Evaluating Gender Bias in AI Applications Using Household Survey Data

Technical Report

October 2023

Seth Goodman, Katherine Nolan, Rachel Sayers, Ariel BenYishay, Jacob Hall, Mavis Zupork Dome, Edem Selormey

AIDDATA
A Research Lab at William & Mary

WILLIAM & MARY
CHARTERED 1693

CDD-GHANA

USAID
FROM THE AMERICAN PEOPLE

DAI
Shaping a more livable world.

Evaluating Gender Bias in AI Applications Using Household Survey Data

Technical Report

October 24, 2023

Seth Goodman[1], Katherine Nolan[1], Rachel Sayers[1], Ariel BenYishay[1], Jacob Hall[1], Mavis Zupork Dome[2], Edem Selormey[2]

[1] AidData, Global Research Institute, William & Mary. Williamsburg, Virginia, US
[2] Ghana Center for Democratic Development. Accra, Ghana

Contacts:

Name	**Email**
Seth Goodman (Technical PoC)	sgoodman@aiddata.wm.edu
Katherine Nolan	knolan@aiddata.wm.edu
Rachel Sayers	rsayers@aiddata.wm.edu
Ariel BenYishay	abenyishay@aiddata.wm.edu
Alexander Wooley (Partnerships PoC)	awooley@aiddata.wm.edu
Mavis Zupork Dome (CDD-Ghana PoC)	m.dome@cddgh.org
Edem Selormey	edem@cddgh.org

AIDDATA A Research Lab at William & Mary

WILLIAM & MARY CHARTERED 1693

CDD-GHANA

USAID FROM THE AMERICAN PEOPLE

DAI Shaping a more livable world.

Table of Contents

Links to Related Documents and Materials

Resource	Description
Executive Summary	Short form overview of major project outputs, successes, challenges, and lessons learned. Intended to be accessible by a broad audience.
AidData Project Page	Dedicated web page on aiddata.org which serves as a centralized access point for project resources and materials.
GitHub Repository	Public repository where all data, code, and technical usage documentation is available.
Local Context Report	CDD-Ghana combined an in-depth literature review and local knowledge to generate a report focusing on gender differences and similarities related to asset acquisition, control, and decision making in Ghana.

Acknowledgements

This work is made possible by the generous support of the American people through the United States Agency for International Development (USAID). The contents are the responsibility of the authors and do not necessarily reflect the views of USAID or the United States Government.

The authors acknowledge William & Mary Research Computing for providing computational resources and/or technical support that have contributed to the results reported within this paper. URL: https://www.wm.edu/it/rc

We would also like to thank the members of the Equitable AI team overseeing the award from DAI and USAID who provided feedback, advice, and other support to our team throughout the project.

Acronyms

Acronym	Definition
AI	Artificial Intelligence
CDD-Ghana	Center for Democratic Development Ghana
CNN	Convolutional Neural Network
CV	Cross Validation
DHS	Demographic and Health Surveys
HoH	Head of household
IWI	International wealth index
ML	Machine Learning
NTL	Nighttime lights
OSM	OpenStreetMap
PCA	Principal component analysis
RF	Random forest
RFR	Random forest regression
RMSE	Root mean square error
VIIRS	Visible Infrared Imaging Radiometer Suite
WI	Wealth index

Introduction

Over the past year, AidData, in partnership with the Center for Democratic Development Ghana (CDD-Ghana), has worked to evaluate the potential of gender bias in wealth estimates generated using artificial intelligence (AI), geospatial data, and USAID's Demographic and Health Surveys (DHS) data. The project leverages AidData's expertise in AI, geospatial data, household surveys, and CDD-Ghana's knowledge of the local context and environments to produce a novel public good that will elevate equitability discussions surrounding the growing use of AI in development.

Funding for the project was awarded through USAID's Equitable AI Challenge - implemented through DAI's Digital Frontiers - which was designed to fund approaches that will increase the accountability and transparency of AI systems used in global development contexts. The project builds upon AidData's broader research initiative on gender equity in development and ongoing AI applications, as well as collaborations between AidData and CDD-Ghana.

Activities spanned two major fronts, leveraging the expertise and resources of both AidData and CDD-Ghana. The first, led by AidData, focused primarily on technical development and analysis of the machine learning models used to estimate wealth and creating a practical and extensible methodology for evaluating potential gender bias. The second, led by CDD-Ghana, incorporated local understanding and engagement to inform development of the machine learning models, and engage with in-country stakeholders and organizations.

The lack of previous research into the role of gender in AI based wealth estimates, combined with unique challenges of the data used, meant that the scope of work was both ambitious and faced numerous uncertainties. Many established approaches for considering gender bias in machine learning training data, or in trained models themselves, could not be directly applied. In addition, incorporating expert knowledge of local conditions was clearly critical from the onset for both producing accurate models and providing opportunities to engage with the population the models are based on and who could be impacted by use of the models.

Our efforts to address these challenges and maintain the standards of a truly equitable AI project ultimately produced valuable insight into the influence of gender on AI based wealth estimates and how gender bias can be evaluated and factored into future work. Additionally, the engagement and interaction with local organizations brought together a diverse set of professionals in Ghana who are linked by the significance of Equitable AI to their work, despite being in industries and sectors that may not typically engage with one another.

By sharing these insights and making our work publicly available and readily accessible - including data, code, documentation, and reports - we aim to encourage and facilitate other researchers and analysts to incorporate more equitable AI based wealth estimates into their work. In this executive summary, we will provide a brief overview of the activities implemented, what we learned, and implications for future work.

In this report, we will first explore the background of related research focused on the use of AI and other methods for estimating wealth, as well as available literature on linkages between assets and gender in developing countries. The Methodology section provides details on the datasets used, the design of the AI models implemented, as well as approaches for household gender classification and producing alternative wealth indices. The Results section breaks down the findings of our analysis into key performance areas which include multiple aspects of model performance along with a comparison of wealth distribution based on alternative wealth indices. In the Discussion, we consider the limitations of the current approach and data, and possible directions for future work. Finally, in the Appendices we provide additional details on specific elements of the analysis, as well as an overview of other elements of the project including in-country engagement, and documentation on the technical materials (data and code) made available for public use.

Background

The significance of outcome measures to monitor development indicators for research and policy applications has been well established in recent literature (Avtar et al., 2020, Allen et al., 2019, Burke et al., 2021). The limited spatial and temporal coverage available from traditional data collection methods such as household surveys can inhibit identifying and addressing critical trends in poverty, health, and other sectors at subnational levels in a timely manner. Over the past two decades, remotely sensed geospatial data such as nighttime lights has been used to fill in existing gaps for metrics including population, poverty, and GDP (Pozzi et al., 2002, Ebener et al., 2005, Henderson et al., 2012, Bennett and Smith, 2017). More recently, researchers have leveraged machine learning methods, satellite imagery, and other geospatial data to produce estimates of development indicators and outcomes including household wealth, infrastructure quality, and crop yield (Jean et al., 2016, Oshri et al., 2018, Lobell et al., 2020).

A subset of machine learning algorithms, known as deep learning, have shown particular promise when used to estimate poverty - primarily based on household assets measured by the Demographic and Health Surveys (DHS) (Burke et al., 2021, Lee and Braithwaite, 2020). While the use of deep learning approaches such as convolutional neural networks (CNNs) have been shown to produce better estimates of poverty than previously established methods leveraging nighttime lights as a proxy, the uptake of deep learning methods can be limited by computational requirements & financial costs, technical knowledge, data accessibility, time to produce, and the interpretability of results (Justus et al., 2019, Christopher Yeh et al., 2021, Lin Htet et al., 2021, Ayush et al., 2020). Recent literature has explored the use of OpenStreetMap (OSM) data paired with a machine learning approach known as random forests (RFs) as an alternative (Tingzon et al., 2019, Lin Htet et al., 2021).

Tingzon et al. (2019) showed that volunteer geographic information (VGI) from OSM on features in the Philippines - including building footprints, road networks, points of interest, and more - could be used alongside nighttime lights data in a random forest regression (RFR) to produce estimates of poverty comparable to deep learning approaches (R2=0.63). Lin Htet et al. (2021) applied a similar approach using OSM features and nighttime lights to estimate poverty in Myanmar with random forests and other machine learning algorithms (RF R2=0.71). Both studies represented poverty using the DHS Wealth Index (derived from household assets), and incorporated over 100 different features based on data from OSM as model inputs.

Despite the extensive coverage of OSM data globally (Barrington-Leigh and Millard-Ball, 2017), variations in coverage can exist across different regions of the world (Neis et al., 2013) and within low and middle income countries (Lloyd et al., 2020). Numerous efforts have been made to assess the accuracy and completeness of OSM (Herfort et al 2023, Zhou and Lin, 2019, Zheng and Zheng, 2014, Tian et al., 2019, Camboim et al., 2015, Barrington-Leigh and Millard-Ball, 2017). A growing body of literature on the use of machine learning to detect building and road features from satellite imagery presents one possible means of supplementing data from OSM (Sirko et al., 2021, Ayala et al., 2021). Although feature detection results may lack valuable classification labels available through OSM, such as building residential status or road speed, there are active efforts to derive similar information using machine learning (Lloyd et al., 2020, Brewer et al., 2021).

In addition to nighttime lights and data on building, road, & other features available from OSM or derived from satellite imagery, remotely sensed datasets with global coverage on climate (Matsuura and Willmott, 2015, Harris et al., 2020), land cover (European Space Agency, 2017), pollution (Gunson and Eldering, 2020), and other topics are available to researchers. A wide range of datasets can be accessed through platforms which abstract data preparation, management and other technical and computational barriers which can limit researchers' ability to use geospatial data (Goodman et al., 2019, Tollefsen Andreas Forøand Strand and Buhaug, 2012). Incorporating these datasets into RF models for estimating poverty may be a low cost approach to improve model performance.

Across approaches for estimating household wealth, the DHS Wealth Index - based on household asset ownership - is one of the most widely used sources of data on wealth and/or poverty. While use of the DHS Wealth Index for AI based wealth estimates has been leveraged in many applications across countries to date, none have explored potential gender bias within these approaches. Research and applications of poverty estimates generated using machine learning have almost exclusively incorporated the full set of DHS households (without gender or other considerations) for training and validating models (see Yeh et al 2020, Tingzon et al 2019). In a recent review of work in this space by Hall et al (2022), no works focused on gender, bias or impacts on potentially vulnerable subpopulation. Although households and individuals sampled by the DHS are representative of the broader population of the countries surveyed, there remains the potential that the asset composition used to generate the wealth index is skewed along gendered dimensions or that wealth estimation models themselves are more effective at generating accurate estimates for one gender over another. This may result in over or under estimation of poverty along gender lines.

In an effort to expand the accessibility and uptake of machine learning methods for estimating poverty, and enable exploring potential gender bias in associated models, we have produced an approach leveraging entirely publicly available data and open source code that can be replicated and modified for a wide range of potential use cases. In this report, we will explore potential gender bias of poverty estimates produced using random forest regressions trained on DHS surveys and a range of geospatial variables.

Literature on Assets-Gender-Wealth Relationships

Wealth can be measured in a variety of ways but it is frequently derived from data collected from household surveys, specifically assets owned by individuals or the household. Available household level data such as the DHS does not normally differentiate between who owns what assets and therefore wealth is measured at a household level regardless of gender. Wealth can be distributed quite differently between men and women even within the same household and therefore the traditional commonly used wealth indicators might not reflect the correct wealth levels for both genders. Below we explore some of the literature around asset ownership and how it can vary between genders.

There is not one strict way to interpret asset ownership or collect asset data for individuals and households. For example, ownership can be challenging to define because the reported owner might be different from the economic owner or from the documented owner of an asset (Doss et al. 2017). Different assets can also be weighted differently as seen with the Sustainable Livelihoods Framework where natural (land and water), physical (agricultural and household durables), financial (cash or savings), human (health, knowledge, and skills), and social (group membership and social networks) all hold different values when it comes to determining wealth (Johnson et al. 2016). These different types of assets can have varying impacts on household wealth and therefore they could have differential impacts on smoothing consumption, or on a household's ability to handle shocks, or a household's ability to increase income (Oladokun and Adenegan 2017). Amongst all of this, asset types and asset preferences can vary both between countries, within countries, and between genders (Oladokun and Adenegan 2017).

There are a variety of ways gender can interact with assets including differences in demographics. For example, in Nigeria, being divorced or widowed, having no or low education, living in a small sized household, being between the ages of 15-24, and living in a household with a female head of household has all been shown to reduce asset ownership. Conversely, having high skilled manual employment and service employment, larger household size, and a higher education level increases asset levels (Oladokun and Adenegan 2017). Wealth levels for men and women can also vary both across and within countries. For example, in Senegal, female-headed households actually have more asset wealth and own more land-wealth than male-headed households (although this is not consistent across the country) (Fisher and Naidoo, 2016). While some gendered wealth findings change between countries, Fisher and Naidoo (2016) find that male headed households have, on average, 13% more asset wealth and 303% more land wealth than women across a number of countries in Africa, Asia, and South America. This finding highlights the need to ensure that different asset wealth is weighted correctly since the divide between men and women in land ownership is much larger than in other asset ownership. Additional cross-nation gendered wealth research has revealed that country level regulations and laws that can also impact gender asset splits, specifically in relation to land ownership. Countries with more gender egalitarian legal regimes have been found to have higher levels of property ownership for married women (Gaddis et al. 2022).

These findings highlight the need to assess gender wealth in a localized geographic way and through more detailed data. Since most of the available country-level household level data does not break assets down by gender, it leaves researchers with using the gender of the head of household for such calculations but this may underestimate female-owned wealth since women in male-headed households also have land and wealth holdings (Fisher and Naidoo 2016).

In order to move beyond using the gender of the head of household in large surveys, more detailed data on asset ownership is needed. In Ghana specifically, the Gender Gap Asset Project[1] has been working on collecting individual-level data to understand women's and

---

[1] https://www.google.com/url?q=https://sites.google.com/view/genderassetgap/home?authuser%3D0&sa=D&source=editors&ust=1673447650003536&usg=AOvVaw0Ki_KFAiSBZ_UrG88_SC0x

men’s asset and property ownership. Their results have been published in several papers that focus on how assets differ between men and women. In Ghana, they find that there is no community property in marriage unless the couple specifically decide otherwise which means men are more likely than women to own the primary residence. Men are also more likely to own a vehicle and a refrigerator (although both vehicle and refrigerator ownership is low) and they are more likely to own all forms of livestock and large agricultural equipment. There is also a high gender gap when it comes to cell phone ownership (42% of men and 15% of women own one). However, women are more likely to own a non-farm business (35% of women, 14% of men). Overall, women own fewer assets than men and additionally, those assets are generally worth less (Doss et al. 2012).

Additional studies have added to this analysis by looking at both asset accumulation and risk preferences through asset ownership in Ghana. Marya Hillesland finds that women hold significantly fewer risky assets than men in Ghana which suggests that there is a difference in relative risk aversion however this could potentially be due to men having more wealth overall than women (Hillesland, 2019). Cheryl R. Doss and her co-authors also find that women are less likely to acquire assets through the market than men and women use a larger number of financial sources to purchase assets such as loans even though it is more difficult for women to receive a loan given that Ghana’s formal lending is concentrated in private banks that favor men (Doss et al. 2019). This is attributed to the fact that women often have more family related financial obligations than men and are therefore less able to save than men (Friendmann-Sanchez, 2006).

Methodology

To leverage geospatial data to produce effective estimates of poverty, we train a series of random forest regressions (RFRs) using varying combinations of features in order to produce wealth estimates at Demographic and Health Surveys (DHS) cluster locations in Ghana. The survey cluster locations and wealth outcome metrics are from the 2014 DHS in Ghana[2]. Building and road footprints, along with other volunteered geographic information (VGI), are retrieved from OpenStreetMap (OSM). Additional geospatial data including nighttime lights are accessed using the free geospatial data repository GeoQuery (Goodman et al., 2019). Varying combinations of the OSM and geospatial variables are explored as input features for the RFR models. The RFRs are optimized over a range of hyperparameter values (settings which influence the behavior of models and are adjusted based on application), and the feature selection used for training is refined based on evaluating feature importance and collinearity of variables. To explore the relationship between model performance, input features, gender, and poverty, we test models utilizing the DHS data both in its raw form (e.g., all households) as well as when subset according to gender based on various gender classification schemes.

Data and Preprocessing

Demographic and Health Surveys

The units of analysis and outcome measure representing poverty are from the DHS rounds conducted in Ghana in 2014. The DHS Wealth Index and related indexes derived from DHS data has been widely used in machine learning based approaches for estimating poverty (Jean et al., 2016, Steele et al., 2017, Tingzon et al., 2019, Lee and Braithwaite, 2020), and are based on household assets such as housing material, sanitation facilities, televisions, and others (Rutstein and Johnson, 2004). DHS surveys are conducted on individual households and aggregated to clusters identified as either rural or urban. To preserve anonymity, cluster locations are randomly displaced by up to 2 km in urban areas and up to 5 km in rural areas (1% of rural clusters are displaced up to 10 km). To account for displacement, urban and rural cluster locations are buffered by 2 km and 5 km respectively to create our units of observation. For each cluster we retain the DHS cluster identifier, the average Wealth Index of the cluster, the cluster longitude and latitude, and the buffered unit of observation.

Using the Ghana DHS survey, Wealth Index values from 11,835 households were produced across 427 clusters. The full set of households/clusters are utilized as the gender agnostic baseline. Gendered subsets of the data based on head of household, household structure, and household asset ownership are produced for comparison and typically cover all clusters for both genders[3].

OpenStreetMap

---

[2] DHS surveys are typically conducted about every 5 years, while in Ghana it has historically been closer to 6 years. The most recent round of the DHS in Ghana is ongoing as of early 2022.

[3] Exact cluster coverage will depend on final classification schemes, but we anticipate few clusters will need to be omitted due to household gender classification.

Data on the location of roads, buildings, points of interests, and other physical features from OSM were obtained using Geofabrik, a repository which provides access to OSM data. Recent OSM data (2021-2022) was used due to substantially reduced coverage existing in earlier years[4]. The groups of raw OSM features (buildings, roads, etc.) contain up to hundreds of distinct, non-standardized labels provided by users to describe features. Examples of building labels include ”medical”, ”clinic”, ”hospital”.

To standardize the wide range of labels, spelling variations, etc. provided by users, we classify features into sub-groups such as ”health” for buildings. For each sub-group, as well as for all sub-groups combined, we generate information such as the count or total area of features within each DHS cluster. The features generated were adapted from approaches utilized previously in related literature (Zhao and Kusumaputri, 2016, Tingzon et al., 2019, Lin Htet et al., 2021). The full list of sub-groups and features generated is listed in Table 1. In total, 71 OSM features are produced for each DHS cluster.

OSM Feature Type	Sub-Groups	Features Generated
Buildings	all, commercial, education, industrial, religious, damaged, government, health, residential, recreational	count, total area, average area, ratio of area to total unit of observation area
Roads	all, major, small, links, nocars, unknown, minor	total length, distance to the nearest road, count of roads segments
Points of Interest	all, food, education, retail, services, entertainment, lodging, landmark, health	count of features
Traffic	all, service, driving, water	count of features
Transport	all, water, air, ground	count of features

Table 1: OpenStreetMap groups and features.

Nighttime Light and Other Geospatial Variables

Publicly available data on nighttime lights and additional geospatial variables are aggregated for each buffered DHS cluster. Nighttime light (NTL) data is available from the Visible Infrared Imaging Radiometer Suite’s (VIIRS) day night band (Elvidge et al., 2017). The raw NTL data is aggregated to the DHS cluster buffers using the mean, median, min, max, and sum of pixels per cluster. Additional geospatial variables related to population, environmental conditions, land cover, and pollution were also prepared using similar aggregation metrics as appropriate. The full list of geospatial variables along with the years of data used are listed in Table 2. Data is extracted from the year prior to the DHS survey (2013) in most cases.

[4] For example, in 2017 approximately 80,000 building footprints had been recorded in OSM for Ghana, while five years later in 2022 over one million building footprints were recorded. This substantial increase is largely tied to growth in OSM user activity rather than actual building development.

Including cluster longitude and latitude to account for unspecified regional trends, a total of 89 features are prepared based on nighttime lights and other geospatial variables. The total number of features exceeds the number of underlying datasets due to underlying data aggregation methods. For example, for each year of land cover data over a dozen features are produced pertaining to each land cover category. Other datasets such as nighttime lights produce multiple features due to multiple methods of aggregating data to DHS clusters (mean, max, sum, etc.). All of the geospatial variables, including nighttime lights, were accessed and prepared via data requests through the free geospatial data platform, GeoQuery (Goodman et al., 2019).

