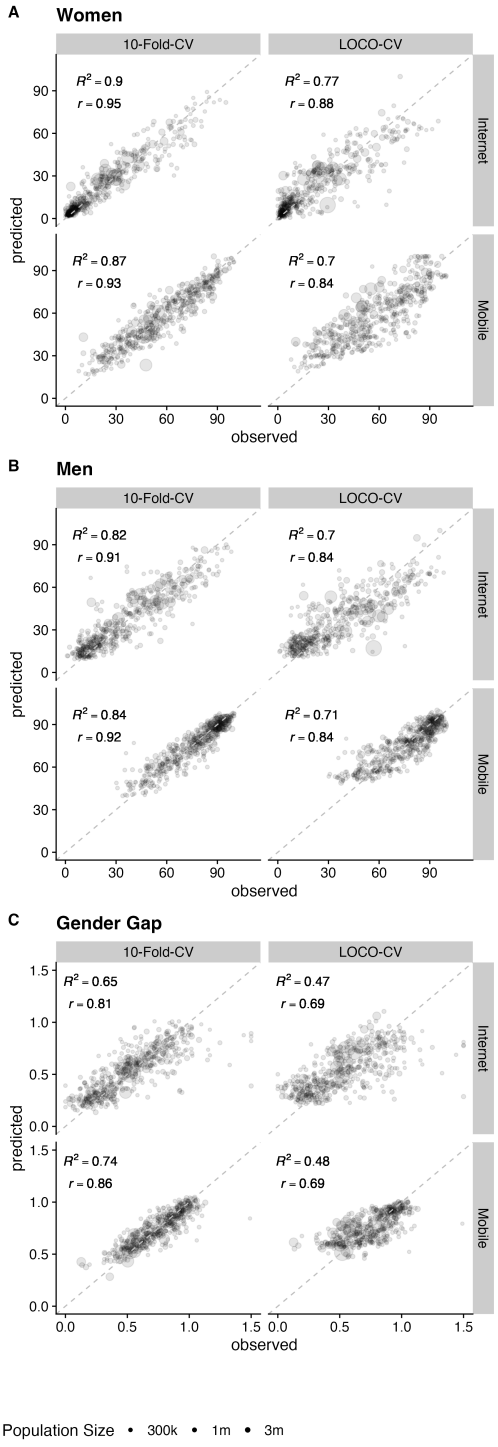


Figure A8: Scatterplot of observed vs. predicted values under 10-fold cross-validation (10-fold CV) and leave-one-country-out cross-validation (LOCO–CV).

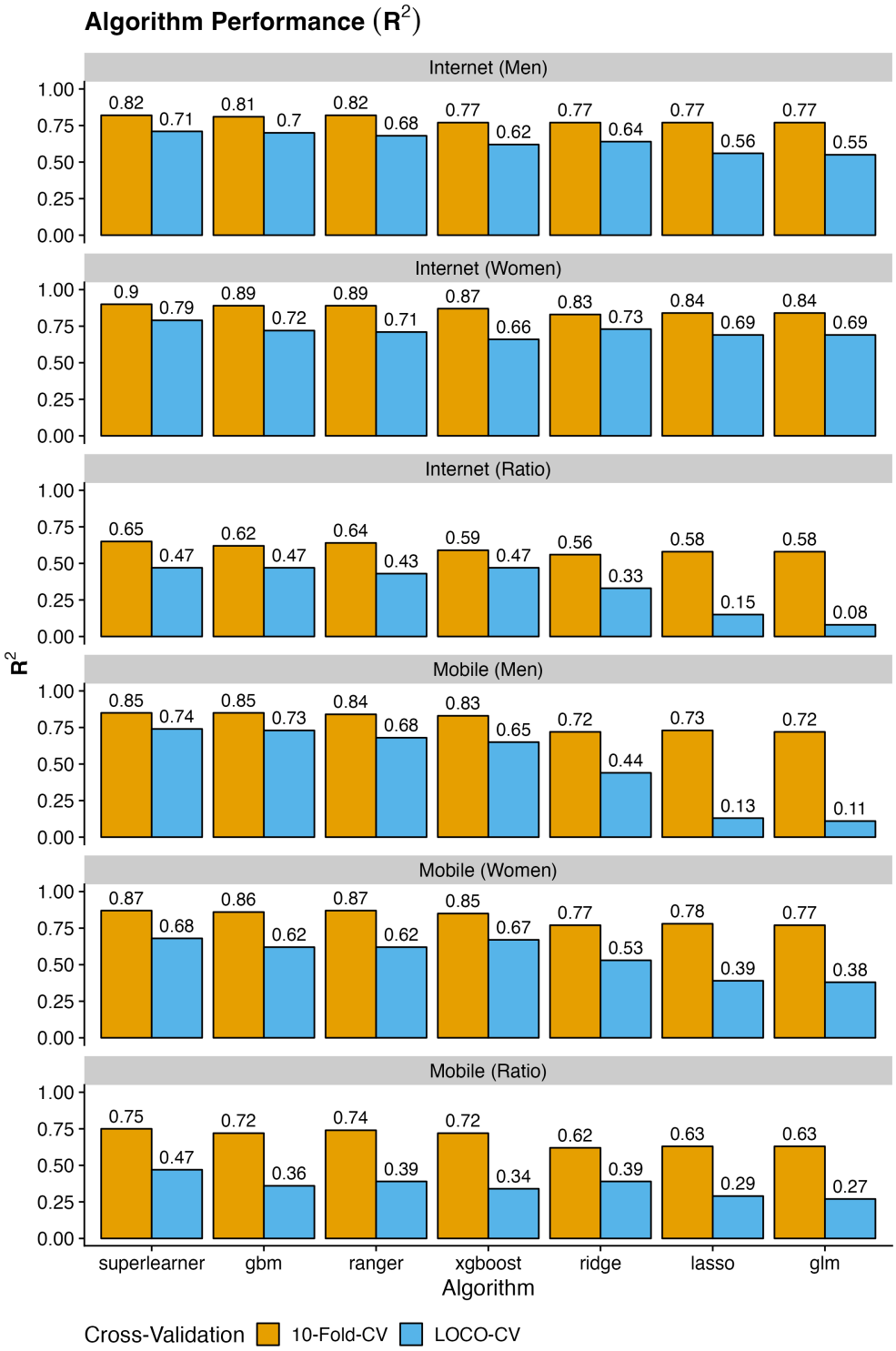


Figure A9: Performance by algorithm, as measured by R^2 (coefficient of determination) under 10-fold cross-validation (10-fold CV) and leave-one-country-out cross-validation (LOCO–CV).