Data	**Year**
Survey cluster locations (Perez-Heydrich et al., 2013)	2014
Nighttime Lights (Elvidge et al., 2017)	2013
Precipitation (Huffman et al., 2014)	2013
Temperature (Wan et al.)	2013
Normalized Difference Vegetation Index (NASA, 2017)	2013
Land Cover (European Space Agency, 2017)	2013
Population (WorldPop and CIESIN, 2018)	2013
Population Density (CIESIN, 2018)	2015
Travel Time to Cities (Weiss et al., 2018)	2015
Elevation & Slope (NASA, 2000)	NA
Distance to Water (Wessel and Smith, 1996)	NA

Table 2: Geospatial datasets and years used in random forest regressions

Random Forest Design

Random forests are an established method for performing classification and regression utilizing an ensemble of decision trees (Breiman, 2001). Random forests have been applied in a wide range of applications incorporating remote sensing and geospatial data (Belgiu and Dragu, 2016), including producing poverty estimates (Zhao et al., 2019). Compared to deep learning approaches, RFs have lower computational costs and require less data and data preparation (Lin Htet et al., 2021). A key distinguishing factor between CNNs or similar deep learning approaches and RFs is that RFs require input features to be defined, while CNNs use inputs (e.g., images) to identify relevant features independently.

Compared to similar methods, such as those incorporating gradient boosting (e.g., gradient boosted trees), or other alternative machine learning algorithms which are less computationally intensive such as support vector machines (SVMs), random forests provide several benefits. Random forest hyperparameters generally require less involved tuning to achieve suitable performance, and interpretation of random forest models and feature importance is more straightforward. In addition, the structure of random forests easily supports parallelization which makes fully utilizing computational resources possible and can significantly reduce run times of models.

Implementation

We utilize the Random Forest Regression (RFR) class from Scikit-Learn due to its accessibility and use in prior research (Tingzon et al., 2019, Zhao et al., 2019). A grid search is used to test a range of values for hyperparameters including number of trees, maximum tree depth, and maximum number of features considered per node[5]. For all RFRs, five-fold nested cross validation is employed to validate model performance. An example set of hyperparameters testing during the grid search is shown in Table 3. The hyperparameters are tested over a range of values that is broadly guided by best practices and previous literature.

RFR Hyperparameter	Values	Relevance
Number of estimators (trees generated per RF)	300, 500, 1000	More trees increase the complexity of the RF and can improve performance, yet will produce diminishing returns at some point.
Max features (considered at a given node for making a split)	"sqrt", 0.33, 0.5, 1.0	A key element of a RF is that not all features are utilized at each split, and features considered are randomly selected. This value defines what proportion of all features at a given node are considered.
Max depth (of an individual tree)	7, 10, 15, 20	Increasing tree depth increases complexity and can refine performance, yet typically will have a limit to avoid overfitting or unnecessarily complicated models.
Minimum samples required to split a node	2, 3, 5	If a node has less than this value it is considered terminal. This can help prevent overfitting when dealing with small leaf sizes and/or limited training data.
Minimum samples required to produce a new leaf node from a split	1, 2, 4	Similar to the minimum sample for splitting, this value defines how many samples must exist in each leaf from a split. It can prevent splits which would result in highly imbalanced leaves.
Error criterion (used for optimization of splits and feature selection)	"Squared_error"	While other options for the error criterion exist, squared error is widely used and effective. It utilizes the variance reduction as feature selection criterion for each split in a tree.

Table 3: Example hyperparameters explored in RFRs

In addition to utilizing Scikit-Learn to generate and train the RFR models, we leverage two tools to support running models and tracking results: Prefect and MLFlow. Prefect is a workflow orchestration tool that enables building, running, and managing flows of code. Use of Prefect enables easily keeping track of model training runs using various training data

[5] Technical documentation of hyperparameters is available at https://scikit-learn.org/stable/modules/generated/sklearn.ensemble.RandomForestRegressor.html

classifications schemes, rerunning existing models when code changes are made, and ensuring all runs complete without errors.

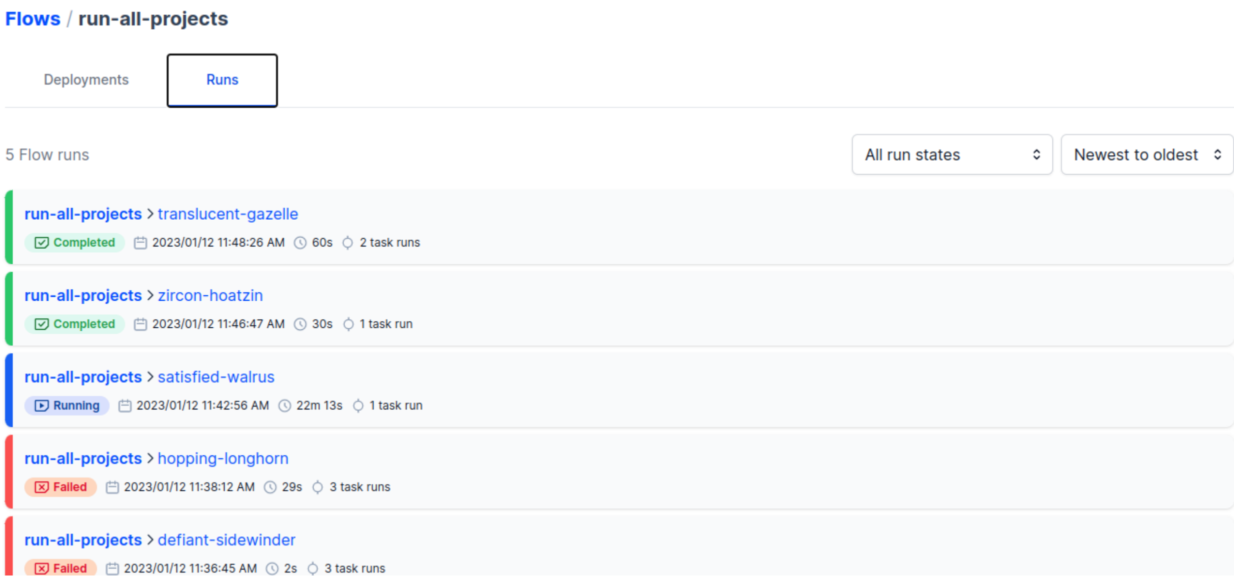


Figure 1: Example of Prefect user interface for managing model runs

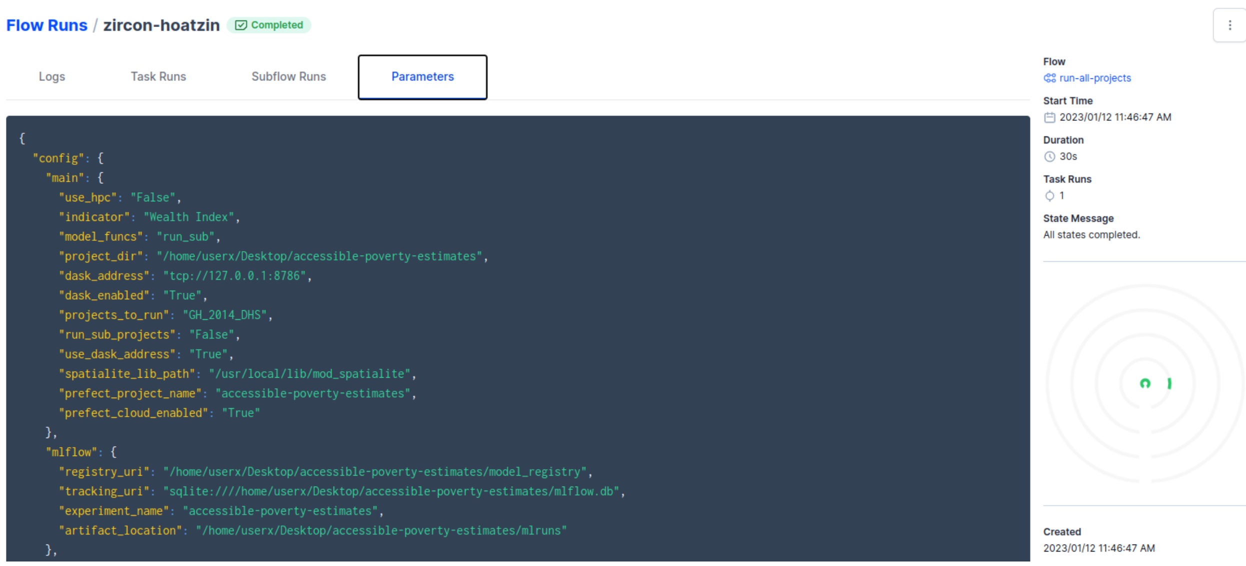
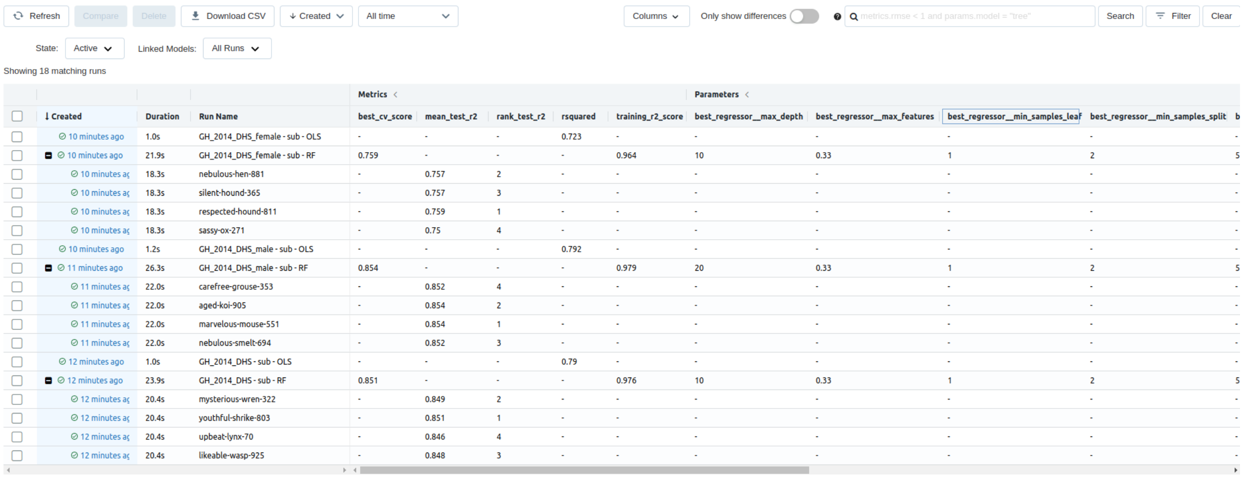


Figure 2: Example of detailed records of configuration used for each run using Prefect

Within Prefect, Dask is leveraged for parallelization of tasks in order to run multiple subsets of the model generation and training process simultaneously and reduce the amount of time required for computation. Beyond the high level workflow management provided by Prefect, Dask provides a separate dashboard which contains more specific information on computational processes and resource usage.

While utilizing Prefect and Dask do not require any additional effort - and can essentially be ignored when all models / code runs smoothly - we also include the ability to run models using simplified and non-parallelized code. This and many other configurations associated with running jobs are controlled by a relatively simple configuration file.

Once the code to generate and train models is run using Prefect, MLFlow enables tracking the results of models. As each model is trained and validated, MLFlow logs the hyperparameters, input data, performance, along with additional outputs to a database and provides a convenient user interface to explore the data. This enables identifying trends across models such as best performing hyperparameters, important model features, along with visualizing detailed plots associated with each model.



			Metrics					Parameters				
Created	Duration	Run Name	best_cv_score	mean_test_r2	rank_test_r2	rsquared	training_r2_score	best_regressor__max_depth	best_regressor__max_features	best_regressor__min_samples_leaf	best_regressor__min_samples_split	b
10 minutes ago	1.0s	GH_2014_DHS_female - sub - OLS	-	-	-	0.723	-	-	-	-	-	-
10 minutes ago	21.9s	GH_2014_DHS_female - sub - RF	0.759	-	-	-	0.964	10	0.33	1	2	5
10 minutes ag	18.3s	nebulous-hen-881	-	0.757	2	-	-	-	-	-	-	-
10 minutes ag	18.3s	silent-hound-365	-	0.757	3	-	-	-	-	-	-	-
10 minutes ag	18.3s	respected-hound-811	-	0.759	1	-	-	-	-	-	-	-
10 minutes ag	18.3s	sassy-ox-271	-	0.75	4	-	-	-	-	-	-	-
10 minutes ago	1.2s	GH_2014_DHS_male - sub - OLS	-	-	-	0.792	-	-	-	-	-	-
11 minutes ago	26.3s	GH_2014_DHS_male - sub - RF	0.854	-	-	-	0.979	20	0.33	1	2	5
11 minutes ag	22.0s	carefree-grouse-353	-	0.852	4	-	-	-	-	-	-	-
11 minutes ag	22.0s	aged-koi-905	-	0.854	2	-	-	-	-	-	-	-
11 minutes ag	22.0s	marvelous-mouse-551	-	0.854	1	-	-	-	-	-	-	-
11 minutes ag	22.0s	nebulous-smelt-694	-	0.852	3	-	-	-	-	-	-	-
12 minutes ago	1.0s	GH_2014_DHS - sub - OLS	-	-	-	0.79	-	-	-	-	-	-
12 minutes ago	23.9s	GH_2014_DHS - sub - RF	0.851	-	-	-	0.976	10	0.33	1	2	5
12 minutes ag	20.4s	mysterious-wren-322	-	0.849	2	-	-	-	-	-	-	-
12 minutes ag	20.4s	youthful-shrike-803	-	0.851	1	-	-	-	-	-	-	-
12 minutes ag	20.4s	upbeat-lynx-70	-	0.846	4	-	-	-	-	-	-	-
12 minutes ag	20.4s	likeable-wasp-925	-	0.848	3	-	-	-	-	-	-	-

Figure 3: Example of database produced by MLFlow for tracking model results

Similar to Prefect, directly interacting with MLFlow is not required, and all outputs are logged and stored into local files and figures which are easily accessible. MLFlow simply makes sorting through the results from many models easier, and helps to identify critical trends in model parameters and performance.

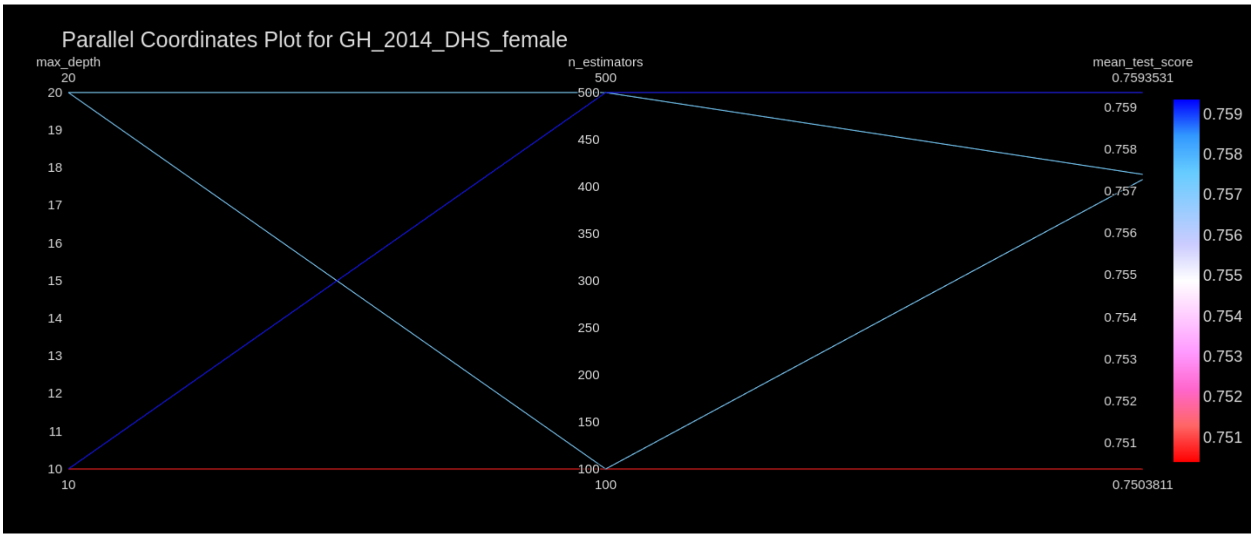


Figure 4: Example of custom interactive model output made available through MLFlow for exploring model hyperparameter performance

Features

Using these tools, we first produce a series of RFR to estimate poverty in Ghana using all household data (i.e., gender agnostic). These RFR models are created using several different combinations of input features. The complete set of features include those from

OSM listed in Table 1, as well as nighttime lights and other geospatial variables listed in Table 2.

Two very simple models are trained to establish a baseline. Nighttime lights is commonly leveraged in linear models to approximate poverty or economic activity and provides an important reference point. In addition, the raw geospatial location (e.g., longitude and latitude values) has the potential to highlight spatial trends driving wealth estimates. Establishing the effectiveness of either NTL or raw location information at predicting wealth will provide meaningful context as additional features are added to the models.

1. Only NTL
2. Survey cluster coordinates only

Next, we train models which are "overloaded" with all features from either OSM or the geospatial variables, or a combination of the two. Many of these features are highly correlated with other features in the set, or are not guaranteed to be generalizable across regions (e.g., certain OSM features are much more likely to be recorded in urban areas than rural areas). In addition, the models resulting from these large sets of features are typically more complex and take longer to train, without resulting in significantly improved performance of more finely tuned models. However, by including all features we can develop an initial impression of model capabilities and begin to assess what features are most important to model performance.

3. All OSM features
4. All OSM features plus NTL
5. All other geospatial variables (including NTL)
6. All geospatial and OSM features

The next step involves subsetting the features to a more practical set for detailed analysis. Features selected for refined models are generally intended to be more easily accessible, available across regions, and available in time series.

For OSM, the criteria for retaining features is whether they could be produced using standard satellite imagery rather than the OSM database. The two main categories of features this results in are based on building footprints and road networks, for which a number of satellite based products currently exist (e.g., Planet Analytic Feeds, Ecopia Tech) or can be produced using a range of machine learning approaches (e.g., SpaceNet competition results and many other academic publications). The resulting OSM features include length of all road segments, distance to nearest road, count of all buildings, average area of all buildings, total areas of all buildings, and ratio of the area of all buildings to total cluster area.

Based on the substantial number of features extracted from GeoQuery across dozens of datasets, time series, and aggregation methods (e.g., min, max, mean, median) the geospatial variables required considerable trimming to produce a practical set of features. The features are refined based on collinearity, data accessibility, and feature importance

(Zhao et al., 2019). Features which have high collinearity, yet relatively lower feature importance (based on models including all OSM and geospatial variables) are then removed.

The selected features represent datasets which are widely available (meaning both publicly available, and having global spatial coverage), available in time series (yet typically data from only the year prior to the DHS survey is used for training), broadly accepted as an accurate and reliable source of data, and could reasonably be associated with wealth or economic conditions.

The resulting subset of geospatial features includes NTL max and median, average elevation, total precipitation (annual), average normalized difference vegetation index (NDVI - a measure of plant greenness), distance to water, travel time to nearest major (population >10k) city, population, land cover associated with urban areas, forests, and cropland, and land surface temperature.

7. The subset OSM features
8. The subset geospatial features
9. The subset OSM features plus all geospatial features
10. The subset OSM features plus only NTL
11. The subset OSM features plus the subset geospatial features

As a final exercise to improve accessibility, and reduce data dependence and model complexity, we train a model using only a minimal set of features. The minimal set of less than ten features (reduced from over 150 features from OSM and geospatial variables total) consists of features which have routinely been found to be important across the above models. Currently, this includes median and maximum nighttime lights, population, urban land cover, longitude, latitude, total length of roads, and ratio of building footprint area to total cluster area. This minimal set of features is intended to provide high levels of performance with reduced model complexity, without overfitting the model for the current use case. This set of features is subject to change as testing and model refinement continues.

12. Minimal set of features

For all of the above features sets we also run a comparable linear regression model alongside the RFR.

Gender Classification

The baseline classification for this work will be **gender agnostic and incorporate data from all households** surveyed. This set of data is reflective of what is widely used in machine learning based approaches to estimating poverty that leverage the DHS WI.

The core comparison of gender specific data will be based on the **gender of the head of household**. Identifying household gender based on the gender of the head of household is the most commonly used approach in literature. While this approach results in roughly two thirds of households being identified as male, the effective sample size for the RFR model is barely impacted since most survey clusters still have some number of female households. The count of households identified for each gender and resulting cluster count is listed in Table 5.

Much smaller counts of gender-specific households within survey clusters has the potential to create noisier data. This is a result of gender classification taking place at the household level, and then being aggregated to the survey cluster level for which geospatial location information is available. We will run additional models as robustness checks in which we artificially limit the number of male households selected for aggregation to the cluster level.

The secondary gender classification is based on **whether a household contains any males**. For the purpose of this work, we limit the age of males considered to those between the ages of 15 and 100. The purpose of this classification is to provide a case where we can eliminate any potential ambiguity between identification of a head of household and the person who owns or controls assets. If no males are in the house, it must be a female household. The tradeoff for gaining additional certainty regarding household gender is that there are fewer female households.

The final set of gender classifications are based on **asset ownership and control**. Gender driven asset ownership and/or control was determined utilizing a combination of information provided by CDD-Ghana in their local context report, and household ownership rates derived from the DHS data. The information provided by CDD-Ghana includes a list of common household assets and whether they are more likely to be controlled (not owned) by one gender over the other. Table S1 in the Supplemental Information contains the list of gendered-controlled assets which overlap with assets used in the DHS WI. The full list of assets as well as a summary of CDD's full report are also provided in the *Local Context Report Summary* section of the Supplemental Information.

Classifying households based on the head of household gender was utilized to generate differential gender ownership rates of assets used in the DHS WI. Assets with highly differential ownership rates between male and female households were cross referenced with the list of gender controlled assets provided by CDD. Four assets associated with male and female households were selected from the overlapping assets. The selected assets, along with ownership rates, are listed in Table 4.

Table 4. Highly gendered assets based on DHS data and head of house households, combined with CDD's findings. Male assets are blue, female assets are red. Percentages reflect the percent of households who own each asset within a category (e.g., all households, female HoH, male HoH).

Asset	**% Ownership Among All HHs**	**% Ownership Among HHs w/ Female HoH**	**% Ownership Among HHs w/ Male HoH**	**Difference (Female-Male)**
Bicycle	30.12	12.10	38.74	-26.64***
Owns land suitable for agriculture	46.18	36.40	50.85	-14.45***

Asset	% Ownership Among All HHs	% Ownership Among HHs w/ Female HoH	% Ownership Among HHs w/ Male HoH	Difference (Female-Male)
Motorcycle/scooter	12.41	3.42	16.71	-13.29***
Bank account	47.61	39.25	51.61	-12.36***
Type of cooking fuel: charcoal	30.00	39.04	25.67	13.37***
Type of toilet facility: ventilated improved pit latrine shared	30.95	36.03	28.52	7.51***
Type of toilet facility: ventilated improved pit latrine	33.74	38.80	31.32	7.48***
Source of drinking water: public tap/standpipe	21.38	26.34	19.01	7.33***

Using the identified assets, four classification schemes were implemented:

1. Any household owning any of the male assets is classified as male. All other households are female.
2. Any household owning any of the female assets is classified as female. All other households are male.
3. Any household owning any of the male assets is classified as male. Any household owning any of the female assets is classified as female. Households which own at least one asset from both the male and female asset lists are included in both the male and female household lists.
4. Any household owning any of the male assets is classified as male. Any household owning any of the female assets is classified as female. Households which own at least one asset from both the male and female asset lists are dropped from both the male and female lists.

Each of the four asset-based classification schemes provide a different way of looking at the data and household gender. An important factor to emphasize is that while differential ownership rates and gendered control of assets may be representative of meaningful gender trends in Ghana, they are not necessarily ideal for classifying households. For example, the most strongly female gendered asset is charcoal cooking fuel, which is over 13% more common in female headed households. Despite that, actual absolute ownership rates for charcoal cooking fuel are 39% for female headed households and 25% for male headed households. Similarly, while 50% of male headed households own agricultural land, so do 36% of female headed households.

The imperfect nature of using assets to classify households is reflected in the resulting count of households classified as male or female using the asset based approaches, as seen in Table 2. When selecting households based on male or female assets, considerably more households are selected for each gender than based on head of household gender. Excluding households which own assets from both categories produces a much smaller subset of households for each gender as would be expected, but limits the cluster coverage which may impact the ability to sufficiently train models. The limitations of these classification approaches, as well as broader limitations and assumptions related to the approaches are discussed further in the subsequent Limitations section.

Table 5. Household counts and cluster coverage based on gender classification strategies

Classification Strategy	# Male Households	# Female Households	# Male Clusters	# Female Clusters
Gender agnostic (all)	11835		427	
Head of Household	8008	3827	427	421
Any Males in Household	8835	3000	427	418
Has Male Assets	9555	2280	427	399
Has Female Assets	5174	6661	411	412
Male vs Female Assets (include overlaps)	9555	6661	427	412
Male vs Female Assets (exclude overlaps)	4479	1585	406	335

Aggregating gender classified households to the cluster level resulted in a total of 13 cluster level datasets. The associated CSV files are available along with this report, included in a ZIP file.

Gender Specific DHS WI Creation

To understand the influence of assets on the DHS Wealth Index, we leveraged documentation on the DHS WI and the DHS WI construction methodology to replicate the wealth index for the 2014 Ghana DHS round. The process of constructing the wealth index produces both the final wealth index values, along with weights associated with each household asset used in the index. The weights are constructed using a principal components analysis (PCA) that allows for differential weights on assets by urban/rural location of the cluster. The construction methodology, however, does not allow for weights of assets to differ by gender of the household.

Gender specific asset importance could be explored by recreating the DHS WI using only subsets of households associated with either male of female headed households. This would, in effect, allow weights on assets to differ by gender of the household. The resulting asset weights and difference in household wealth index values can then be used to gain insight into what assets drive the DHS WI for different subsets of the surveyed population, and how resulting variation in the DHS WI values may influence wealth estimation models.

International Wealth Index

The International Wealth Index (IWI) is a simpler, standardized wealth index that can be generated for households across countries and surveys using data on only twelve assets (seven consumer durables, access to two public services, and three housing characteristics). Unlike the PCA-based DHS WI, the IWI uses fixed weights and can be produced far more easily. We have generated a Python script that can be utilized to generate the IWI for all households in the 2014 Ghana DHS, and can be adapted to other surveys/rounds by simply remapping a subset of the asset variables where necessary.

The standardized nature and fixed weights of the IWI provide the benefit of broader compatibility across countries and potentially reduced chances of country-specific asset bias influencing the index. However, the use of fixed weights could potentially ingrain any overarching gender bias which exists in the smaller set of assets/weights. We will explore differences between IWI values and the DHS WI values for households to consider how the IWI might be leveraged for more effective and equitable AI approaches for wealth estimation.

Results

In this section we explore the analysis and findings, based on the background and methodology established in the previous sections. The findings detailed here will cover five areas: 1) model performance, 2) feature importance, 3) recreating the DHS WI, 4) exploring an alternative wealth index, and 5) using gender specific models for cross-gender estimates. In the subsequent discussion section we will cover critical limitations of the work conducted to date, and directions for future work. Additional figures, tables, and other referenced material will be included in the Appendix. Other content referenced, such as code, has been made publicly available as part of the outreach and dissemination efforts and is accessible via links provided.

The first section on model performance will evaluate the effectiveness of models produced for each gender by exploring several components of model creation. First we will look at the impact of hyperparameters that define how the model is created and trained. Next we will consider what features (independent variables) provided to the models are most useful and practical. Then we will explore different gender classification approaches and how they impact model performance. As a subset of gender classification, we will also test if imbalances in the number of gender specific households within each cluster used to train models can impact performance. Finally, for a refined set of models we will consider what features were most important to the model creation and how those features varied between genders.

In the second we will utilize models trained on gender specific data to produce wealth estimates using data from the other gender. Section 3 will explore the underlying creation of the DHS WI and the relevance of specific assets by recreating the DHS WI using data for each gender. Then, in section 4 we will consider the use of an alternative standardized wealth index which may be less subject to survey specific data biases. Finally, we will highlight some critical limitations of the current work and present potential directions for future work.

Model Performance

We subset the 2014 Ghana DHS data into 13 sets of household data (Table 1a) based on our gender classification criteria which include a gender agnostic approach using all household data, as well as male and female subsets resulting from six different gender classification approaches for households. Not all sets of data are used in each portion of the model performance analysis, yet the core household gender classification based on head of household gender is included in every analysis as a consistent point of reference, as it is the most commonly used gender classification approach seen in research literature.

The gender of a household is an imperfect definition, as there is no absolute gender for a single household in most cases. In most households, assets and overall wealth are reflective of household members of both genders. Household member specific assets and wealth are not available however, and would likely be difficult to collect, so we attempt to produce meaningful approximations of “household gender” using these approaches.

Table 6a. Overview of all data classification strategies to be used for running models.

		Male	Female
Classification Strategy	*Gender agnostic (all)*	1	
	Head of Household Gender	2	3
	Any Males in Household[6]	4	5
	*Household Owns Male Assets**	6	7
	*Household Owns Female Assets**	8	9
	Male vs Female Assets (include overlaps)	10	11
	Male vs Female Assets (exclude overlaps)	12	13

A critical consideration when applying different classification approaches is the resulting sample sizes in terms of both households counts and cluster coverage. The associated household and cluster counts for each classification approach are detailed in Table 6b and will be referenced in the subsequent analyses.

Table 6b. Household counts and cluster coverage based on gender classification strategies

Classification Strategy	**# Male Households**	**# Female Households**	**# Male Clusters***	**# Female Clusters***
Gender agnostic (all)	11835		427	
Head of Household	8008	3827	427	421
Any Males in Household	8835	3000	427	418
Has Male Assets	9555	2280	427	399
Has Female Assets	5174	6661	411	412
Male vs Female Assets (include overlaps)	9555	6661	427	412
Male vs Female Assets (exclude overlaps)	4479	1585	406	335

Hyperparameters Search

The initial component of model performance we explored was the impact of hyperparameters. Hyperparameters define how the random forest and trees within the forest are constructed and optimized, and can affect a model's practical performance. Poorly tuned hyperparameters can result in models that are unable to effectively learn from the training data to make accurate estimates, or models which are overfit to the training data and fail to be generalizable.

[6] Households with at least 1 adult male (>16 years old) are classified as male, while only households with no adult men are classified as female.

To assess the impact of hyperparameters on models to estimate poverty, we tested a broad range of hyperparameters associated with random forests including the number of estimators (trees generated per RF), maximum features (considered at a given node in a tree for making a split), the maximum depth (of an individual tree), the minimum samples required to split a node, and the minimum samples required to produce a new leaf node from a split. Hyperparameters were tested separately across models trained on data from male headed households and female headed households.

Four combinations of feature sets (variables used for training) were tested to evaluate models under different conditions. The feature sets used included 1) only nighttime lights (NTL) metrics (mean, min, max, etc.), 2) all OSM features, 3) all geospatial features, and 4) a subset of features from OSM and the geospatial variable. Additional details on the feature sets and further testing of their impacts are explored in a subsequent section.

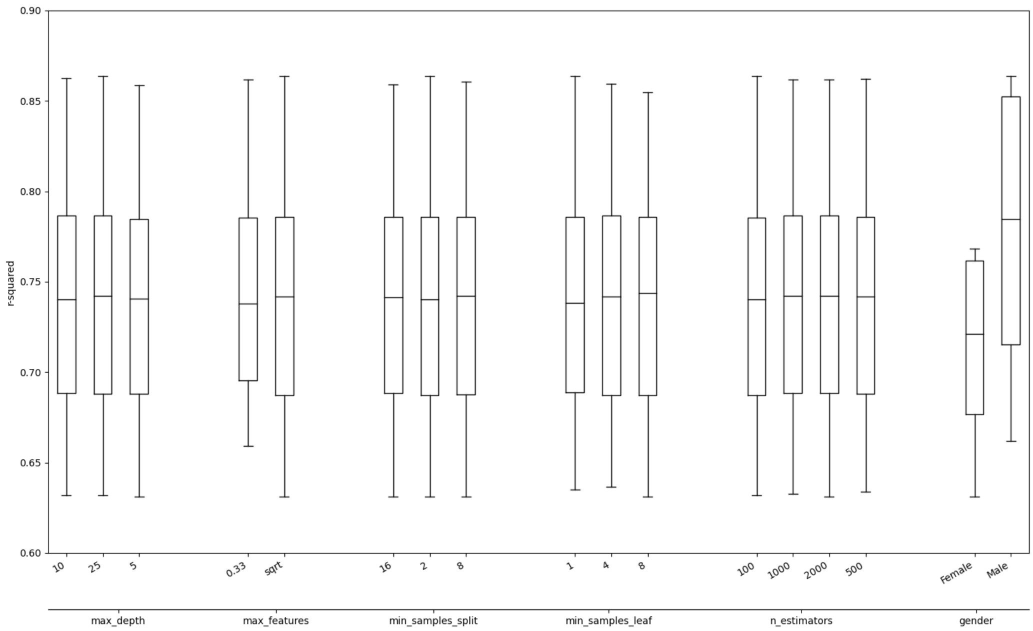


Figure 5. Boxplots of the performance (R^2 values) across all male HoH and female HoH models based on a range of hyperparameters. Includes boxplots with results of all runs aggregated by HoH gender.

Results from the hyperparameter search revealed that there were relatively minor differences in model performance (based on model R^2 value[7]) across the hyperparameters when considering the full set of models tested (Figure 5). Some of the extremes (minimum or maximum R^2 across tests) for a given hyperparameter varied, but median values were fairly consistent. There was however a clear shift in the performance of models using male household data compared to models using female household data. On average, across the range of hyperparameters male models were around 10% better than female models.

Looking at hyperparameter impacts within each set of gendered models separately did not reveal any notable gender specific variations. Male household models consistently produced a median R^2 value of around 0.79 (Figure 6) while female household models consistently produced a median R^2 value of around 0.72 (Figure 7).

[7] R^2 was used as it is most frequently used across literature on AI based wealth estimates.

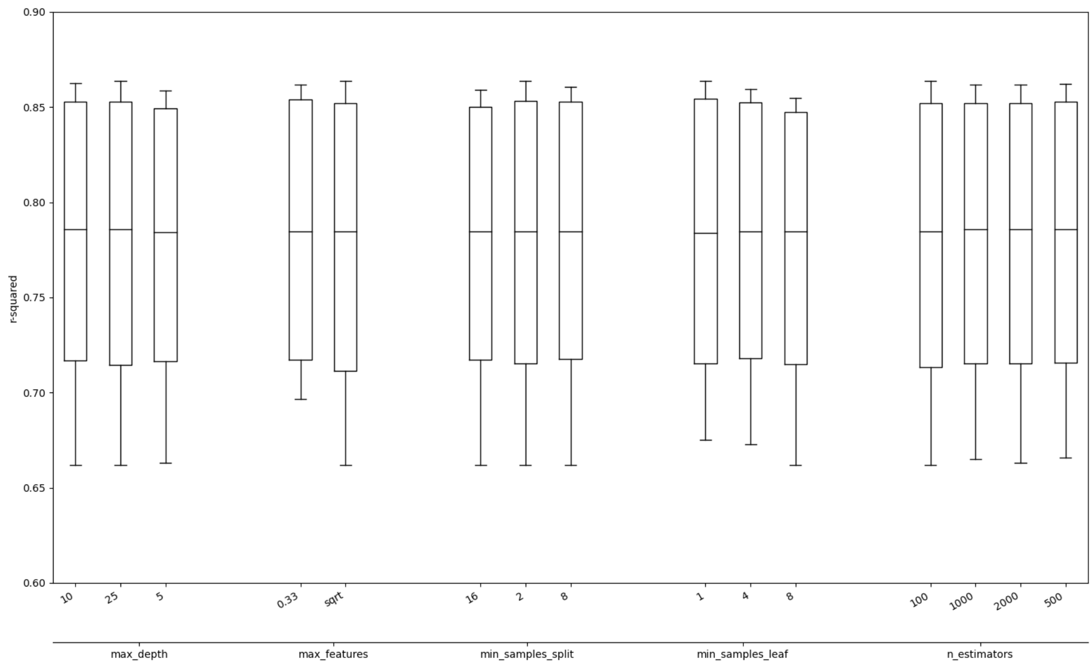


Figure 6. Boxplots of model performance (R^2 values) using varying hyperparameters for only male HoHs.

An additional set of hyperparameter tests were performed which incorporated data from an alternative household gender classification approach. In addition to the previous models which classified gender based on the head of household, we included models which classified gender based on whether any males were present in the household. The hyperparameters used in this set of tests were narrowed slightly from the original range of hyperparameters based on minor differences in performance detected in the first set of tests.

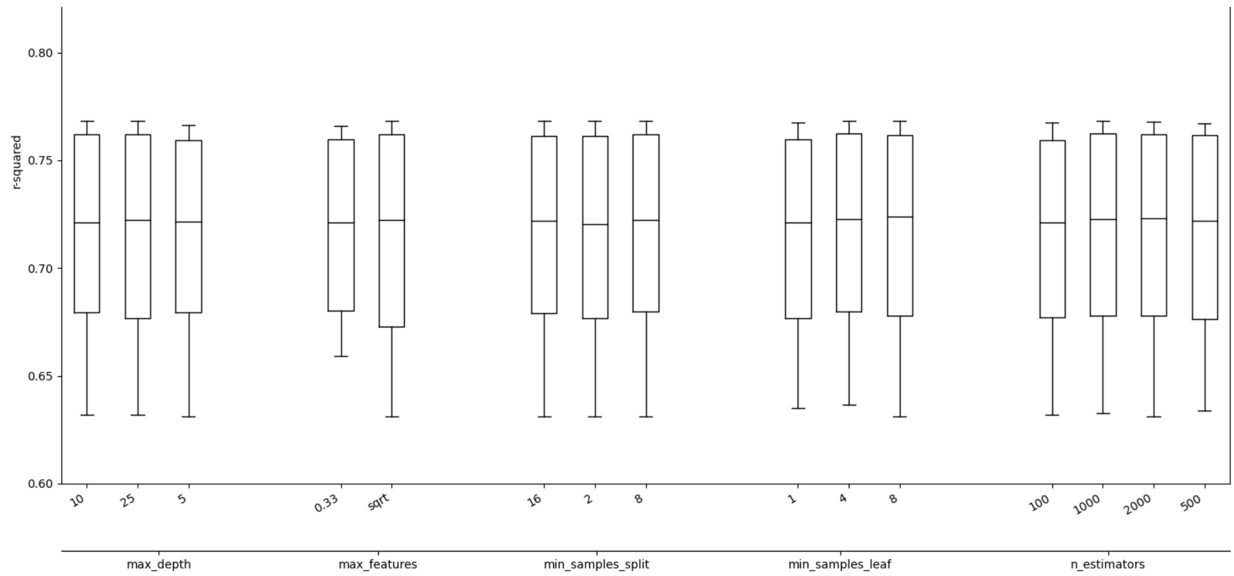


Figure 7. Boxplots of model performance (R^2 values) using varying hyperparameters for only female HoHs.

The results again indicated no major shifts in performance due to hyperparameters. Differences between the male and female household models were also similar to the first set of tests. Models using the presence of males to classify household gender slightly underperformed relative to models based on the head of household gender on average across all hyperparameter tests (Figure 8).

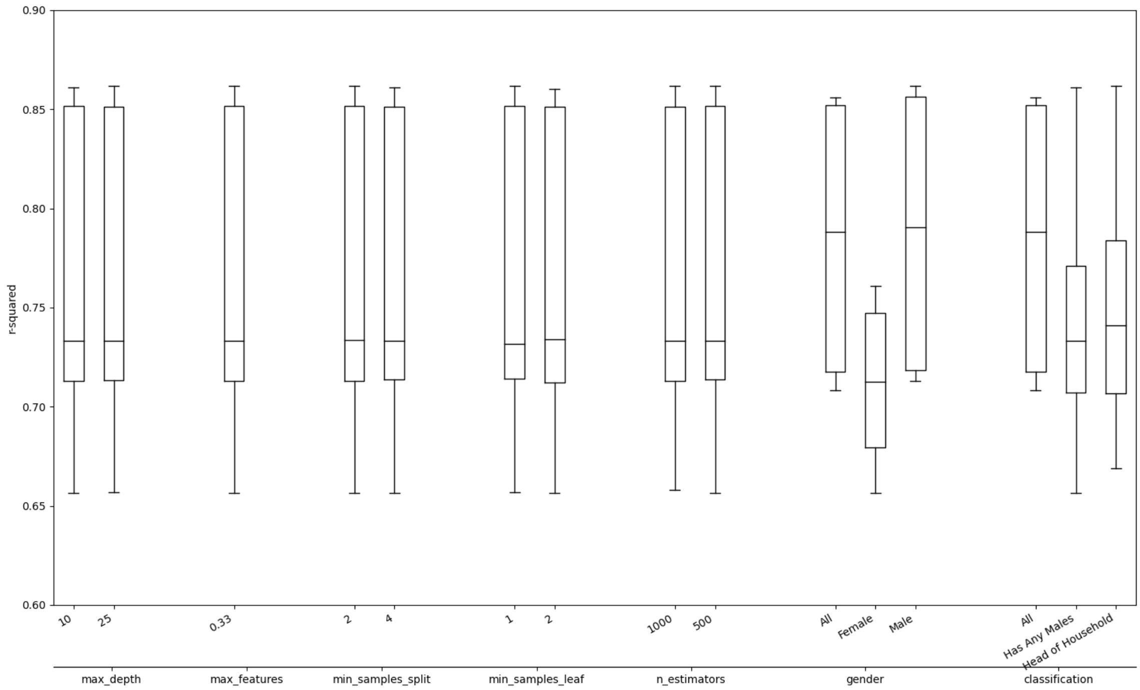


Figure 8. Boxplots of model performance (R^2 values) using refined hyperparameters and incorporating an additional gender classification approach based on male presence in a household.