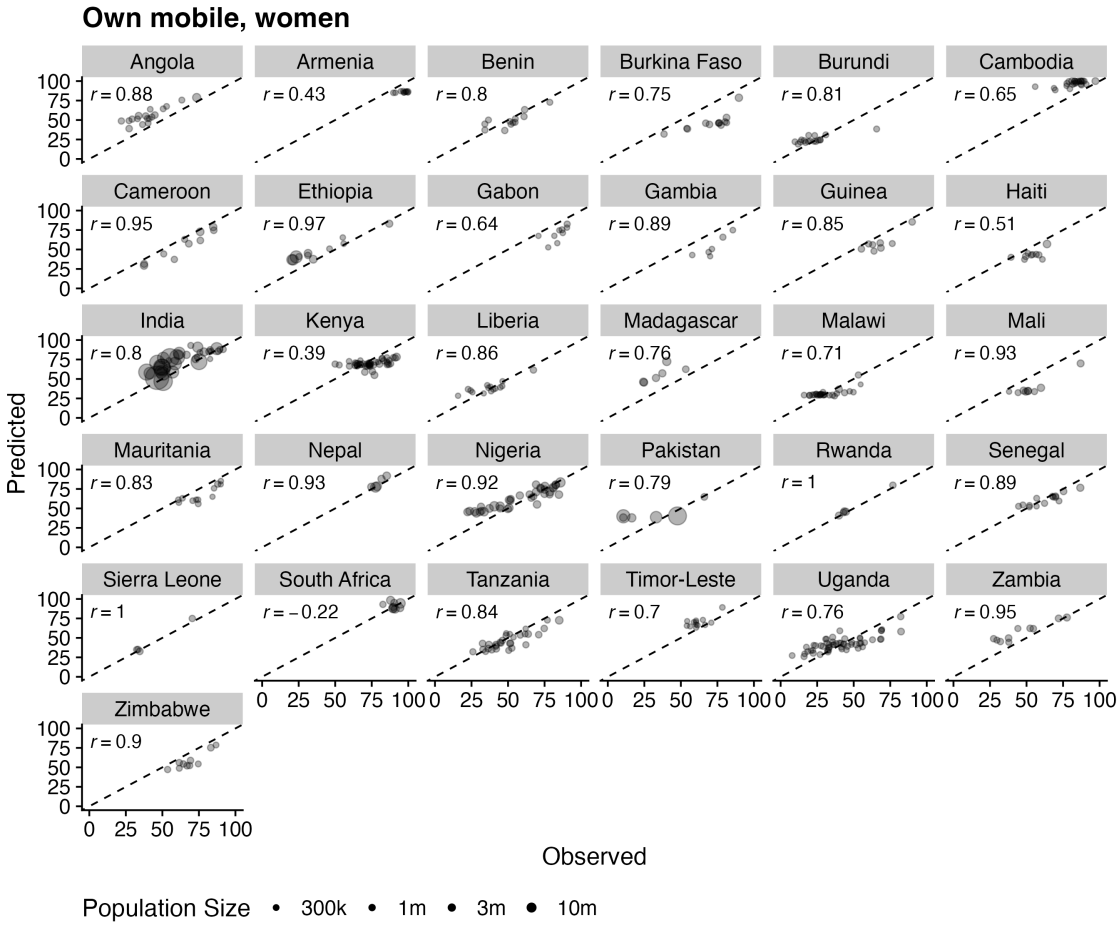


Figure A10: Scatterplot of observed vs. predicted mobile phone ownership for women by country.

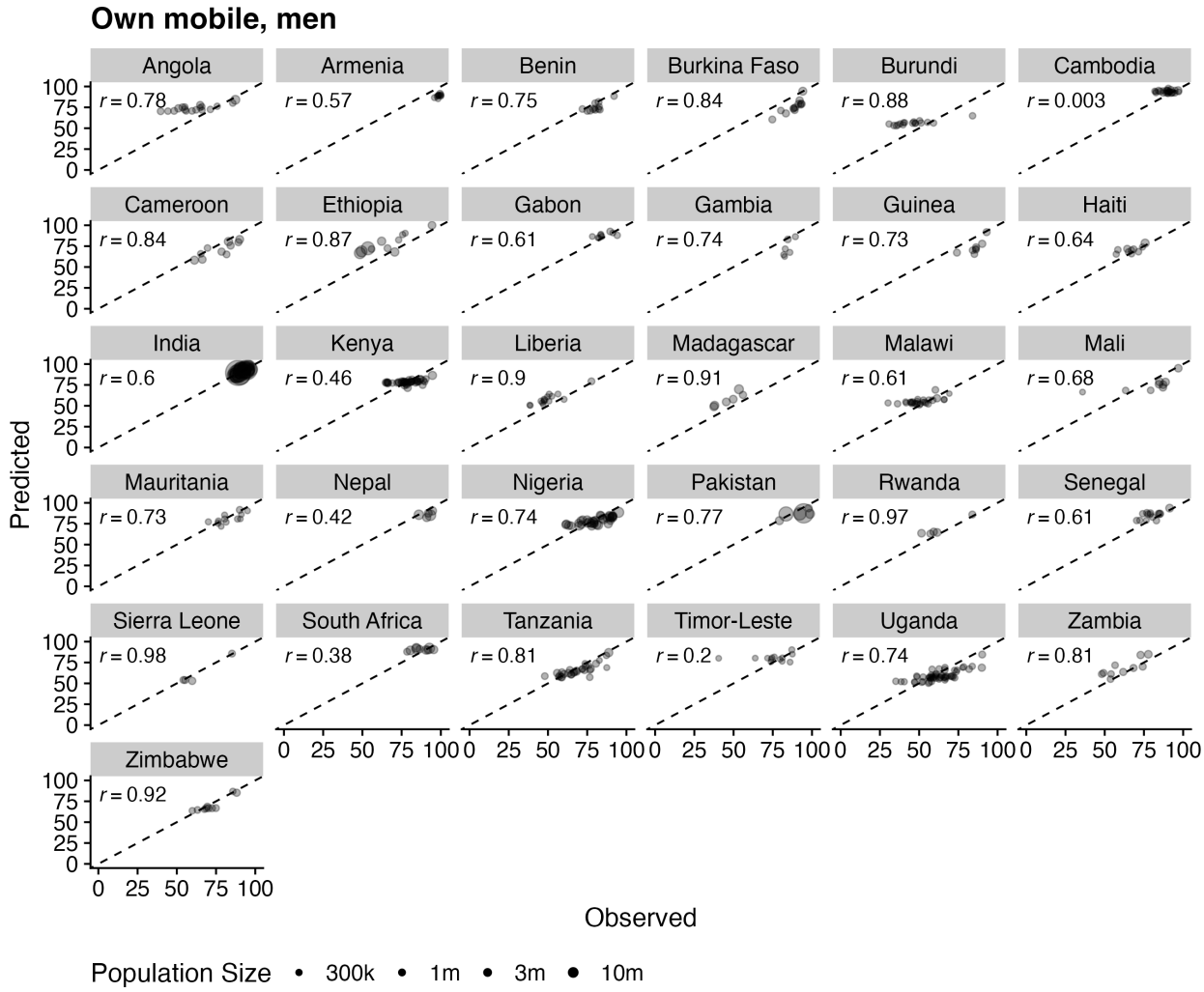


Figure A11: Scatterplot of observed vs. predicted mobile phone ownership for men by country.

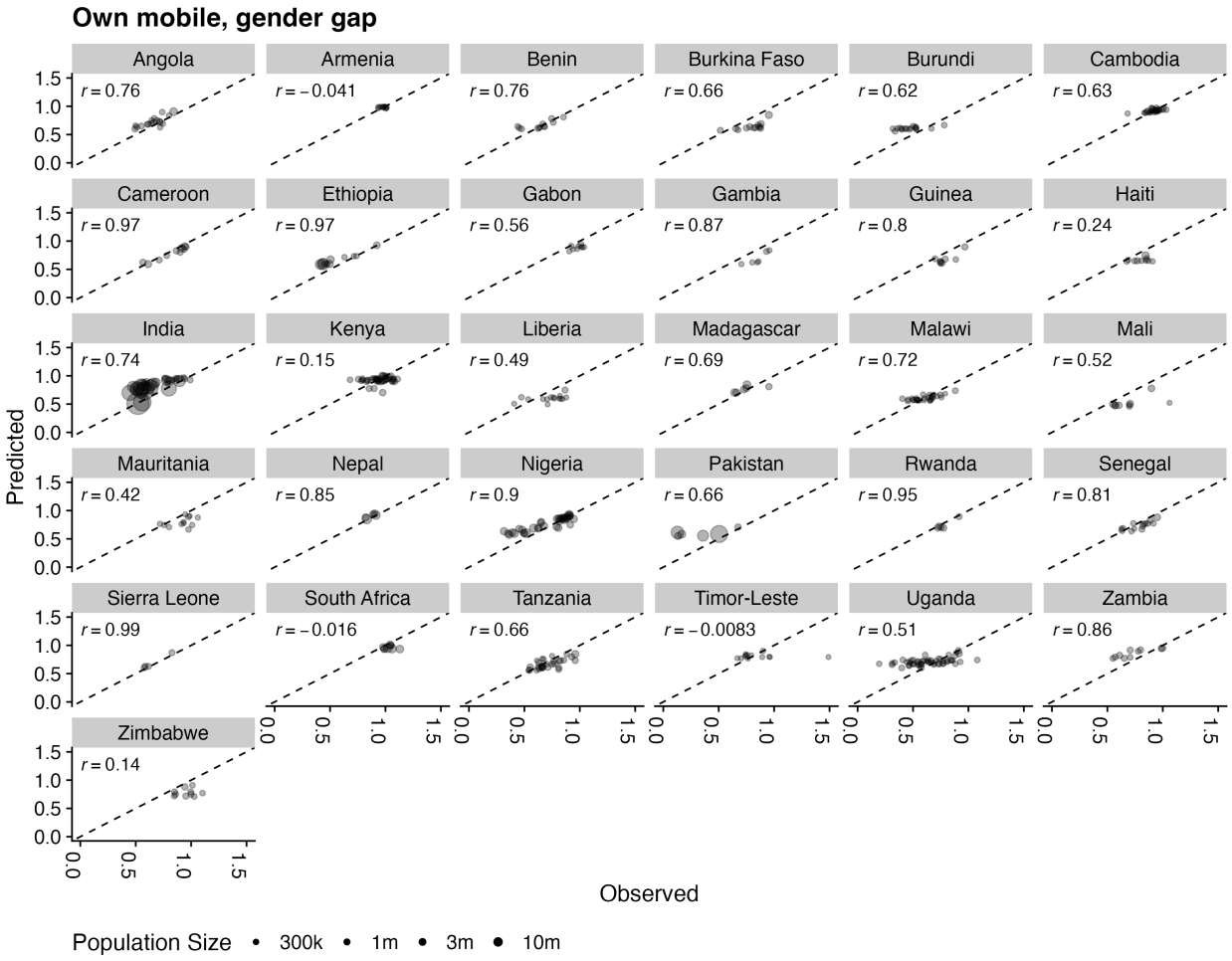


Figure A12: Scatterplot of observed vs. predicted mobile phone ownership gender gaps for women by country.