The specific variations across gender and classification approach are illustrated in Figure 9. Male household models - both based on head of household gender and presence of males in household - consistently outperformed female household models and showed very similar performance to the models based on all household data. Female household models based on presence of males slightly underperformed female household models based on head of household gender. The reduced performance may be the result of fewer households having no men than households headed by women. The impact of household sample size on model performance will be further explored in a later section.

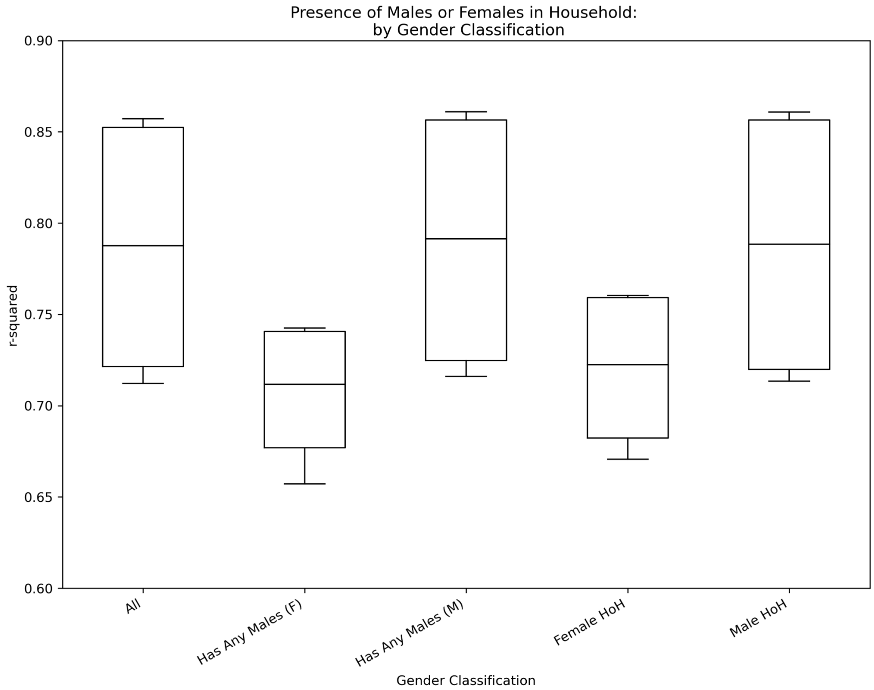


Figure 9. Boxplots of model performance (R^2 values) using refined hyperparameters broken down gender classification approach and gender.

Based on all findings from the hyperparameter search, an optimal set of hyperparameter values were selected to be used in the remaining analysis.

Table 7. Optimal hyperparameters

Hyperparamer	**Optimized Value**
Number of estimators	500
Max features	0.33
Max depth	10
Min samples split	2
Min samples leaf	1

Feature Selection

A total of eleven different feature sets, utilizing over 100 distinct features, were tested across models for both male and female households. Features are derived from spatial information sources such as satellite imagery, sensors, and other Earth observation data, along with community driven records of spatial features such as roads and buildings from OpenStreetMap. All features are produced using spatial data integration methods based on the locations of the DHS clusters connected with households and their associated DHS WI.

Table 8 describes each feature set and the variables it contains. The exact features associated with each feature set will be made publicly available as part of the GitHub repository with the final report. All features are spatially extracted for areas corresponding to the DHS survey clusters which are constructed as either 2km or 5km buffers (urban or rural) to account for spatial anonymization of the data to protect respondents.

Table 8. List of feature sets used for training poverty estimation models.

Feature Set Name	Description of Features
loc	Contains only the longitude and latitude of the cluster
ntl	Features describing nighttime lights within the cluster (min, mean, max, mean, median, and sum)
all-geo	Contains over 30 geospatial variables pulled from AidData's GeoQuery platform. These include a broad range of features on land use, climate and the environment, populations in the area, and more. Also includes the ntl features.
all-osm	75 features derived from OpenStreetMap data consisting of the counts of different types of buildings, roads, points of interest, as well as other traffic and transport infrastructure (e.g., bus stops).
sub-geo	Contains 14 geospatial variables that were correlated with the DHS WI and were seen to be important features in early tests. Highly correlated features from within all-geo were all trimmed to keep only the most important among them. Features include distance to water, NDVI, NTL max and median, temperature and precipitation, travel time to cities, urban, cropland, and forest land cover, elevation and population. Also include the base longitude and latitude features.
sub-osm	A subset of features based on OpenStreetMap data which reflect features that could be derived directly from satellite imagery (i.e., using machine learning). Includes the total and average area of all buildings, the count of buildings, and ratio of building area to total area, the length of all roads, and distance to nearest road.
sub-osm-all-geo	Combines the sub-osm and all-geo features
sub	Combines the sub-osm and sub-geo features
all-osm-ntl	Combines the all-osm and ntl features
sub-osm-ntl	Combines the sub-osm and ntl features
all	Combines all-osm, all-geo, ntl, and loc features

The results of all models using each feature set, without yet considering gender, provide insight into the usefulness of the different features (Figure 6). The best performing (most effective at allowing the model to estimate wealth) feature set is "all-geo" using all the geospatial features aside from the OSM features. Including every feature (all geospatial plus OSM) performs only slightly worse. The slight reduction when adding more features may reflect a reduced ability of the RF to be created using an excessive amount of features that do not add much value. The combined subset of geospatial features (sub-geo) and subset of geospatial and OSM features (sub-osm) performs nearly as well as the all-geo, indicating the performance is likely highly dependent on a handful of key features (see "sub" description in Table 8).

The primarily OSM based feature sets (all-osm and sub-osm) lag notably behind the use of only geospatial features, yet receive significant improvements by incorporating NTL features (all-osm-ntl and sub-osm-ntl). NTL features alone enable reasonable model performance (ntl). Use of only the location of clusters by providing the longitude and latitude as features (loc) served as a simple baseline for model performance.

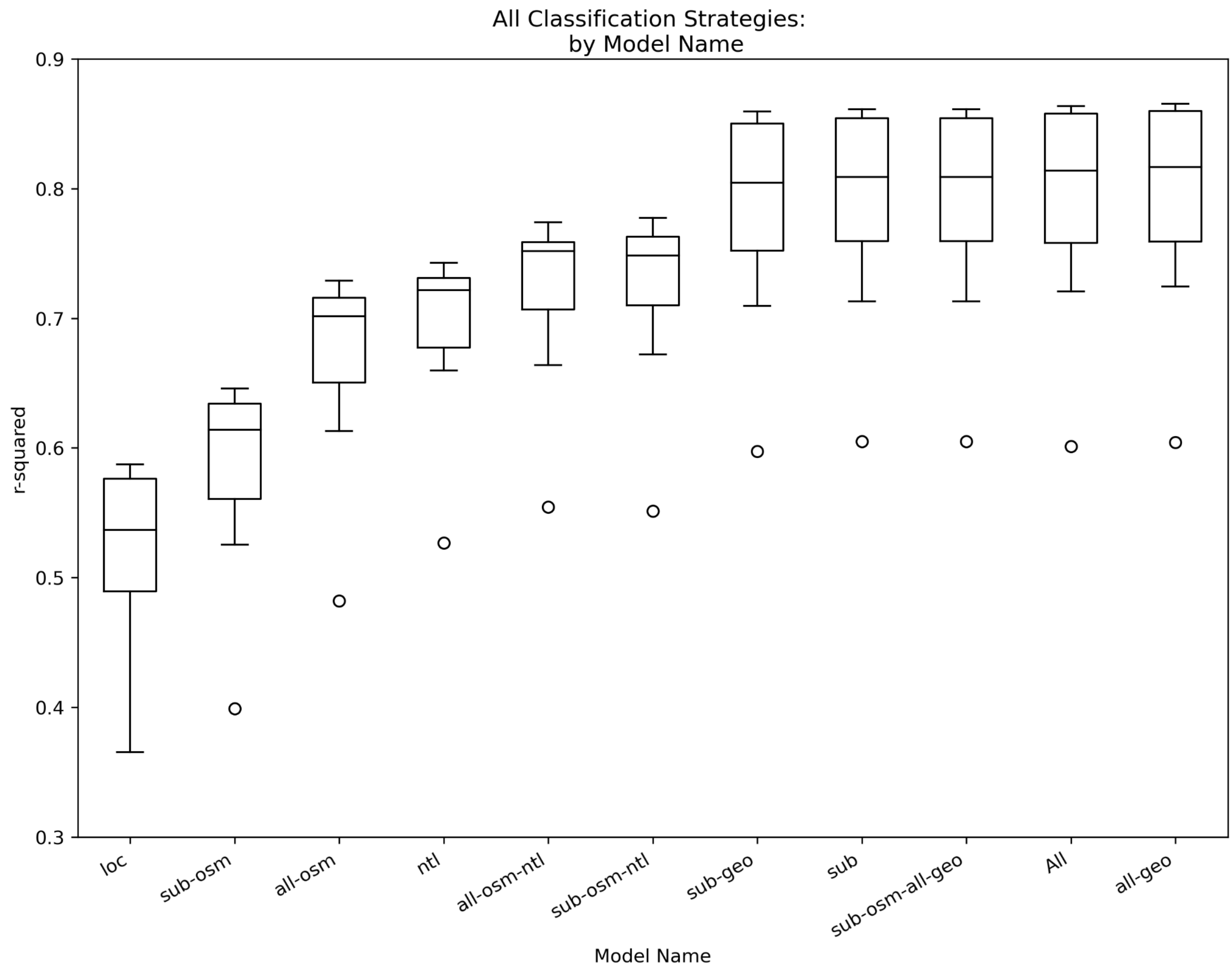


Figure 10. Boxplots of model performance (R^2 values) across all models (gender agnostic, male, and female) broken down by feature set used for models.

The feature set results are then broken down by gender and we can compare performance of models using all household data, male household data, and female household data. The mean R^2 values across all models for each gender classification (including all classification approaches in Table 6b) for each feature set were recorded (Figure 11). Male household models were very similar in performance to gender agnostic models across all feature sets. Female household models were consistently lower? by about 0.1 R^2.

While no deeper gender related trends emerged from this analysis, it did enable us to refine our feature selection. The most efficient feature set (i.e., fewest features for best performance) was clearly the subset of geospatial and OSM features. Efficiency is not directly relevant to the analysis, but will facilitate use of these methods in further applications by reducing dependency on additional datasets. For the remaining analysis we will prioritize utilizing the subset of geospatial and OSM features (sub) along with features sets for NTL, all OSM, and all geo as robustness checks to continue to monitor for any potential variations in performance across features as different facets of the model performance are explored.

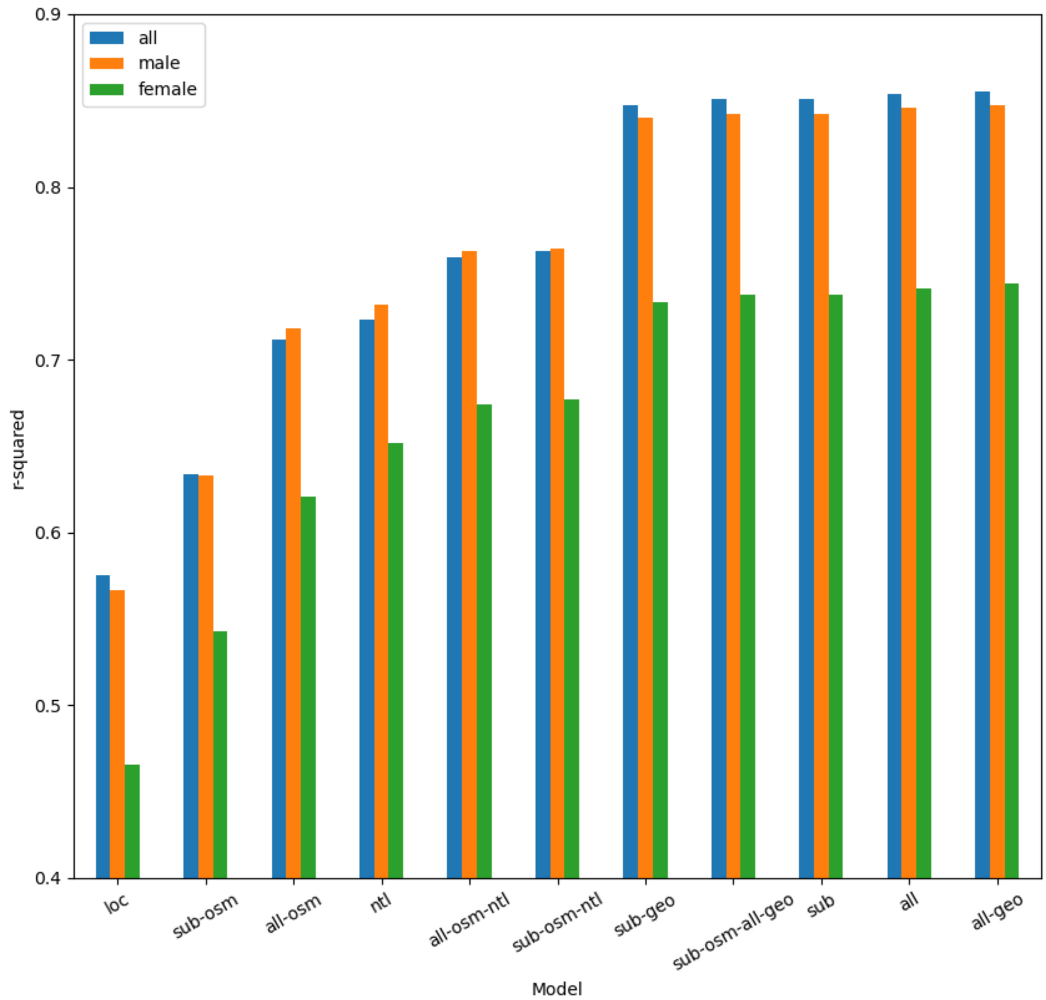


Figure 11. Bar chart of model performance by gender based on the mean R^2 of all models using each feature set.

Gender Classification

The next component of the analysis focused on comparing all gender classification approaches. A total of 6 classification approaches (male and female subset for each) were explored in addition to the baseline gender agnostic approach (Table 6a). Models for each classification approach and gender were run using all feature sets described in Table 8 with the optimal hyperparameters identified from the hyperparameter search.

The result of training and validating models across gender classification approaches revealed that female classified models consistently underperformed male classified models (Figures 12a, 12b).

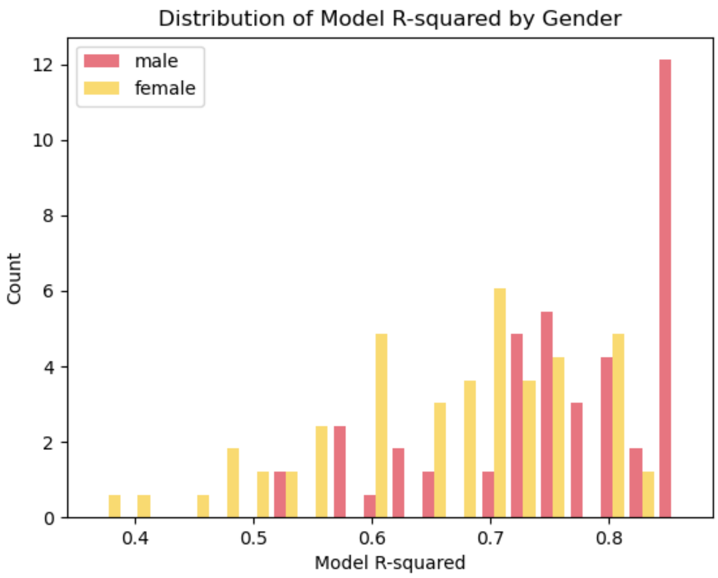


Figure 12a. Distributions of model performance (R^2 values) across all male and female models tested.

Interestingly, although classification based on male presence in the household would seem to be the more reliable predictor of “true positives” for female households (i.e., it may miss many households led by women since they include men - the “false negatives” - but it will accurately capture every female led household without men) it does not result in a better performing model. The female models based on head of household gender outperform female models based on male presence, but by a narrow margin.

The best performing female models are those based on households selected based on ownership of “female” assets as identified in the gender classification process (see Methodology) leveraging information on the local context of gendered asset ownership and DHS asset ownership trends (Table 6b). It is critical to note though, that this is also the most broad classification of female households, and results in the largest household count. While the female asset classification is not necessarily the best classification of female households, as many male households likey own at least one of the same assets, it may be better reflective of a household’s overall economic status that is more common among female households than male. This would suggest that the model is more accurately able to estimate wealth within a more narrow band of the economic spectrum that is more correlated with female households (based on our classification approaches).

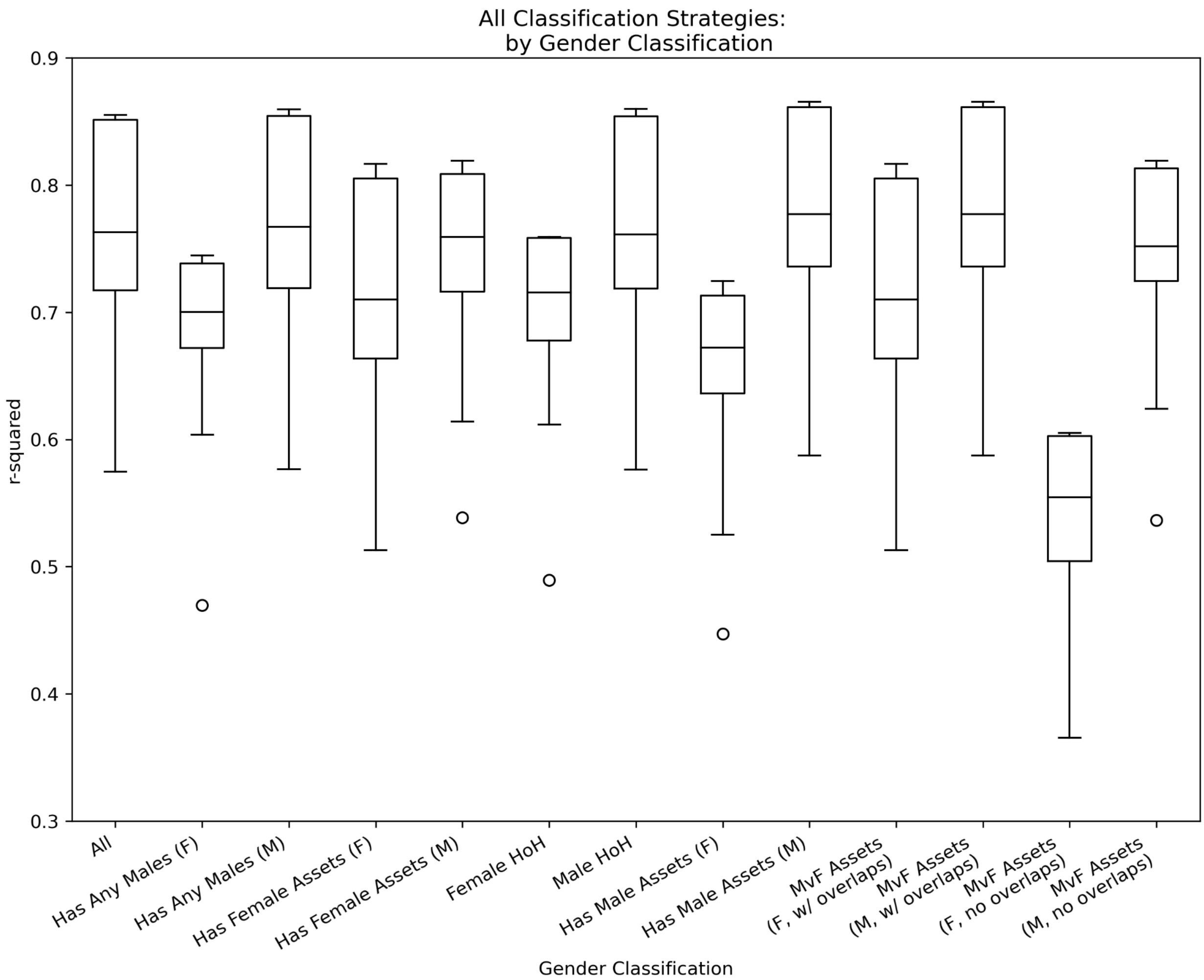


Figure 12b. Boxplots of model performance (R^2 values) across all gender classification approaches for each gender.

A notable finding based on the performance of models is the relationship between the household count associated with each gender and classification approach combination. In general, female household counts are considerably lower than male households counts (Table 1b). While male household counts vary across classification approaches, and are on average less than 70% of the total household count, the male models consistently achieve performance comparable to the gender agnostic model.

Conversely, even when female household counts slightly exceed male household counts for a classification approach, the male model performs better (see “has female assets” in Figure 12). In addition, female models typically have larger differences in household counts (fewer) compared to equivalent male models. These lower household counts are also accompanied by clear drops in model performance for female models. The relationship between household counts and performance, and whether it convolutes the ability to detect gender bias, is critical to understand. The next portion of the analysis will dive further into the impact of variable household counts within clusters resulting from the different gender classification approaches.

Gender Sample Sizes

One of the major caveats of the initial analysis of gender classification approaches is that the number of households (and in some cases the number of resulting clusters) for each gender / classification combination varies. To evaluate whether imbalance between the household and/or cluster count for genders impacts model performance, we conducted two tests. First, we compare unbalanced and balanced models based on head of household gender. Next, we compare performance when artificially reducing the size of gender agnostic data from all households.

Data for models are balanced at the cluster level prior to cluster aggregation by randomly dropping households within a cluster for the gender with a larger number of households in that cluster. For example, assume using unbalanced datasets cluster A has 30 male households and 20 female households, and cluster B has 22 male households and 26 female households. After balancing, 10 male households would be randomly dropped from cluster A so that there are 20 male and female households. Similarly cluster B would drop 4 female households so that there are 22 male and female households.

Balancing the data for models based on head of household gender showed a notable decrease in performance in the balanced models (Figure 13). Male models were impacted the most by the balancing (household counts reduced), yet there were some clusters in which there were more female households. As a result, there is a clear decrease from the unbalanced male HoH model to the balanced male HoH model. There is also an equivalent decrease for the female models, albeit much smaller. The decrease in the balanced models does reduce some of the discrepancies between male and female model performance, but there is still a clear gap between the gendered models.

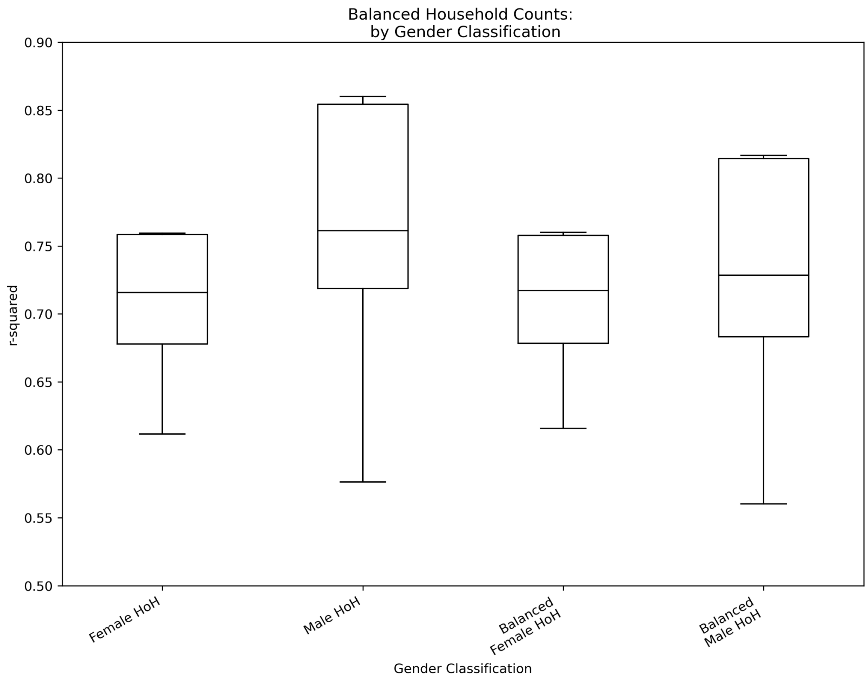


Figure 13. Boxplots of model performance (R^2 values) across models based on head of household gender using both balanced and unbalanced datasets.

To more broadly explore how the number of households within a cluster impacts model performance we artificially reduced the number of households in clusters for gender agnostic models. The size of the original, unaltered clusters (containing all households, labeled “all” in figures) count varies but averages around 28 households. The original clusters were artificially reduced by randomly dropping households at two different levels. The “medium” clusters were created by reducing household counts to 19 households, while the “small” clusters were created by reducing to 9 households each.

The performance of the “medium” models showed a relatively small decrease compared to the original models. However, the “small” model showed a much larger decrease in performance (Figure 14). The disproportionately large decrease seen in the “small” model may be indicative of a critical threshold in the amount of household data necessary to reasonably reflect the household level characteristics (i.e., wealth) of a given cluster.

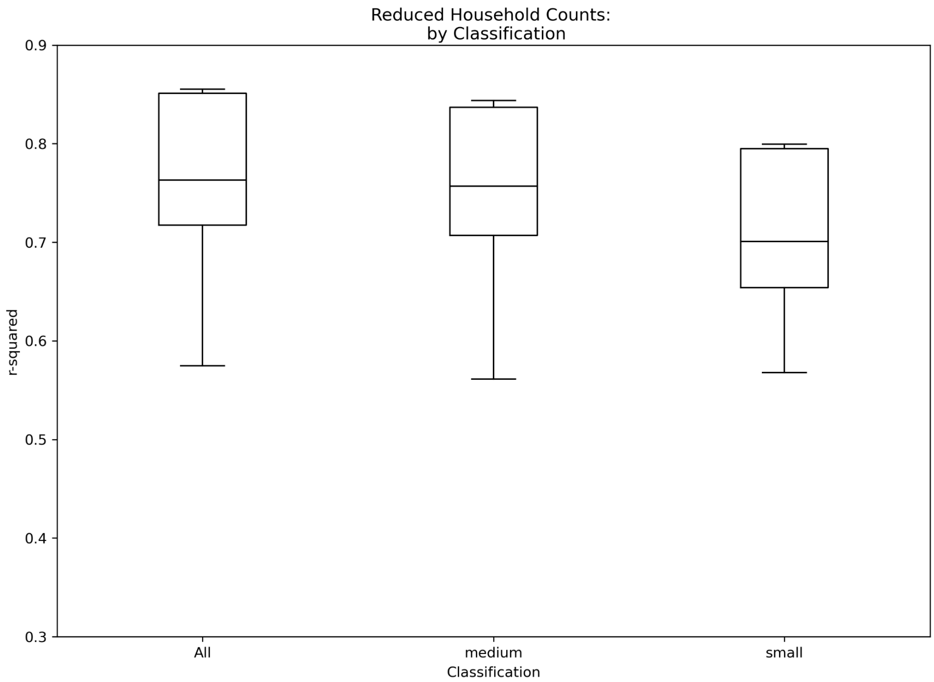


Figure 14. Boxplots of model performance (R^2 values) across gender agnostic models from all households in which cluster data was artificially reduced by varying amounts.

Feature Importance

The importance of features provided to the random forest model was also assessed. The metrics underlying feature importance reflect how useful a given feature is in constructing decision trees within the random forest. Usefulness itself is determined based on reducing “impurity” within a tree, or how well the tree is able to split the data at a given node using the features provided. More broadly, features with a high importance are more useful in explaining wealth variation across the data.

Feature importance was calculated for a subset of models run for both male and female classified households. The resulting importance metrics were then converted into boxplots for each gender (Figures 15 and 16). NTL median value, urban area coverage, and population count were the top three most important features for both male and female households, although urban coverage (which reflects the amount of land in the cluster that contains urban areas) had slightly greater importance for male households.

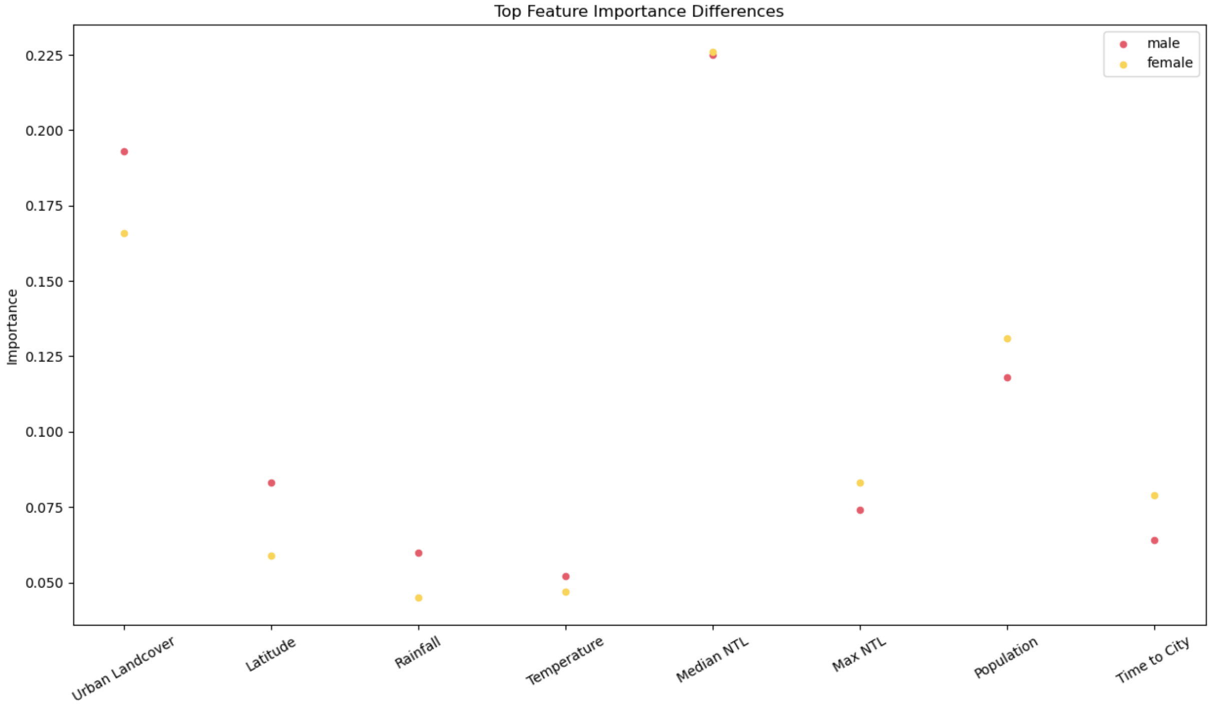


Figure 15a. Differences in feature importance between male and female models for top features. Features further left indicate greater importance in male models, while features further right indicate greater importance in female models. The y-axis indicates the absolute feature importance and distance between red (male) and female (yellow) points for each feature reflects the magnitude of the feature importance difference.

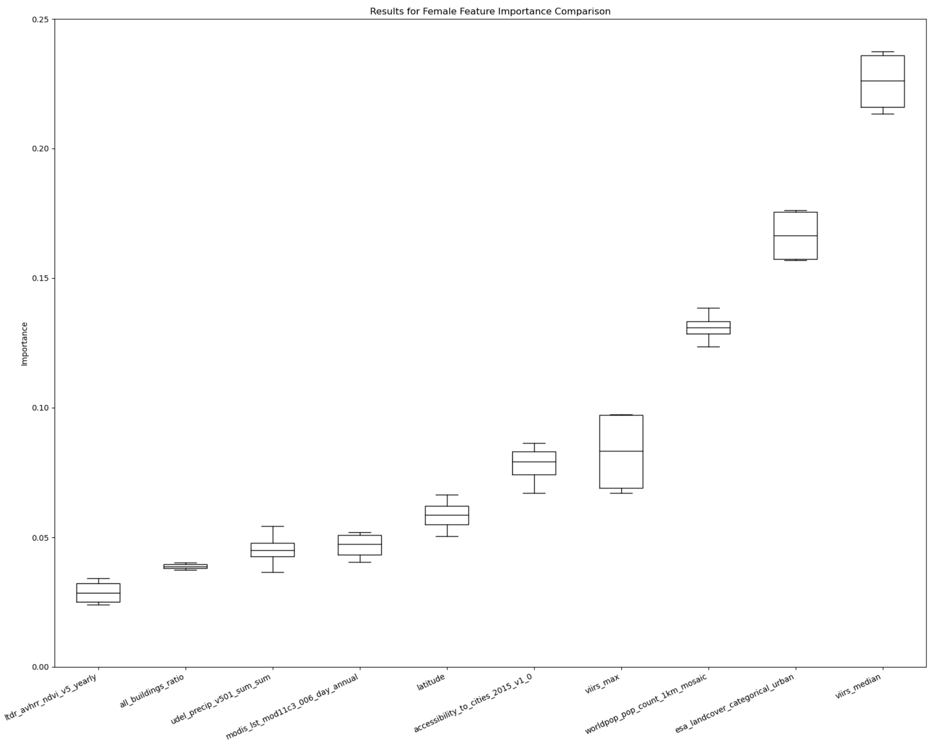


Figure 15b Feature importance plot for female household models.

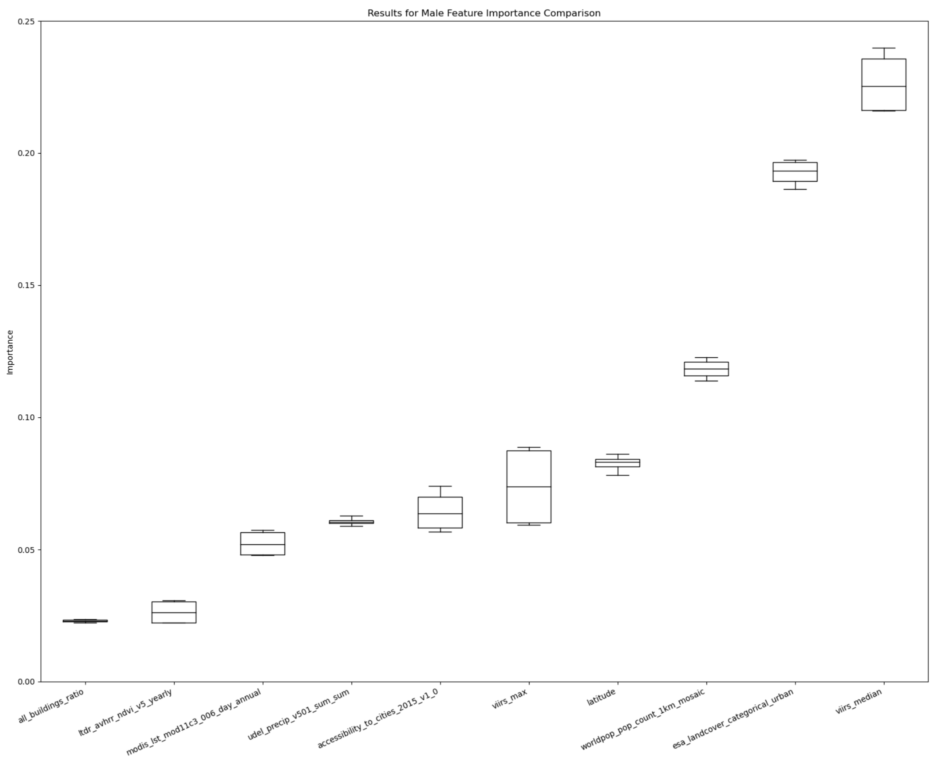


Figure 16. Feature importance plot for male household models.

While the remaining differences were relatively minor, there are two of note. First, accessibility to major cities was more important when explaining wealth within female household models (closer equates to wealthier). As women are less likely to own vehicles or other modes of transportation, and may be depending upon men for transport, access to cities could provide access to resources to support their household and/or jobs, both of which would allow them to increase wealth and assets. Second, precipitation was more important for male households. This may be reflective of men's larger role in the agricultural sector and dependence on crop yield for income.

Predictions

The performance of models as evaluated through training and validation, along with associated characteristics (hyperparameters, feature importance), which we explored in the previous sections provide valuable insights into potential gender biases when estimating poverty. Another tool for understanding bias within models is utilizing trained models to estimate poverty levels on gender-specific subsamples. In the training and validation explored in the previous section, models were trained and validated on data using the same gender classification approach. I.e., a subset of training data classified as female based on the HoH gender was validated using a subset of data which was also classified as female based on the HoH gender that had been set aside. Here, however, we aim to compare the predictive ability of models trained on one gender subsample when tested on a different gender.

To test the cross-classification predictive ability, we utilize models trained on clusters using 1) all households, 2) only male headed households, and 3) only female headed households to produce predictions using data from each of the classification approaches. The R^2 value is calculated based on the predicted DHS WI value and the true value for each (Table 9).

Table 9. Model prediction results.

R^2		**Data Trained On**		
		All	**Male**	**Female**
Data Used For Prediction	**All**	*0.973*	*0.969*	*0.897*
	Male	*0.956*	*0.974*	*0.829*
	Female	*0.913*	*0.881*	*0.958*

The resulting R^2 values should be considered only for the purposes of comparing relative effectiveness when used across the classification approaches, and not as an absolute metric of performance of these models. This caveat is critical, as in some ways the training data and validation data overlap. As a general rule, data should never be used in both training and validation when assessing a model for future usage. The current application is somewhat unique in that there is a limited and fixed number of clusters used across all models, while the outcome metric (the DHS WI value) associated with each cluster is produced as an aggregate of households within the cluster.

The set of households associated with each cluster varies based on the gender classification approach and specific gender and produces distinct outcome metrics for each model. Given that the cluster location and associated input features are the same yet the outcome metrics vary, this is a bit of a gray zone regarding training/validation overlap and the implications. I.e., A female model is trained for Cluster A on households 1, 2, 3, while a male model is also trained for Cluster A on households 4, 5, 6. Within each of those models, typical training/validation rules still apply (i.e., 85% of clusters are randomly selected for training, and the remaining 15% are set aside for validation). Ultimately, the results of this exercise should be considered only as an additional data point in the broader analysis of gender bias, and not as robust evidence of bias.

The results of the prediction accuracy largely reflect trends in the previous sections showing a stronger connection between gender agnostic models and male specific models, and a disconnect between both and female specific models. The model trained on all households has a decrease in R^2 3.5x greater when applied to data using only female households compared to when applied to data using only male households (decrease from 0.973 to 0.956 vs 0.973 to 0.913). Similarly, models trained on gender specific data show notable decreases in R^2 when applied to data for the opposite gender. Models trained on male data however, do better at prediction using the gender agnostic data than models trained on female data.

DHS WI PCA Asset Weights

The DHS WI is generated based on the use of Principal Component Analysis (PCA) and regression models. PCA involves taking a large number of variables and breaking them down into a reduced number represented by a set of principal components (e.g., first principle component, second, etc.). For the DHS WI, PCA is first used to determine weights which reflect the importance of all assets the DHS includes in the WI. The weight value is calculated based on how much each asset contributes to the first principal component, and reflects the amount of variance in the data the asset explains. I.e., assets will be weighted heavily if they explain variance within the population. If almost everyone or almost no one owns an asset, that asset will not be weighted heavily. The weights are then used in a linear regression model to produce the DHS WI value for each household based on asset ownership.

Because of how PCA is used to calculate the DHS WI, the DHS WI itself is dependent on the asset ownership characteristics of the population it is being calculated for. It is therefore possible that minority groups or other subpopulations are not accurately represented by the DHS WI that was designed based on the population's characteristics as a whole. To explore how well the DHS WI represents male and female led households, we calculated the DHS WI separately for male and female led households.

Based on the DHS WI construction and methodology documents released by the DHS, we attempted to replicate the DHS WI for the 2014 Ghana DHS. As the documentation is generalized and exact construction approaches can vary for specific surveys, we were not able to perfectly replicate the original DHS WI. The final recreation of the DHS WI had 93.7% of households within the same quintile as the original DHS WI (Table 10). The Stata code utilized to recreate the DHS WI will be released publicly on GitHub.

Table 10. Transition matrix comparing quintiles of original DHS WI vs recreated DHS WI.

Count		Original DHS WI					
		1	**2**	**3**	**4**	**5**	***Total***
Recreated DHS WI	**1**	2442	78	0	0	0	*2520*
	2	63	2227	68	0	0	*2358*
	3	0	98	2385	74	0	*2554*
	4	0	0	98	2098	83	*2279*
	5	0	0	0	101	1935	*2036*
	Total	*2505*	*2403*	*2551*	*2270*	*2018*	*11747*

The methodology for recreating the DHS WI was then also applied to gender-specific subsets of the data based on the gender of the head of household. When looking at female headed households, the DHS WI produced using data from all households will often cause households in the lower quintiles to be classified as wealthier than in the DHS WI based only on female households (Table 11a). These differences - over 20% of households in the female-based DHS WI having a lower WI value in the original DHS WI (vs <5% having a greater WI value) - are substantial enough that they could alter the relationship between model input variables (i.e., geospatial characteristics) and the wealth index, resulting in noticeable impacts on the models trained using the data.

Table 11a. Transition matrix for female households comparing quintiles of DHS WI created using data from all households vs DHS WI created using data from only female headed households.

Count		WI Using Female Households					
		1	**2**	**3**	**4**	**5**	***Total***
WI Using All Households	**1**	462	0	0	0	0	*462*
	2	460	397	4	0	0	*861*
	3	0	387	622	74	0	*1083*
	4	0	0	92	632	79	*803*
	5	0	0	0	3	591	*594*
	Total	*922*	*784*	*718*	*709*	*670*	*3803*

Comparing the equivalent recreation of the DHS WI for male headed households, the differences versus the DHS WI created using data from all households are much smaller (Table 11b). The notable differences are due to underestimating the wealth of male households in the lowest quintiles in the original DHS WI compared to the male household based DHS WI.