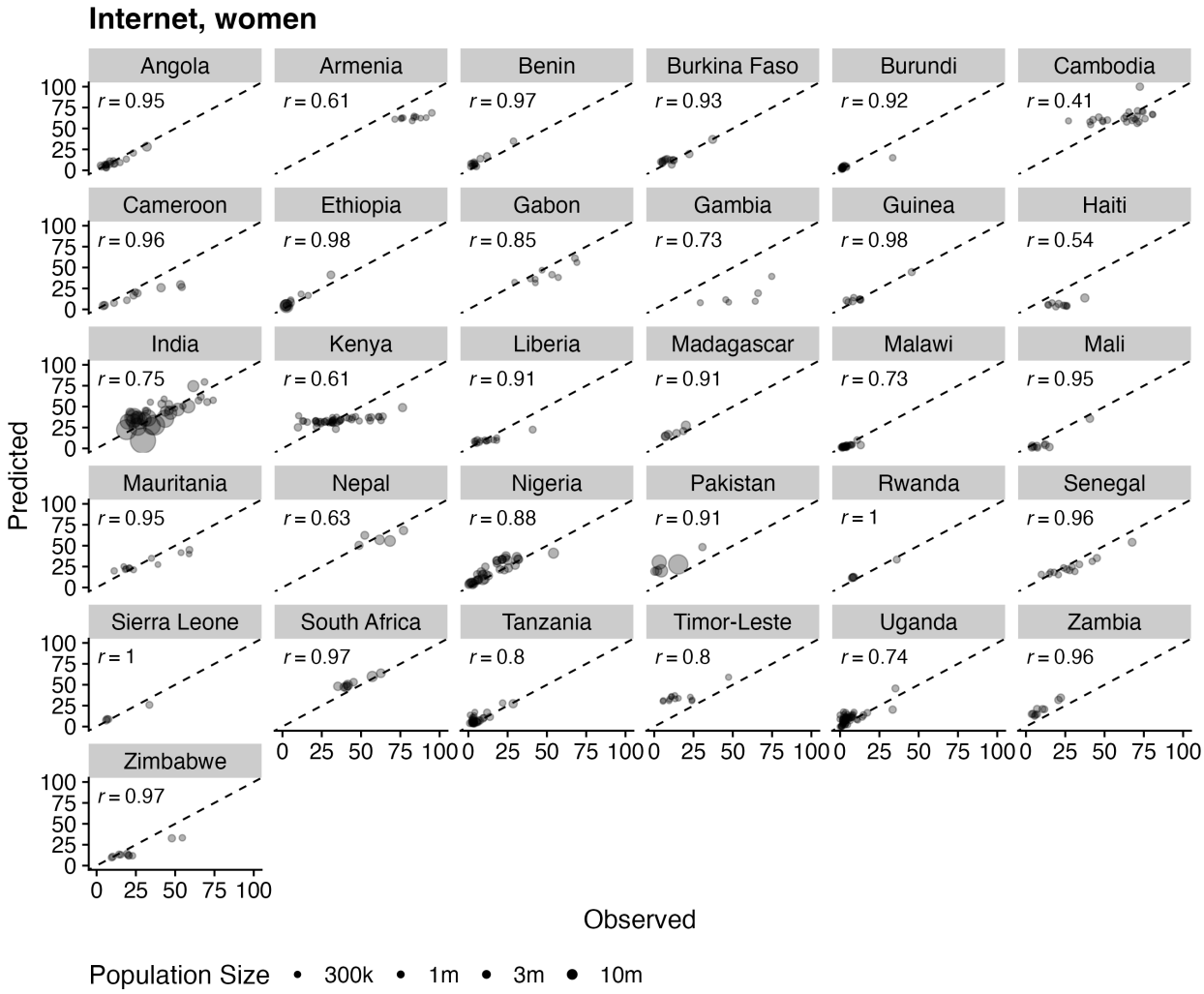


Figure A13: Scatterplot of observed vs. predicted internet adoption for women by country.

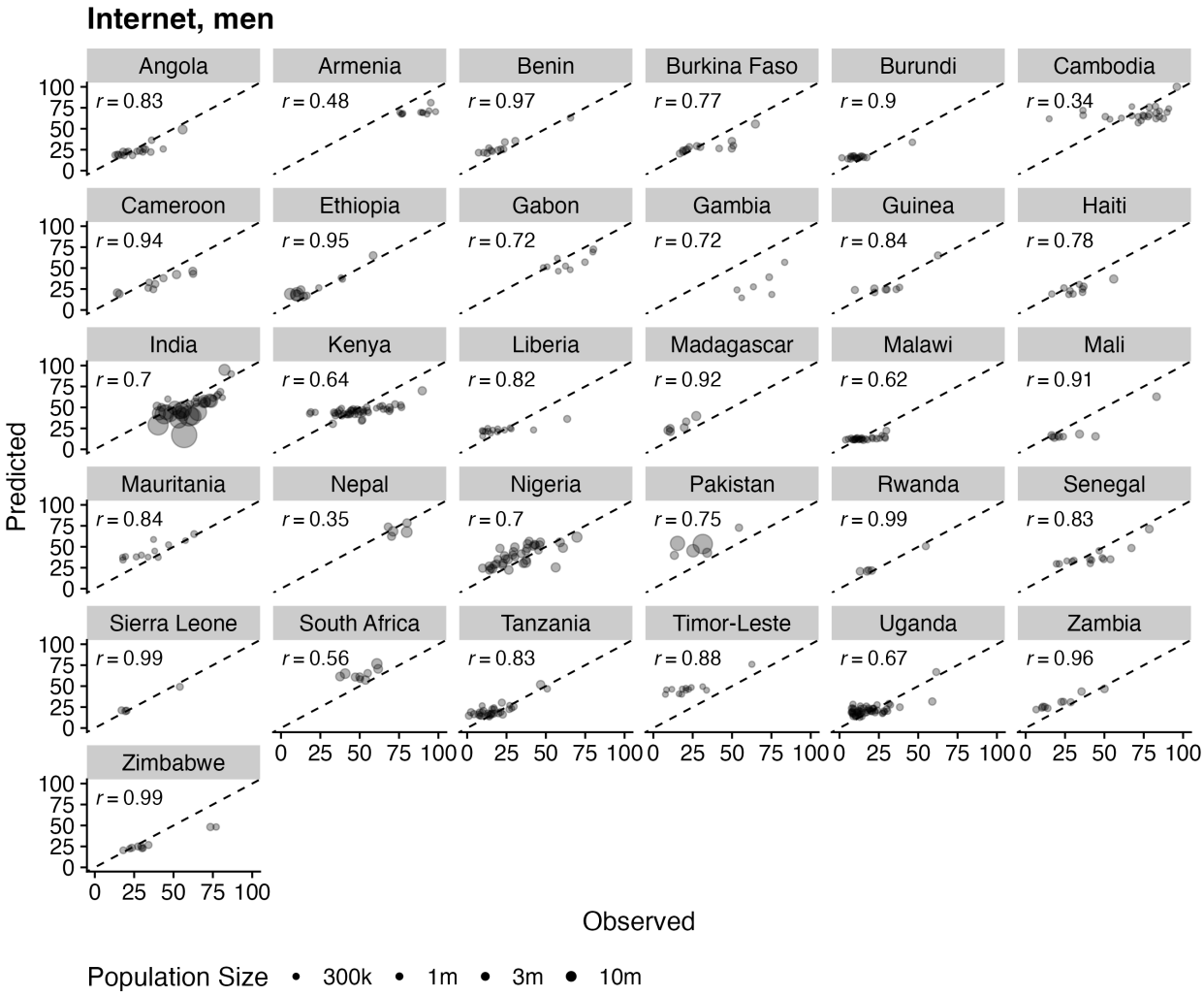


Figure A14: Scatterplot of observed vs. predicted internet adoption for men by country.