Table 11b. Transition matrix for male households comparing quintiles of DHS WI created using data from all households vs DHS WI created using data from only male headed households.

Count		**WI Using Male Households**					
		1	**2**	**3**	**4**	**5**	***Total***
WI Using All Households	**1**	1735	323	0	0	0	*2058*
	2	0	1221	276	0	0	*1497*
	3	0	2	1402	67	0	*1471*
	4	0	0	56	1409	11	*1476*
	5	0	0	0	75	1367	*1442*
	Total	*1735*	*1546*	*1734*	*1551*	*1378*	*7944*

Although less substantial than the changes seen when calculating the female DHS WI (impacted male households in lower quintiles are <10% of data vs >20% for females), there is still the potential that these differences would alter relationships between model input variables and the wealth index, and thus impact the overall model.

Equivalent transition matrices were also generated based on recreating the DHS WI using the alternative gender classification approach of whether any males were present in the household (no males in household indicates a female household), and showed similar patterns (Appendix Tables A1 and A2).

Using the PCA weights, we further analyzed the differences underlying the DHS WI created based on all households, male headed households, and female headed households. For each, we ranked assets based on the PCA weights (larger weights equals lower rank number). We then selected all assets which were ranked in the top 20 in each of the three DHS WI recreations (all, male, female), and calculated the difference in weight between the female and male DHS WI weights (Table 12).

Table 12. Assets in the Top 20 of all households, male HoH, and female HoH based PCA.

	PCA Rank	<	<	**PCA Weight**	<	<	<
name	**all**	**male**	**female**	**all**	**male**	**female**	**diff F vs M**
Mobile telephone	18	19	12	0.152	0.144	0.177	0.034
Watch	11	14	9	0.174	0.164	0.197	0.033
Wall clock	9	9	8	0.190	0.186	0.199	0.013
Type of cooking fuel: LPG	6	6	4	0.208	0.204	0.215	0.011
Refrigerator	3	3	2	0.232	0.228	0.237	0.009
Cabinet/cupboard	8	8	10	0.190	0.187	0.196	0.009
Bank account	7	7	6	0.203	0.201	0.209	0.008
Source of drinking water: sachet water	10	12	11	0.174	0.172	0.178	0.006
Color television	1	1	1	0.248	0.247	0.252	0.005
Bed	16	17	15	0.155	0.153	0.157	0.004
Video deck/DVD/VCD	5	5	5	0.213	0.214	0.215	0.000
Access to internet in any device	15	16	14	0.159	0.159	0.158	-0.001
Computer/tablet computer	12	11	13	0.172	0.173	0.166	-0.007
Type of cooking fuel: wood	2	2	3	0.236	0.239	0.225	-0.013
Electricity	4	4	7	0.219	0.224	0.201	-0.023

Fifteen total assets were found in the top 20 weights across the PCA for all, male, and female households. The assets with the largest differences in weights between the female and male PCA were mobile telephone, watch, and electricity. Mobile telephone and watch assets were lower ranked (larger weights) in the female PCA, while electricity were lower ranked in the male PCA.

Examples of assets which were consistently less meaningful to the PCA include: use of wood or wood planks for flooring, roofing, or walls; coal cooking fuel; drinking water from a protected spring; and black & white televisions.

It is important to note that weight in the PCA is not inherently indicative of ownership rates of an asset. As the PCA weight is reflecting the extent to which a given variable (asset) helps explain variance in the data, larger relative weights (i.e., larger differences in weights between the genders, not larger absolute weights for a given gender) can seem to be tied to either very low or very high ownership rates.

For example, if every single female household owns asset X, and no male household owns asset X, asset X would be very useful for determining what gender a household is, but would not be useful for assessing variation in the wealth (or any other characteristics) of female households. Since every female household owns asset X, it cannot explain any variance in the data for female households.

As a result, assets with large weights for each gender relative to the opposite gender can end up being assets with higher ownership rates for the opposite gender. For example, ventilated improved pit latrines have large PCA weights for male households relative to female households, despite being one of the assets most commonly associated with female headed households (Table 13). Despite lower levels of use by male headed households, ventilated improved pit latrine may be strongly associated with only a subset of male headed households (e.g., poorer male households), yet common across all female households, and thus more effective at explaining variance in the male household data compared to the female household data.

Table 13. Assets most influential in Male HoH PCA (vs Female HoH PCA)

	PCA Rank	<	<	**PCA Weight**	<	<	<
name	**all**	**male**	**female**	**all**	**male**	**female**	**diff F vs M**
Type of toilet facility: no facility/bush/field	13	10	18	0.170	0.177	0.147	-0.030
house	69	65	97	0.028	0.038	0.004	-0.034
Type of toilet facility: ventilated improved pit latrine	50	40	76	0.056	0.072	0.017	-0.055
Type of toilet facility: ventilated improved pit latrine shared	52	42	83	0.051	0.068	0.010	-0.058
Bicycle	39	33	108	0.071	0.094	0.001	-0.093

Conversely, bicycles have relatively high levels of male household ownership, but low levels of female household ownership, and still have a high weight in the male household PCA (Table 13). In this case, bicycles have a very low weight for women since far fewer female households own bicycles, and a relatively high weight for men since many, albeit not all, male households own them.

Table 14. Assets most influential in Female HoH PCA (vs Male HoH PCA)

	PCA Rank	<	<	**PCA Weight**	<	<	<
name	**all**	**male**	**female**	**all**	**male**	**female**	**diff F vs M**
Radio	28	30	17	0.111	0.100	0.148	0.048
Mobile telephone	18	19	12	0.152	0.144	0.177	0.034
Watch	11	14	9	0.174	0.164	0.197	0.033
Motorcycle/scooter	89	98	58	0.008	0.003	0.034	0.031
Main floor material: cement	36	38	28	0.086	0.078	0.108	0.030

The variation in asset weights between male and female household based PCA does not lend itself to intuitively drawing specific insights about asset roles in households, yet comparing the broader set of asset weights derived from gender-specific household data with the original DHS WI asset weights can provide insight into potential gender bias. Using the asset weights and ranks produced from data on the male and female headed households, we calculate the discrepancy between the asset weights and ranks produced from data on all households. This is measured by the root mean square error (RMSE) (Table 15). RMSE is an approach commonly used for evaluating the quality of predictions made using a model against the true values, but can be leveraged more broadly, such as here, to explore how closely two sets of values align. Lower values of RMSE indicate more similarity between the two sets of values. The results indicate that RMSE for female households is approximately three times greater than for male households. This would suggest that the DHS WI created for female households is considerably more different from the default DHS WI than the DHS WI for male households.

Table 15. Overview

RMSE	**PCA Weights**	**PCA Ranks**
All vs Male	0.00494	3.67486
All vs Female	0.01407	12.19829
Female:Male Ratio	2.846	3.319

Alternative Wealth Index

The DHS WI is widely used to assess economic conditions in households based on asset ownership information collected from surveys. However, the DHS WI values are uniquely generated for each survey and therefore cannot be compared across countries or over time. An inherent characteristic of the DHS WI is that each time it is generated, the underlying weights or importance of assets used to create the DHS WI can vary.

Given the methodology behind the DHS WI construction and the results seen in the previous section exploring recreating the DHS WI for gender specific subsets of a DHS survey, it is possible that gender specific asset bias within the surveyed population may influence the DHS WI. An alternative approach to estimating household wealth is the International Wealth Index (IWI) created by the Global Data Lab. The IWI utilizes a standardized formula to calculate wealth from assets that was derived from data across many countries and surveys. While it is still possible that broad asset bias may exist within the IWI, the more standardized nature of the IWI may make it less likely that country specific conditions will influence (and potentially bias) the index. The creation of the IWI itself will still be relevant and potential biases would need to be considered (as an extreme example: if one of the fixed assets included in the IWI have restricted access for women in a country).

Code to calculate the IWI for the 2014 Ghana DHS was implemented using a Python script based on the IWI methodology provided by the Global Data Lab. The script will be made publicly available for replication and use as a template for easily adapting and applying the IWI to other countries and surveys. The resulting IWI was compared to the existing DHS WI utilizing a transition matrix based on households being broken down into quintiles for each wealth index. Transition matrices were generated based on using all households, only male headed households, and only female headed households.

The results indicated that for over 50% of households - regardless of gender - there was disagreement on the wealth index quintile when comparing the DHS WI to the IWI. Notably, female headed households were far more likely to be classified in a poorer quintile using the IWI than male headed households. As seen in Table 16, over 35% of female headed households were in a poorer quintile when using the IWI. Conversely, while there were similar overall levels of disagreement between the DHS WI and IWI for male headed households, the male headed households were slightly more likely to be in a weather quintile using the IWI.

Table 16. Shift in wealth quintiles using IWI compared to DHS WI.

Classification	Poorer quintile	Same quintile	Wealthier quintile
All	27.93%	47.05%	25.02%
Male head of household	24.39%	46.90%	28.71%
Female head of household	35.35%	47.35%	17.30%

In addition, female-headed households were considerably more likely to be severely reclassified (i.e., quintiles changed by more than one) using the IWI, while male-headed households were slightly more likely to be severely reclassified as wealthier using the IWI. A summary of quintile reclassification based on the transition matrices across all households, female headed households, and male headed households is detailed in Table 17. The complete transition matrices for all households, male headed households, and female headed households can be seen in Appendix Tables A3, A4, and A5 respectively.

Table 17. Quintile difference from DHS WI to IWI (values are percent of households)

difference	all	female	male
-3	0.14%	0.13%	0.15%
-2	4.04%	5.51%	3.33%
-1	23.75%	29.71%	20.90%
0	47.05%	47.35%	46.90%
1	20.23%	15.34%	22.56%
2	3.86%	1.78%	4.86%
3	0.82%	0.16%	1.14%
4	0.11%	0.03%	0.15%

Another way to visualize the differences between the DHS WI and the IWI is using a scatter plot where each point represents a household, with the IWI on the y-axis and the DHS WI on the x-axis (Figure 17). Both indices have been normalized for comparison. Points below the diagonal line can be interpreted as being classified as poorer by IWI than by DHS WI. The larger number of female households in red below the diagonal line represent the IWI classifying more households as poorer than the DHS WI. The orange trend line represents the female household data and emphasizes the divergence of the IWI from the DHS WI for poorer households in particular.

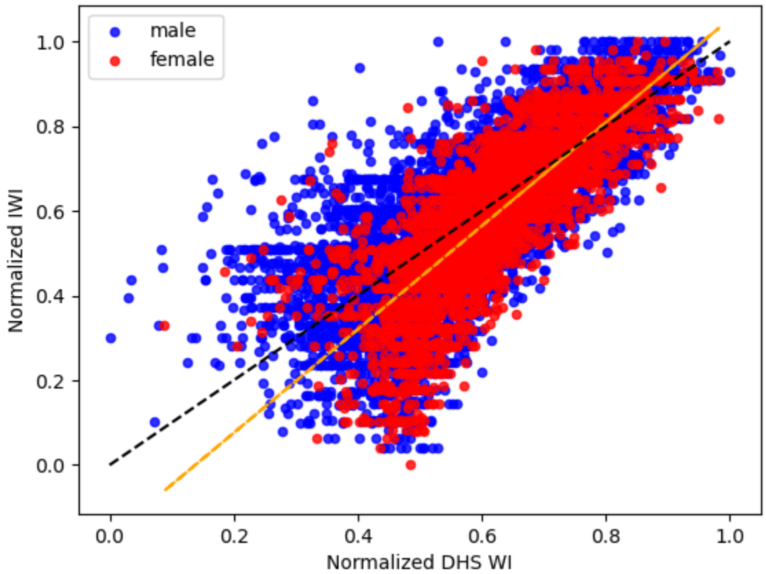


Figure 17. Comparison of DHS WI and IWI for male and female households.

A key caveat to both the gender specific DHS WI and IWI presented here is that we do not have an absolute or ground truth metric to say whether they are more accurate. However, alignment of both the gender specific DHS WI and IWI in classifying lower percentile female households as poorer than in the original DHS WI would suggest further research into the application of wealth indices for accurately classifying households is warranted. A key question for future research to explore is whether a standardized cross-country and cross-survey approach would serve to minimize specific biases such as those unique to gendered asset ownership in Ghana. Unfortunately, a deeper exploration of the true accuracy (ground-truthed) and merits of both the DHS WI and IWI (as well as other wealth index approaches) are beyond the scope of this work.

There is at least one notable caveat to these findings. The standardized approach for calculating the wealth index used by the IWI may reduce country / survey specific gender bias that the DHS WI may be subject to, yet increases dependency on a fixed and reduced set of assets for calculating the IWI. If ownership patterns of these fixed assets are gendered and not reflective of a gender-agnostic measure of wealth, the IWI may have inherent biases itself. For example, a bicycle is one of only seven consumer durables considered by the IWI, and is significantly more likely to be owned by male headed households in Ghana.

Discussion

The findings from this report are broadly indicative of a discrepancy in the effectiveness of ML based poverty estimates between male and female households. However, there are key limitations in what can be conclusively determined from the current analysis. In this section we will highlight a number of these limitations and directions for future work to help address them.

Limitations & Potential Solutions

Household Gender Classification

Approaches to classifying households by gender will inherently be limited along certain dimensions, as the majority of households consisting of a family are typically influenced in varying ways by both the men and women. The aim of the gender classification approaches, and the subsequent models/analysis in which were incorporated, are therefore not intended to provide precise gender statistics, but explore the potential for variations in trends which can be seen when looking at the data through different approaches. The classification approaches we have selected each provide a practical lens on household gender that may be useful in research or applications:

- **Gender of the head of household** reflects the mostly widely used approach in gender studies and therefore will be most directly comparable to existing work.
- **Whether a household contains any males** is a far more restrictive classification approach, yet guarantees that households classified as female could not possibly have male influence regardless of cultural norms or differences in how individuals view/define the head of the household.
- **Asset ownership and control** are unlikely to provide truly accurate classifications of households, yet isolating specific assets would potentially allow us to understand the effects of specific assets (of which ownership may be notably gendered) on wealth estimates based on asset ownership.

While each gender classification approach may provide specific value to more broadly understanding potential gender bias in AI-based wealth estimates, it is still critical to understand the limitations of each of these approaches as the limitations provide critical context relevant to how results may be interpreted.

A fundamental concern is that classification based on the gender of the head of household may obscure specific household dynamics related to key assets used in creating the DHS WI. For example, a household led by an older male may be practically dependent on younger female family members. Conversely, negative conditions individually faced by a female head of household which impact her asset acquisition (e.g., lower wage than a man in the same position) may be obscured by assets purchased by male members of her household. Similarly, information on cultural/regional trends related to who heads a household may be obscured or missing from a simple “head of household” based gender classification.

Another prominent concern is the implications of household gender classification on sample sizes. For example, restricting female households to those without any males greatly reduces the number of households classified as female, and also may not accurately reflect typical living conditions for women in Ghana. Use of specific assets ownership for gender classification can also result in either very limited or very inflated household counts based on excluding or including overlaps (i.e., a household may own both male and female assets).

To further complicate asset based classification, assets which are highly gendered in terms of absolute ownership rates may still have very high ownership for both genders. E.g., 20% more male headed households own a radio than female headed households, yet over 50% of female headed households still own a radio. While there is a large difference in gender-based ownership, it is not practical to use this asset for classification purposes.

Determining whether assets are gendered is difficult in itself and is based on two components. The first component is based on ownership rates related to the gender of the head of household. This is inherently limited as assets which exist within a household may be specifically owned or controlled by someone other than the head of household. Control, as it relates to asset ownership and decision making, is also difficult to uniformly define across households. The second component is based on CDD-Ghana's local context report related to gender-specific control of assets. Unfortunately no explicit definition of control/ownership is likely to apply to all cases. Within CDD's report, control tends to be associated with decision making, and in general, decision making is tied to who purchased/acquired the asset. But at the same time, the practicalities of most households means items may be purchased as a household while either the man/woman tends to be in control of them.

The ability to accurately classify household gender and gender specific association of asset ownership or control is highly dependent on the data and format in which data is collected from existing surveys. More granular surveys in which asset ownership and control, as well as subsequent wealth metrics, are at the individual level rather than the household level could help improve our ability to train AI models and evaluate the effectiveness of AI models at estimating gender specific wealth. However, the realities of household dynamics will likely mean that it is impossible to truly isolate individual wealth within a household. As such, improvements to approaches to accurately reflect gender driven wealth trends within the limitations of the available data will be key to advancing this line of work.

Gendered Sample Sizes

As indicated above, the size of gender specific samples used in models is highly dependent upon the household gender classification approach and findings indicate that there is a relationship between the number of households in a cluster and model performance. While tests revealed that this relationship does not account for the full disparity in model performance between male and female households, further work is needed to better understand the implications. The initial tests we conducted artificially reducing household counts within clusters could be expanded further across all gender classification approaches.

In addition to testing actual model performance, it may be possible to derive insight from the underlying distributions of the DHS WI within clusters of varying household sizes.

For instance, male household models may perform better in general because their DHS WI values are more closely correlated with full population DHS WI values. Reducing the population size past critical thresholds for female households may simply no longer reflect a consistent set of conditions for the population that the model is able to reasonably adapt to.

In addition to the limitations of our classification approaches, other limitations or assumptions exist in order to practically leverage the data available. Notably, clusters may have few households for minority gender. Across the classification schemes, particularly in the more extreme/restrictive approaches, certain survey clusters may end up with very few households (i.e., as few as 2) for a particular gender. This has the potential to create noisy data which is heavily dependent on conditions of a small sample of households not truly representative of the broader population (e.g., all female headed households in Ghana).

Feature selection and importance

Differences in feature importance between male and female models were relatively minor, yet are potentially connected to key local context conditions. Additional research is warranted to determine if the importance of certain features are in fact meaningful and how they reflect actual conditions in Ghana.

As an example, one feature which showed greater importance for male headed households was precipitation. As this could be reflective of men being more involved in the agricultural sector, one approach to test this would explore if models trained on a subset of male households in areas with greater agricultural activity showed even greater reliance on the precipitation feature. Similar tests could be applied for other features such as accessibility to cities, which was more important in female household models.

In addition, it would be worth exploring whether other geospatial features not included in this project could be incorporated into the models to improve performance - both in general, or for gender specific models.

Generalizability of models and use for predictions

One of the challenges in training and validating models within a single country is the limited amount of data. This is particularly notable when working with DHS data as the households are aggregated to clusters. Using only 427 samples (households clusters) to train and validate a model is difficult, and significantly impairs the ability to set aside extra data for an additional round of out of sample testing. Related work in this space has overcome this issue while simultaneously producing more generalizable models by including data from multiple countries. Typically models will be trained and validated on data from all but one country, and then tested on the remaining country's data. When exploring gender bias, there is of course the potential that this approach may overshadow country specific gender conditions to some extent.

Expanding on use of alternative wealth indices

We found that there are differences in the weight of specific assets considered in the original DHS WI and recreations of the DHS WI for gender specific populations. As explained earlier in the methodology section, this is because the DHS WI is dependent upon the asset ownership of the population upon which it is built. To further evaluate the difference between the original DHS WI and a gender specific WI, we would need to implement a similar analysis as we used to evaluate model performance for the original DHS WI in this project, but using the gender specific DHS WI as a comparison.

Similar to the gender specific DHS WI, we found that the IWI produces a sufficiently different classification of household wealth that warrants further analysis. The IWI could likewise be used to assess model performance and then conduct a comparison with models based on the DHS WI. In addition, other alternative wealth indices could be considered for inclusion in the comparison.

Conclusion

Equitable AI is still a young and evolving focal area in which there is still much to learn, particularly within specific topic areas. As such, the insights and experiences from practical applications and research of Equitable AI provide incredible value to the broader community to build upon. No known research has previously explored the relationship between gender and AI based estimates of wealth, or even considered the potential approaches for evaluating the performance of AI models for subsets of populations traditionally used to train wealth estimation models. Beyond the technical findings, the broader lessons learned from this project - understanding what worked and what did not - along with how to conceptualize and address application specific challenges, can hopefully help both encourage and facilitate future work around Equitable AI and the use of AI based estimates of wealth.

Our current research has indicated that AI models trained on female household data underperform relative to models trained on male household data, yet there are many aspects left to explore. An important area for consideration is whether current household gender classification is appropriate, and, more broadly, whether future surveys can be improved to assess gender-specific wealth and support gender-specific AI applications in general. Understanding what drives the differences between models trained on male and female data is also important - can other geospatial data features used in model training improve the performance of female models? Future research might also explore the possibility of utilizing wealth indices other than those produced by the DHS. The methods and code we have produced aim to provide an accessible approach for others to continue exploring these questions and others related to the role of gender in AI wealth estimation models.