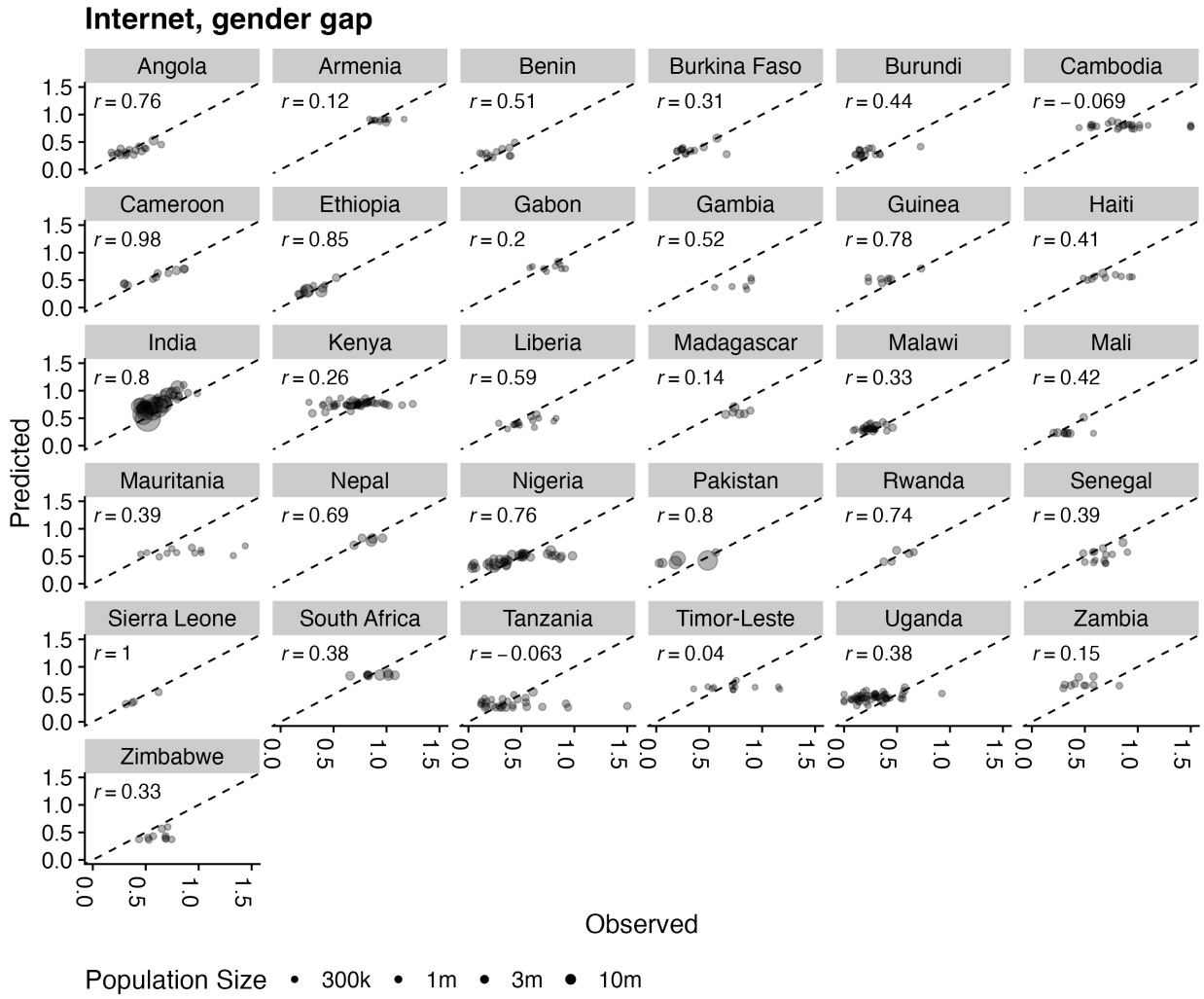


Figure A15: Scatterplot of observed vs. predicted internet adoption for women by country.

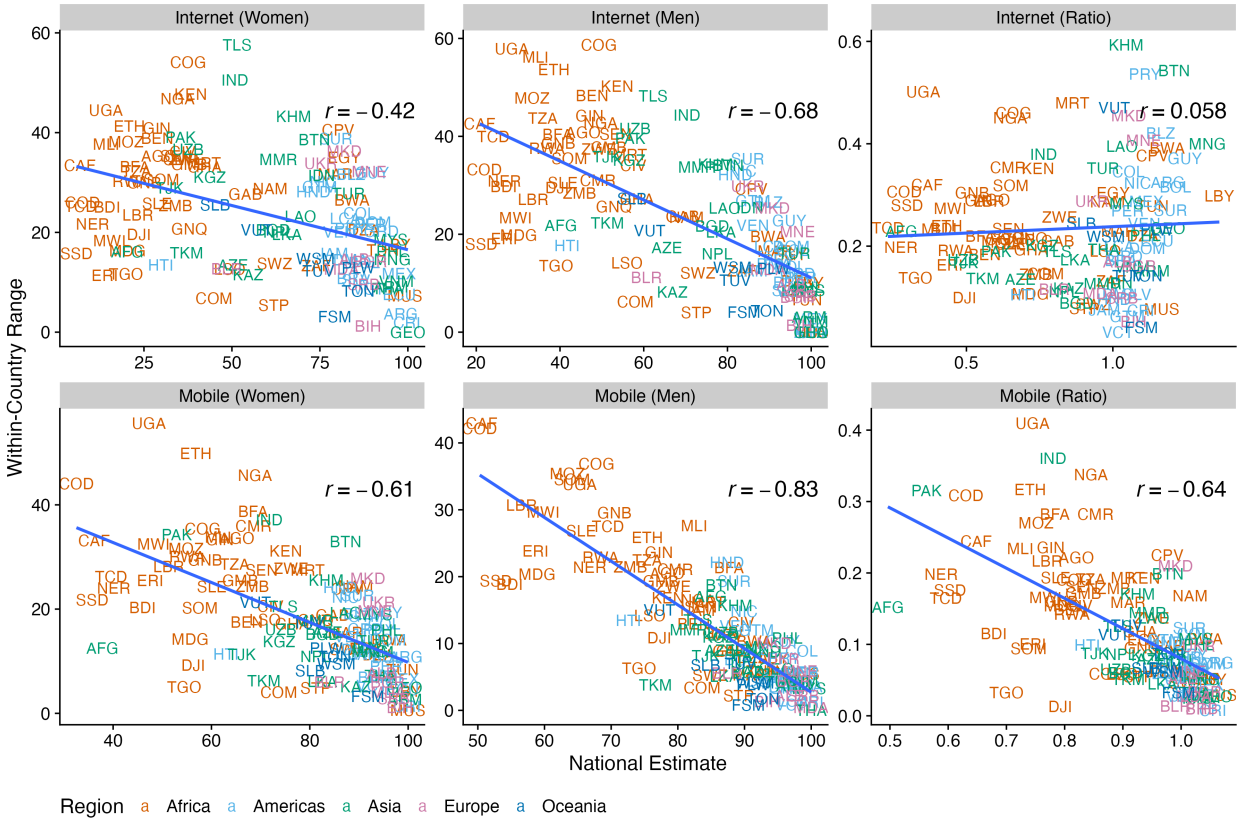


Figure A16: Relationship between national-level estimates and the within-country range between the top and bottom region for our six outcomes of interest. Labels show the ISO-3 country codes.

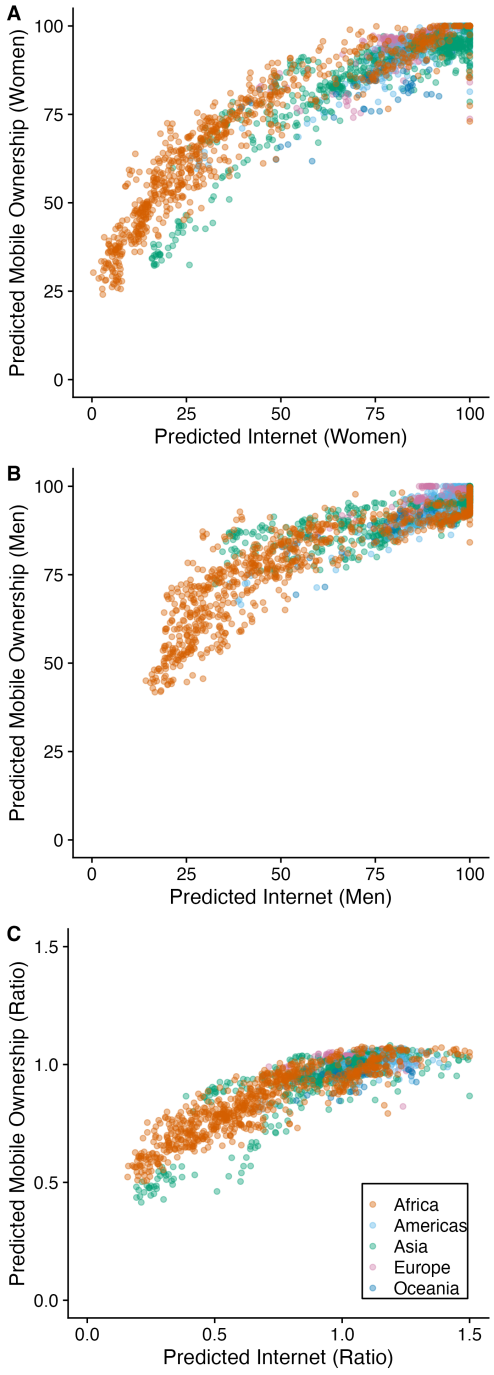


Figure A17: *(A)* Comparison of predicted internet penetration and mobile phone penetration for women. *(B)* Comparison of predicted internet and mobile phone ownership for men. *(C)* Comparison of predicted internet gender gap and predicted mobile gender gap.

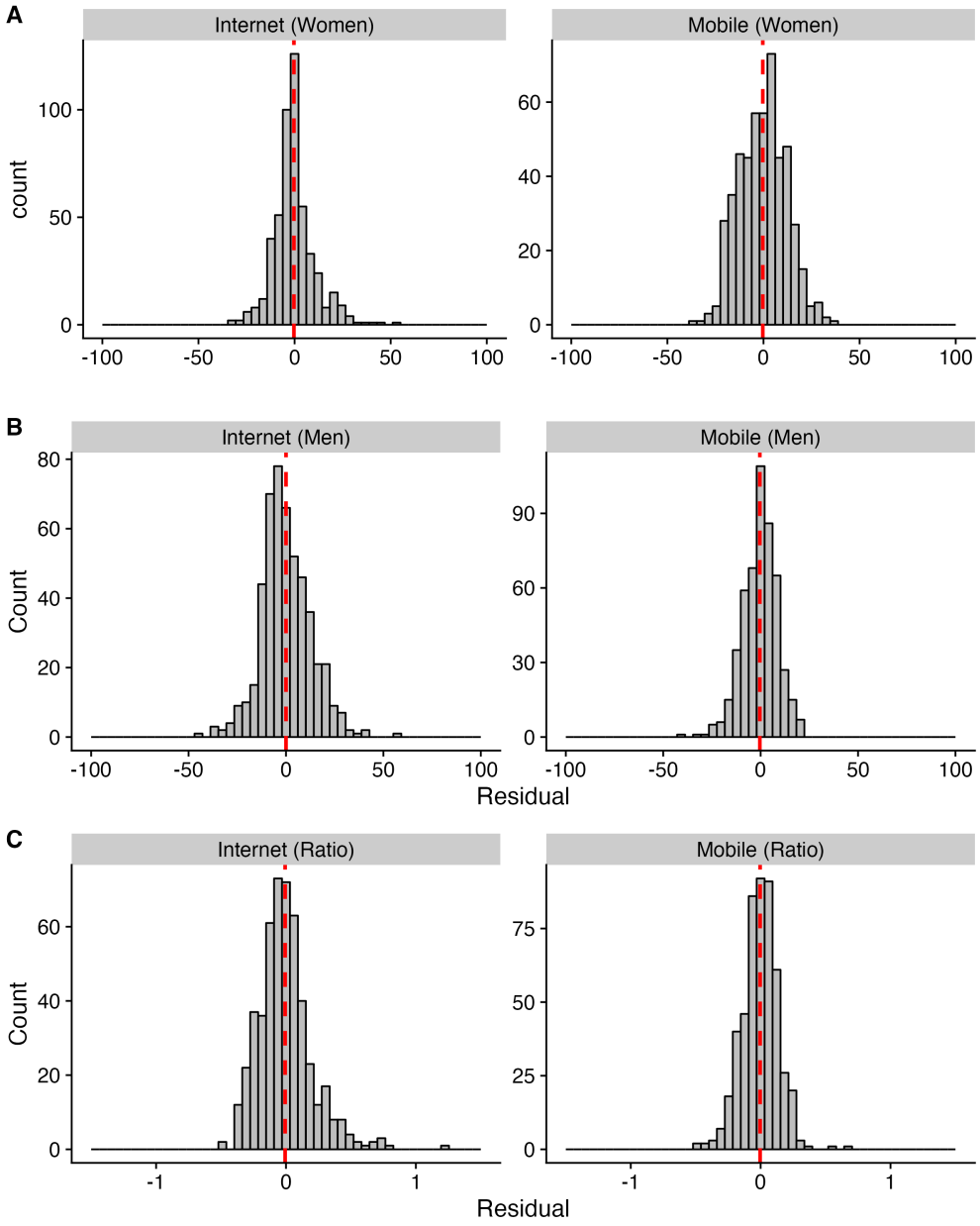


Figure A18: The residual (observed - predicted) from leave-one-country out cross-validation for all subnational regions with available ground truth data. Red dashed line shows mean or residual values.