References

Data & Methods

Allen, C., Metternicht, G. & Wiedmann, T. Prioritising SDG targets: assessing baselines, gaps and interlinkages. Sustain Sci 14, 421–438 (2019). https://doi.org/10.1007/s11625-018-0596-8

Avtar, Ram et al. “Assessing sustainable development prospects through remote sensing: A review.” Remote sensing applications : society and environment vol. 20 (2020): 100402. doi:10.1016/j.rsase.2020.100402

Ayala, C.; Sesma, R.; Aranda, C.; Galar, M. A Deep Learning Approach to an Enhanced Building Footprint and Road Detection in High-Resolution Satellite Imagery. Remote Sens. 2021, 13, 3135. https://doi.org/10.3390/rs13163135

Ayush, Kumar, et al. ‘Generating Interpretable Poverty Maps Using Object Detection in Satellite Images’. CoRR, vol. abs/2002.01612, 2020, https://arxiv.org/abs/2002.01612.

Barrington-Leigh, Christopher, and Adam Millard-Ball. ‘The World’s User-Generated Road Map Is More than 80% Complete’. PLOS ONE, vol. 12, no. 8, Public Library of Science, 08 2017, pp. 1–20, https://doi.org10.1371/journal.pone.0180698.

Belgiu, M., Dr˘agu, L., 2016. Random forest in remote sensing: A review of applications and future directions. ISPRS Journal of Photogrammetry and Remote Sensing 114, 24–31. doi:10.1016/J.ISPRSJPRS.2016.01.011.

Bennett, Mia M., and Laurence C. Smith. ‘Advances in Using Multitemporal Night-Time Lights Satellite Imagery to Detect, Estimate, and Monitor Socioeconomic Dynamics’. Remote Sensing of Environment, vol. 192, 2017, pp. 176–197, https://doi.org10.1016/j.rse.2017.01.005.

Brewer, Ethan, et al. ‘Predicting Road Quality Using High Resolution Satellite Imagery: A Transfer Learning Approach’. PLOS ONE, vol. 16, no. 7, Public Library of Science, 07 2021, pp. 1–18, https://doi.org10.1371/journal.pone.0253370.

Breiman, L., 2001. Random forests. Machine Learning 45, 5–32. doi:10.1023/A:1010933404324

Burke, Marshall et al. “Using satellite imagery to understand and promote sustainable development.” Science (New York, N.Y.) vol. 371,6535 (2021): eabe8628. doi:10.1126/science.abe8628

Camboim, S.P.; Bravo, J.V.M.; Sluter, C.R. An Investigation into the Completeness of, and the Updates to, OpenStreetMap Data in a Heterogeneous Area in Brazil. ISPRS Int. J. Geo-Inf. 2015, 4, 1366-1388. https://doi.org/10.3390/ijgi4031366

CIESIN, 2018. Gridded Population of the World, Version 4 (GPWv4): Population Density Adjusted to Match 2015 Revision UN WPP Country Totals, Revision 11. URL: https://sedac.ciesin.columbia.edu/data/set/gpw-v4-population-density-adjusted-to-2015-unwpp-country-totals-rev11, doi:https://doi.org/10.7927/H4F47M65.

Defourny, P. (2017): ESA Land Cover Climate Change Initiative (Land_Cover_cci): Land Cover Maps, v2.0.7. Centre for Environmental Data Analysis, 7/2017"

Ebener, S., Murray, C., Tandon, A. et al. From wealth to health: modeling the distribution of income per capita at the sub-national level using night-time light imagery. Int J Health Geogr 4, 5 (2005). https://doi.org/10.1186/1476-072X-4-5

Elvidge, C.D., Baugh, K., Zhizhin, M., Hsu, F.C., Ghosh, T., 2017. VIIRS night-time lights. International Journal of Remote Sensing 38, 5860–5879. URL: https://doi.org/10.1080/01431161.2017.1342050, doi:10.1080/01431161.2017.1342050

European Space Agency, 2017. ESA Land Cover. URL: http://maps.elie.ucl.ac.be/CCI/viewer/index.php.

Goodman, Seth, et al. ‘GeoQuery: Integrating HPC Systems and Public Web-Based Geospatial Data Tools’. Computers & Geosciences, vol. 122, 2019, pp. 103–112, https://doi.org10.1016/j.cageo.2018.10.009.

Hall, Ola, et al. ‘A Review of Explainable AI in the Satellite Data, Deep Machine Learning, and Human Poverty Domain’. Patterns, vol. 3, no. 10, 2022, p. 100600, https://doi.org10.1016/j.patter.2022.100600.

Harris, I., Osborn, T.J., Jones, P. et al. Version 4 of the CRU TS monthly high-resolution gridded multivariate climate dataset. Sci Data 7, 109 (2020). https://doi.org/10.1038/s41597-020-0453-3

Henderson, J. Vernon, Adam Storeygard, and David N. Weil. 2012. "Measuring Economic Growth from Outer Space." American Economic Review, 102 (2): 994-1028.

Herfort, B., Lautenbach, S., Porto de Albuquerque, J. et al. A spatio-temporal analysis investigating completeness and inequalities of global urban building data in OpenStreetMap. Nat Commun 14, 3985 (2023). https://doi.org/10.1038/s41467-023-39698-6
Neal Jean et al. ,Combining satellite imagery and machine learning to predict poverty. Science 353,790-794(2016).DOI:10.1126/science.aaf7894

Huffman, G., Bolvin, D., Braithwaite, D., Hsu, K., Joyce, R., Xie, P., 2014. Integrated Multi-satellite Retrievals for GPM (IMERG), version 4.4

Neal Jean et al. , Combining satellite imagery and machine learning to predict poverty. Science 353, 790-794 (2016). DOI:10.1126/science.aaf7894

D. Justus, J. Brennan, S. Bonner and A. S. McGough, "Predicting the Computational Cost of Deep Learning Models," 2018 IEEE International Conference on Big Data (Big Data), Seattle, WA, USA, 2018, pp. 3873-3882, doi: 10.1109/BigData.2018.8622396.

Lee, Kamwoo, and Jeanine Braithwaite. ‘High-Resolution Poverty Maps in Sub-Saharan Africa’. World Development, vol. 159, Elsevier BV, Nov. 2022, p. 106028, https://doi.org10.1016/j.worlddev.2022.106028.

Lin Htet, Nyan, et al. ‘Machine Learning Approach with Multiple Open-Source Data for Mapping and Prediction of Poverty in Myanmar’. 2021 18th International Conference on Electrical Engineering/Electronics, Computer, Telecommunications and Information Technology (ECTI-CON), 2021, pp. 1041–1045, https://doi.org10.1109/ECTI-CON51831.2021.9454768.

Lloyd, Christopher T., et al. ‘Using GIS and Machine Learning to Classify Residential Status of Urban Buildings in Low and Middle Income Settings’. Remote Sensing, vol. 12, no. 23, 2020, https://doi.org10.3390/rs12233847.

Lobell, D.B., Azzari, G., Burke, M., Gourlay, S., Jin, Z., Kilic, T. and Murray, S. (2020), Eyes in the Sky, Boots on the Ground: Assessing Satellite- and Ground-Based Approaches to Crop Yield Measurement and Analysis. Amer. J. Agr. Econ., 102: 202-219. https://doi.org/10.1093/ajae/aaz051

NASA, 2000. Shuttle-Radar Topography Mission. URL: http://www2.jpl.nasa.gov/srtm/.

NASA, 2017. Long Term Data Record. URL:
https://ltdr.modaps.eosdis.nasa.gov/cgi-bin/ltdr/ltdrPage.cgi?fileName=
13 products.

Neis, Pascal, et al. ‘Comparison of Volunteered Geographic Information Data Contributions and Community Development for Selected World Regions’. Future Internet, vol. 5, no. 2, 2013, pp. 282–300, https://doi.org10.3390/fi5020282.

OCO-2 Science Team/Michael Gunson, Annmarie Eldering (2020), OCO-2 Level 2 bias-corrected XCO2 and other select fields from the full-physics retrieval aggregated as daily files, Retrospective processing V10r, Greenbelt, MD, USA, Goddard Earth Sciences Data and Information Services Center (GES DISC), 10.5067/E4E140XDMPO2

Oshri, Barak, et al. 'Infrastructure Quality Assessment in Africa Using Satellite Imagery and Deep Learning'. Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, ACM, 2018, https://doi.org10.1145/3219819.3219924.

Perez-Heydrich, C., Warren, J.L., Burgert, C.R., Emch, M.E., 2013. Guidelines on the use of DHS GPS Data: DHS Spatial Analysis Reports 8. Technical Report. USAID. URL:https://dhsprogram.com/pubs/pdf/SAR8/SAR8.pdf

Pozzi, Francesca, et al. 'Modeling the Distribution of Human Population with Nighttime Satellite Imagery and Gridded Population of the World'. Earth Observation Magazine, vol. 12, 01 2003, pp. 24–30.

Rutstein, S.O., Johnson, K., 2004. The DHS Wealth Index. URL: https://dhsprogram.com/publications/publication-cr6-comparative-reports.cfm.

Sirko, Wojciech, et al. 'Continental-Scale Building Detection from High Resolution Satellite Imagery'. CoRR, vol. abs/2107.12283, 2021, https://arxiv.org/abs/2107.12283.

Steele, J.E., Sunds\oy, P.R., Pezzulo, C., Alegana, V.A., Bird, T.J., Blumenstock, J.E., Bjelland, J., Engø-Monsen, K., de Montjoye, Y.A., Iqbal, A.M., Hadiuzzaman, K.N., Lu, X., Wetter, E., Tatem, A.J., Bengtsson, L., 2017. Mapping poverty using mobile phone and satellite data, in: Journal of the Royal Society, Interface.

Tingzon, I., et al. 'MAPPING POVERTY IN THE PHILIPPINES USING MACHINE LEARNING, SATELLITE IMAGERY, AND CROWD-SOURCED GEOSPATIAL INFORMATION'. The International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences, vol. XLII-4/W19, 2019, pp. 425–431, https://doi.org10.5194/isprs-archives-XLII-4-W19-425-2019.

Tollefsen, A. F., Strand, H., & Buhaug, H. (2012). PRIO-GRID: A unified spatial data structure. Journal of Peace Research, 49(2), 363–374. https://doi.org/10.1177/0022343311431287

Wan, Z., Hook, S., Hulley, G., . MOD11C3 MODIS/Terra Land Surface Temperature/Emissivity Monthly L3 Global 0.05Deg CMG V006 [Data set].

Weiss, D.J., Nelson, A., Gibson, H.S., Temperley, W., Peedell, S., Lieber, A., Hancher, M., Poyart, E., Belchior, S., Fullman, N., Mappin, B., Dalrymple, U., Rozier, J., Lucas, T.C., Howes, R.E., Tusting, L.S., Kang, S.Y., Cameron, E., Bisanzio, D., Battle, K.E., Bhatt, S., Gething, P.W., 2018. A global map of travel time to cities to assess inequalities in accessibility in 2015.Nature 2018 553:7688 553, 333–336. URL: https://www.nature.com/articles/nature25181, doi:10.1038/nature25181.

Wessel, P., Smith, W.H., 1996. A global, self-consistent, hierarchical, high-resolution shoreline database. Journal of Geophysical Research B: Solid Earth 101, 8741–8743. URL: https://www.researchgate.net/publication/248796490_A_global_self-consistent_hierarchical_high-resolution_shoreline_database, doi:10.1029/96JB00104

Willmott, C. J. and K. Matsuura (2001) Terrestrial Air Temperature and Precipitation: Monthly and Annual Time Series (1950 - 1999), http://climate.geog.udel.edu/~climate/html_pages/README.ghcn_ts2.html

WorldPop, CIESIN, 2018. Global High Resolution Population Denominators Project. doi:https://dx.doi.org/10.5258/SOTON/WP00647

Yeh, C., Perez, A., Driscoll, A. et al. Using publicly available satellite imagery and deep learning to understand economic well-being in Africa. Nat Commun 11, 2583 (2020). https://doi.org/10.1038/s41467-020-16185-w

Yeh, Christopher, et al. 'SustainBench: Benchmarks for Monitoring the Sustainable Development Goals with Machine Learning'. arXiv [Cs.LG], 2021, http://arxiv.org/abs/2111.04724. arXiv.

Zhao, L., Kusumaputri, P., 2016. OpenStreetMap Road Network Analysis for Poverty Mapping.

Zhao, X., Yu, B., Liu, Y., Chen, Z., Li, Q., Wang, C., Wu, J., 2019. Estimation of Poverty Using Random Forest Regression with Multi-Source Data: A Case Study in Bangladesh. Remote Sensing 11, 375. URL: http://www.mdpi.com/2072-4292/11/4/375, doi:10.3390/rs11040375

Zheng, Shudan, and Jianghua Zheng. 'Assessing the Completeness and Positional Accuracy of OpenStreetMap in China'. Thematic Cartography for the Society, edited by Temenoujka Bandrova et al., Springer International Publishing, 2014, pp. 171–189, https://doi.org10.1007/978-3-319-08180-9_14.

Zhou, Qi, and Hao Lin. 'Investigating the Completeness and Omission Roads of OpenStreetMap Data in Hubei, China by Comparing with Street Map and Street View'. CoRR, vol. abs/1909.04323, 2019, http://arxiv.org/abs/1909.04323.
Tian, Y.; Zhou, Q.; Fu, X. An Analysis of the Evolution, Completeness and Spatial Patterns of OpenStreetMap Building Data in China. ISPRS Int. J. Geo-Inf. 2019, 8, 35. https://doi.org/10.3390/ijgi8010035

Classification / Asset Literature

Behrman, J.A. "Women's land ownership and participation in decision-making about reproductive health in Malawi." *Popul Environ* 38 (2017) 327–344. https://doi.org/10.1007/s11111-017-0272-4

Doss, Cheryl R. et al. The Rural Gender Asset and Wealth Gaps: Evidence from Ghana, Ecuador, Uganda and Karnataka, India. Bangalore: Indian Institute of Management Bangalore. 2012.

Doss, Cheryl R. et al. "Gendered Paths to Asset Accumulation? Markets, Savings, and Credit in Developing Countries." *Feminist Economics* 25.2 (2019) 36–66. Web.

Doss, Cheryl and Kieran, Caitlin and Kilic, Talip. Measuring Ownership, Control, and Use of Assets (July 2017). World Bank Policy Research Working Paper No. 8146, Available at SSRN: https://doi.org/10.1596/1813-9450-8146

Fisher B, Naidoo R (2016) The Geography of Gender Inequality. *PLoS ONE* 11 (2016) 3. e0145778. https://doi.org/10.1371/journal.pone.0145778.

Friedmann-Sanchez, Greta. 2006. "Assets in Intra Household Bargaining Among Women Workers in Colombia's Cut-Flower Industry." *Feminist Economics* 12(1/2): 247-69.

Gaddis, Isis and Lahoti, Rahul Suresh and Li, Wenjie, Gender Gaps in Property Ownership in Sub-Saharan Africa (August 30, 2018). World Bank Policy Research Working Paper No. 8573, Available at SSRN: https://ssrn.com/abstract=3246145

Hillesland, Marya. "Gender Differences in Risk Behavior: An Analysis of Asset Allocation Decisions in Ghana." *World Development* 117 (2019) 127–137. Web.

Johnson et al. 2016. "Gender, Assets, and Agricultural Development: Lessons from Eight Projects." *World Development* 83 (2016) 295-311. https://doi.org/10.1016/j.worlddev.2016.01.009

Miedema et al. "Women's empowerment in East Africa: Development of a cross-country comparable measure". *World Development* 110 (2018) 453-464. https://doi.org/10.1016/j.worlddev.2018.05.031.

Oladokun, Yetunde Olasimbo Mary and Adenegan, Kemisola Omorinre. "Level Asset Ownership and control among Women in Nigeria." *Asian Journal of Agricultural Extension, Economics & Sociology* 21 (2017) 1-11; Article no. AJAEES.35943

Peterman et al. "Women's Individual Asset Ownership and Experience of Intimate Partner Violence: Evidence from 28 International Surveys." *AmJPublicHealth* 107 (2017) 747-755. Doi: 10.2105/AJPH.2017.303694.

Pyburn, Rhiannon, and Anouka van Eerdewijk (editors). 2016. *A different kettle of fish? Gender integration in livestock and fish research*. LM Publishers, Volendam.

Raj, Anita. “Gender Empowerment Index: a Choice of Progress or Perfection.” *The Lancet Global Health* 5.9 (2017) 849–850. Web.

Tasciotti, Luca and Wagner, Natascha. “How much should we trust Micro-data? A comparison of the Socio-demographic Profile of Malawian Households using census, LSMS, and DHS data.” *The European Journal of Development Research* 30 (2018) 588–612. https://doi.org/10.1057/s41287-017- 0083-6; published online 30 March 2017

Appendices

A1. Asset Based Gender Classification

Table S1. Gendered assets based on control, defined by CDD's local context report. Limited to assets which overlap with assets used in DHS wealth index.

Assets	Mode of Acquisition	Control
Agric suitable lands	Inheritance	Male
Motorcycle/ Scooter	Self-purchased/ inheritance	Male
Car/ Truck	Self-purchased/ inheritance	Male
Boat with a motor	Self-purchased/ inheritance	Male
Music player	Self-purchased	Male
Computer/Laptop	Self-purchased	Male
Photo camera	Self-purchased	Male
Animal-drawn cart	Self-purchased	Male
Bicycle/ Tricycle	Self-purchased	Male
Video deck/DVD/VCD	Self-purchased	Female
Sewing machine	Self-purchased	Female
Type of cooking fuel	Self-purchased	Female
Refrigerator	Self-purchased	Female
Freezer	Self-purchased	Female
Washing machine & Dryer	Self-purchased	Female

Table S2. Highly gendered assets based on DHS data and whether any males are in the household, combined with CDD's findings. Note: this used males aged 15-49 rather than 65 as discussed in the main section of the report.

Asset	**% Among All HHs**	**% Among HHs w/ No Males**	**% Among HHs w/ 1+ Males**	**Difference (Female-Male)**
Bicycle	30.12	13.96	38.01	-24.05***
Motorcycle/scooter	12.41	3.63	16.70	-13.07***
Bank account	47.61	39.31	51.67	-12.36***
Owns land suitable for agriculture	46.18	40.80	48.80	-8.00***
Type of cooking fuel: charcoal	30.00	35.42	27.35	8.07***
Source of drinking water: public tap/standpipe	21.38	25.22	19.50	5.72***
Type of toilet facility: ventilated improved pit latrine	33.74	37.27	32.01	5.26***
Type of toilet facility: ventilated improved pit latrine shared	30.95	34.31	29.31	5.00***

Table S3. Omitted assets

Asset	**% Among All HHs**	**% Among HHs w/ Female HoH**	**% Among HHs w/ Male HoH**	**Difference (Female-Male)**
Radio	66.40	52.26	73.15	-20.89***
Video deck/DVD/VCD	35.41	26.52	39.66	-13.14***
Main wall material: cement	51.58	57.02	48.99	8.03***
Main roof material: asbestos/slate roofing sheets	61.72	67.13	59.13	8.00***

We did not include assets which although having a strong difference in ownership between male and female households, either A) conflicted with CDD-Ghana's findings on gendered asset control (e.g., video deck) or B) were very common in all households (e.g., radio). Notable examples of omitted assets are included in Table S3.

A2. Alternative DHS WI Transition Matrices

Table A1. Transition matrix comparing quintiles of DHS WI created using data from all households vs DHS WI created using data from households with no adult males.

Count		**WI Using Female Households**					
		1	**2**	**3**	**4**	**5**	***Total***
WI Using All Households	**1**	298	0	0	0	0	*298*
	2	365	292	4	0	0	*661*
	3	0	294	476	94	2	*866*
	4	0	0	75	455	96	*626*
	5	0	0	0	0	445	*445*
	Total	*663*	*586*	*555*	*549*	*543*	*2896*

Table A2. Transition matrix comparing quintiles of DHS WI created using data from all households vs DHS WI created using data from households with at least 1 adult male.

Count		**WI Using Male Households**					
		1	**2**	**3**	**4**	**5**	***Total***
WI Using All Households	**1**	1985	237	0	0	0	*2222*
	2	0	1494	203	0	0	*1697*
	3	0	1	1647	40	0	*1653*
	4	0	0	61	1585	7	*1653*
	5	0	0	0	82	1509	*1591*
	Total	*1985*	*1732*	*1911*	*1707*	*1516*	*8851*

A3. IWI Transition Matrices

Table A3. Transition matrix for all households between IWI and DHS WI

All Households		IWI Quintile				
		0	1	2	3	Q
DHS Wealth Index Quintile	0	50.57%	35.74%	9.80%	3.34%	0.55%
	1	38.32%	32.66%	22.14%	6.13%	0.76%
	2	10.77%	26.32%	40.64%	18.88%	3.38%
	3	0.34%	4.94%	29.91%	40.43%	24.38%
	4	0.00%	0.38%	4.48%	24.21%	70.93%

Table A4. Transition matrix for male households between IWI and DHS WI

Male Head of Household		IWI Quintile				
		0	1	2	3	4
DHS Wealth Index Quintile	0	47.25%	37.47%	10.81%	3.86%	0.62%
	1	35.09%	30.34%	25.66%	7.85%	1.06%
	2	8.44%	23.07%	42.03%	22.13%	4.33%
	3	0.40%	5.26%	27.58%	41.24%	25.52%
	4	0.00%	0.36%	4.27%	24.56%	70.80%

Table A5. Transition matrix for female households between IWI and DHS WI

Female Head of Household		IWI Quintile				
		0	1	2	3	4
DHS Wealth Index Quintile	0	65.80%	27.83%	5.19%	0.94%	0.24%
	1	44.07%	36.78%	15.86%	3.06%	0.24%
	2	14.08%	30.92%	38.67%	14.29%	2.04%
	3	0.23%	4.39%	33.95%	39.03%	22.40%
	4	0.00%	0.42%	4.96%	23.37%	71.25%

A4. Local Context Report Summary

In this section we provide a summary of the local context report assembled by the Center for Demographic Development based in Accra, Ghana. The full report can be accessed online here.