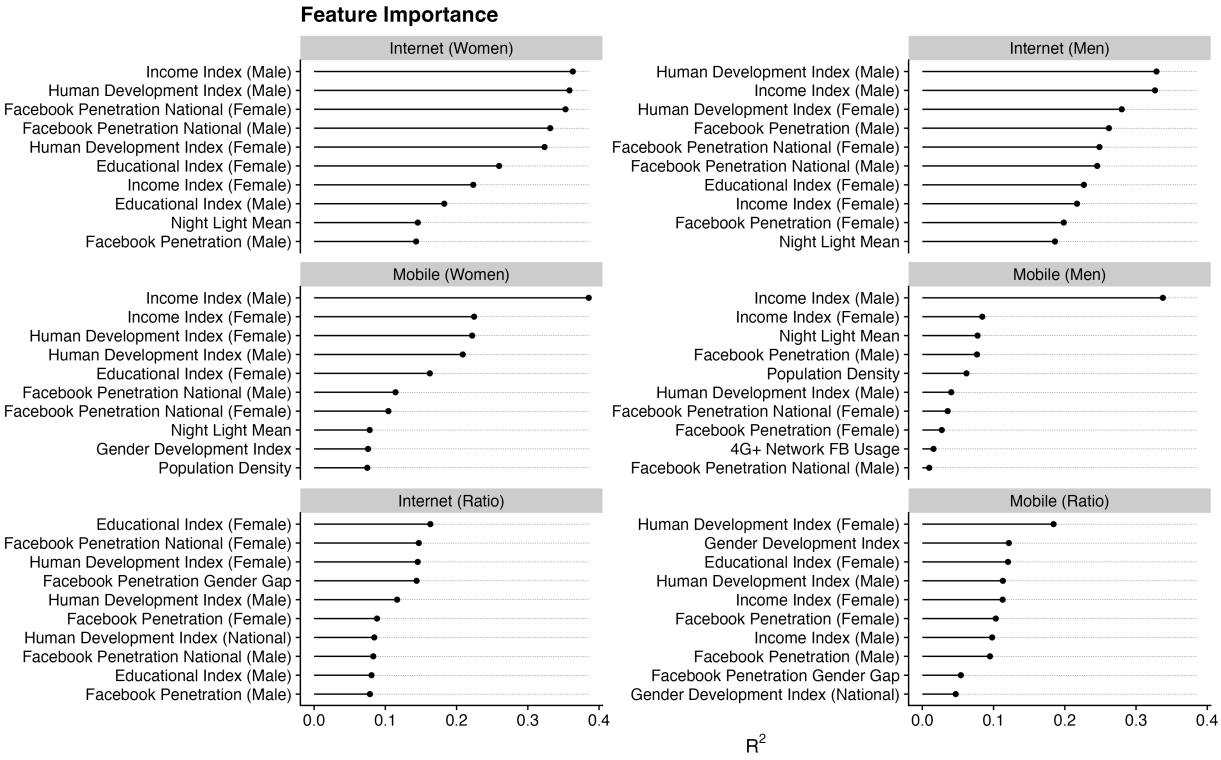


Figure A19: Feature importance for each outcome of interest.

E REFORMS Checklist

The REFORMS checklist, developed collaboratively by 19 researchers across the social, computer, and physical sciences, is a resource for promoting transparency and reproducibility in machine learning science (Kapoor et al., 2023). We complete this checklist to promote methodological transparency.

E.1 Study Goals

1a. *State the population or distribution about which the scientific claim is made.*

This study makes scientific claims about levels and gaps of digital adoption in 2,158 subnational regions in 118 different low- and middle-income countries (LMICs).

1b. *Describe the motivation for choosing this population or distribution.*

We focus on LMICs settings as overall levels of digital penetration are low and gender gaps disfavoring women are high. We focus on the first subnational level as subnational variation represents a substantial component of overall inequality, especially in countries with the lowest levels of development.

1c. *Describe the motivation for the use of ML methods in the study.*

We use machine learning methods to predict levels and gaps of internet adoption and mobile phone ownership by gender. Machine learning is an appropriate choice for this task because our primary focus is on predictive accuracy rather than model interpretation. Additionally, machine learning algorithms are better suited for handling the complex interactions and non-linear relationships among features.

E.2 Computational Reproducibility

2a. *Describe the dataset used for training and evaluating the model and provide a link or DOI to uniquely identify the dataset.*

The dataset used to train and evaluate the model contains columns corresponding to model features or outcomes and rows corresponding to different subnational units. The

dataset is available from:
https://doi.org/10.17605/OSF.IO/5E8WF. For more information on how each feature was constructed, see Appendix A.

2b. *Provide details about the code used to train and evaluate the model and produce the results reported in the paper along with link or DOI to uniquely identify the version of the code used.*

To train our machine learning algorithms, we used an ensemble Superlearner (Phillips et al., 2023). We first separate models for each of our six outcomes separately using the SL3 package in R (Coyle et al., 2021). All code to train these models is available in our replication package:
https://doi.org/10.17605/OSF.IO/5E8WF.

2c. *Describe the computing infrastructure used.*

All computations were carried out on 2023 MacBook Pro with an Apple M2 Pro chip, 16GB memory, and Sonoma 14.1 operating system.

2d. *Provide a README file which contains instructions for generating the results using the provided dataset and code.*

A README file containing instructions for generating all estimates, figures, and tables presented in the paper is available from: https://doi.org/10.17605/OSF.IO/5E8WF.

2e. *Provide a reproduction script to produce all results reported in the paper.*

The replication scripts for generating all figures and code in the paper is available here:
https://doi.org/10.17605/OSF.IO/5E8WF

E.3 Data Quality

3a. *Describe source(s) of data, separately for the training and evaluation datasets, along with the time when the dataset(s) are collected, the source and process of ground-truth annotations, and other data documentation.*

The data for this study come from several different sources. First, our ground truth data came from a series of 31 Demographic and Health Surveys (DHS). The survey datasets were obtained directly from the DHS Program's website on March 1, 2024. To construct our features, we combine data from the following sources: Worldpop, NASA Earth Observation Group, Global Data Lab, and the Facebook Marketing API. For more details on data processing, see Appendix A.

3b. *State the distribution or set from which the dataset is sampled (i.e., the sampling frame).*

The dataset used for our analysis is structured where each observation (row) corresponds to a different subnational unit. Each feature (column) corresponds to a characteristics of each subnational unit. The key rates of digital adoption in the dataset are calculated based on microdata from DHS surveys, which usually use a stratified two-stage cluster design (Program, 2023).

3c. *Justify why the dataset is useful for the modeling task at hand.*

This dataset is useful for modeling subnational levels and gaps of internet and mobile phone adoption as it includes at the admin-1 level (1) ground truth measures of digital adoption and (2) a carefully curated set of predictors associated with digital adoption and gaps.

E.4 Data Pre-processing

4a. *Describe whether any samples are excluded with a rationale for why they are excluded.*

We only make predictions for LMICs where Facebook MAU data is available. No Facebook MAU counts are available for the following countries: American Samoa, China, Cuba, Fiji, French Southern Territories, Kosovo, Marshall Islands, Mayotte, North Korea, Papua New Guinea, Russia, Réunion, Saint Helena, Ascension, and Tristan da Cunha, Seychelles, Sudan, Western Sahara.

4b. *Describe how impossible or corrupt samples are dealt with.*

There are no impossible or corrupt samples.

4c. *Describe all transformations of the dataset from its raw form to the form used in the model, for instance, treatment of missing data and normalization—preferably through a flow chart.*

For most subnational units in our analysis, we have no missing predictors. When data is missing, we impute using the median value. An overview of data processing is available in Appendix A.

E.5 Modeling

5a. *Describe, in detail, all models trained.*

We fit six different ensemble Superlearner models, one for each outcome of interest. Each ensemble Superlearner algorithm combines multiple predictions from a library of individual machine learning algorithms. Specifically, we include the following individual machine learning algorithms in our ensemble library:

- Generalized Linear Model (GLM)
- Lasso Regression
- Ridge Regression
- Elastic Net Regression
- Polynomial Spline Regression
- Random Forests
- Gradient Boosted Machine (GBM)
- Extreme Gradient Boosting (XGB)

5b. *Justify the choice of model types implemented.*

We chose to use an ensemble Superlearner model to enhance predictive accuracy by leveraging the strengths and smoothing over limitations of each individual model. The choice of individual machine learning algorithms was driven by their ability to handle various data characteristics, such as non-linear relationships, interactions, and high-dimensional data. We selected a diverse set of machine learning algorithms for our ensemble library to capture a wide range of data characteristics, including non-linear relationships, interactions, and high-dimensional features.

5c. *Describe the method for evaluating the model(s) reported in the paper, including details of train-test splits or cross-validation folds.* The models were evaluated using two types of cross-validation:

- **10-fold cross-validation**: The data were randomly split into ten folds, with nine folds used for training and one for testing, repeated for each fold. This gives a sense of how the model would perform for countries where we have some subnational ground truth estimates of adoption.
- **Leave-one-country-out cross-validation (LOCO-CV)**: Data from one country were held out at a time, the model was trained on the remaining countries, and predictions were made for the subnational units in the held-out country. This process was repeated for each country, ensuring that no data from the test country influenced the training. This strategy is more conservative, and gives insight into how the model would perform on a country where we have some subnational ground truth estimates of adoption.

5d. *Describe the method for selecting the model(s) reported in the paper.*

We report both the performance of the ensemble Superlearner algorithms and the individual machine learning algorithms in Figure A9. However, in the main body of the text, we only show predictions from our best-performing Superlearner algorithm. The ensemble Superlearner algorithms is a weighted combination of predictions from each individual machine learning algorithm, and has the best overall performance across all indicators. Weights are calculated using a non-negative least squares (NNLS) regression meta-learner and shown in Appendix C.1.

5e. *For the model(s) reported in the paper, specify details about the hyperparameter tuning.*

Hyperparameter tuning was performed ad-hoc for each individual model within the Superlearner framework using cross-validation. Specific tuning processes were applied as follows:

- Random forests: Number of trees
- Gradient boosting machines: Learning rate, number of trees, and tree depth

5f. *Justify that model comparisons are against appropriate baselines.*

Our model baseline is a generalized linear model (GLM), fit within our Superlearner framework. This simple model is a standard and appropriate baseline that allows us to assess the performance of our ensemble superlearner algorithm.

E.6 Data Leakage

6a. *Justify that pre-processing and modeling steps only use information from the training dataset (and not the test dataset).*

Pre-processing and modeling steps were strictly limited to using information from the training dataset. The LOCO-CV approach ensured that no data from the test countries were included in the training process, maintaining strict data separation.

6b. *Describe methods used to address dependencies or duplicates between the training and test datasets.*

Dependencies and duplicates were managed by ensuring that each country's data were treated independently during the LOCO-CV process. This approach inherently avoids any overlap or dependencies between training and test datasets.

6c. *Justify that each feature or input used in the model is legitimate for the task at hand and does not lead to leakage.*

All features used in our models were carefully selected to avoid data leakage. None of the features in our models are proxies for the outcome, and were all measured independently of the outcomes.

E.7 Metrics and Uncertainty

7a. *State all metrics used to assess and compare model performance. Justify that the metric used to select the final model is suitable for the task.*

For our primary model performance metric, we use the coefficient of determination, R^2. This metric is valuable because it quantifies the proportion of variance in the

dependent variable explained by the model, providing an absolute measure of model fit. Additionally, we use the mean absolute error (MAE) to assess the average magnitude of prediction errors, offering a direct interpretation of prediction accuracy in the units of the dependent variable. Finally, we use the Pearson correlation coefficient (r), which measures the linear correlation between observed and predicted values, showing how well the predicted values follow the trend of the observed ground truth data. These metrics collectively ensure a comprehensive evaluation of the model's performance, capturing both fit quality and prediction accuracy.

7b. *State uncertainty estimates and give details of how these are calculated.*

To estimate uncertainty for each subnational unit, we regress the absolute residuals against all observable variables for units with ground truth data. We use non-negative least squares regression to ensure estimated absolute residuals are positive. We then use this model to predict the absolute residual size for all subnational units.

7c. *Justify the choice of statistical tests (if used) and a check for the assumptions of the statistical test.*

We do not conduct any statistical tests.

E.8 Generalizability and Limitations

8a. *Describe evidence of external validity.*

Our models are calibrated using ground truth data from DHS surveys. It is possible that countries that do not have DHS surveys differ in important ways from the countries that do have DHS surveys. Given the availability of ground truth coverage from 31 different countries, this is unlikely to drastically affect our external validity. Further, a comparison with an independent ground truth estimates of digital adoption from MICs surveys revealed strong and robust correlations in countries with no DHS surveys.

8b. *Describe contexts in which the authors do not expect the study's findings to hold.*

In settings where digital adoption is especially high, the relationship between Facebook use and internet use might be more heterogeneous. Our models' performance in high-

adoption settings is still reasonably accurate, but it performs best at estimating internet adoption levels in lower penetration settings. This is reflected in the higher uncertainty estimates in high-adoption settings.

In addition, if the correlation between Facebook penetration and digital adoption weakens in the future, the effectiveness of our methods may decline. However, the rise of other social media platforms could provide alternative or complementary data sources, if these user counts become publicly available.