Overview. Assets accumulation and control within households (usually consisting of a husband, wife, children, and extended family) is varied across genders. More often than not, the local conditions and norms favor the men. In classifying households and assets control based on male-headedness and female-headedness, scholars have held almost common grounds in their discussions.

- Male-headed households are more likely to accumulate more assets than female-headed households.
- Livelihoods in a male-headed household are expected to be better than that of a female-headed household because of the ability of the male-headed household to acquire more assets relative to that of the female-headed household.

Decision making. Traditionally, men are considered the head of the household.

- Paying the bride price gives men control of decision-making in almost all aspects of the household.
- In most cases, the husbands are more likely to be older than their wives and to some extent, exercise authority over their wives, including asset accumulation and control.
- Wage earnings also contribute to the dynamics of decision-making over assets. Even when a wife earns more a husband is often the head of household and controls assets.
- Scholars argue that men are more likely to claim ownership of assets than women, and thus, they have greater control.
- Men are more likely to put up their assets for collateral than women would or are able to and thereby more likely to make decisions over assets than women.

Religion. The majority of the populace follow religions in which men's power over women is deeply entrenched.

Marriage. While single women are more independent in decision-making on asset accumulation, married women require the consent of their husbands in most cases to make final decisions on asset control.

- The situation worsens for married women when they do not come into the marriage with assets
- Both men and women generally accept status-quo of male dominance. Women can influence household decisions, yet may ultimately face varying levels of punishment for going against their husbands.

Urban vs Rural. Urban households are more likely to own other properties and businesses and can accumulate more durable financial assets than rural areas because they generate more income.

- Rural households usually prefer to accumulate productive assets than to have financial assets, and have a higher share in residential property, agricultural land (excluding family land), farm equipment, and livestock,

Assets Acquisition. Assets are mainly acquired through two principal means within the Ghanaian setting: inheritance and self-purchasing.

- Inheritance:
    - Inheritance turns out to be the most common form of acquiring assets in the Ghanaian setting, often granting equal opportunity to all, irrespective of their gender.
    - Under customary law, Ghana has matrilineal and patrilineal family systems.
        - Inheritance regimes can be broken down by region based on the dominant ethnic group.
    - Customary law disallows spouses from claiming each other's assets, and can be biased against women.
    - Laws can create a leeway for the man's family to deprive the woman of some portions of the assets/properties acquired in the marriage.
    - In multigenerational families, women are most likely to be handicapped in accumulating and controlling assets.
        - usually passed through the male lineage.
    - Land is the most common asset acquired through inheritance.
        - The accumulation and control of land varies across gender and regions.
        - overall men often possess more land such as agricultural parcels, and may benefit from greater ownership rights
        - Women, particularly in North, can be dependent on men for access to land
        - Women in matrilineal communities have more land access and control than those in patrilineal civilizations
    - The ceremonial practice for marriage has turned out to be the commonest means of asset accumulation by women – through dowry payment (often considered a woman's inheritance).

- Self-purchase/ self-acquired assets
    - Largely influenced by education and income level with regards to what is owned and quantities owned as well as location.
    - The positive effect of education on asset ownership is stronger for females than males.
    - Income has a stronger positive effect on asset ownership for men than for women.
    - Income has a stronger positive effect in rural households and education has a stronger positive effect in urban households.
    - With comparable low levels of education, male heads were relatively better than the female heads. Possibly due to favor of investing in male education over female education

Education, Employment, and Income. Males are more economically active, more likely to be employed with wages, earn more per month, and less likely to be unemployed. There are also higher rates of male literacy and comparatively higher levels of education

- Has been argued that men have more potential to accumulate assets than women based on the various employment fields
    - More men in skilled agricultural work results from the men's control of farmlands.
- Women (and rural residents) hold fewer active financial accounts vs men (and urban residents)
    - Affects their purchasing power and the ability to invest in assets since they are not exposed to bigger financial resources to provide for these assets (e.g. loans, mortgages)
    - Culminates in their weaker financial strength and ability to secure financial and durable assets over time.
- Scholars establish a positive correlation between education and asset accumulation.
    - People with little or no education are most likely to venture into the informal sector.
    - Far more women are service/sales workers with the motive of raising core household income rather than accumulating lasting assets.

List of selected assets, mode of acquisition, gender control

Assets	**Mode of Acquisition**	**Control**	**Included in DHS WI**
Agric suitable lands	Inheritance	Male	Y
Greenhouse	Inheritance	Male	
Building	Self-purchased/ inheritance	Male	
Motorcycle/ Scooter	Self-purchased/ inheritance	Male	Y
Car/ Truck	Self-purchased/ inheritance	Male	Y
Boat with a motor	Self-purchased/ inheritance	Male	Y
Music player	Self-purchased	Male	Y
Computer/Laptop	Self-purchased	Male	Y
Photo camera	Self-purchased	Male	Y
Animal-drawn cart	Self-purchased	Male	Y
Bicycle/ Tricycle	Self-purchased	Male	Y
Television	Self-purchased	Male	Y
Exotic dogs	Self-purchased	Male	
Satellite dish	Self-purchased	Male	
Bluetooth music-playing device	Self-purchased	Male	
Bed	Self-purchased	Male	
Art (paintings)	Self-purchased	Male	
Gym equipment	Self-purchased	Male	
Gold and Silver Jewelry	Self-purchased/ Inheritance	Female	
Video deck/DVD/VCD	Self-purchased	Female	Y
Sewing machine	Self-purchased	Female	Y
Type of cooking fuel	Self-purchased	Female	Y
Refrigerator	Self-purchased	Female	Y
Freezer	Self-purchased	Female	Y
Washing machine & Dryer	Self-purchased	Female	Y
Cabinet/ cupboard	Self-purchased	Female	
Cooking stove (gas/electric)	Self-purchased	Female	
Exotic potted plants	Self-purchased	Female	

A5. Identifying Local Organizations

CDD-Ghana identified 20 organizations in Ghana working in areas dealing with 1) development, 2) research and technology, and 3) gender. These organizations were identified as having potential projects which involved the use of AI, geospatial data, household data, or gendered data, as well as analysis or other work in the same areas. They therefore could provide insight into practical in-country applications and needs related to AI based estimates of wealth, as well as the broader state of equitable AI and gender related developments within Ghana.

Table A5.1 - List of Identified Organizations

	Development - National and Subnational
1	Ghana Statistical Service
2	Ghana Health Service
3	National Development Planning Commission
4	Innovations for Poverty Action (IPA)
5	National Population Council
6	The Millennium Development Authority (MiDA)
7	United Nations Development Programme (UNDP)
	Research & Tech Institutions
8	Ghana Tech Lab
9	Institute of Statistical Social and Economic Research, ISSER
10	National Information Technology Agency
11	AIMS Ghana (African Institute for Mathematical Sciences)
12	KNUST Responsible AI Lab
13	Kofi Annan International Peacekeeping Training Centre
14	PurpleDot Limited (Private)
15	Ashesi University (Private)
16	Penplusbytes (NGO)
	Gender Specific Development
17	United Nations Population Fund (UNFPA)
18	NETRIGHT – GHANA
19	ACT Foundation
20	Women in Tech Africa

Out of the 20 organizations identified, CDD in partnership with AidData completed 6 interviews with a subset of the organizations that were A) identified as having the most relevant work overlap with this project and B) were available for interviews. These 6 interviews were with Responsible AI Lab (RAIL), Ghana Statistical Service (GSS), the National Development Planning Commission (NDPC), Ghana Tech Lab, National Population Council (NPC), and Purple Dots Ltd. The interviewed organizations work with a variety of different tools on a wide range of different projects. Many of the organizations are very interested in utilizing AI more in their work and several of them are currently utilizing AI programs to assist users on the ground.

For example, RAIL is using AI to enhance services for victims of GBV/IPV and they have also produced a tool that helps palm sellers (predominantly women) verify the oil they sell in the market. Other organizations such as Ghana Tech Lab and Purple Dot Ltd. are working on AI tools for health care. The biggest challenge organizations listed as issues are obtaining low-cost or free data and the purchase and maintenance of the technology needed to run these AI programs. These organizations highlight the presence of active organizations working in AI in Ghana with the experience and interest to collaborate on future AI projects in Ghana.

A6. Local Workshop

On Wednesday August 16th, CDD-Ghana hosted nearly 40 participants in person for the workshop held at their offices in Accra, Ghana. In addition to the live portion of the workshop, a live stream was provided using a combination of YouTube and Zoom. The live stream was attended by approximately 20 virtual participants.

Participants ranged from students at local universities such as University of Ghana to staff from government agencies. There was also a mixture of local nonprofits, development organizations, and private companies. In person attendance was fairly balanced between genders, with over 40% of the participants being women.

Following opening remarks by Dr. Edem Selormey, Director of Research at CDD-Ghana, were the main presentation sessions by the Ghana Statistical Service and AidData. Dr. Peter Takyi Pepra, the Director of Field Operations & Coordination for the 2022 Ghana Demographic and Health Survey, presented a first look at takeaways from the latest round of DHS data in Ghana. Dr. Rachel Sayers of AidData then presented on the findings of our work within the Equitable AI Challenge on evaluation gender bias in AI applications using household survey data.

EQUITABLE AI PROJECT AGENDA - EVALUATING GENDER BIAS IN AI APPLICATIONS DISSEMINATION WORKSHOP Date: 16th August, 2023 Time: 9:30 a.m. – 1:15 p.m		
09:30 am - 10:00 am	**Arrival and Registration**	All
10:00 am – 10:10am	**Opening Remarks**	Edem Selormey (PhD), Director of Research, CDD-Ghana
10:10 am – 11:10 am	**Presentation:** • Latest Ghana Demographic and Health Survey (DHS) Findings • Evaluating Gender Bias in AI Applications Using Household Survey Data: Findings	Dr Peter Takyi Peprah, Dir. Field Operations & Coordinator, 2022 GDHS. Ghana Statistical Service Rachel Sayers (PhD), Research Scientist, AidData
11:10 am – 12:05 pm	**Panel Discussion**	Panel Members: Selaseh Pashur Akaho, Statistician, GSS Dr. Rita Udor, Gender Inclusivity Officer, RAIL- KNUST; and Deborah Dormah Kanubala ML Researcher, Saarland University Germany
12:05 pm – 1:00 pm	**Plenary**	Moderator
1:00 pm – 1:10 pm	**Closing Remarks**	Mavis Zupork Dome, Research Coordinator, CDD-Ghana

Workshop agenda

An expert panel then discussed the work presented within the context of their own experience and with a broader lens towards the state of AI, gender, and Equitable AI within Ghana. Panelists included Selaseh Pashur Akaho, a statistician from the GSS; Dr. Rita Udor, a gender inclusivity officer from the Responsible AI Lab at the Kwame Nkrumah University of Science and Technology; and Deborah Dormah Kanubala, a machine learning researcher at Saarland University in Germany. The diverse specialization of the panel of Ghanaian researchers ranged from survey statistics, to gender inclusivity, and applied machine learning and AI.

Following the panel discussion, and a short break due to a power outage, the workshop resumed with a plenary discussion which opened the floor to the audience. The conversation touched upon the role of Equitable AI and poverty estimates in Ghana, as well as the broader context of gender bias, the future of AI, and the importance of ongoing discussion and efforts to ensure no minority groups are harmed by advancing technologies. Participants were clearly deeply engaged and passionate about the topic, yet also raised critical realities that may impede the advancement of Equitable AI in the short term.

> *"Until we alter our gender perspectives as a people, we are likely to influence AI models to exhibit biases. We must ensure that the development of AI models does not negatively impact minorities."*

In particular, participants highlighted the influence of biased gender norms within the country that must be addressed, and the risk of inherent bias being reflected in AI applications. The discussion raised fundamental issues concerning people in Ghana surrounding whether a nation struggling with more basic issues is ready for AI, and whether AI risks making things worse. One participant questioned the extent to which AI might result in job loss, and whether it would disproportionately impact women. Panelists emphasized the potential of AI to improve lives, and that while it may change what jobs look like, it is unlikely that many will lose jobs.

The conversation was ongoing as the workshop passed the allocated time, and Mavis concluded the event with reflections upon the project and collaboration, and the future of Equitable AI.

Photos & Recordings

The primary photos from the photographer at the event are available on CDD-Ghana's Flickr. A recording of the event has been made available on YouTube.

Media Coverage

Social Media

AidData and CDD-Ghana's communications teams were actively promoting the workshop across major social media platforms. While there was limited anticipated engagement on these platforms, posts leading up to the event were viewed by hundreds of users.

AidData produced multiple posts on Twitter [1, 2, 3] as well as on Facebook and LinkedIn. Similarly, CDD-Ghana produced multiple posts on Twitter [1, 2, 3] and Facebook.


CDD-Ghana
@CDDGha

From today's discussion:
Until we alter our gender perspectives as a people, we are likely to influence AI models to exhibit biases. We must ensure that the development of AI models does not negatively impact on minorities.
#aiapplications
#genderbias
#equitableai
#WomensWealth

DAI and 5 others

10:14 AM · Aug 16, 2023 · 211 Views

In addition, a couple of other organizations posted or reposted on Twitter after the event and related meetings.These included the [International Centre for Evaluation and Development] and [Development Impact West Africa].


ICED
@ICED_THINKTANK

So many opportunities for capacity strengthening to support individuals' development:

Today, our staff member participated in and captured this dissemination workshop on #AI & #genderbias by @CDDGha, with experts like Dr. Edem Selormey (@WatchWomanEdem) & Dr. @PeterPeprahGSS

William & Mary and 3 others

2:48 PM · Aug 16, 2023 · 62 Views


Development Impact West Africa (DIWA)
@diwa_gimpa

DIWA meets @AidData, a research lab based at the @williamandmary in Virginia, USA to discuss partnership arrangements for the evaluation of interventions in #Gender Equity in urban areas within the Sub-region.

7:50 AM · Aug 15, 2023 · 395 Views

Local Media

Nine representatives of local media organizations were in attendance. The media organizations included the Ghana News Agency, Ghana Times, the Daily Graphic, the

Ghana Broadcasting Company, Ghana TV, TV3, and 3 News. Notably, the media was almost equally represented by men and women.

Immediately following the workshop, Rachel and Mavis engaged in a Q&A session with the media that spanned a range of topics, and resulted in local publications by the GNA, All Africa, and Graphic Online, as well as coverage during a segment on Ghana Broadcasting Company (GBC) Radio . The following morning, Rachel and Mavis were also interviewed on the Ghana Broadcasting Corporation's Uniiq FM Breakfast Show in an effort to engage a broader local audience by leveraging radio.

The media articles did somewhat focus more on the broader discussion around AI and bias than on the specific elements of our project's application, and some specific paraphrasing of speakers was extrapolated to apply to much more generalized ideas than was perhaps intended. Overall, coverage was positive and reflects both public interest in concerns around AI and the importance of raising awareness in local areas where AI may impact communities.

Reflections

The event went very smoothly overall, even despite a power outage, and there was very good attendance by interested and engaged participants and media. The most resounding success of the workshop was the level of engagement and discourse between incredibly knowledgeable panelists and presenters, and a clearly passionate and informed audience.

A survey was provided to all participants to explore their thoughts on the workshop, the content, and discussions. A limited number of participants responded to the survey, but overall their responses were positive. All 7 respondents said they would recommend the workshop to others, and 6 out of 7 said the workshop improved their understanding of the topic. There were some responses indicating that the workshop may have been slightly longer than needed, yet when asked to rate the workshop overall, all responses were positive.

Follow Up Opportunities

While there were few organizations or participants who specifically leverage AI based poverty estimates, many were active in various work streams involving household survey data, gender data, and/or AI in general. A key short term focus of our follow up engagement with participants will be to provide everyone on the registration list with the final report and other dissemination materials.

The Responsible AI Lab (RAIL) at the Kwame Nkrumah University of Science and Technology (KNUST) is one of the few groups in Ghana focusing directly on Equitable AI that was identified during our organization interviews earlier in the project, and they were very well represented by our panelist Dr. Rita Udor. We will aim to stay in contact with Dr. Udor, and explore collaboration opportunities with RAIL-KNUST.

While some groups like RAIL-KNUST were expected as strong candidates for follow up engagements, our other panelist Deborah Dormah Kanubala is someone we did not initially know of as she is an active graduate student, yet is clearly well versed in not only the technical aspects of AI, but also cognizant of the challenges of pursuing Equitable AI applications.

Other familiar organizations such as the Ghana Statistical Service (GSS) are also likely candidates for follow up on any work involving the Ghana DHS. With the forthcoming release of the 2022 Ghana DHS, Dr. Peter Takyi Pepra's presentation was both timely and helped to establish a direct relationship within the GSS to reach out to in the future.

Many participants were deeply interested and engaged in the topic, and we will express our interest in general follow up opportunities when distributing the final report and other material from the project to workshop participants. However, a few participants indicated deeper knowledge and applied considerations that are worth exploring directly. One example includes a researcher who expressed practical concerns around the lack of connection between many generic presentations on AI methods with practical follow up that can be adopted by real world projects within Ghana.

A7. Code and Other Replication Materials

All code and finalized data, aside from data retrieved directly from the DHS survey, along with documentation are made publicly available through a GitHub repository (or "repo"). GitHub is a widely used platform for sharing projects containing code and data, and provides functionality that supports users who wish to contribute to existing projects or "fork" projects to adapt them for their own applications. In addition, a repo maintains a history of all contributions and changes made to a project, as well as numerous other tools for including documentation (or "ReadMes") alongside code and data, and facilitating sharing and replication of work.

The GitHub repository contains the base code and documentation for preparing data and training models to generate wealth estimates using random forests. The code can be extended and applied to any particular application of generating wealth estimates, such as for different countries or exploring arbitrary subsets of surveyed populations. Different survey data or input features can also be used with minimal to moderate modifications to the code, depending on the potential data formats and preprocessing. Configuration settings and additional analysis code specific to the Equitable AI work is included alongside the base code.

Data processing and model training to replicate workflow for this project are detailed in the main ReadMe within the repository. By following the instructions within, and utilizing the "eqai_config.ini" configuration file as specified as directed, users can reproduce the outputs and findings of our work. In addition, for users who wish only to explore the metrics produced across the wide range of models generated without actually having to run all the models, we provide a database consisting of all model outputs. The database, which can be accessed as described in the ReadMe, is also accompanied by model files which will enable users to directly load a pretrained model.

Beyond the core code and files to prepare data, train models, and store model outputs, we have also provided numerous additional files and outputs associated with our analysis of the models and metrics. The contents of the "equitable-ai" subfolder in the repo enabled users to replicate all figures, tables, descriptive statistics provided in our reports and findings.

Within the "equitable-ai" folder, there are several subfolders.

- *dhs_docs* - provides all documentation associated with the 2014 Ghana DHS round
- *figures* - Contains all figures produced for the Equitable AI reports and other products
- *gendered_dhs_wi_comparison* - Stata code and ReadMe for reproducing the DHS WI, and for generating gender specific versions of the DHS WI
- *dhs_wi_misc* - This folder was not used for any outputs or analysis, but - as it may be relevant to others - it contains a Python based script for reproducing a more primitive version of the DHS Wealth Index (purely PCA based, without additional refinement steps used in the Stata code following the official DHS WI construction methodology).
- *iwi_comparison* - Contains Python code for 1) generating the International Wealth Index for the 2014 Ghana DHS round (along with instructions for running and

adapting to other surveys), and 2) comparing the IWI with the DHS WI for households in the 2014 Ghana DHS round.

- *pca_weights* - Contains code and outputs for comparing the weights associated with assets generated during the creation of the DHS WI. This compares the original DHS WI asset weights with the gender specific DHS WI asset weights.
- *predictions* - We provide Python code to load a specific trained model and then run all 2014 Ghana DHS households through the model to estimate their wealth.

All code has been tested and run on Debian and Fedora based Linux distributions, but can reasonably be expected to run with most Python installations given proper environment setups (as detailed in main ReadMe). Any issues, suggestions for improvements, or contributions can be added directly via the GitHub repo.